discovery lessons

Senior Managers	
Metadata Creators	✓
Developers	✓
Suppliers	

Entities and Authorities

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

Identifying entities, such as people and places, and describing them in a consistent manner (both within single datasets and between datasets from multiple sources) is a well known challenge. This Guide outlines the issues relating to using 'authority' data to identify and describe entities in metadata, captures the related outcomes from the eight Discovery projects and makes clear recommendations for others wishing to publish open, reusable, metadata.

The Problem Space

One of the driving concepts of the Discovery initiative is that compelling services can be built of aggregations of open metadata from across Libraries, Museums and Archives. When aggregating data from sets of heterogenous resources, identifying concepts which are either similar or identical is a significant challenge. These concepts may represent 'aboutness' of a resource (such as subject headings), or people, organisations, places, dates or other entities with some specific relationship to the resource.

Libraries, Museums and Archives have traditionally approached this problem in several ways including:

- Following community rules when creating headings (textual strings) for entities (e.g. NCA Rules[1]; AACR2[2])
- Using local or community wide authority files or controlled headings lists

While these approaches to can mitigate the problem of identifying when the same entities are being referenced, variation in use both within and across communities means that they cannot be relied upon, even within a single collection, to ensure that the same entity is identified in the same way consistently.

Outcomes

A simple approach to reducing the issues of aggregation is to copy existing metadata descriptions from other sources. This approach is well known in libraries and is a natural consequence of cataloguing cooperatives and copy cataloguing. However, the growing open data environment creates more opportunities to utilise existing records and to compare across data sources. The Jerome project at the University of Lincoln[3] adopted this approach drawing on records from Open Library[4] to both enhance their local records and to flag potential errors (where details varied across records) to library staff at the University. It should be noted that this approach relied on existing shared identifiers (ISBNs) to match bibliographic records across sources.

Those project based on Linked Data tended to consider identifying and matching entities not just at the level of a resource (such as a book, manuscript or object) but also related entities such as 'people', 'subjects' and 'places'.

1 http://anws.llgc.org.uk/ncarules/title.htm
2 http://www.aacr2.org
3 http://jerome.blogs.lincoln.ac.uk
4 http://openlibrary.org

The first step in many cases is local identification, and sometimes standardisation, of entities. Work in the Sussex Archive Linked Data Application (SALDA) project[5] resulted in the use of authorised forms of names as well as a controlled terms list for subjects, and an overall update of cataloguing procedures.

The Cambridge University Library Cambridge Open Metadata (Comet) project[6] found that by using some simple normalisation routines on name entities, multiple strings identifying the same entity could be linked to a single identifier. While this approach is unlikely to be 100% successful due to the amount of variation possible in textual strings describing entities, in the Comet project it allowed for a large number of entities to be matched automatically.

Both SALDA and Comet went further than establishing local identifiers for entities, and matched some entities to other Linked Data sources. Of particular significance is the use of VIAF (Virtual International Authority File)[7] by both projects, which means that they can share common identifiers for people, despite having different underlying data formats and cataloguing rules. The use of common identifiers for common entities across Libraries, Museums and Archives would be a significant step towards easier aggregation and discovery.

The Open Metadata Pathways (OMP) project[8] took this process of exploiting external identifiers the next logical step and produced tools to embed the use of these into the metadata creation process. Notably OMP drew on non-traditional 'authorities' such as GeoNames[9] for places.

It seems clear that the ability to use data from third parties easily and without impediment - that is open both by license and accessibility - was key across all of these approaches. The evidence from the eight projects also suggests that while Linked Data is not a requirement for establishing local identifiers and exploiting external identifiers for entities, the Linked Data approach encourages this explicitly, and provides clear mechanisms of doing this.

Recommendations

- Establish globally unique identifiers for entities in your data
- Use community-wide identifiers where possible and appropriate
- Consider the use of identifiers from outside the community where possible and appropriate
- Build the use of unique entity identifers into the metadata creation process
- Do not rely on text strings to uniquely identify entities

Key Discovery Projects

- Comet - http://cul-comet.blogspot.com
- Jerome - http://jerome.blogs.lincoln.ac.uk
- OMP - http://openmetadatapathway.blogspot.com
- SALDA - http://blogs.sussex.ac.uk/salda

5 http://blogs.sussex.ac.uk/salda
6 http://cul-comet.blogspot.com
7 http://viaf.org
8 http://openmetadatapathway.blogspot.com
9 http://www.geonames.org

Other Related Work

- Linked and Open Copac and Archives Hub (LOCAH) - http://blogs.ukoln.ac.uk/locah/
  A project which addresses how Linked Data can support cross-domain discovery of L&A material and beyond
- Linking University Content for Education and Research Online (LUCERO) - http://lucero-project.info/lb/
  Project to publish Linked Data at the Open University (including bibliographic data) and to provide guidance to those wishing to do the same
- British Library Linked Open BNB - http://www.bl.uk/bibliographic/datafree.html
  The British National Bibliography as Linked Data
- Virtual International Authority File (VIAF) - http://viaf.org/
  A project aiming to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web
- Library of Congress Linked Data - http://id.loc.gov
  A variety of key authority and term lists as Linked Data