Using Dublin Core Application Profiles to Manage Diverse Metadata Developments

Robina Clayphan
The British Library
Tel: +44 1937 546969
Fax:+44 1937 546586
robina.clayphan@bl.uk

Bill Oldroyd
The British Library
Tel: +44 1937 546856
bill.oldroyd@bl.uk

Abstract:

This paper discusses the use of Dublin Core application profiles at the British Library as part of a resource discovery strategy. It shows how they can be used to control the proliferation of metadata formats in digitisation activity and provide interoperability at a high level between diverse legacy systems. A technical architecture is described. This allows the use of Dublin Core based metadata to support cross-searching of multiple disparate databases.

Keywords:

Dublin Core application profiles, British Library, metadata formats, interoperability, resource discovery strategy, SRU, Z39.50, gateway.

1. Introduction

The British Library (BL) successfully operates under a strong corporate identity and, from the outside, may appear to be a monolithic organisation. Internally, it is an organisation still in the process of a successful evolution from the many smaller entities that have come together over the years to form it. Some of these started as entirely separate organisations, such as the National Sound Archive, and others were largely autonomous centres of expertise following the dictates and practices of their discipline. Indeed, until a few years ago, the concept of the BL as a monolith was not even true in terms of its buildings, having been housed in seventeen separate locations around London and Yorkshire. These diverse origins mean that the British Library has inherited many different legacy systems and bibliographic formats. The data contained in the latter is of considerable worth, representing many years of expert thought, and the associated systems continue to support services targeted at specific user groups.

A more recent addition to this mix has been the output of the many digitisation activities that the BL has undertaken. In the early stages these activities could be characterised as fairly local undertakings – possibly a department of the library putting a database on-line, or working with a university department to digitise some material of significance to a particular body of researchers. The result has been a mixture of customised datasets and standalone systems developed before interoperability became the watchword for digital information.

The following sections of this paper discuss some aspects of a resource discovery strategy aimed at controlling the proliferation of metadata formats in digitisation activity and bringing some degree of harmonisation to the legacy systems. It goes on to demonstrate how the use of a common format and a simple internet based systems architecture can be used to support cross-searching of multiple disparate databases.

2. Digitisation Projects and the Use of Application Profiles

Like most other major cultural institutions, the British Library is undertaking large-scale digitisation of selections of its resources. Many of the resulting digitised images or sounds do not have an existing

catalogue record, either because a record simply did not exist in machine readable form or the item is an image or a detail from a larger resource. As a rule the timescales, budgets and human resources do not permit the routing of the metadata creation through normal MARC21 cataloguing channels. In addition, in many cases, these projects are externally funded and undertaken collaboratively with non-library partners where the use of the MARC format would be inappropriate. Historically there was, and to some extent still is, a tendency for such projects to invent their own metadata format to fit with the particular aims and constraints of the project, the principal factor being the material type in addition to the more pragmatic ones mentioned above.

Typically these projects will buy their own content management systems (CMS) and produce a self-contained system housing both the metadata and the digitised objects. For access there will be a link from the BL web pages to a customised search interface and from there users will be able to find and view the resources. The main drawback is that these resources are not searchable in conjunction with other BL resources - there is no cross-searching of these databases. The BL web pages show the existence of a rich selection of digitised material but from a search of the integrated catalogue most of these resources are invisible. In addition, the customised nature of the metadata means it is not interoperable, making it difficult to harvest for sharing internally or externally and necessitating the development and maintenance of a mapping for each format.

Clearly there is a need to control this proliferation of non-standard metadata formats whilst at the same time retaining an adequate degree of flexibility to meet the needs of particular resource types and the functionality of their associated services. These requirements, control and flexibility, can be seen as necessary twin strands in the resource discovery strategy that is currently evolving. The underlying architecture is not yet in place but is likely to incorporate portal functionality of some description and this will need to be able to read these various datasets. So although the resource discovery strategy is not yet fully defined, one or two basic principles have been articulated and are being acted upon. Chief amongst these is the following:

Where metadata is being created outside the main BL systems, any content management system or database must be capable of generating and exporting a file of coherent descriptive records in a format conforming to one of the supported BL standards. The minimum requirement is to supply records in compliance with the BL Application Profile (B-LAP) encoded in XML, the B-LAP being a Dublin Core-based application profile.

3. Development of the B-LAP

The development of the B-LAP was an attempt to identify a core of metadata terms that had broad applicability across the many resource types in the library. It was based on the Dublin Core Library application profile (DC-Lib) (1), The European Library (TEL) application profile for objects (2) (itself based on DC-Lib) and, finally, the profile that had been used in an early BL digitisation project called Collect Britain (3) (itself based on DC terms (4)). In the end four namespaces were incorporated, DCMI, TEL, MODS (5) (for a couple of in terms from DC-Lib) and some terms that will have to be incorporated into a BL namespace as they have been locally coined. As well as declaring which terms are included, the profile provides some usage rules. The profile is intended to provide a core that whilst probably not meeting all the requirements of any particular project, specifies the nucleus around which tailored profiles can be developed thereby ensuring that at this core level, the metadata will be interoperable. It also ensures that a mapping to this core set is created at the outset of the metadata definition activity of each project.

So far, five initiatives have developed a profile based on the B-LAP and are at different stages of implementation: newspaper article digitisation, conference and journal articles, the archival sound digitisation project, BL web pages and an administrative metadata set to provide an audit trail for deaccessioned material. Each of these has used a subset of B-LAP terms and additionally may have:

added a few terms from another namespace. For example the profile for the web pages has incorporated some terms from the UK government metadata set (6) as they are mandatory for all government agencies.

coined a new term where it could not be found in another namespace – which will be registered in a BL namespace. For example the term PlaceofPublication.

Where a term is found to be applicable to more than one application profile it can be incorporated back into the BL application profile. One such example has been identified to date – DateReceipt.

3.1. Documenting the Application Profiles

All the application profiles are documented according to the CEN DC Application Profile Guidelines produced as a CEN Workshop Agreement in 2003 by the MMI-DC Workshop. (7)

For a few years prior to the production of this document, creators of application profiles attempted to document them in a useful fashion but, in the absence of guidance, had invented a wide range of presentation

formats. This made any kind of direct comparison difficult. The CEN guidelines examined twenty-two existing application profiles and distilled their salient features into a format that is as concise as possible but as detailed as is sometimes necessary to support their various uses. Their modest aim was to provide developers with a normalised and readable view of DC-based metadata models.

An application profile is a declaration of which terms are used in any particular application and may give a customised definition and include some usage rules for that particular application. They do not declare any new terms – these must already exist in a namespace somewhere, ideally uniquely identified with a permanent URI.

3.2. A British Library Namespace

In using DC-based application profiles the BL has coined a few terms that do not exist in any other namespace. This means we should establish a BL namespace and provide the necessary URIs for our locally coined terms. We are currently moving towards this so will need to consider the implementation of a registry system and the best way to offer the necessary degree of permanence.

4. Handling Legacy and Customised Datasets – Themed Collections

Apart from the datasets being created as part of digitisation projects there are other bodies of customised data. The first type are the significant datasets contained in major legacy systems such as MOLCAT (8) – the Manuscripts Catalogue. The second type are those datasets being created or enhanced as part of curatorial initiatives. The concepts of controlling further proliferation of datasets, retaining flexibility and enabling future interoperability can be adapted to these areas as well.

4.1. Legacy systems.

As mentioned at the start of this paper, many of the departments of the BL started life as autonomous bodies which have come together over many years to form the single organisation of today. The consequence of this is the existence of many diverse datasets which are nonetheless very rich and have functionality geared to providing services for specific target user groups. Several problems arise from this situation: there is the actual technical difficulty of trying to maintain elderly legacy systems and make them interoperate with the current BL systems on newer platforms; the amount of human resource required to support this; the familiar issue of the metadata not being interoperable.

These datasets are being examined and where appropriate will be migrated into the integrated library system, if possible by converting the data to the MARC21 format. In the meantime, or if MARC21 is not an appropriate format, the data will be loaded into the Themed Collections system. The Themed Collections system is being developed to enable the storage of all such legacy datasets within one single database. Within this, each dataset will be separately indexed and still be searchable as an individual collection. Ultimately a common data input tool will be provided that will be configured to offer cataloguers the terms appropriate to record creation for each specific collection.

It is proposed to use the B-LAP to underpin the provision of search functionality across all the datasets within the Themed Collections system. A subset of the terms used in the collection-specific datasets can be mapped to the generic terms used in B-LAP. Indexing these terms will allow a search across the full range of collections and from the results list a user will be able to link through to the full description held in the individual collection dataset.

4.2. Customised Datasets

Curatorial staff acquire funding to create databases relating to specialist material. Generally speaking the metadata to be created will not fit in the MARC21 format as it will contain a great deal of contextual material about the resources being described. For example, alongside details of an object, details of the organisation that produced the object may be recorded, accompanied by information about the city during the time when organisation was located there. In addition, the databases are not being created by cataloguing staff but by curators with specialist knowledge of the resources. It is not possible to constrain such endeavors to a limited set of descriptive terms and ultimately Themed Collections should be able to offer a range of standard but customisable formats for these purposes.

What is possible however, is to constrain the choice of system used for the creation and storage of the metadata and to require that some consideration is given at the outset to how the metadata will interoperate with other BL data. At the very least, a mapping must be produced to B-LAP and by this means it should be possible to avoid adding to the number of databases with problematic access. Once these datasets can be incorporated into Themed Collections the mapping will enable cross-searching from within that system.

5. Applying and Managing Multiple Metadata Schemas

It is recognised that for the time being at least we live in a world where multiple metadata formats exist and that these are probably not going to go away. It will be necessary to manage a range of formats and essential that this range is limited. To make managing this easier, the relationships between them need to be made explicit. The range of supported metadata schemas has not been fully defined yet. Certainly one will be the full MARC21 catalogue records that currently exist. Another will be the B-LAP together with its related application profiles from digitisation work. It is possibly useful to visualise a range of metadata standards lying along a spectrum from the richest to the most minimal. The rich standards of MARC21 and EAD can be seen at one end with MODS, ONIX and DC ranged along the line towards the minimal end. Which standard is employed in the creation of the metadata for any particular set of resources is likely to vary according to a set of criteria including:

- the funding available
- the perceived value of the resource
- the time available
- the expertise available
- the purposes to which the resources will be put
- the purposes to which the metadata will be put

All the above criteria are significant but the last one provides an indication of how the metadata can be managed. To what purposes is the metadata going to be put? For example – for providing rich, value-added descriptions of complex objects EAD may be employed; for co-operating with partners from outside the library sector it may only be appropriate to supply simple DC metadata for OAI harvesting; for storing objects in a Digital Object Management system, possibly minimum metadata combined with intelligent technology.

What can be envisaged is moving the metadata up and down the spectrum, transforming from one format to another according to whatever purpose is being served. It must be recognised that in the transformation from richer to poorer and back there will be loss of data.

Critical to moving up and down this spectrum is the existence of crosswalks between the schemas and, ideally, the use of XML encoding. With the latter comes the possibility of using XSL stylesheets to generate of the different views as required. If all the different schemas can be expressed in XML the transformations become relatively simple to accomplish. This notion can be characterised as "getting the data onto the XML bus" and, with the existence of a lossless representation of MARC21 in XML, can incorporate traditional library data alongside the more emergent formats.

6. The current systems architecture

In order to be useful, an application profile which is to be applied across an organisation's metadata requires a technical implementation that will take into account the complete database infrastructure. A profile which only applies to new metadata applications is likely to be taken up slowly: an approach which may undermine the whole concept. The British Library has devised a short-term tactical approach which makes much of the library's metadata available in the application profile. This includes data in both new and legacy systems.

As mentioned earlier, the British Library is a partner in the European Library (9). The European Library portal (10) sets out to provide a common interface to the bibliographic databases of Europe's national libraries. There are currently 10 partners providing access to their databases through the portal. As a result of the TEL-ME-MOR project (11), this number will soon increase with the addition of national libraries of the new member states of the European Union.

The European Library portal requires that partners provide access to their databases using the SRU protocol (12) and the presentation of their metadata records in XML using the European Library Application Profile (2). The SRU protocol is based on the http internet protocol. A database is identified by a URL which in turn is extended by parameters for the search query and other controls. The result of the search and the resulting records are returned as XML. This standard interface applied across potentially hundreds of databases has enabled the European Library portal to be implemented as a Javascript and XSLT application running in the user's web browser rather than at a central server. This is a novel architecture (13).

SRU is a modern protocol resulting from the recent developments of Z39.50 known as ZiNG. Until recently SRU has not been widely implemented. In order to help partners participate in the European Library service the British Library developed a prototype SRU/Z39.50 gateway (14) to translate between the two protocols. This allows partners that provide a Z39.50 service to conform to the European Library requirements without further technical development.

The SRU/Z39.50 gateway, and developments of it, form a technical infrastructure which has been used to implement the British Library Application Profile

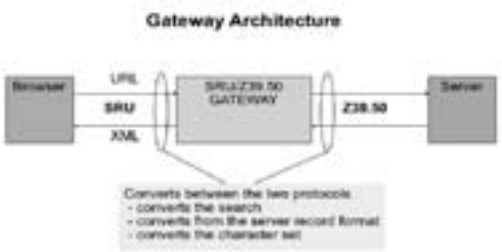

Figure 1. The basic function of the gateway.

across most of the library's bibliographic databases, and as we shall discover, external databases as well.

The gateway performs a number of functions. The most obvious function is the conversion between SRU and Z39.50. Each search is a discrete and separate transaction and is equivalent to one Z39.50 session: perform the search and present the specified records.

However it is the additional functions that are important. First, the incoming SRU query is converted to the query language, character set and index names used by the various Z39.50 target. Within the limitations of the underlying Z39.50 databases this provides a uniform method of searching.

Then, on output, the native format returned by the target is converted: this includes conversions from Marc21, Unimarc, Sutrs, XML, or MarcXML to the British Library Application Profile. In addition it is also possible to extend the metadata by the addition of standard data to every record, such as rights data, the creation of new presentations of the data, for example an OpenURL, the generation of identifiers and predetermined URLs to create links to other resources such as thumbnails and digital objects, and the addition of data arising from code conversions and database lookup.

As a result of the development of the gateway it is now possible to search and present records in the British Library Application Profile from:

British Library Integrated Catalogue (15) and other bibliographic databases which run on an Ex Libris Aleph system.

British Library Sound Catalogue (16) which runs on a Sirsi Unicorn system.

Databases using the **Cheshire 2** database software.

British Library Electronic Table of Contents Database (ETOC) which runs on a bespoke implementation of the BRS database.

The gateway can, with permission, also provide records in the British Library Application Profile retrieved from the catalogues and databases of other libraries, for example COPAC (17), the Bibliothéque nationale de France, the Swiss National Library and the Library of Congress. Although this feature isn't currently used, it shows that data integration can extend outside an organisation's own metadata if this is required.

In addition to the main bibliographic databases, the British Library has created a considerable amount of metadata describing digitised items and images. As a rule the databases used for these services do not provide an interface that can be easily accessed using the gateway approach. In these cases the metadata is transferred from the local system and stored in an SRU-capable database. The conversion from the local format, usually a set of database tables, to the British Library Application Profile is partly carried out in the process of loading the records and partly at the time the records are retrieved. For example this approach has been adopted for the Collect Britain (3) and Illuminated Manuscript (18) databases.

It does not require a great leap of imagination to see that the principal of the gateway could be applied to convert between SRU and other protocols. At present the most useful conversion being between SRU and other http and XML based search services. Such services are frequently used for new internet-based search applications. So far the British Library has developed gateways which interface with the following services :

Google Search Appliance (19), which provides the search function used by the British Library's web site.

PubMed databases made available through the NCBI Entrez utilities interface (20).

Ex Libris X Server interface, which will in future remove the need to use Z39.50 to access the databases held in Aleph.

As a result of these three related developments it has been possible to make much of the British Library's metadata available in the Application Profile by a simple method based on a widely used network protocol.

One reason for describing this architecture as short-term is that perhaps the principal benefit it provides to the British Library is the ability to prototype services. The simple access method and common XML metadata format are easy to use in modern web browsers and programming languages. Also, as the conversion between formats is carried out when records are retrieved it is possible to experiment without changing the underlying databases. The expectation is that prototyping and experimentation will lead to more substantial plans to provide operational services along the lines of the development of the Themed Collections.

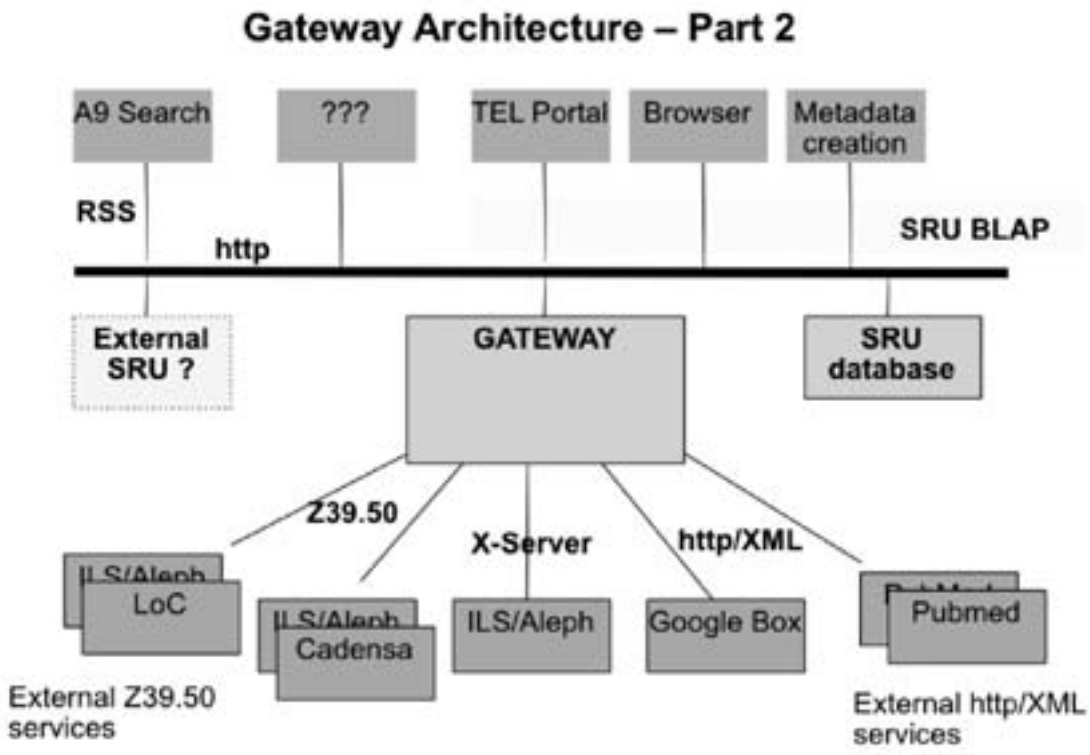

Figure 2. The fully developed gateway architecture (as at April 2005).

7. Some Example Applications

This architecture has two main aims. The first is to provide a single interface across a diverse set of databases and second to make the databases available to a range of user interface systems.

Figure 2 illustrates a more complete architecture for the gateway and the following describes some example prototypes that have been created.

Portals. The European Library portal is a sophisticated application, but it is possible to create simpler, more lightweight portal applications that run in the user's web browser. Searches can be easily formulated in a web form and the resulting responses formatted for display using an XSLT and CSS stylesheets. This approach provides a simple common interface for searching and metadata display from all the library's databases. As the search and record display is generated in the browser it is possible to use standard functions such as bookmarking and search bars to enhance the user interface. If a database of collection descriptions is provided, a user may use the common interface to search and select collections and services before performing more detailed searches in a specific collections.

Example: http://herbie.bl.uk:9080/cgi-bin/blils.cgi?query=dickens *Examine the source of the displayed page to see the underlying XML representation of the metadata.*

Global Search. A search across the library's different databases can be created simply by adding the user's query to a set of pre-determined database URLs. This can take the form of a simple XML document which can be processed by Javascript or XSLT to perform the search and display the records from all the chosen databases.

Example: http://herbie.bl.uk:9080/cgi-bin/blils.cgi?query=dickens *Examine the source of the page to see how the search is applied to several databases.*

Bookmarklets. These are simple one line Javascript applications which can be stored as a bookmark. Perhaps the most famous bookmarklet in the library world is the LibraryLookUp bookmarklet (21). A user displaying a web page for a book on the Amazon site can use the bookmarklet to jump directly to a search for the book in their local library OPAC. This technique has been adapted to work in a slightly different way with the British Library databases. A user may mark any text on a web page and then by

clicking the bookmark, perform a search in one of the library's catalogues.

Example : http://herbie.bl.uk:9080/lookup.html *Load and use the bookmarklet : BLClick .*

URLs provide metadata. A URL containing a search query or the control number of a metadata record provides metadata that can be used in internet aware applications and web pages. Digitisation projects frequently require the use of metadata extracted from a main cataloguing database, usually using the British Library Application Profile. The metadata editing software for these projects can generate URLs to automatically derive skeleton records for the digitised items.

OpenSearch. A9.com provides a user-driven portal to web searches and selected XML based search services using a specification called OpenSearch. Although this service is not based on an SRU response but instead uses RSS, it is a good example to show how the use of http based protocols can create opportunities for new types of service integration. A simple change was made to the gateway to accept the OpenSearch query and generate the RSS output.

Example : http://opensearch.a9.com/

There is considerable potential to discover new ways of integrating data across the internet. The use of http based searching and XML output offers considerable potential to integrate bibliographic metadata with services such as social bookmarking and Google Maps.

8. Conclusion

The BL has a legacy of many diverse databases and this is coupled with the need to find a realistic solution to metadata creation in on-going digitisation work. The establishment of a generic DC-based application profile opened the possibility of customised but interoperable profiles for projects as well as offering the potential to cross-search other BL datasets. The exploitation of a new SRU protocol to search for and present records, used in conjunction with the profile, has provided the BL with this uniform method of access across its collections. Simple prototype systems demonstrate the advantages of this approach and, because they leave the underlying databases unchanged, allow the experimentation that will lead to more substantial plans for operational services.

9. References

1. DC Library Application Profile http://dublincore.org/documents/2004/09/10/library-application-profile/
2. The European Library Application Profile for Objects http://krait.kb.nl/coop/tel/handbook/metadata_handbook.html
3. The Collect Britain website http://www.collectbritain.co.uk/
4. DCMI Metadata Terms http://dublincore.org/documents/dcmi-terms/
5. Library of Congress: Metadata Object Description Schema. http://www.loc.gov/standards/mods/
6. UK Government: eGovernment Metadata Standard V3 http://www.govtalk.gov.uk/schemasstandards/metadata_document.asp?docnum=872
7. CWA14855 - Dublin Core Application Profile guidelines http://www.cenorm.be/cenorm/businessdomains/businessdomains/isss/activity/wsmmi.asp
8. British Library Manuscripts Catalogue http://www.bl.uk/catalogues/manuscripts/
9. The European Library. http://www.europeanlibrary.org/
10. The European Library portal (beta version). http://www.theeuropeanlibrary.org/portal
11. TEL-ME-MOR. http://www.telmemor.net/
12. The Z39.50 Maintenance Agency. ZiNG/SRW/SRU web pages. http://www.loc.gov/z3950/agency/zing/srw/
13. T. van Veen and B. Oldroyd. Search and retrieval in the European Library: a new approach. In: *D-Lib Magazine*, vol. 10, n. 2 February 2004. http://www.dlib.org/dlib/february04/vanveen/02vanveen.html
14. British Library SRU/Z39.50 gateway software. http://herbie.bl.uk:9080
15. British Library Integrated Catalogue. http://catalogue.bl.uk
16. British Library Sound Catalogue. http://www.bl.uk/catalogues/sound.html
17. COPAC http://copac.ac.uk/
18. British Library Catalogue of Illuminated Manuscripts. http://www.bl.uk/catalogues/illuminatedmanuscripts/welcome.htm
19. Google Search Appliance. http://www.google.com/enterprise/gsa/index.html
20. NCBI Entrez interface. http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html
21. John Udell: LibraryLookup. http://weblog.infoworld.com/udell/stories/2002/12/11/librarylookupGenerator.html