discovery project review

Discovering Babel

http://blogs.oucs.ox.ac.uk/martinw

Discovering Babel 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

Discovering Babel aimed to make the digital resources in the Oxford Text Archive easier to discover for potential users. This project aimed to enhance the OTA resource discovery mechanisms of the OTA in line with good practice in the research community and with the Discovery initiative[1].

Institution	University of Oxford
Responsible group	The Oxford Text Archive[2] (OTA) is managed by the University of Oxford Computing Service (OUCS)
Capacity	OTA has technical capability through OUCS. Experience of the emerging web discovery and aggregation environment has been developed through its participation in the Arts and Humanities Data Service[3] (AHDS - up to 2009) and in the global linguistics and digital humanities research communities.
Data Scope	The Oxford Text Archive is contains literary and linguistic resources in electronic form. Many of the resources are the outputs of projects funded by the British Academy, AHRC[4] and JISC.
Data Scale	The dataset contains approximately 1400 electronic resources, with standards-based TEI[5] encoded metadata forming part of each resource file.

Mechanics

Metadata enhancement

The metadata descriptions of OTA electronic resources inform potential users about the resource, including its title, a summary of the content, where the electronic resource came from (its provenance), technical formats, types of annotation, any restrictions on use, etc.

The source metadata is encoded in the XML file containing the resource itself, according to the guidelines of the Text Encoding Initiative (TEI). In the area of literary and linguistic computing, the TEI Guidelines are a widely recognised reference point and standard for the encoding of data and metadata. The metadata for OTA resources is therefore in the form of a TEI Header.

The work on making this metadata more visible, and on transforming it into other formats revealed areas where it was necessary to update the existing metadata. For example, the language of a resource was missing in some cases, usually where the language was English, and was perhaps considered the default value in the past – a common issue when opening up metadata.

1 http://discovery.ac.uk
2 http://ota.ox.ac.uk
3 http://www.ahds.ac.uk
4 http://www.ahrc.ac.uk
5 http://www.tei-c.org/index.xml

Technical approach

The technical approach was focused on making the OTA catalogue data available in new ways, which involved two areas of automation:

- **Making the catalogue records available** to be collected by online resource discovery services
- **Transforming the metadata** into a variety of different formats for the different services of importance to the research community

Making the records available

Before Discovering Babel, the OTA metadata was available only in abbreviated form in the catalogue list on the website, and on the web pages for each resource, or in full when a user downloaded the resource. An important step towards wider discovery, especially through aggregations, was therefore to make the full metadata available for online services to harvest it, using the most widely used protocol for this purpose, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[6].

This required the following steps:

- Add the appropriate Apache and Perl modules to the web server to allow OAI-PMH queries to the new web service;
- Implement crosswalks (using XSLT) from the metadata in TEI Header format to the formats with traction in the research community: i.e. currently Dublin Core with Open Language Archives Community (OLAC) extensions, with a view to adding the emerging CLARIN Component Metadata Infrastructure element set (CMDI);
- Register as a metadata provider with relevant aggregators;
- Set up procedures to ensure the ongoing availability, persistence, maintenance and updating of the OAI-PMH service with records harvested from http://ota.oerc.ox.ac.uk/oai2/XMLFile/ota/oai.pl.

Transforming the metadata to different formats

OTA initially wrote the crosswalks using XSLT 2.0 using the saxonb-xslt package, but found that the performance was too slow for the harvesting services. The team therefore ported the code to XSLT 1.0 using the xsltproc package on Ubuntu, which enabled the harvesters to operate. OTA plans to investigate these issues further with other CLARIN centres to see if future improvements to the performance can be achieved.

OTA considered the further option of moving to a servlet-based solution – for example, a Java-based OAI implementation (jOAI) deployed on a Tomcat Server. However it was assessed that this would raise an additional maintenance overhead, with risks to robustness and sustainability.

Reusability

OTA chose to make metadata available in a number of formats via OAI-PMH, to fit the expectations and requirements of a number of harvesters relevant to the linguistic research field, notably in Europe (CLARIN[7]) and in North America:

- OLAC[8] – as used by the North American research community including such as University of Virginia and BYU; this requires Dublin Core, with OLAC extensions.
- CMDI[9] – OTA will provide CMDI metadata for the CLARIN aggregator, but this format has not yet

6 http://www.openarchives.org/pmh
7 http://www.clarin.eu/external
8 http://www.language-archives.org
9 http://www.clarin.eu/cmdi

achieved sufficient maturity and stability. In the meantime, the CLARIN aggregator is harvesting OLAC metadata, and in this way they are presenting OTA resources in the Virtual Language Observatory service[10].

- The project also automated production of RDF so the records can be published as Linked Data. However, the target community is currently focused on OLAC and CMDI and therefore Linked Data does not have traction with the key aggregators, notably CLARIN. Furthermore, the establishment of the authority terms essential to make effective use of linked data is subject to current ISO work, though CLARIN is already using ISO Concepts Registry and Language codes[11].

Impact

Licensing

In order for users to be able to find, evaluate and reuse the resources, good descriptions of their nature and context are necessary. It is usual in the domains using language resources for the descriptions to be made freely available, but usually there is not a specific and clear statement of the terms under which they are made available.

The project recognised that, in order to maximise use, it is better to assign a specific open access licence to all metadata records. The licensing decision had to cover two scenarios:

- The datasets to which the metadata refer are not owned by the OTA, but the OTA has permission to make the resources available subject to a user licence, which restricts use to exploitation for the purposes of education and research. In these cases, because some OTA resources are TEI XML documents, with the metadata embedded in the header, it was necessary to apply a single licence to both metadata and data. For this purpose OTA has selected the Creative Commons Non-Commercial Share-Alike approach.[12]
- In cases where just the metadata is available, for example as a catalogue or to metadata harvesters, OTA will apply the least restrictive Creative Commons CC0 licence.[13]

Business Benefits

The Research Community

The research community wants to enhance discovery and to enable reuse, recognising benefits that will arise from addressing core challenges:

- How can researchers make their language resources easier to discover and use, minimising duplication of effort, so others do not have to spend time and resources on creating their equivalent resource?
- Can ease of discovery and access contribute to testing, refining and building on research results, and is necessary for the verification of research findings and interpretations in many scientific domains?
- How can users be helped to find corpora and other digital language resources? (That is unlikely to involve a one-stop-shop as the CLARIN and OLAC services both have mass).
- Can the content of language resources be described in ways that help users to compare them, and find the right ones for their research?
- Once we discover what we want, how can we make it easier to use and to combine language resources? For example, can we create virtual research environments for corpus users? This goes beyond the discovery of resources to providing services and tools.

10 http://www.clarin.eu/vlo
11 http://www.lat-mpi.eu/latnews/tag/cmdi
12 http://creativecommons.org/licenses/by-nc-sa/2.5
13 http://creativecommons.org/publicdomain/zero/1.0

Tools are typically limited to sets of resource fragmented across local silos – exemplified in the UK by services offered by the OTA and the University of Leeds (IntelliText).[14]

Enhancements such as those achieved by Discovering Babel and aggregations such as the CLARIN VLO will not only facilitate discovery, but should also be key steps towards enabling enhanced access, where standards-conformant web services can perform operations on them, so they can be deployed in virtual research environments, which will almost certainly be distributed and gadget-enabled, open and social.

Finally, it should be emphasised that these open developments based on harvesting and perhaps eventually enriched as linked data will be particularly important for interdisciplinary research, where the community aggregation may not hold the same position. The OTA approach recognises that Linked Data is likely to form part of the longer-term solution. However the current lack of taxonomies is an inhibitor, though essential authorities are under development, as noted above.

The University

These changes to the OTA technical set up will help to connect together more services, potentially lowering maintenance costs, sharing facilities, spreading expertise and enabling new services. Furthermore this infrastructure is hospitable to the types of inter-disciplinary and international partnership that is central to research. Such models are applicable elsewhere in the university. Deploying these technologies will therefore help the university to acquire skills and processes that will be transferable to other projects and disciplines.

Outcome

OTA has implemented and embedded sustainable procedures to ensure that:

- Each resource in the collection is assigned an http URI
- Each URI is registered as a persistent identifier with the EPIC Handle Service
- Each URI is resolved to a machine processable resource with relevant metadata
- Crosswalks for metadata are implemented to DC with OLAC enhancements, CMDI and RDF

Metadata is licensed as openly as possible, in line with the Discovery principles[15].

Sustainability

OTA has considered sustainability in terms of the service, the data and the wider mission.

Recognising that the biggest risk is usually the sustainability of the outputs, the project worked on embedding its work in ongoing production services which are part of Oxford's core institutional IT infrastructure. Furthermore, as a direct result of being harvested and exposed within the CLARIN aggregation, the data itself is now more strongly embedded in the global research infrastructure. From this experience, the project suggests a sustainability checklist:

Who is responsible for

- The overall service
- The infrastructure on which the service depends
- The technology including code and scripts
- Human resources (server maintenance, user support)

14 http://corpus.leeds.ac.uk/it

15 http://discovery.ac.uk/principles

What will the situation be in 1, or 2, or 5, or 10 years time?

- What happens if you (or other key persons) leave or take on a different role?
- What happens at the end of the current round of funding?
- Would it be better to move the service to another institutional home?

The project emphasises that this work is not an end in itself, bearing in mind that these developments open up the challenges, as set out above, of integrating the metadata, the data and the necessary tools in the wider and necessarily international virtual research environment.

Lessons Learned

- Adding value to the research community and the effectiveness of the individual researcher are the compelling business drivers
- International considerations may therefore 'trump' local UK strategies in terms of such as the place and the nature of aggregation – though the Discovery principles should stand good.
- The latest technology versions are not necessarily the best when taking account of such as performance factors within complex service environments (such as harvesting).
- Sustainability of repeated technical and editing processes needs to be considered from the outset of enhancement.
- The real goal of Discovery may not be to enable aggregation (which is nevertheless a useful tactic in the current environment) but to optimise 'exposure' of metadata and data for an increasingly distributed set of purposes.

See Also

- CLARIN, the Common Language Resources & Technology Initiative - http://www.clarin.eu
  The CLARIN Virtual Language Observatory is collecting and making available in a single place the information about language resources from all around Europe - http://www.clarin.eu/vlo
- OAI-PMH, the Open Archives Initiative Protocol for Metadata Harvesting -
  See the OAI-PMH 'Beginners Guide' at http://www.oaforum.org/tutorial
- OLAC, the Open Language Archives Community - http://www.language-archives.org