discovery project review

AIM25 Open Metadata Pathway

http://openmetadatapathway.blogspot.com

AIM25 Open Metadata Pathway 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 OMP project was designed by AIM25 to test the value of Linked Open Data in enhancing the archival cataloguing process (particularly by embedding indexing based on authority files, such as Open Calais[1] and the UK Archival Thesaurus(UKAT)[2]), and therefore improving the user experience by speeding up archivists' ability to deliver indexed cataloguing thus improving web-based discovery and enhancing term based searching and browsing.

Partnership	AIM25 (www.aim25.ac.uk) provides electronic access to collection level descriptions of the archives of over 100 higher education institutions, learned societies, cultural organisations and livery companies within the Greater London area.
Responsible group	Kings College London (KCL) and the University of London Computer Centre (ULCC) are leading the OMP project.
Capacity	AIM25 has a long track record in the development of its shared archival services, including collection descriptions and associated thesauri and authority records. ULCC has a technical team experienced in working with web services, Linked Data, automated data enhancement and associated user interface design.
Data Scope	The project is addressing the complete AIM25 dataset of collection descriptions.
Data Scale	The dataset contains descriptions of collections from 123 archives in the London area.

Mechanics

Aims

OMP set out to provide cataloguing and browsing interfaces which:

1. Extend the usability of existing offerings
   a. Eliminate the need for archivists to use mark-up
   b. Integrate the indexing process with the metadata recording process
2. Make use of semantic annotation within the cataloguing template
   a. Analyse the textual input against authoritative external and internal sources
   b. Suggest and record indexing terms derived from the analysis

[1] http://www.opencalais.com

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

3. Use Linked Data to enhance the user experience of the AIM25 website
    a. Mark-up the indexed terms within the ISAD(G)[3] display
    b. Add further indexed terms rigorously constructed according to NCA rules and/or derived from UKAT
    c. Provide links to related services based on the semantic properties of the terms

Metadata enhancement

The AIM25 archival descriptions are based on ISAD(G) fields and are stored in EAD format[4]. The records have been enriched through lookup links with thesauri and authorities, notablyUKAT, MeSH[5] (medical), Library of Congress[6] and Getty[7]. Given this baseline, the OMP project addressed two key metadata considerations:

- Which ISAD(G) fields to include in linked data indexing? Focusing only on 'Scope' and 'Content' would be too limiting bearing in mind the use of the 'Administrative & Biographical' and 'Related Records' fields in some catalogues. The solution was to allow archivists to configure the process by selecting the relevant fields for their catalogue from a template.
- How to enable dynamic indexing? Archivists need to select terms from one or more authorities in a real time manner rather than as a secondary lookup process. The solution was to investigate Linked Data vocabularies, notably Gate[8] and Open Calais, leading to the selection of the Open Calais service (OCS)[9] for the project.

Technical approach

URI Scheme and RDFa

The transformation of the catalogue to linked data format required the adoption of a suitable URI scheme, which has been based on the LOCAH model. This involved defining a data namespace for AIM25, binding to http://data.aim25.ac.uk. These URIs are then used in identifier attributes for EAD elements and thence readily transformed into an RDFa format for the Web-based HTML rendering of the AIM25 catalogues. Publication of linked open data is in SKOS format.

Linked Authority Data

The project added the dynamically generated output from OCS, leading to a modified version of EAD with the OCS output embedded in the content. Browser-side scripting to the original HTML pages was required to highlight terms identified by Open Calais in the user interface.

The above uses the OCS dynamically. Similarly, it would be highly beneficial if other authorities such as Library of Congress Subject Headings were also available as a service to be incorporated in this manner for editing, validation and discovery.

Data Entry

For archivists, the process of linked data indexing is embedded in the standard cataloguing process. A mouse 'rollover' feature allows users to match a term in a tick box drop down menu, connecting to one or more external authority services including GeoNames[10] (e.g. Is 'London' the capital of the UK, a place in Canada, an author or part of a corporate name?). Triples can also be interrogated in rollovers in order that the editing archivist might validate or clarify these entities. The same technical approach is also use to validate the results and to enhance the user display and navigation.

[3] http://www.icacds.org.uk/eng/standards.htm

[4] http://www.loc.gov/ead

[5] http://www.nlm.nih.gov/mesh

[6] http://id.loc.gov

[7] http://www.getty.edu/research/tools/vocabularies/index.html

[8] http://gate.ac.uk

[9] http://www.opencalais.com

[10] http://www.geonames.org

The automated enhancement of the existing data entry processes for AIM25 involves dispatching the EAD-compliant data entered to Open Calais and returning the data, with enhanced mark-up, for checking by the submitter. Importantly, this process can be repeated through periodic data dumps into Open Calais or automated calendared refreshes.

Architectural Considerations

The OMP team paid considerable attention to the scalability, robustness and extensibility of the service and its underlying architecture.

- AIM25 is using the free Open Calais web service[11], which allows 50,000 transactions per day per user and far exceeds loadings in the archival domain.
- When processing the very largest text blocks, there are challenges with the performance of the client machine, suggesting that more of this post-response processing should be pushed over to the server.
- A move to more server-side processing would also improve extensibility of the framework.

Reusability

Being based on ISAD(G) fields and the EAD data format, the software solution for cataloguing, validation and browsing has significant potential for reuse beyond the AIM25 catalogue environment.

The AIM25 data itself has always been reusable, as illustrated in the existing relationships between AIM25 collections data and such as the Archives Hub[12] and the Women's Library[13]. The introduction of Linked Data with integrated open licensing has expanded the potential through machine-to-machine discovery, leading to increased access and reuse.

Impact

Licensing

Choice

An open licensing scheme will be adopted by the partner institutions prior to the Linked Data service launch.

Linked Data requires licensing appropriate to its intended uses. As the compilation of databases is not regarded as a creative act under at least US law, the OMP team took a view that the Creative Commons licence may not be most appropriate for licensing open linked data. Instead, the Open Data Commons licences[14] appear to be an appropriate choice. The Open Database License (ODbL) is the preferred option as it allows sharing of the data provided it remains attributed (like ODC-By) and any adaptations are distributed under the same licence.

Implementation

A statement is being linked to all descriptions reflecting AIM25 as the origin of the data, that AIM25 is a partnership in which contributing members have rights, but that the data is otherwise freely accessible for reuse. The statement will reference the use of an Open Data Commons licence.

Thesaurus support must be addressed. AIM25 expects to gain approval for inclusion in an open licence in this context of UKAT (which is based on UNESCO), MeSH, plus Gay and Lesbian and other vocabularies developed by partner organisations. There are also downstream challenges, including how to implement and assert relationships between AIM25 entities and other open datasets, covering such as matching entities.

[11] http://www.opencalais.com/documentation/calais-web-service-api

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

[13] http://www.londonmet.ac.uk/thewomenslibrary

[14] http://opendatacommons.org/licenses

Business Benefits

Linked Data Benefits

A key question for the OMP archivists' focus groups was whether the improvements offered through Linked Data sufficient to justify the extra effort required from staff? The project has begun the process of demonstrating the value of Linked Data approaches in a number of ways:

- Improved the front-end user experience in terms of discoverability and browsing.
- Increased return on investment in cataloguing (speeding up cataloguing, enabling archivists to locate and link information including existing authority records, such as places from GeoNames[15]).
- Enhanced ability to justify expenditure on services and resource development (improved web-hits and connecting with heavily used services).
- Exposure of information to novel and different uses (improved interoperability with other domains and services, including commercial services).

Supporting Workflows

To achieve those goals, archivists need practical tools that they can plug into their workflows without fuss. Linked Data provides the opportunity for more automation and speedier indexing.

Archivists are also more likely to embrace Linked Data if they can painlessly re-index their current content or validate metadata created out of mass digitisation and optical character recognition (OCR), automatically enhancing the records against Open Calais and other specialised and global authorities.

Enhancing Discovery

The potential for cross-domain services is a further attraction. AIM25 archives are often embedded in libraries, museums and galleries, generating both institutional and user expectations of seamless discovery, fuelled by the emergence of cross-searching tools.

- The use of Linked Data ontologies within the system will provide opportunities to associate AIM25 records automatically and intelligently with other information resources.
- It will allow other information resources to locate and link to archives information in AIM25, enhancing discovery, and supporting the aggregation of AIM25 data into dynamic searches and aggregators across the sector.
- At the user interface, roll-over features and drop downs make Linked Data of mainstream value rather than a preserve of technical experts.

Measures

Feedback from archivist focus groups suggests that the editing and validation interface will reap the desired benefits. The automated processes devised by ULCC are currently running without additional staffing overheads and unnecessary cross-checking. Initial evidence suggests that the exposure of Linked Open Data to the web has generated increased access to AIM25 records.

ach of these measures needs to be validated with further evidence in coming months through the ongoing commitment of AIM25 to the collection of management information, which will be made available to the community in support of this work.

[15] http://www.geonames.org

Outcome

The OMP project proposed to investigate the technical possibilities and operational benefits of converting the AIM25 collection descriptions to Linked Data.

It has in fact been possible to develop a sufficient level of automation and integration that the approach has been applied to the complete working service for back end cataloguing, for public access and for open data reuse. This system can be supported by ULCC, the AIM25 technical partner at minimal additional cost.

The project has gone a long way to establishing that Linked Data enhancement works well when it is aligned with the grain of professional practice, embedded within existing cataloguing workstreams, adding supplementary metadata that builds on the work practices of archivists rather than imposing a new layer of work

Consequently it is a realistic expectation to link AIM25 archive descriptions with other services so users can connect and aggregate, 'mix and mash' this data in a richer context, cross-walking with content generated and used by wider audiences

The consortium is therefore expected to adopt the approach later in 2011.

Lessons Learned

- A shared aggregation service such as AIM25 provides a powerful platform for developing and validating practice.
- Automation is essential for economic and time efficient conversion to linked data, including embedding in standard archival workflows.
- Assessment of benefits is feasible but will require monitoring over an extended period.
- Input from archive vocabularies such as UKAT is needed to enrich the Open Calais corpus with archival terminology.
- Service based access to thesauri and authority files, such as Library of Congress Subject Headings[16] and the UK NRA codes, would add considerable value to the OMP approach.
- The design of the cataloguing interface and validation model may be reusable in the wider archival community.

See Also

- LOCAH – http://blogs.ukoln.ac.uk/locah/about
  URI scheme developed for the Archives Hub
- Open Calais – http://www.opencalais.com
  Linked data toolkit from Thomason Reuters
- SALDA – http://blogs.sussex.ac.uk/salda
  An alternative approach to enabling archival indexing

[16] http://www.loc.gov/aba/cataloging/subject