“Past present and future: developments in British Library cataloguing"

By Chris Martyn (British Library System Projects Co-ordinator)

With contribution by Alan Danskin (British Library Bibliographic & Metadata Standards Co-ordinator)

The British Library

The British Library (BL) is only 36 years old. It was created by the combination of a number of existing organisations:

- British Museum Department of Printed Books
- National Lending Library for Science and Technology
- India Office Library and Records
- National Reference Library of Science and Invention
- National Central Library
- British National Bibliography
- National Newspaper Library
- British Institute of Recorded Sound.

Each of these organisations had (and for a long time retained) its own identity and its own collections, catalogues and systems. Indeed, it is only since the physical occupation of the site in St. Pancras, London, in 1997 that a corporate identity has really begun to emerge. The physical integration of the library anticipated the integration of the collections and the systems that enable collection management and resource discovery.

Integration of data

Before the integration of mainstream print material in English and Western European languages by the late 1990s, collections did their own cataloguing, to their own individual standards, on their own dedicated systems. Additionally, brief bibliographic records were created in bespoke non-MARC formats on dedicated acquisitions systems. Bibliographic records therefore exhibited a high degree of diversity, e.g. collection-specific data, different standards applied, historical variances (e.g. many pre-AACR records). There were 25 separate locations for bibliographic data (acquisitions systems, cataloguing systems, products, BL catalogues), which in the main were stand-alone systems.

The British Library implemented the Aleph Integrated Library System (ILS) in 2004. It replaced the various legacy systems for mainstream acquisitions and cataloguing and the OPAC. As part of the data migration to the Aleph ILS legacy records were merged as far as possible and data cleaning was carried

out. 75.3 records were migrated in total and 12 million records were converted to MARC21 from UKMARC.

The main British Library bibliographic database remains a complex environment, containing records for:

- Different media: maps, music, print, electronic, microform
- Catalogues: pre-1975; Science; Document Supply; Asia Pacific African – data within these extremely varied
- Products & services: British National Bibliography (BNB), ISSN UK Centre
- Purposes: acquisitions; to state that the BL has decided not to select a work.

There is still much duplication of legacy records and issues with data, which are being addressed by a variety of means.

Trends and the BL response

The main long-term trends from 1990 onwards which have affected British Library cataloguing have included:

- Increasing intake through legal deposit (2008-09 was the highest-ever recorded intake of legal deposit monographs)
- Declining cataloguing resources
- Distributed cataloguing i.e. re-use of bibliographic records wherever possible, sharing our data
- A gradual move to e-media (though there has not yet been a corresponding decline in intake of print monographs – the opposite is true)

E-media material catalogued by the BL currently includes covers voluntary legal deposit and some digitised collection items such as some antiquarian books. Full legal deposit of e-media (eJournals and newspapers in the first instance) will require new technologies, workflows and additional resource. These are under investigation.

In addition to implementing an integrated library system the BL has managed its response to these trends by moving to international standards so as to increase cataloguing efficiency. It is recognised that harmonisation is key to achieving this:

- UKMARC, the UK national format which the BL used prior to 2004, and legacy BL systems hindered effective data interchange

- The move to MARC21 and the subsequent move to a full NACO (name authority records) and SACO (subject authority records) contribution have greatly increased interoperability of data and the potential for deriving records from elsewhere. It is difficult to determine exactly, but BL original cataloguing now accounts for less than half of all bibliographic cataloguing
- Since 2004 the BL has conformed to the Library of Congress Rule Interpretations
- The BL can now consider using German/Austrian records as they are now in MARC21, although policy discrepancies require investigation.

Data interchange

Effective data interchange is of critical importance in an era of declining resources and increasing intake. Examples of how the BL re-uses and distributes data include:

- The creation of bibliographic records for legal deposit monographs, which is distributed among the BL and the other 5 legal deposit libraries in the UK and Ireland
- Pre-publication records for UK-imprint monographs created by an outside agency, which are loaded to facilitate claiming
- Processing staff at the BL use Library of Congress records and records obtained via Z39:50 wherever possible and make appropriate amendments to save cataloguing effort. The Library of Congress is the primary external source for monograph records
- LC records are used for automated batch upgrading of pre-publication records
- The BL has several national library databases as Z39:50 targets (e.g. Hungarian National Library) but has been unable to use German/Austrian records in the same way in the past due to the format differences (MARC/MAB). The BL can now consider using these as they are now in MARC21, although policy discrepancies require investigation
- The BL extracts data from its main bibliographic database for various products (e.g. BNB).
- Institutions/utilities receive datasets (e.g. BNB) extracted from the BL database, e.g. Blackwell, OCLC. Subscription model in some cases
- Users can download individual MARC records from the OPAC
- The BL will be offering full free download access to MARC records via Z39:50
- The BL provides access to the ESTC (English Short Title Catalogue) union catalogue via the BL website. Data is exchanged daily with the US agency in California.

Record content

In current processing there are different cataloguing levels depending on the category of material. The emphasis is on data content rather than the quantity of fields. The record content can be very different depending where in the workflow the record is. For example acquisitions records can be very sparse, containing only control number, title, publisher, author, and minimal fixed-field info, while a full-level record must contain full range of notes, authority-controlled headings, and subject headings (including

Dewey 22 classification for UK imprint and science material). By the end of the workflow the record will have been upgraded to the appropriate level, manually or automatically.

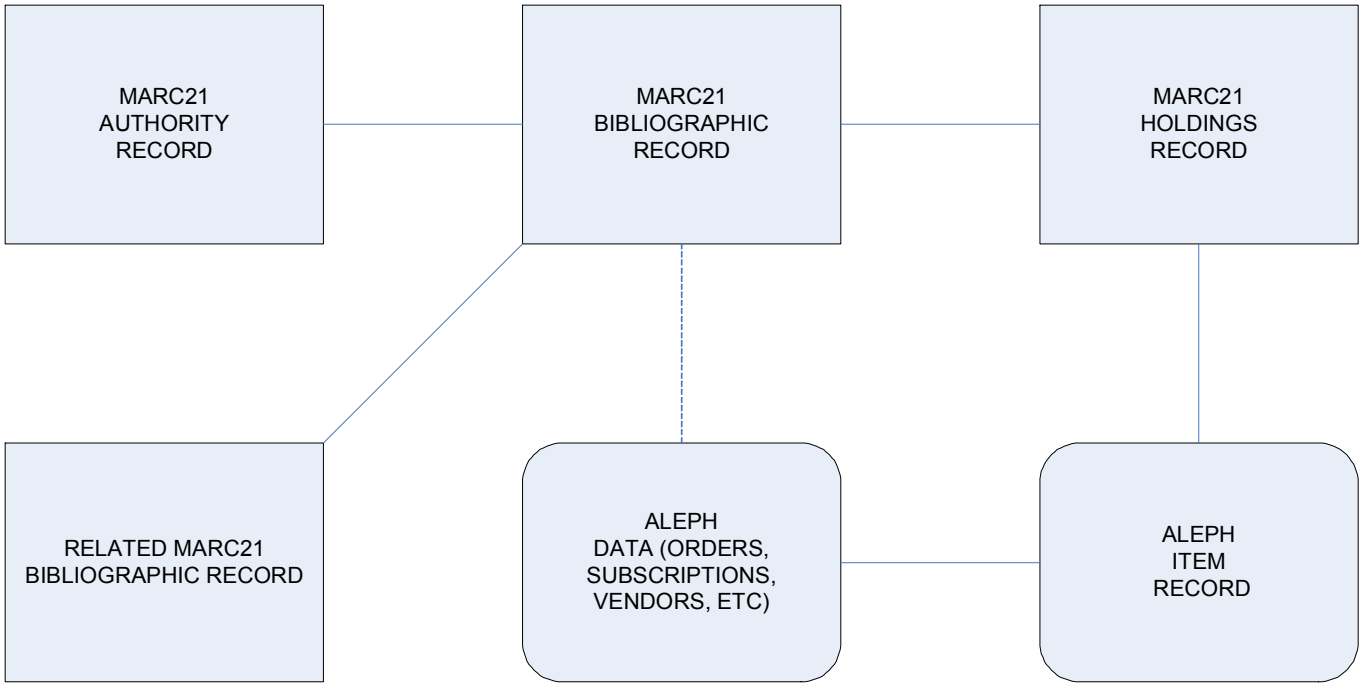
Most records must eventually contain Library of Congress Subject Headings (LCSH) and authority control. The definition of the BL Standard Record can be found here: http://www.bl.uk/bibliographic/catstandards.html.

Additionally, the BL has many local requirements which cannot be accommodated in MARC, such as copyright information and some data elements in legacy data. This information is held in local fields. The BL also includes value-added data (which is not relevant for itself) such as the British National Bibliography (BNB) number for the BNB product. Additional data is added to the holdings record if the item has unique characteristics (such as being damaged, or having been signed by the author), in accordance with the new British Library policy for Copy-Specific information. This information is visible to the public via the OPAC.

What content records should have has been the subject of a number of surveys over the years. When asked, users have wanted more information rather than less. These studies have informed decisions on what data to include in BL records. Statistics for the number of searches on the BL OPAC also show that adequate record content is important to users. The most popular searches in 2007-08 in descending order were: keyword, ISBN, year, title, shelfmark, publisher, subject, format. Adequate record content is also required for exchange of data; there must be enough metadata present to be able to make an authoritative match.

Aleph data structure and record linking

Linking between records of various types is fundamental to Aleph functionality, and to properly express the work as held by the BL. In Aleph the MARC bibliographic record is at the heart of the data structure. MARC authority and holdings records are linked to it, as are the various non-MARC Aleph records which contain the administrative data associated it (e.g. orders, subscriptions, items which represent physical copies). The links between these records can be expressed as follows:



The different types of data (bibliographic, authority, holdings, authority, Aleph) are held in dedicated Aleph databases. All records are linked via control number except for authority records. In this case the link is via the heading - if the heading is in the same form, the link is made to the bibliographic record. This is the way Aleph operates. Therefore, the BL does not use the $0 authority control number link which was implemented in MARC21 for the German and Austrian conversion to that standard.

Authority records aside, Aleph links MARC records via a dedicated Aleph field (LKR). MARC linking fields (e.g. holdings 004, bibliographic 7xx) have no functionality in Aleph, but must be present as records are exported.

Links between records (e.g. related bibliographic records) are kept relatively simple in the BL implementation of Aleph. Related serial records (e.g. continuations) are related, and the BL Map Library creates linked analytic records representing images within a work, but records for multivolume works and monographs in series are not generally linked, either in MARC terms or using the Aleph linking functionality.

From the present to the future

New cataloguing environment

There is now a totally new context for resource discovery. This environment continues to evolve to be more and more Web based. The range of information carrier which needs to be catalogued is much wider. The depth and complexity of the content of the resources being catalogued has increased. Metadata is now created by a wider range of personnel: not only by skilled professional cataloguers, but by support staff, non-library staff, and also publishers - who have a wider range of skill levels. Some are

using structures other than the MARC format for records – like using Dublin Core for some resources. Although the British Library's main metadata format remains MARC21 and its cataloguing rules AACR2, some categories of material have remained outside these standards, for example manuscripts and sound recordings. Metadata schemas new to the BL, such as MODS and METS are beginning to be employed to create records.

We also now have access to descriptive data for resources in digital form (for example, descriptive data for books available from many publishers in ONIX, publisher-supplied metadata for eJournals in a variety of schemas) – this is information we can capture for our bibliographic records. These new metadata schemas have explicit links to the digital content they describe, facilitating resource discovery of content

Much of this metadata is created with purposes other than resource discovery in mind. When libraries reuse such metadata they are really repurposing it. We need tools that enable this to be done as efficiently as possible, and for digital materials to enable users to retrieve the content the metadata describes.

To support this new environment, a number of new developments are either being planned for or are already being incorporated in the BL. In terms of the cataloguing standard, Resource Description and Access (RDA) is the proposed replacement for the Anglo-American cataloguing Rules (AACR2). The BL and other English-speaking national libraries responsible for developing RDA plan to test and evaluate RDA during 2010, with a view to implementing RDA in 2011 (at the earliest).

The Digital Library is a core part of the BL vision for the future. The infrastructure to support this is currently being developed. New XML-based metadata standards for digital materials are being actively investigated and in some cases already being used, including:

- Metadata Encoding & Transmission Standard (METS) - for management, access and use of digital materials
- METS, Metadata Object Description Schema (MODS) and other XML-based schema - ingest of e-materials and data interchange.

The BL is already using MODS and METS for some categories of digital material, e.g. eJournals. The BL has already mapped relevant MARC21 data elements to MODS/METS - for eJournals, Aleph data in MARC21 is used as the basis for the MODS record. The proposed eJournals workflow will involve converting publisher-submitted metadata to BL standards, using the OmniMark content processing language. New systems are being developed to work alongside (and in some cases integrated with) Aleph, e.g. e-ingest and digital preservation systems.

Although there are some issues with MARC in this new environment, the BL may not move completely away from MARC21. Although it may move to an XML-based format (not yet specified) for storage, input and display, it may still be used for record exchange. It is unlikely that all the BL's customers will abandon MARC21.

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is being used to facilitate data harvesting, for example to harvest data for UKPMC (UK PubMed Central), and from the BL's web archive to the Primo resource discovery interface. Its use will no doubt be expanded.

The BL is also investigating alternatives to the traditional OPAC for resource discovery. User expectations are shaped more by experience of Google and Amazon than cataloguing practices. The resource discovery interface needs to exploit the underlying syndetic strength of the catalogue to deliver users the results they want through a clear and understandable interface. The BL's test interface, Primo, is discussed later.

Resource Description and Access (RDA)

In tandem with these developments in encoding metadata are developments in the cataloguing rules used to create metadata. The aim is to simplify the cataloguing code and to establish it as a content standard for resource description for various metadata schema, and to encourage its use worldwide.

A new standard is needed:

- that will be more consistent across the various types of content and media, and that demonstrates the commonalities of different types of resources
- to address current problems with rules in AACR2, such as with GMDs (general material designators) and for cataloguing digital materials
- to change the approach to cataloguing, to get back to more principle-based rules that build cataloger's judgment and are simple to use
- that will encourage the application of the FRBR (Functional Requirements for Bibliographic records) and FRAD (Functional Requirements for Authority Data) data models.

While developed for use in English language communities, RDA can also be used in other language communities – it is anticipated that other countries will translate it and adjust its instructions to follow preferred language and script conventions – just as there are now many translations of AACR2. Options are also available to allow for use of other languages and scripts, other calendars, other numeric systems, etc.

RDA implementation scenarios

RDA has three implementation scenarios:

Scenario 1

- RDA data stored in relational or object-oriented database structure that mirrors the FRBR and FRAD conceptual models
- Descriptive data elements stored in records that parallel the primary entities in the FRBR model: work records, expression records, manifestation records, item records
- Data elements used for access point control are stored in records that are centred on the primary entities in the FRAD model: persons, families, corporate bodies, etc.
- Relationships between primary FRBR entities reflected through links from one record to another

Scenario 2

- Bibliographic and authority files linked. Bibliographic record also contains links to authority records for persons, families, corporate bodies, etc., associated with the work, etc., embodied in the resource described

Scenario 3

- Bibliographic and authority files not linked. Access points using the preferred name or title for the person, etc., stored in the bibliographic record along with the descriptive data

In Scenarios 2 and 3:

- RDA data stored in database structures conventionally used in library applications
- Data stored in bibliographic, authority, holdings records. Additionally in Scenario 2, the bibliographic record also contains links to authority records for persons, families, corporate bodies, etc., associated with the work, etc., embodied in the resource described
- Descriptive data elements stored in bibliographic records
- Variant names and other data used for access point control stored in authority records

Scenario 1 is the most different from current record structures, as it requires that different records express different aspects of the item being described (e.g. the work, the expression, the manifestation). These separate records are brought together “on the fly” to provide the description of the work. There are no bibliographic records as we know them.

RDA and library systems

Liaison with system vendors on how RDA can be supported and/or integrated is already happening. Aleph already accommodates Scenario 2, which the BL plans to implement. BL systems will be changed to accommodate new MARC fields/subfields and GMDs. With Scenario 2 there will also be

retroconversion issues, for example changing “Dept” to “Department” and Bible New and Old Testament (from “N.T” and “O.T” to New and Old Testament respectively). However, the links between records which are required for this scenario already exist in Aleph. In the first instance, the BL plans to have an implementation environment within Aleph, including all new MARC coding, to test workflows and system aspects.

To move to Scenario 1 would require an enormous amount of retrospective amendment to bring the BL’s existing records into line, and would require wholesale changes to current cataloguing.

Resource discovery

Effective resource discovery depends on adequate metadata. It is not possible to discover all BL resources through the Web OPAC, as the ILS does not contain metadata for all BL collections. The metadata is also not granular enough to provide access to certain items, for example article-level access is not possible. OPACs are relatively user-unfriendly (e.g. the BL’s data is difficult to interpret and interface has issues). Users expect a “google-like” search experience, wherein keyword searching via a simple interface retrieves relevance-ranked results independent of geographical location (i.e. federated searching). The BL is presently attempting to achieve this by means of the Ex Libris Primo product, which is available on the BL website as a beta test resource.

Primo will provide access in an integrated way via a “google-like” interface to those collections already on Aleph and to collections which are not able to be represented on Aleph (i.e. not in MARC), for example manuscripts on the IAMS (Integrated Archives and Manuscripts System) and recorded sound, and will facilitate the retrieval and display of materials not previously accessible via the Web OPAC (e.g. full text of journal articles). Data is harvested from different systems (e.g. Sound Archive, BL catalogue, IAMS), in some cases using OAI-PMH. It is also planned to use Primo to facilitate resource discovery of digital objects themselves such as digitised newspapers and books, and eJournal articles. The fact that Primo is FRBR-ised means that it can display related items in a more user-friendly way than the OPAC (subject to the constraints of BL data). Primo also supports such Web 2.0 functions as user reviews and user tagging.

The BL is also currently working through the implications of searching across its catalogues, full text of digital items, the BL website, and other resources. This is a huge undertaking, but will potentially give access to BL resources in an integrated way.

The general trend is towards federated searching. The BL offers this to a limited extent via Primo, and WorldCat provides access to resources across many library catalogues. However, there is evidence to suggest that people will search Google for a resource and not a library catalogue at all, even if they are in a library! How can libraries ensure their collections are discoverable? There are interesting times ahead!