discovery project review

Jerome

http://jerome.blogs.lincoln.ac.uk

Jerome 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

Before successfully bidding for JISC funding, Jerome was an ongoing 'unproject' (an unofficial project with no dedicated resource) which aimed to 'reimagine library services'. Jerome built on previous projects at the University which had successfully exploited a specific technology stack. Post JISC funding, Jerome returns to it's unproject status. Jerome is both the name of the project, and the name of the application and service created as part of the project. This document refers to 'the Jerome project' and 'the Jerome application' where it is necessary to differentiate these aspects.

Institution	The University of Lincoln is a new university (est. 2001) spread across four campuses, with approximately 10,000 students, the vast majority of these being undergraduate students. The institution is currently introducing a 'Student as Producer' principle across the university (http://studentasproducer.lincoln.ac.uk).
Responsible group	The Jerome project is a collaboration between the Library and ICT at Lincoln. During the project a new technology group has been formed at Lincoln called LNCD (http://lncd.lincoln.ac.uk). Key staff from the project are members of this new group.
Capacity	The team at the University of Lincoln does not have significant resource in terms of numbers of staff or direct funding, but the staff involved have a great deal of technical expertise and commitment.
Data Scope	The project addresses a wide range of library resources available to students and staff at the University of Lincoln, covering the records in the library catalogue, e-journal records, the contents of the institutional repository and journal table of contents made available via the JournalToCs service (http://www.journaltocs.hw.ac.uk).
Data Scale	The dataset contains over 265,000 records

Mechanics

Formats

The key data format for Jerome is JSON, with the RIS metadata format being used to guide decisions on what metadata to include - the basic principle being that it should be possible to create a citation or populate citation management software from the Jerome record. It was also felt that this represented the key information necessary for the majority of 'discovery' needs. JSON fitted with the underlying technology being used to build the Jerome application and other University of Lincoln services (see below), as well as being a simple format that was easy to work with.

The underlying data store used by the Jerome application is MongoDb, a document-oriented database. MongoDb stores data as JSON-style documents, but the Jerome application has been built with the ability to deliver data in multiple formats[1]. During the project JSON, RDF/XML and RIS were supported, and MARC could have been added in theory. Additionally, the 'schemaless' nature of MongoDb means that extending the metadata stored would be relatively trivial, with no need to re-work the existing data or code.

1 http://jerome.blogs.lincoln.ac.uk/2011/07/31/the-jerome-resource-model

Currently the data used by the Jerome application does include identifiers for any entities except the item/record. In contrast to the Cambridge Open METadata (COMET)[2] project, this means that a straight RDF representation of the existing data does not have URIs for entities such as people, places or subjects. While this could be addressed, it suggests a different set of priorities and mindset when approaching the data.

Note that, following the end of the project, support for formats other than JSON has been temporarily withdrawn, but will be restored at a later date.

Technologies

The main underlying technologies used by the Jerome project are[3]:

- MongoDb (Document oriented storage, using JSON-style documents[4])
- Sphinx (Search Server[5])

The choice of these technologies was down to a variety of factors. Local expertise and use of the technologies in other projects informed the decision, as did a match to user requirements and the ability to support very high performance[6]

While the technologies used by the project are interesting, and certainly the use MongoDb is breaking new ground in the sector and reflects some wider interest in this approach (e.g. at The Guardian[7]), overall the technology was seen as an adjunct to delivering an excellent user experience, and was not 'the point' of the project.

Discussion at the Jerome/COMET 'hack day' event suggested that it might be relatively easy to re-purpose aspects of the Jerome application to provide a search interface for other data stores and data formats[8].

Enhancement

The Jerome project didn't use 'authorities' in the traditional library sense. However, it did demonstrate a very interesting use of external sources, specifically the Open Library[9], to enhance records.

In the Jerome application, where a record has an ISBN it is looked up on the Open Library using the Open Library API. If there is a matching record in the Open Library data it is used to either enhance the Jerome record, or to trigger workflows. For example:

- Subject keywords from the Open Library are always added to the Jerome record.
- Title strings from the Open Library are compared to the title in the Jerome record (with some simple normalisation rules applied). If the strings are similar then the Jerome title is retained, but if there are differences then the record is flagged with the aim that these flags will eventually drive reports for cataloguers to check records.

While Open Library is currently used, the mechanism could be easily extended to other sources - as discussed at the Jerome/CUL-Comet meeting, the COMET data could be one such source.[10]

2 http://cul-comet.blogspot.com
3 http://jerome.blogs.lincoln.ac.uk/2010/07/30/the-slides
4 http://www.mongodb.org
5 http://sphinxsearch.com/about/sphinx
6 http://jerome.blogs.lincoln.ac.uk/2010/07/23/engage-ludicrous-speed
7 http://nosql.mypopescu.com/post/4961835847/mongodb-at-guardian-co-uk
8 http://jerome.blogs.lincoln.ac.uk/2011/08/10/jeromecomet-hack-day-fun-in-the-fens
9 http://openlibrary.org
10 http://jerome.blogs.lincoln.ac.uk/2011/08/10/jeromecomet-hack-day-fun-in-the-fens

Usability

The Jerome project team believed that making data 'open' was not just about licensing, but also about providing usable data[11]. This focus led to the decision to support JSON and RIS ahead of MARC, although the question of whether to support MARC was discussed[12].

The Jerome project also looked at some innovative approaches to the search interface, most notably the 'mixing desk'[13] interface in the Jerome application which allows users to amend their search based on the record source/type.

Impact

Licensing

The data in the Jerome application is generally licensed as CC0. Data from the JournalTOCs service was licensed as CC-BY on the understanding this was a requirement from JournalTOC, although there is a lack of evidence that this is the case. While some of the library catalogue records will have been sourced from third parties (notably the British Library), these are being enhanced via Open Library and also transformed from MARC format to JSON as described above (and not currently made available as MARC).

Benefits

The Jerome application will be used to power a new 'reading list' solution for the University of Lincoln (replacing an existing commercial offering).

Further possibilities for exploiting the Jerome application as part of a 'production service' are being discussed. These include:

- As a tool to help improve metadata (e.g. via reports for cataloguers as described above).
- As a 'kiosk interface' for resource discovery for use in the library (i.e. an on-site OPAC replacement)
- To provide personalisation features for resource discovery

Business Case

The decision to invest in the Jerome project comes from some key institutional drivers. Notably the 'Student as Producer' ethos and the newly formed LNCD. It also links in with Lincoln University's 'open data' work[14].

The project team emphasised that improving the 'search experience' for students was a key driver.

Outcome

Jerome was an existing 'un-project' before receiving JISC funding, and returns to this status post-funding. As such it has a low level of resource within the institution. However, the team believe that by embedding the use of the Jerome application into production services, such as the Reading List system mentioned above, sustainability will be increased.

11 http://jerome.blogs.lincoln.ac.uk/2011/07/27/following-in-jeromes-footsteps

12 http://jerome.blogs.lincoln.ac.uk/2011/04/21/three-quarks-for-muster-marc

13 http://jerome.blogs.lincoln.ac.uk/2011/03/27/mixing-it-up

14 http://data.lincoln.ac.uk

Lessons Learned

- Document based storage and NoSQL approaches have significant potential for library and related data
- Given a committed team of experts, small amounts of resource can produce innovative services.
- Existing identifier systems (such as ISBN) can be used to link library entities without using Linked Data with RDF.
- Taking a non-Linked Data approach may reduce the priority of linking entities within datasets, especially where there are not existing identifiers
- Linking across datasets can normalise data and offer quality assurance mechanisms for local data.
- For some UK HE institutions there are few licensing issues associated with publication of bibliographic metadata.

See Also

Related projects / developments:

- COMET - http://cul-comet.blogspot.com
  Alternative approach to publishing bibliographic metadata
- VuFind - http://vufind.org
  Open Source 'next generation' library search interface
- XC Extensible Catalog - http://www.extensiblecatalog.org
  Open source software suite for libraries including user interface, metadata management and library system connectivity tools