discovery project review

SALDA

http://blogs.sussex.ac.uk/salda

SALDA 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

SALDA (Sussex Archive Linked Data Application) aimed to extract metadata records for the Mass Observation Archive from the University of Sussex Special Collection's Archival Management System (CALM) and convert them into Linked Data that would be made publicly available. The data and the knowledge gained will be fed into The Keep project[1] which brings together East Sussex Record Office, Brighton and Hove Council and the University of Sussex Special Collections under one roof in a purpose built archive repository requiring a unified discovery system.

Institution	University of Sussex
Responsible group	University of Sussex Special Collections is part of the library service
Capacity	Whilst the library team has key technical competencies, they had no prior experience of creating Linked Data and the associated mechanics of data transformation, URI allocation and discovery services
Data Scope	The Mass Observation Archive (MOA)[2] specialises in material about everyday life in Britain. It contains papers generated by the original Mass Observation social research organisation (1937 to early 1950s), and newer material collected continuously since 1981. The data used in this project refers specifically to the original phase of Mass Observation (1937-1950s). The MOA is in the care of the University of Sussex.
Data Scale	The dataset contains almost 67000 encoded archival descriptions (EADs).

Mechanics

Metadata enhancement

Enhancement

When Sussex imported the MOA into its CALM database in summer 2009 most information went into the title field, so there were no separate fields for 'date' or 'description' and they don't use access points. Logically speaking, this is not a good starting point for generating Linked Data so some ingenuity was required to determine an efficient and affordable way to augment the content with access points and to highlight key elements such as dates.

However, there are some associations implicit in the MOA data, and there are some 'hooks' in the data, which can provide the basis for generating explicit associations in the RDF data. The team identified 28 names out of the data in authorised form using National Register of Archives[3] rules, and around 100 keywords that appeared in the data and covered subjects from 'air raids' to 'sex', including places and organizations, events and wider concepts like 'class', 'family' and 'education'.

A second approach has been to scan the content of some EAD elements for words or phrases that can be mapped to specific entities (concepts, persons, organisations, places). This takes advantage of the fact that the MOA collection has a fairly well defined context.

[1] http://www.eastsussex.gov.uk/leisureandtourism/localandfamilyhistory/esro/thekeep/default.htm

[2] http://www.massobs.org.uk

[3] http://www.nationalarchives.gov.uk/nra/default.asp

Guides

As a result of this work, an unexpected result of the SALDA project was a review of the University of Sussex archival cataloguing procedures and the following guides were produced:

- CALM_ISADG_Collection level – maps the required ISAD(G)[4] fields to the CALM fields with guidelines on how to populate the fields and indicating the fields required for export to EAD
- Cataloguing procedures component level – guidelines for completing component level records in CALM

Vocabularies

The MOA data now references terms from (amongst others) the following RDF vocabularies:

- http://purl.org/dc/terms/
- http://xmlns.com/foaf/0.1/
- http://www.w3.org/2004/02/skos/core#
- http://www.openarchives.org/ore/terms/
- http://linkedevents.org/ontology/
- http://data.archiveshub.ac.uk/def/

Technical approach

To take EAD from the Calm archival system and turn it in to discoverable and reusable Linked Data, the project exported EAD/XML from CALM and then followed a 4-step approach, using open source tools:

1. Develop XSLT script, working with Pete Johnston of Eduserv[5]
2. Use Saxon to transform the EAD/XML in to RDF/XML
3. Convert the RDF/XML to RDF/N-Triples using Raptor
4. Use the Pynappl interface to upload this to the Talis Platform.

Chris Keene from the University of Sussex library team captured the sense of adventuring in to the unknown in his blog post: "This is a big step. Until now our process looked something like this: Export EAD data > send it to someone else > Magic > Linked Data." The process that unfolded turned out to be efficient, effective and, most important, sustainable within the Sussex team.

Process

Step 1: Develop the XSLT script. Pete Johnston of Eduserv drew on his work in the LOCAH project[6] (which is transforming EAD data aggregated by the JISC funded Archives Hub[7] into Linked Data) to provide the 'magic' part, much of the complexity being hidden in an XSLT script to process XML in to an RDF schema.

This provided significant opportunities to consider metadata enhancements that would benefit both the source catalogue and the resultant Linked Data, as explained above.

Step 2: Convert EAD/XML to RDF/XML using Saxon HE XSLT (Java) to do the transformation. This is a black box operation using software that is easy to download and setup.

Step 3: Convert the RDF/XML in to the alternative RDF N-Triples format (and also Turtle) using the Raptor RDF parser. Whilst the most common method of writing RDF is using XML, it is very verbose and can be difficult to read. N-Triples is considered easier to read as each line contains a self-contained Triple (i.e. its subject, predicate and object, mostly expressed as URIs). This also allows the data to be split into smaller files for uploading to the Talis Platform.

[4] http://archiveshub.ac.uk/isadg

[5] http://www.eduserv.org.uk

[6] http://blogs.ukoln.ac.uk/locah

[7] http://archiveshub.ac.uk

Step 4: Upload N-Triples files to the Talis Platform, which is a well-established Triple Store. While you can run your own Triple Store, this provides a stable, robust and quick solution. You interact with the Platform with standard HTTP Requests; GET, POST, DELETE etc. However an interactive command prompt front end allows you to simply specify the store you wish to work with, authenticate, and then use commands such as 'store filename.rdf' to upload data. A simple script can upload all the data to the Platform.

To sample the results, you can test the Sparql interface at: http://api.talis.com/stores/massobservation/services/sparql

Try a query such as:

```
SELECT * WHERE {

?a ?b <http://data.lib.sussex.ac.uk/archive/id/concept/moa/religion>

}
```

Reusability

The Process – Pete Johnston of Eduserv has reflected on how the SALDA use case has opened up new possibilities for the transformation code developed for the Archives Hub in the LOCAH project: "What is interesting is how we've "specialised" the fairly "general" LOCAH approach, based on "local knowledge" of specific characteristics of the MOA data. While it's perhaps premature to draw general conclusions from this single case, I should probably think about how this is reflected in the [LOCAH] transformation process."

The Data – The resulting Linked Data is re-usable in a number of ways, as raw RDF and through the Sparql end-point provided by the Talis Platform. The opportunities for re-use form part of the benefits case developed by the archive and its local partners in the Keep project.

Impact

Licensing

The Sussex team chose the MOA catalogue data for its initial demonstrator of linked data as they were confident of its provenance so could make it freely available under the ODC-PDDL license[8]. This should allow the greatest flexibility for people wanting to use the data and fits the Discovery principles[9].

Certainty of provenance being within the organisation (or a known 'friendly' party) is more typically the case with archival descriptions than library catalogues. This makes it much easier to commit to licensing as open data than in cases where there may be other parties who could theoretically step forward to assert rights at any time. In this respect Sussex was in the same position as the Oxford Text Archive[10] and the AIM25 consortium[11] and unencumbered with the other considerations set out in those case studies.

Business Benefits

The benefits of the first steps of any institution involving open licensing and linked data are hard to quantify. The Sussex team has focused on three core benefits:

Enabling discovery and reuse – For the MOA, like any specialist archive of research value and public interest, the potential uses of the data outside of the archive reading room are most attractive. They have identified potential for mobile applications and contributions to other projects – for example, Brighton and Hove Museums already contribute to Culture Grid.

[8] http://opendatacommons.org/licenses/pddl/1-0

[9] http://discovery.ac.uk/principles

[10] http://ota.ox.ac.uk

[11] http://openmetadatapathway.blogspot.com

Establishing the agenda – The success of this project (both the visible outcomes and the delivery to time and to budget) also means that open data has been established on the strategic agenda in the university library.

Enhancing descriptions – The MOA recognised the need to future proof its data so that it can be efficiently and flexibly exported, transformed or mapped across to other systems. The Linked Data transformation process enabled core weaknesses, which are typical of a catalogue used in isolation, to be identified and addressed. The project therefore compiled cataloguing guidelines to ensure that all collection level records are ISAD(G) compatible and that certain fields are always populated in component records.

Outcome

The SALDA Project has produced the following:

- The full catalogue data of the Mass Observation Archive is now available on the Talis Platform licensed under ODC-PDDL.
- Simple text search – http://api.talis.com/stores/massobservation/items
- Sparql interface – http://api.talis.com/stores/massobservation/services/sparql

The SALDA XSLT stylesheet is here licensed under modified BSD licence

- The direct link to the SALDA produced data from the Mass Observation Archive – http://data.lib.sussex.ac.uk/data/mass-observation
- You can download the data in RDF
- The dataset is registered on CKAN

Sustainability

Strategy

Sussex is going to look at its collections and make a priority list of ones where the catalogue data could be turned into Linked Data by considering:

- Whether the data can be licensed under ODC-PDDL
- What changes / additions we need to make to the data and it's structure
- The potential uses / benefits

Web pages for open data have been created on the library website to keep open data on the agenda, reflecting the strategic goals of the Library e-strategy 'Search and discovery 2011-2015'.

Metadata

Archival metadata changes and is used in different ways than metadata relating to such as publisher products in library catalogues. Consequently transformation to other formats is unlikely to require frequently repeated iteration. Nevertheless, the transformation processes have been designed with efficient updating in mind.

Linked Data Structures

The project made key decisions about the sustainability and future uses of the data with sustainability in mind. SALDA used 'Designing URI Sets for the UK Public Sector'[12] as a guide for creating URIs. The choice of the stem of http://data.lib.sussex.ac.uk keeps a simple URI, and allows Sussex to merge in other datasets in to the same 'pool'.

[12] http://www.cabinetoffice.gov.uk/resource-library/designing-uri-sets-uk-public-sector

Platform

The hosting of the http://data.lib.sussex.ac.uk namespace does not present any difficulties for the university. Meanwhile the decision to use the Talis Platform addresses challenges of scale and availability for what will hopefully be a growing volume of linked open data released by Sussex.

Lessons Learned

Project Manager Karen Watson was asked what advice she would give if others wanted to take a similar approach:

- **Get your data ready –** "Regardless of whether Linked Data is the future, we are working on our catalogue data to make it more structured and therefore ready to export to other formats and more portable. This is time consuming, so there are benefits in a targeted and time limited process, as was imposed by this project."
- **Are you in a position to licence your data? –** "We chose the MOA catalogue data for our initial demonstrator as we were confident of its provenance so we could make it fully open and available under ODC-PDDL. This hopefully will allow the greatest flexibility for people wanting to use the data and fits the JISC Discovery principles."
- **Find out about other similar projects –** "We anticipated the value of our SALDA blog posts to anyone wanting to do a similar project. Likewise, we followed in the footsteps of the LOCAH project and were able to use their stylesheet[13] and experience in transforming archival data into Linked Data."
- **Find real examples of Linked Data in human readable format –** "This will enable you to show stakeholders and colleagues what it is that you are on about. At the beginning, no one (management, colleagues, friends) will know what you are talking about when you mention Linked Data. I use the BBC wildlife pages[14] and how they link to Animal Diversity Web[15]."

Some human readable examples of the SALDA data are at:

- http://data.lib.sussex.ac.uk/archive/doc/person/nra/harrissonthomas1911-1976anthropologist
- http://data.lib.sussex.ac.uk/archive/id/archivalresource/gb181SxMOA1

See Also

- **LOCAH project –** http://blogs.ukoln.ac.uk/locah/about/
- **Talis Connected Commons –** http://www.talis.com/platform/cc/
- **Pete Johnston's posts on indexing and transformation –** http://blogs.sussex.ac.uk/salda/2011/05

[13] http://data.archiveshub.ac.uk/xslt/ead2rdf.xsl

[14] http://www.bbc.co.uk/nature/life/Lion

[15] http://animaldiversity.ummz.umich.edu/site/index.html