This article was published in:
Catalogue and Index, no.146, Winter, pp. 1-6

DEMYSTIFYING METADATA

Ann Chapman

Abstract
This paper considers what metadata is and its relevance to the library community. A brief look at the MARC bibliographic format as a metadata type is followed by descriptions of the ONIX book trade standard, Dublin Core metadata, Encoded Archival Description, the Collection Description schema and the MARC21 Community Information Format. The paper concludes by considering the future possible use of each metadata type in libraries.

Introduction

What's your reaction to the word metadata? 'It's technical so I'll leave it to others.' 'Can't see why we need it.' 'I've been told we'll be using it and I haven't a clue what it's about.' So for anyone feeling confused or intimidated, this paper gives a general overview of metadata, setting out what it is, how you use it, and why and when you use it.

So what is metadata? Among the varying definitions of metadata, the simplest one is that it is 'structured data about resources'. So library catalogues are a form of metadata, as are abstracting and indexing services (structured data about articles), finding aids (structured access to archival repositories) and documentation of museum contents. Community information metadata may be a recent term, but directories of institutions, organisations and societies, and local communities have been published for many years. So it is clear that metadata has been around for a long time – but not known by that name.

However, holding metadata in electronic form is a more recent development and a variety of metadata formats have been created to hold information about all sorts of resources. Metadata formats are ways of structuring content, and this content is held by and delivered by one or more carriers. The markup languages are one set of carriers and can hold a variety of metadata formats. In contrast, MARC provides both carrier (the MARC record) and content, supporting only the MARC format.

The Markup Languages

As several of the formats described below use markup languages, I will outline some points about SGML, XML and HTML, the markup languages before moving on to the metadata formats themselves. *SGML* – Standard Generalised Markup Language – controls the formatting of electronic documents for publication, either as printed output or electronic display. *XML* – Extensible Markup Language – is sometimes described as 'next generation' SGML, with more features and flexibility. *HTML* – Hyper Text Markup Language – is a subset of SGML and controls the display of web pages (font size and type, background and text colours, use of bold and italic and page layout) and is often referred to as the 'lingua franca' of most pages on the Internet.

The markup languages share a common structure. The text of a document is divided up into a number of *elements* and *sub-elements*, each of which is named and allocated a start tag and an end tag, defined by paired angle brackets. Tag names

can be single letters , abbreviated words <li>, single words <title> or several words run together without spaces <PublisherName>. The actual data text sits between the start and end tags; e.g. *<title> Demystifying metadata </title>*. A paragraph within a document would be preceded by and followed by . The structure also contains sub-divisions of some elements, such as ordered (i.e. numbered) lists and unordered lists. Within the list element, the actual list items are repeated instances of <li> ; this is sometimes referred to as 'nesting'. For clarity, it is standard practice to use separate lines for each element and where one element contains sub-elements.

 1. Apologies
 2. Minutes of previous meeting
 3. Matters arising

The MARC format

The MARC format is a metadata format we are familiar with and so provides a useful reference point for the other formats described in this paper. The format is a structure for holding defined pieces of information about a variety of items of intellectual content (books, serials, articles, musical scores, maps, sound recordings, computer files, etc.).

The MARC format divides the data held into groups of elements: coded data (0XX fields), main entry (1XX), title, edition, and publication data (2XX), physical description (3XX), series entries (4XX), notes of various types (5XX), subject access entries (6XX), added entries (7XX), added entries for serials (8XX) and references and local fields (9XX). Numeric and alpha tags identify the fields and sub-fields: for example 245 holds title information, with $a holding title proper and $b holding sub-title. Many of the fields are of variable length but there are also a number of fixed length fields, which hold data that only exists in a specific number of characters (e.g. ISBN) or where a set of codes has been drawn up (e.g. the three letter codes for country of publication). Although the format uses numeric and alpha tags, many displays of records and input forms replace the tags with words. So an OPAC user is likely to see Author rather than 100 and Title rather than 245.

As each format is described, it will become apparent that there is a substantial core of elements which occur in all formats, supplemented by a smaller number of elements specific to a particular format. For each format, there is an illustration of an example record. Due to limitations of space, the examples are not full records and many other elements could have been included.

ONIX

ONIX is a group of product information standards being developed by the book trade primarily as a method of passing information from publishers to Internet booksellers. It was developed by Book Industry Communication (BIC) in the UK and the Book Industry Study Group in the US and funded by the Association of American Publishers. It aims to deliver rich product information: in addition to basic bibliographic data there can be reviews, abstracts and summaries, and tables of contents and it can include links to images (cover images, video and audio files about the author and/or content, sample illustrations and text, etc.). Use of the format has been slower than hoped for, but many publishers can now export records in this

format. A survey[1] has looked at possible uses of ONIX by libraries and it has been mapped to the MARC format.

The development of ONIX concentrated initially on books, with a first version available in 1999 and Release 2.0 available in 2001. The standard will be extended to progressively cover other media, and the draft *ONIX for Serials* was released in early 2002. Other draft standards in preparation are *ONIX for Subsidiary Rights* and *ONIX for Videos and DVDs*.

The ONIX standards use XML as the carrier for the format and the individual data elements each have an XML reference name and tag. ONIX information about a book is contained in a group of records within an XML file: a message header containing data about the sender and addressee, a product record and main series and sub-series records.

The records consist of a number of elements, many of which are familiar – identifiers, authors, title, edition, language, subject, audience, descriptions, publisher, dates – though the sub-elements may differ from that in a MARC record. For instance there are a number of Title elements – distinctive title, title without prefix, subtitle, translation title. Author elements hold not only the name, but also details of any professional position, institutional affiliation and biographical notes. The Subject elements hold codes for subjects taken from the BIC Standard Subject Categories; this set of terms is based on an analysis of the way subjects are typically grouped in bookshops. There are separate Date elements for announcement, publication and copyright. There are a number of additional elements of specific interest to the book trade – dimensions, suppliers, availability and promotions. While libraries may include height and number of pages, book trade dimensions data also includes weight, as this will affect carriage costs. Availability and promotions cover when the title is available, minimum order numbers, discounts, and promotional deals. And in contrast to library practice of including series information in the item record, ONIX attaches series and sub-series records to the item record.

An example ONIX record
<ISBN> 0123456789 </ISBN>
<DistinctiveTitle> Alice in Wonderland </DistinctiveTitle>
<Contributor>
<ContributorRole> Author </ContributorRole>
<PersonNameInverted> Carroll, Lewis </PersonNameInverted>
</Contributor>
<PublisherName> Collins </PublisherName>
<PublicationDate> 2000 </PublicationDate>

Dublin Core Metadata Element Set

This is more familiarly known as Dublin Core and sometimes as DC. It was developed as a means of simple resource discovery for the Internet and HTML is the most common carrier in use. It was a design principle to keep the element set small and basic Dublin Core has just fifteen elements. Some of these elements can be sub-divided by using qualifiers to make an element more specific: For example Description element qualifiers are *Description:tableofContents* and *Description:abstract*. For some elements, existing encoding schemes are used – LCSH, MeSH, DDC, UDC and LCC can all be used in the subject element. A level of obligation is set for each element, mandatory, mandatory if applicable or optional.

In practice, the basic Dublin Core element set has proved too restrictive and a number of application profiles (AP) are being developed. A Library AP is in development, with a first draft in 2001 and work on a second draft is in process. Application profile development starts by reviewing definitions of the basic elements and adding to these if necessary to give more guidance on usage within a specific domain or sector. The basic set of elements is also reviewed and additional elements may be proposed – such proposals are kept to the minimum. If a suitable additional element exists in another application profile, this is used, perhaps with additional qualifiers, to retain consistency.

The basic Dublin Core elements are Title, Creator (e.g. author), Subject, Description, Publisher, Contributor, Date, Resource Type (nature or genre of resource), Format (extent and medium), Resource Identifier (unambiguous reference such as DOI or ISBN), Source (another resource from which this one is derived), Language, Relation (is version of x, replaced by x, part of x, etc.), Coverage (spatial or temporal coverage of the resource) and Rights. As an example of qualifiers, the Date element can be Date:created, Date:valid, Date:available, Date:issued, or Date:modified.

Example Dublin Core record
<Title> Alice in Wonderland </Title>
<Creator> Lewis Carroll </Creator>
<Subject> <LCSH> Fiction </LCSH> </Subject>
<Publisher> Project Gutenberg </Publisher>
<Date> 2000 </Date>
<Format> ASCII file via FTP </Format>
<Identifier> http://prmo.net/pg/..... </Identifier>
[*This item is an electronic text file available for download from the Project Gutenberg site.*]

Encoded Archival Description

The archives domain was slower than the library community to use machine readable records and it was 1993 that saw the start of a project to develop a standard for machine readable findings aids, with version 1 released in 1998. This standard, the Encoded Archival Description, or EAD, is increasingly used by archival repositories world wide. The initial decision was to use SGML as the carrier for the format, but XML became available during the development period. It was therefore decided to make the format XML compliant so that either carrier could be used.

The archives and library domains have different approaches to documenting their collections, with libraries focusing on individual items and archives focusing on the relationship of items in a hierarchical arrangement and this is reflected in the format. Although standardizing systems and terms within the archives domain is on-going, and there is still a great deal of local practice, a set of internationally agreed terms is used in the format for the various levels.

Example of an archive repository
Repository: County Records Office
Management group: Ecclesiastical records
Management sub-group: Anglican Church records
Fonds: Parish of St. X, Somewhere (other fonds for other parishes)
Series: Baptism registers (other series Marriage registers, Burial registers, PCC minutes)
File: Baptism register 1875-92 (other files for other register volumes)
Item: a single entry in the register

While it is not possible to include a complete MARC record in EAD, some MARC elements can be embedded. Software can then be used to generate a skeletal MARC record by combining the individual elements and exported into a catalogue, where it can be enhanced with additional data.

EAD contains two segments with an optional third segment, which when present is located between the two main segments. Segment 1 (*eadheader*) contains much of the bibliographic information about a finding aid (e.g. a catalogue, a list, an inventory, etc.) and is held in a prescribed order. The optional segment (*frontmatter*) allows an alternative presentation of the eadheader data to suit local practice. Segment 2 (*archdesc*) contains information about the body of archival materials described by the finding aid.

The eadheader is divided into *eadid*, a unique identifier for the finding aid (e.g. ISBN if a published catalogue), *filedesc*, which is further subdivided to hold details about the author, title, sub-title, edition and publisher, and *profiledesc*, which includes details such as the language of the finding aid.

The archival description (archdesc) is divided into *level* (management group, fonds, series, etc.), *did* (descriptive identification – physical description, location, etc.), *admin info* (scope of collection, arrangement, biographical notes), *dsc* (description of subordinate components), *add* (adjunct descriptive data), *odd* (other descriptive data), and *controlled access headings*. When MARC elements are included they are identified by the use of the tag 'encoding analog'.

Example EAD record

```
<ead>
<eadheader>
<eadid> 0123456789 </eadid>
<filedesc>
        <titlestmt>
                <titleproper> Pitman Shorthand Collection Catalogue </titleproper>
                <author> Ann Chapman </author>
        </titlestmt>
        <publicationstmt>
                <date> 1990 </date>
                <publisher> Bath University Library </publisher>
        </publicationstmt>
</filedesc>
</eadheader>
<archdesc> collection
        <did>
                <abstract> A collection of materials in and about shorthand collected
                by Sir Isaac Pitman </abstract>
        </did>
        <controlaccess> <subject encodinganalog="MARC650"> Shorthand
        </controlaccess>
</archdesc>
</ead>
```

Collection Description

While the archival domain has always provided information about collections, libraries have concentrated on providing access to items. However, recent

developments have meant that libraries are now paying more attention to identifying and describing collections. Firstly, there are increasing implementations of the Z39.50 protocol that enables a user to cross-search a number of collections. The search is more effective and faster if search targets are restricted to a small number of collections that are most likely to hold relevant materials. Secondly, the Full Disclosure initiative has drawn up a national strategy for retrospective conversion and retrospective cataloguing. In order to achieve this in the most effective way, it is important to identify which collections do not have machine-readable records, and the priority level for that collection, in terms of international, national, regional, or local significance in specific subject areas and material types. Finally, a wider range of individuals are seeking information on collections of materials for a range of purposes – work, study and leisure interests.

In May 2002, a schema – or set of elements – to describe collections was developed by UKOLN. So far there have been three implementations of the schema. In summer 2001 UKOLN used these elements in an Access database format for the Research Support Libraries Programme (RSLP) as an administrative support for programme projects, which often included retrospective cataloguing work. Then in spring 2002 UKOLN worked with Samsara Research to produce a web-based version using an SQL database for the Reveal project. The Cornucopia database, developed by System Simulation, for museums collections is also based on the schema and uses an SQL database.

In the collection description schema the elements are termed attributes. These attributes can be divided into five groups: general attributes, subjects, dates, associated agents and external relationships. While the RSLP version uses the schema as originally devised, both Cornucopia and Reveal have added a small number of extra attributes.

General attributes are title, identifier, description, strength, physical characteristics, language, type, access control, accrual status, legal status, custodial history, notes and location. Some of these terms will be unfamiliar. *Strength* indicates subject depth and stock policy; in the Conspectus collection assessment tool Music 1 would be allocated to a library containing minimal stock in this area while Music 4 would indicate a collection containing research level material, with older material retained for historical research and appropriate new materials acquired. Attributes within the *type* element identify type of collection, curatorial environment, content and policy/usage. *Accrual status* indicates whether a collection is still being added to (e.g. open or closed) and by what method (e.g. purchase, donation).

Subject attributes allow collection records to be indexed for searching. Available attributes are concept (i.e. topic), object, name, place and time. Classification numbers, subject headings and keywords can all be used to record the subject attributes.

Two types of date information can be recorded in the date attributes. The first date attribute is accumulation – the period over which a collection has been built up. The second attribute is contents – the time period within which all the materials in the collection were created or published.

Agent attributes record information (name and contact details) relating to people or corporate bodies associated with the collection in a variety of roles – creator of the collection, owner of the collection or administrator of the collection.

The final set of attributes is for relationships between the collection being described and other resources. A collection might be divided into a number of sub-collections. Additionally there may be catalogues, indexes and finding aids for the collection, and associated collections or publications. As most of these resources would qualify for collection description records in their own right, it is possible to link the related records. If the records are held in a web-based implementation it is possible to include URLs as links, allowing the user to move, for example, from the collection record to an OPAC.

Example Collection Description record
Title: Pitman Collection
Strength: Shorthand – national collection
Phys. Desc: Printed texts and manuscripts
Lang: English, Spanish, Esperanto, ...
Access: Written request to the Librarian, University of Bath
Accrual: open, passive, deposit
Location: The Library, University of Bath
Subject: Shorthand
Subject: Sir Isaac Pitman
Owner: Pitman Publishing Company
Catalogue: University of Bath OPAC

MARC21 Community Information Format

Not only do individuals want to find out about items in a range of media, the collections that hold them, and the catalogues and finding aids that help access them, they also want to find information about their local community. While the phrase 'local community' is most likely to mean the geographical area in which you live, and the organisations and activities within that area, it can also be applied to organisations. So these records could describe the different parts and activities of a school, college or university, a large firm, a scientific research body, a hospital trust, a local authority or a professional body.

This metadata format brings us back full circle in that it is part of the MARC21 family of formats and many features will be familiar. As with other MARC21 formats, a Community Information format record contains a leader, fixed fields and variable fields and the same basic structure of numeric and alpha tags.

The leader section of the record identifies whether the record is describing an *individual* (perhaps the local MP or the Chief Executive of a firm), an *organisation*, a *programme* (a lecture series, literacy classes or a festival), an *event* (a public meeting or a workshop) or '*other*' (this is usually some sort of facility, e.g. a swimming pool).

The fixed fields contain a variety of coded information. Field 007 contains information about disability facilities – provision of ramps, lifts, specialist equipment, staff help and if staff can use sign language. Field 008 has information on special aspects – is childcare available for a programme, are there opportunities to act as a volunteer or can you be supported by a volunteer.

Variable fields show obvious similarities to fields in the bibliographic format. The 1XX fields are for name of society, organisation, individual, programme, etc. The 2XX fields are for title, address, telephone and email details and opening hours, while 3XX holds physical description data including details of rooms and equipment for hire. Series data – mostly used for events – is held in 4XX and notes are held in 5XX, this

includes details of publications issued. The 6XX fields are for subject access, 7XX are for added entries and 8XX for other variable fields, including URL's for web pages.

Example Community Information record
110 $a CILIP
245 $a CILIP HQ
247 $a LA HQ ***$f*** 19?? – 2002
270 $a 7 Ridgmount St., London, WC1E 7AE ***$k*** 020 7255 0505 ***$m*** info@cilip.org.uk ***$r*** 9am to 6pm
311 $a Ewart Room ***$d*** seats 50 ***$g*** £100 per day
312 $a Overhead projector ***$f*** £10 per day
581 $a Library + Information Update
856 $a http://www.cilip.org.uk/

Fit for purpose

This brief trip around the metadata world has concentrated on MARC Bibliographic, ONIX, Dublin Core, EAD, Collection Description and Community Information. These are not the only formats but others are focused on more specialist areas. Some information could be held in more than one format: a web resource could have Dublin Core metadata and an MARC record. So what's the future for libraries?

I think it likely that MARC will be around for some time yet, given the large proportion of libraries using it and the fact that the format has a structure and procedure for continuous assessment and development in MARBI. However, it is possible that the format might move to using other carriers. This might be putting MARC records into XML wrappers, or mapping the MARC tags to XML tags and using XML as the carrier. The likelihood is that movement here will be linked to development in library management systems and how they manage MARC data.

It is difficult to predict the actual use of ONIX by the book trade. In the short term, it is most likely that libraries will encounter ONIX when acquiring electronic publications – you may be offered an ONIX record along with the title (requiring conversion into MARC for you to use it) or the publisher may generate a MARC record from the ONIX record. If ONIX records come into general use, acquisitions departments may be faced with the need to convert records, or library suppliers may offer such a service. At present there is no advantage to libraries to moving to ONIX records for cataloguing purposes.

Some search engines do use Dublin Core to index sites. These tend to be specialist applications looking for high quality sites. Libraries and their parent institutions and organisations should be considering adding DC metadata to the header part of HTML pages, and using the metadata in the search options offered on their sites.

How much libraries will come into contact with EAD depends on the closeness of the relationship between a library and any archives either locally or part of its parent body. Embedding MARC elements does provide a way of collecting basic data for a MARC record for finding lists, catalogues and inventories of an archive, which can then be added to the library catalogue.

The use of Collection Description records is likely to increase in the future. You may be asked to contribute records for your library to a multi-partner project or initiative. Or large libraries with many special collections may consider using such records themselves.

The use of Community Information records is also likely to increase, as libraries move to MARC21 format and, where supported by suppliers, can use the Community Information format. A database of such records could be a useful service for the library to propose to its parent body as a way of supporting the body both internally and as a way of outreach.

The one thing that does seem certain is that libraries will need to become familiar with more than just MARC. It is likely that each library will some of the other metadata formats, but which ones and for which purpose will depend on their requirements for their particular user base. But if you can work in MARC format, you too can learn to use the other metadata formats, since you are already familiar with the concepts and principles behind metadata.

References

1. Burton, Celia / *Investigation into the feasibility of using ONIX International as a standard for bibliographic data transmission between the book trade and libraries in the UK*. Prepared for BIC with funding from the British National Bibliography research Fund, 2001.

Glossary

DDC: Dewey Decimal Classification
DOI: Digital Object Identifier
FTP: File Transfer Protocol
Fonds: a single collection or archive
LCC: Library of Congress Classification
LCSH: Library of Congress Subject Headings
MeSH: Medical Subject Headings
PCC: Parochial Church Council
SQL: Standard Query Language
UDC: Universal Decimal Classification
URL: Uniform Resource Locator – the address of a web page

Web URLs

Community Information at London Borough of Barnet libraries
http://www.libraries.barnet.gov.uk/
Click on Search Web OPAC
Guest login
Choose InfoLINK Community Information
Try 'scouts' 'painting' 'arts'
Conspectus
http://www.rlg.org/conspechist.html
Cornucopia
http://www.cornucopia.org.uk/
Dublin Core
http://dublincore.org/
Element set: http://dublincore.org/documents/dces/
Qualifiers: http://dublincore.org/documents/dcmes-qualifiers/
Reveal – expected launch in autumn 2002
Will be hosted on National Library for the Blind site http://www.nlbuk.org/

This paper is based on a presentation given at the CIG conference, Newcastle, 16-17 April 2002, and repeated at staff development sessions: 15th May 2002 for University of Bath with the University of West of England and 2nd July 2002 for University of Westminster.

Ann Chapman is Research Officer for Bibliographic Management at UKOLN.