discovery lessons

Senior Managers	
Metadata Creators	✓
Developers	✓
Suppliers	

Metadata Formats

The Discovery programme has been funded by JISC to improve access to collections that support research and education. This document is part of a series that describes the lessons from 8 JISC projects funded under the Discovery programme in 2011 to explore open metadata for libraries, museums and archives. More information about the projects can be found at: http://www.jisc.ac.uk/whatwedo/programmes/inf11/infrastructureforresourcediscovery.aspx
The other documents in the series can be found at: http://discovery.ac.uk

Scope

To be able to reuse data appropriately, it is necessary to understand it. This Guide outlines the issues relating to publishing metadata in understandable formats, captures the related outcomes from the eight Discovery projects and makes clear recommendations for others wishing to publish open, reusable, metadata.

The Problem Space

Metadata is only usable by others if there is a common understanding of what each piece of metadata represents. That is to say, it is clearly important to understand that 'Jane Eyre' is the title, not the author, of a book.

The Discovery initiative encourages the open publication of metadata, with the intention that it will be reused by others, and so the question arises of how a third party understands the metadata available.

Traditionally libraries, museums and archives have relied on using well established metadata standards to ensure that metadata can be easily understood within the relevant community. In each community, there tends to be an overlapping set of standards which together define the structure of a metadata record, the information contained within the record, and how that information should be represented.

Within libraries the MARC21[1] standard defines how a bibliographic record should be encoded, while standards such as AACR2[2] relate to how specific pieces of information should be represented - such as how names should be formatted. In archives EAD[3] and ISAD(G)[4] serve similar purposes. Practice in museums is more varied[5], but standards such as CDWA[6], SPECTRUM[7] and the CIDOC-CRM[8] start to offer some unified practice across the museum and gallery community. "Metadata for All" by Elings and Waibel[9] is recommended reading for those wishing to get a good overview of the relevant standards and how they interact.

While effective when working within a community, these standards are not as effective in offering a shared understanding of metadata when working across and outside community boundaries, and when describing a wide range of different types of item.

1 http://www.loc.gov/marc/bibliographic
2 http://www.aacr2.org
3 http://www.loc.gov/ead
4 http://www.icacds.org.uk/eng/standards.htm
5 http://www.pro.rcip-chin.gc.ca/normes-standards/guide_normes_musees-museum_standards_guide/metadonnees-metadata-eng.jsp
6 http://www.getty.edu/research/publications/electronic_publicatio ns/cdwa/index.html
7 http://www.collectionstrust.org.uk/index.cfm/collection-management/spectrum
8 http://www.cidoc-crm.org
9 http://firstmonday.org/article/view/1628/1543

Outcomes

Simplicity

One approach to delivering understandable metadata is to keep it simple. The Jerome project at the University of Lincoln[10] took this approach, offering a simple set of metadata in JSON format. Decisions on which metadata elements to user were informed by the RIS format used by a variety of 'reference management' software packages[11].

This approach has the advantage of being relatively easy to work with for third-parties, even if they are not very familiar with library metadata, and simple data may well serve a large number of possible use cases for the data. However, it also loses some of the detailed information that might be typically recorded in a library catalogue record as well as contextual information that is derived from cataloguing standards.

Just a very few examples of information that might be lost when moving from detailed library catalogue records to simpler representations include: the physical characteristics of the item which might be useful for specific enquiries or preservation contexts; the concept of transcribed information (e.g. the place of publication in a library catalogue record is often a transcription of text printed in the item, with special syntax to indicate corrections or additions by the cataloguer); variations and translations of the title; country (as opposed to more specific place) of publication.

Linked Data and RDF

The approach taken by the majority of projects was to publish data as Linked Data[12] in RDF (Resource Description Framework)[13]. While there are a wide range of reasons projects chose to take this approach, there are specific advantages to this approach in terms of offering understandable metadata.

The nature of RDF means it is possible to use a mixture of different vocabularies (or 'ontologies') to describe things within the same data set. This means that where specialist properties are required, which might only be understood within a community or niche area, those can be used alongside more general ontologies that are more broadly used.

For example libraries, archives and museums may commonly want to describe people and organisations. Traditionally they would do this within the confines of their own, community specific, metadata schemas and formats. However, having adopted RDF and Linked Data mechanisms they can use the 'FOAF' (Friend Of A Friend) ontology. FOAF was originally created to describe relationships between people, but since it had to describe basic properties of people, such as names, to achieve this, it has become widely adopted by the Linked Data community as a standard way of describing people. This opens the possibility of libraries, archives and museums sharing a common way of describing people (in terms of metadata elements), while still using specialist metadata elements where necessary.

This ability to reuse ontologies is exemplified across the projects using a Linked Data approach. The following ontologies were used across several projects including Comet[14], SALDA[15] and OMP[16]:

- DCMI Metadata Terms (http://dublincore.org/documents/dcmi-terms) for describing a range of basic properties relating to objects in museum, library and archival collections
- FOAF (http://xmlns.com/foaf/spec) for describing people and organisations
- Bio (http://purl.org/vocab/bio/0.1) for describing biographical information about people
- SKOS (http://www.w3.org/2004/02/skos) for describing lists of controlled terms, thesauri and authority files
- RDFS (http://www.w3.org/2000/01/rdf-schema), specifically the 'label' property which is "a human-readable name for the subject"

10 http://jerome.library.lincoln.ac.uk/about
11 http://www.adeptscience.co.uk/kb/article/FE26
12 http://linkeddata.or
13 http://www.rdfabout.com/
14 http://cul-comet.blogspot.com
15 http://blogs.sussex.ac.uk/salda/
16 http://openmetadatapathway.blogspot.com/

Also noteworthy for bibliographic data is Bibliontology (http://bibliontology.com) used by the Comet project, and extensively in the publication of bibliographic Linked Data elsewhere.

Other projects used more specialist ontologies, for example the use of the MuSim ontology by the CORE project[17]. The MuSim or Similarity Ontology[18] was initially developed to express similarity between musical items (such as tracks or artists). However, the ontology was designed in such a way that it could be used to describe similarity more generally and CORE used this to express computed similarity between items in institutional repositories.

Where a suitable ontology is not already defined, new ontologies can be defined, and RDF has mechanisms for both defining ontologies and relating ontologies to each other. The mechanisms for defining ontologies, and the identification of properties within these ontologies by URIs sometimes lead to RDF being defined as[?] 'self-describing'[19]. However, when defining new ontologies there is a risk that without broad adoption, at least within a relevant community, the data published will remain hard to exploit by others.

Identifying appropriate ontologies to use remains closer to an art than a science, although there are some attempts to document starting points for research[20] and to offer catalogues of ontologies[21].

Recommendations

- Consider both your immediate target audience and possible broader audiences when deciding metadata formats for publishing data.
- Look at common practice both inside and outside the relevant communities to identify suitable metadata formats.
- When publishing Linked Data, use existing ontologies wherever possible.

Key Discovery Projects

- CORE - http://core-project.kmi.open.ac.uk
- Jerome - http://jerome.blogs.lincoln.ac.uk
- Comet - http://cul-comet.blogspot.com

Other Related Work

- Schema.org - http://schema.org
  A joint initiative between the major web search engines to enable structured metadata to be embedded in HTML documents
- Linking University Content for Education and Research Online (LUCERO) - http://lucero-project.info/lb
  A project to publish Linked Data at the Open University (including bibliographic data) and to provide guidance to those wishing to do the same
- Bibliographic Knowledge Network - http://bibserver.okfn.org
  A collection of tools and initiatives to enable the publication of open bibliographic data in a simple way

17 http://core-project.kmi.open.ac.uk
18 http://purl.org/ontology/similarity
19 http://www.w3.org/2001/tag/doc/selfDescribingDocuments.html#RDFSection
20 http://www.w3.org/wiki/Ontology_Dowsing
21 http://schemapedia.com