Abstract

In October 2015 the British Library published its Collection Metadata Strategy. This paper reviews the reasons for developing a metadata strategy. The paper will focus on the challenges (familiar to most libraries) that have prompted the British Library to articulate a holistic vision of its collection metadata as a key corporate asset, including:

- The extension of legal deposit to non-print media and the growth of bulk digitisation have created demands for more metadata and more flexible metadata. Metadata is no longer only for discovery, it is also critically important for the automation of workflows and processes.
- The switch of content from physical to digital carriers challenges distinctions which are embedded in our workflows and the metadata standards we use.
- Changing government policy and user expectations that challenge traditional business models for dissemination and re-use of metadata.

The paper will include an update on progress achieved so far and consider the lessons learned to date.

Introduction

The British Library is the National Library of the United Kingdom. It is one of 5 Legal Deposit libraries. There is a legal requirement for publishers to deposit publications with the library within one month of their publication; since 2013 the library has been empowered to receive non-print publications, including digital resources and to archive the UK Web domain. The Library was created by Act of Parliament in 1973 by merging several existing libraries and agencies, including the British Museum Library, the British National Bibliography, the National Lending Library, the Patent Office Library, the India Office Library and Records, the National Sound Archive and most recently the Public Lending Right. Each of these institutions had its own collections, housed in its own building, curated by its own staff and with its own catalogues and databases. With the passage of time, most of these historic silos have been eroded and the Library’s assets, including the collection, buildings and staff, brought under centralised management.

Why a collection metadata strategy?

Convergence

The implementation of Aleph, the Library’s first true Integrated Library Management System, in 2004, resulted in the migration of approximately 13 million bibliographic records from multiple catalogues and bibliographies into a single production data base. However this undertaking only integrated metadata for published resources, including books, sheet music, maps and serials. Metadata describing manuscripts and archives, sound recordings and journal articles, which now total well in excess of 40 million records, were not in scope for Aleph and continue to be managed in separate repositories.

- Aleph 500 ILS for published resources
- ETOC for articles (In-house database)
- IAMS for Manuscripts and archives (in-house Integrated Archives and Manuscripts System)
- SAMI for Sound and Moving Image (SirsiDynix Symphony ILS)

Each of these repositories constitutes a distinct silo, with its own standards, workflows and associated staff. From an operational perspective this is inefficient and inflexible; for example, staff cannot be easily transferred between workflows to relieve short term bottlenecks.
More seriously, this leads to fragmented services, including a complex discovery layer, which in turn makes it difficult for users and staff to understand the collection and find the resources or information that they need. Many of the platforms we use to create, store and disseminate our metadata are reaching end of contract or end of life and may need to be replaced. So there is an opportunity to take a more holistic view of the functions our metadata needs and develop a more coherent architecture.

Hidden metadata=hidden collections

The user may also assume that everything we have is recorded in our catalogues. This is not the case. There remain substantial pockets of “hidden metadata”. Although most of our catalogues of publications have been converted to MARC there are still about 2 million catalogue records to be converted, split more or less evenly between publications and archival/Mss collections. In addition, there are substantial backlogs of uncatalogued materials.

Changing expectations

Our legacy metadata is being asked to do things that it was not designed to do. For example, when Panizzi began work to create the British Museum catalogue, he did not think it necessary (in most cases) to record the language of the content, because this would be self-evident from the description. When these records were transcribed and retrospectively converted to MARC there was therefore no information to put in the MARC 21 008/35-37 language bytes. Consequently, if someone wants to know how many books the Library has in Arabic or wishes to create a visualisation of the collection by language or create a data set of records for resources in European languages, comprehensive answers cannot be provided.
The preceding examples are also indicative of the growing significance of collection metadata as data: raw material for new research or management information. The Library has traditionally provided bibliographic records to libraries, publishers and other third parties in MARC but the technical difficulties of using MARC data put them beyond the reach of all but the most committed researchers. There is also increased interest in sharing metadata for other types of resources, including archival collections and manuscripts. Central government has become increasingly interested in the economic potential of digital information, including metadata, and public bodies are encouraged to open up their data for reuse.

The digitization of content erodes many of the practical justifications for historic distinctions between collections and users reasonably expect to be able to search holistically, regardless of content or carrier types.

In 2013 legal deposit was extended to non-print media. Non-print media include digital books and journals, but also the UK web domain. Processing digital resources creates new challenges of scale, granularity and volatility. Our expectations were that as publishers were switched from print to electronic legal deposit the increase in the intake of electronic resources would be balanced by a reduction of the print. In fact, there has been little fall off in print but even with only a small number of publishers switching to digital, the digital intake for 8 months was equivalent to about half of the print intake for a whole year.

It is clear that once the digital tap is open, pretty much everything the publishers issue will flow through it, including back catalogue:
From a national collection development perspective this is a good thing; from a collection processing perspective a 50% increase in intake within 8 months poses an interesting challenge.

**What is the collection metadata strategy?**

Collection metadata is defined as structured data that supports collection management, preservation and accessibility. Metadata recorded by the Library for other purposes is out of scope. Collection metadata should be managed as an asset, like the collections or the library staff and estate.
The strategy provides us with a framework to manage that asset better and to unlock more value from it for our users, our staff and our stakeholders.

“Our vision is that by 2020 the Library’s collection metadata assets will be comprehensive, coherent, authoritative and sustainable, enabling their full value to be unlocked for improved content management, greater collaboration and wider use of the collection.”

If we unpack this further:
- **comprehensive** means that we should have metadata for everything in the collection;
- **coherent** means that silos will be broken down to provide the ability to search across the whole collection;
- **authoritative** means that the metadata we create is accurate and trustworthy;
- **sustainable** means that we are looking for long term solutions to long standing problems quick wins.

To realise these aspirations we set the following broad objectives

- Drive efficiencies in the creation, management and exploitation of collection metadata to support delivery of the Library’s strategic priorities and programmes
- Improve the Library’s return on investment in its collection metadata assets by ensuring their long term value is maintained for future activities
- Open up more of the Library’s collection metadata to improve access to Library content and promote wider re-use

To manage delivery of the objectives we set out a five year plan, which is flexible enough to respond to changing needs and priorities. The plan identifies the strategic priorities under each of these broad objectives and sets milestones and targets for delivery.

![Figure 5: Spider diagram illustrating strategic plan, 2015-2020](image-url)
In addition to the deliverables identified by the plan, there are performance indicators to measure progress. The strategy is aligned with the British Library’s corporate strategy, Living Knowledge and collection metadata is fundamental to realising its aspirations.

How did we create the strategy?

Historically, responsibility for the Library’s collection metadata has been diffuse, with authority distributed across different collections products and services. A key element in the strategy has been to develop the Collection Metadata team, headed by Neil Wilson, as the acknowledged point of responsibility. This builds on the team’s well-established provision of MARC based bibliographic services by expanding its scope to responsibilities to include archives and sound recordings.

Each of these teams has a clearly defined area of responsibility, but they also work closely together to deliver solutions to problems. The Collection Metadata team is responsible for delivery of the strategy. A parallel structure has been put in place to provide strategic oversight and expert advice.
This means that collection metadata has a champion on the Library’s Strategic Leadership Team. Among SLT’s responsibilities is the review of new business cases, all of which now have to explicitly consider metadata requirements and impacts, thus ensuring that Collection Metadata is able to influence new initiatives. The structure also provides formal channels through which internal stakeholders are able to input to the development and delivery of the strategy and provides enhanced channels for communication.

Figure 7: Internal Stakeholders

The strategy was has been disseminated internally through a comprehensive communication plan. General information has been communicated through staff talks and our internal newsletter. The Collection Metadata wiki provides guidance about the services we offer and how to contact us. More focused information has been provided by engaging with specific stakeholders to develop policies and procedural documentation. The strategy is available to external stakeholders on our Website.

Figure 8: External Stakeholders
Where are we now?

Library’s fiscal and planning year ends on 31st of March, The objectives for 2014/15 were completed and we are still on track, approaching the midpoint of 2015/16. The following progress summary is organized according to the priorities and is not intended to be comprehensive.

Strategic Priority 1: Drive efficiencies in the creation, management and exploitation of collection metadata to support delivery of the British Library’s strategic priorities and programmes

- **New e-publisher metadata assessment & ingest processes**
  * Publisher and aggregator data is highly variable, even if it is received in a standard format such as ONIX. Collection Metadata has developed processes to assess the quality of metadata so that this can be taken into account when prioritising publishers for transition to e-legal deposit. Ideally, the publisher’s metadata will be compatible with the generic transformation. If not, a bespoke transformation may have to be developed using XSLT to convert the metadata provided from ONIX, Dublin Core, or csv, to MARC 21 or MODS.

- **Automated enhancement processes for e-books and Western European Languages**
  * Our long-standing “batch upgrade” process to match Cataloguing in Publication records with full records without cataloguer intervention has been adapted to create new automated workflows for WEL material (for which no CIP data is usually available) and e-books. This is still being tested, but we expect the processes to go live by November 2016.

- **Spread sheet data capture & crowd source workflows**
  * We have helped develop tools to lower barriers to productivity of staff on fixed term contracts. These include a spread sheet based tool to guide input and support selection of authorised access points from NACO and FAST and XSLT transformations. This reduces the training overhead, because staff do not have to learn to use Aleph or other complex systems.

- **FAST consultation**
  * We have consulted on the possibility of using FAST more extensively in future as a potentially more efficient indexing system than LCSH.

Strategic Priority 2: Improve the Library’s return on investment in its collection metadata assets by ensuring their long term value is maintained for future activities

- **Documentation**
  * It is essential that we have a clear picture of our metadata assets. We have created registers of existing assets, including hidden metadata, to ensure that information about the metadata is accurate and current. This supports tracking of progress and other management information.

- **Metadata impact assessment for business programmes**
  * To manage workload metadata impact assessments have been built into the Library’s business planning process and we have recently instituted a work request system so that we can plan priorities and allocate resources more effectively.

- **Data enhancement**
  * Enhancement of the quality of metadata will be essential to realise the aspirations the library has to improve resource discover. Projects are underway to extend coverage of DDC and to improve identification of Works.
Strategic Priority 3: Open up more of the Library’s collection metadata to improve access to Library content and promote wider re-use

- **Free data**
  - In recent years the Library has sought to remove barriers to access to our data by exposing MARC data via Z39.50 and removing licence restrictions from non-MARC data.
  - The Z39.50 service is being used by more than 1550 users in 177 countries. It is particularly valuable to professionals in singleton posts and small institutions, such as school libraries.
- **Development of .csv “researcher format”**
  - We recognised that the formats in which metadata was available were themselves a barrier to researchers, who may not have the tools or expertise to interpret MARC or RDF. In 2015 we began to make metadata available in the so-called researcher format. The records can be imported from our downloads page in .csv (comma separated value) format. This format enables manipulation of the data using desktop applications, such as Excel or Open Refine. At present, the datasets are selected by collection metadata staff but we are testing a tool that would allow researcher to define their own datasets.
  - We have developed the capability to integrate data from both the printed collections and the MSS collection into the researcher format.
- **Linked Open Data**
  - We have worked with Fujitsu Ireland and TSO to improve our understanding of how the Linked Open Data BNB is being used. We want to know who is using our data. What data are they using? How can we optimise publication of the data?
  - We have enriched LOD BNB by the addition of 745,000 International Standard Name Identifiers. A project is currently underway to accommodate pre-publication information (CIP) within the model.

This implementation Roadmap for 2016 gives an overview of the activity.
Conclusion

Implementation of the strategy is still in the early stages. In addition to the progress already discussed, the strategy is already having a significant impact. The visibility of collection metadata has been significantly enhanced within the organization. This means that we are consulted much earlier in projects than would have been the case in the past and are able to influence rather than react to decisions. We have seen a substantial increase in demand for our input and we have introduced a work request system to manage it openly and efficiently. We have also been able to demonstrate the need for additional resource to deal with the increasing workload.

Establishing the need for a collection metadata strategy its articulation were difficult tasks that are now beginning to pay dividends. Specifically it has opened up the question of how we can take advantage of future systems replacements to converge our legacy metadata. Convergence of the metadata silos is a precondition for eroding the longstanding silos that hamper discovery of the collection and impede the efficiency of processing.

References

*Unlocking the value: the British Library’s Collection Metadata Strategy, 2015-2018*

*Collection Metadata Strategy Roadmap, 2015-2018*

British Library. Free Data Services
http://www.bl.uk/bibliographic/datafree.html

British Library. Downloads
http://www.bl.uk/bibliographic/download.html