discovery project review

COMET

(Cambridge Open METadata)

http://cul-comet.blogspot.com

Comet was 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

Background

The Comet project built on earlier work by CUL as part of the JISC Open Bibliography project, under which CUL released around 130,000 records. The drive to release 'open data' was to a large extent driven by requests from academics for this to happen - this was the basis of the JISC Open Bibliography project led by an academic unit at Cambridge. This approach to openness is recognised as a catalyst for change in the CUL strategic framework which states 'Open is the New Normal'.

Institution	Cambridge University Library (CUL) is a legal deposit library that both directly serves the needs of the staff and students of the University of Cambridge, and also a global community of users as a world-class research library.
Responsible group	The CUL-Comet project was delivered by a small team of staff drawn from CUL and the Centre for Applied Research in Educational Technology (CARET) at the University of Cambridge.
Capacity	While CUL is a large library organisation with a budget in excess of £10m per annum the team that delivered CUL-Comet was small, perhaps especially so given the scale of the overall operation.
Data Scope	The project released a subset of bibliographic records from the CUL Catalogue, including locally created records and records originally sourced from a wide range of external organisations including Research Libraries UK (RLUK) and OCLC.
Data Scale	Overall the project released data representing approximately 2.2 million MARC records.

Mechanics

Formats

From the very start, the Comet project aimed to release data as RDF[1] with the development of tools and procedures to convert bibliographic data from MARC format to RDF being a major focus of the project.

As with similar projects Comet found that there was a lack of clear best practice for mapping MARC to RDF. This led to the creation of a mapping based on local requirements and drawing on the experience of other similar projects (more details below).

The transformation of records from one format to another, and particularly the difficulty of exactly recreating the original MARC records from the Comet RDF representation, proved helpful when negotiating with third party record suppliers regarding releasing data derived from their records under an open license.

[1] http://www.w3.org/RDF

Technologies

Comet made use of the ARC2 PHP library[2], which can utilise MySQL as an RDF store. The project decided to use this latter function rather than investing in the installation of dedicated triple- or quad-store software. The project team believe that for relatively small collections this would be adequate, although had concerns about it's scalability to collections the size of the CUL.

Comet developed two key tools (written in Perl and using the MARC::Record module):

- MARC Record sorter (A script to sort records by ownership, see licensing section below)
- MARC21 to RDF Conversion Utility (see below for more details)

The tools used and developed by Comet are based on Apache, MySQL, PHP and Perl, which are widely used in the UK HE sector. All of the tools would be straightforward to install given familiarity with these underlying technologies.

Enhancement

As noted above, there is no established best practice for transforming MARC (or other common library bibliographic metadata formats) to RDF, and different projects have taken different approaches[3,4,5]. However, there are common Linked Data ontologies that are being used across a variety of projects and Comet used many of these include:

- Dublin Core Terms[6]
- The Bibliographic Ontology (also known as Bibliontology or Bibo)[7]
- SKOS (Simple Knowledge Organization System)[8]
- FOAF (Friend of a Friend)[9]

The project also used some less commonly used ontologies for a few specific data elements, notably two (as yet) unratified library specific ontologies:

- RDA Group 1 Element Vocabulary[10]
- ISBD elements[11]

Perhaps the most notable aspect of the use of these library-specific ontologies are that they were required for less than a handful of data elements, all others being covered by more commonly used ontologies.

Entity identification and linking

As Comet was focused on producing Linked Data, a key aspect of the project was assigning URIs to entities in the bibliographic data (such as 'people', 'organisations' and 'subjects').

URIs for 'records' were created using an identifier from the original catalogue record.

URIs for other entities such as 'people' and 'subjects' were formed by creating a unique value based on the relevant text strings, with punctuation removed (as they found cataloguing practice regarding punctuation to be particularly variable). This approach meant that where identical, or near-identical, strings were used for an entity in the catalogue record (such as author name), the same URI would be used, and so a local 'authority' established. The original text string was retained as a 'label' so that it would be possible to see all variations used across the catalogue.

2 https://github.com/semsol/arc2
3 http://blog.libris.kb.se/semweb/?p=7
4 http://www.bl.uk/bibliographic/pdfs/datamodelv1_01.pdf
5 https://wiki1.hbz-nrw.de/display/SEM/Converting+the+Open+Data+from+the+hbz+to+BIBO
6 http://dublincore.org/documents/dcmi-terms/
7 http://bibliontology.com/
8 http://www.w3.org/2004/02/skos/
9 http://xmlns.com/foaf/spec/
10 http://metadataregistry.org/schema/show/id/1.html
11 http://metadataregistry.org/schema/show/id/25.html

In the case of Library of Congress subject headings, where possible entities were then linked to the relevant authorised headings on http://id.loc.gov, thus linking the Comet dataset into the wider Linked Data environment. Unfortunately the Library of Congress did not release its Name Authority File as Linked Data until shortly after the end of the Comet project, as this would have also been a target for the project to link to.

Comet also links to other vocabularies supported by http://id.loc.gov such as Country and Language codes.

Work was also started linking to OCLC data in the form of FAST (subject headings) and VIAF (names), with a sample of 10,000 records.

Usability

Linked data

The use of Linked Data may be seen by some to be a barrier to usability[12]. However, the use of common ontologies for the vast majority of data published by Comet opens up usability across the Linked Data community, and the clearly documented approaches to URI formation would allow others to easily map entities from their own data to Comet entities.

Comet also supports two specific approaches which increase (re-)usability. Firstly, Comet supports the ability to search the data by some text strings, avoiding the need to understand SPARQL to query the data[13].

Secondly, the approach of assigning URIs based on record identifiers, had the side effect of delivering a usability benefit. Given a 'record' URI, it is possible to retrieve representations and data from other library systems that share that identifier. This opens up the possibility of exploiting aspects of the Linked Data representation such as SPARQL and 'follow your nose' linking, while retrieving records in MARC XML from the Aquabrowser SRU interface[14]. It should be noted that this relationship isn't currently expressed in the RDF representation and requires knowledge of the URI structure to exploit it.

Impact

Licensing

Ownership and rights

Comet did a considerable amount of work to examine the source and ownership of MARC21 records[15]. Due to the nature of the CUL collection, the records in the CUL catalogue have been derived from a wide variety of sources over a long period of time. While this includes a significant number of locally created records, there are also a large number of records in the collection that have been supplied under specific contracts or agreements.

Where possible data has been published using the ODC-PDDL license[16], however the project recognised this may not always be possible, and developed a work process and analysis tool to help split batches of MARC21 records into datasets to enable different licenses to be applied[17].

The project has recorded agreements with different suppliers[18] regarding the licensing of records, which establishes some important precedents, and will be extremely valuable to other institutions wishing to openly license their bibliographic records.

12 http://electronicmuseum.org.uk/2010/03/22/linked-data-my-challenge

13 http://data.lib.cam.ac.uk/faq.php

14 http://www.lib.cam.ac.uk/api

15 http://cul-comet.blogspot.com/p/ownership-of-marc-21-records.html

16 http://opendatacommons.org/licenses/pddl/1.0

17 http://cul-comet.blogspot.com/2011/07/where-exactly-does-record-come-from.html

18 http://cul-comet.blogspot.com/p/ownership-of-marc-21-records.html

Technical approach

Comet has created 'named graphs' (which group together data) and applied a license to each 'graph'. This relationship is expressed in RDF and so is machine-readable. An example can be seen at http://data.lib.cam.ac.uk/context/dataset/cambridge/bib

Benefits

Comet was partially a reaction to direct demand from academic staff and also in line with institutional strategic direction. However, the immediate direct benefits for users are not clear. There is clear interest in machine interfaces to the CUL catalogue as demonstrated by the development of an iPhone application by a student using (non-RDF) APIs made available by CUL[19].

Business Case

Despite academic demand, the business case specifically for a Linked Data version of the CUL bibliographic data is as yet unproven.

Outcome

Comet is part of a movement towards the use Linked Data by libraries, and the work carried out is certainly cutting edge. The Linked Data environment continues to change and develop quickly, which suggests that while the technical development and Linked Data tools published by Comet will certainly offer shortcuts for other institutions carrying out similar projects, the work and tools relating to licensing will be of more long term benefit than the Linked Data developments.

Lessons Learned

- Suppliers of bibliographic records to libraries are often happy for data derived from those records to be published under an open license.
- Format matters to record suppliers, and there is less resistance to publishing data openly if the original records cannot be recreated easily in the original format from the published data.
- A Linked Data approach encourages the identification of entities within bibliographic data rather than simply 'records'.
- It is possible to publish Linked Data using standard software and platforms.
- The use of textual strings and the variations in cataloguing data make entity identification and linking both internally and externally challenging.
- Given a committed, expert, team, small amounts of resource can produce innovative services.
- Using common record identifiers can enable bridging between Linked Data and more traditional representations of the same bibliographic items.

See Also

- Jerome - http://jerome.blogs.lincoln.ac.uk
  Alternative approach to publishing bibliographic metadata
- Linked Open BNB from the British Library - http://www.bl.uk/bibliographic/datafree.html
  Alternative modelling of library data, with overlapping data
- Lucero - http://lucero-project.info/lb
  Publication of linked, open, institutional data from the Open University

[19] http://itunes.apple.com/us/app/ucam-library-search/id459882806?ls=1&mt=8