

Hybrid Archives

Survey of existing OAI-PMH/Z39.50 tools

Document Details

Document type: Project Report
Author: Tony Austin
Draft / Version: Version 1.0
Date of edition completion: 17/06/2004

Contents

Contents	2
Overview.....	3
Case Studies	6
AHDS cross search catalogue	6
OAster	6
HEIRPORT/CIE/Arena	7
Archives Hub	7
Cross-Domain Resource Discovery Project ("Cheshire")	7
Discussion	8
Bibliography.....	8
URLs.....	9

Overview

The Open Archives Initiative OAI-PMH¹ and Z39.50² protocols were both conceived within the library community in order to facilitate the sharing of metadata and data therein described. Sharing or interoperability of metadata requires conformance to shared schema. The most useful schema are likely to be international standards as these will be widely supported. Libraries have a number of standards for cataloguing published materials; for example, MARC³ or MACHine Readable Cataloguing. MARC is community specific. Recent developments aimed at using OAI-PMH and Z39.50 for wider cross-community interoperability have largely looked toward the Dublin Core Metadata Initiative⁴ (DCMI). Unqualified Dublin Core (DC) is the default schema for OAI-PMH in that support is required for conformance to the protocol. Similarly, the Bath⁵ (at functional level C) and CIMI⁶ Z39.50 profiles, both international standards, require support for unqualified DC as a record syntax.

DC has not been free from adverse criticism having been consistently described as simplistic; however, as Weibel *et al* (1997) note it has been '*aimed at intermediate precision and high interoperability*'. More specific weaknesses have also been suggested by Godfrey Rust, for example, who sees DC as fundamentally flawed as well as being simplistic especially in terms of rights management (1998). Semantically; however, it is the generality and simplicity of DC that facilitates interoperability between diverse data sets. Whilst its nature precludes the presentation of complex data structures it has seen increasing take up for resource discovery. Thus there is a trade off in terms of quality of data versus interoperability.

The advantages and disadvantages of using a relatively simplistic schema were clearly demonstrated in a recent project to develop a Common Information Environment⁷ (CIE) demonstrator (Austin *et al*, forthcoming). The project involved harvesting DC metadata using both OAI-PMH and Z39.50. OAI-PMH harvested datasets were set up as localized Z39.50 targets so that they could be searched in conjunction with remote targets. The demonstrator was developed from an existing platform, the Historic Environment Information Resources Portal⁸ or HEIRPORT (Austin *et al*, 2002, Pinto *et al*, 2002). HEIRPORT is based on two Java Application Programming Interfaces (APIs); Zava, a Z39.50 server or target, and Zavax, a Z39.50 client or portal. The system requires conformance to a number of international standards including the Bath and Cimi Z39.50 profiles and hence to DC. By definition CIE was aimed at a broader community than HEIRPORT. With relatively little effort the project was able to point at existing Bath (functional level C) compliant Z39.50 targets such as the Archives Hub⁹. Unqualified DC metadata harvested from existing OAI-PMH repositories was more problematic. For example, the date element

```
<dc:date>2001-10-10</dc:date>
```

is meaningless without a qualifier indicating 'last modified', 'record created', etc. Similar problems exist with other elements in that, for example., dc:creator needs qualifying with a role. There were further problems with data mappings. One repository contained 47,000 (out of 53,000) records with the same title making it impossible to meaningfully search on this element. The dc:title element had been generated by concatenating two fields, archaeological period and object type, from an underlying relational database of early medieval coins with the consequence that most titles read 'Medieval Penny'. The underlying data would have allowed the construction of more meaningful title elements, for example, 'silver penny of Ecgberht (765-780)'. The repository is currently being rethought as part of the JISC funded Harvesting the Fitzwilliam Project¹⁰.

Some of the above problems can be avoided or at least overcome with Z39.50 in that qualifiers held in a target dataset can be generated as part of the results set through suitable queries

```
SELECT CREATOR||': '||CREATOR_ROLE FROM CREATOR_TABLE WHERE <conditional clause>
```

The above is plausible as the underlying data is unchanged. This would not be the case with OAI-PMH where the harvested dataset would contain concatenated elements (this differs from constructing a single element such as title where this does not exist in the original dataset).

Some care has been taken to describe the problems that might arise from the use of OAI-PMH and/or Z39.50 as it will affect the rest of the survey. A number of generic points can be made

- Many tools default to the Unqualified DC standard
- Using such tools should increase opportunities for interoperability but reduce ease of resource discovery
- However, unqualified DC is insufficient for OAI-PMH harvested data (outside of specific contexts such as library data where date is known to = publication date and creator = author)
- Many tools support multiple schema but the need for an extended schema will complicate installation and configuration. It may also decrease opportunities for interoperability unless used within a closed context
- An extended schema will also increase the complexity of data mapping

It is also important to directly compare and contrast the protocols under consideration

**OAI
Advantages...**

Installation and configuration should be simpler

Overheads on data provider server is minimal

Disadvantages...

Data unlikely to be current as harvested at intervals

There is no mechanism for dealing with records deleted by a data provider apart from a complete re-harvest

Loss of service if data server is down (ie centralised service)

Centralisation means more records to search for each query as compared to a distributed model

May be copyright issues in that a non-local copy of data is made

**Z39.50
Advantages...**

Data is always current as querying sources directly

Distributed data sources linked to simultaneous searching should be more efficient (ie less records to search at each source)

There will still be a service even if one or more data server is down

Disadvantages...

Relatively complex installation and configuration

Target server need relatively robust system (server, database, etc)

Development of an extended schema within an OAI-PMH implementation may reduce its major advantage where installation and configuration are seen to be more straightforward. Conversely the development of Search/Retrieve Web Services or SRW¹¹ protocols may simplify the creation of gateways. SRW is built out of Z39.50 but effectively simplified through not retaining all the features of the latter (Storey, 2004).

Many toolkits are built using open source software such as Java or Perl and as such should be platform independent; however, the toolkits by definition need to interact with other applications such as a database which may have very different requirements in terms of protocols and data formatting. Most OAI toolkits are pre-configured for the more popular database packages such as Microsoft[®] Access or MySQL as is the case with OCLC's OAICat implementation. It became a major task to get this to work with an Oracle[®] database during a recent project (for example; Storey, 2003, 11). This is not a problem with Z39.50 as an abstract layer sits between Z39.50 requests and the database server.

Even a basic install of an OAI repository will require a wide skill base to be available both in terms of familiarity of technologies to be employed and the semantics of any standards or schema employed. For instance a default repository install will require a mapping of the underlying database to DC and an understanding of the query language used to extract data. A basic understanding of XML and related technologies, the implementation language and installation procedures will also be required. Z39.50 server installations can be more complex in that adherence to any number of standards is required in the form of profiles, attribute sets, record set syntaxes, tag libraries, etc, and metadata schema referenced therein; however, much of this can be pre-configured into a package. In being part of a distributed system a deeper understanding of network technologies and database connectivity will be required. It is possible to install such systems at remote sites where required expertise is not present locally. The main problems relate to security and levels of access. In general root or systems administrator privileges should not be required as long as underlying application software is in place, for example, a Java platform may be required.

Case Studies

The intention here is to present a small range of comparative use scenarios that have been implemented outside of mainstream bibliographic data handling using OAI and/or Z39.50/SRW protocols. Three criteria; implementation costs, interoperability and ease of resource discovery, have been subjectively graded for the purposes of comparison.

AHDS cross search catalogue

The collections level metadata underlying the AHDS cross search catalogue¹² was harvested using OAI from the various subject centres that comprise AHDS. Thus each data centre set itself up as an OAI-PMH v. 2 repository with the AHDS using the popular OAICat¹³ harvester. In response to the perceived shortcomings of unqualified DC, the default metadata schema for OAI, and to accommodate mapping to a shared schema for subject centre descriptions the AHDS developed a Common Metadata Format. CMF grew to a large number of elements in order to encompass data centres using widely different schemas locally (Polfreman, 2004). In addition to the problems already noted above in that that OAI-PMH repository implementations are often specific in terms of database support subject centres had to set up support for CMF.

Whilst this should be seen as an innovative project in moving away from using the default unqualified DC support supplied with repository implementations its development of a localised schema will limit its usefulness in terms of interoperability. It does; however, make the point that that OAI can be set up to work with other schema. It is also noted that CMF can be mapped to DC and the Research Support Libraries Programme (RSLP) Collection Description schema¹⁴ for OAI harvesting by other organisations.

Implementation costs: medium

Interoperability: low

Resource discovery: high

OAlster

The OAlster¹⁵ project of the University of Michigan Digital Library Production Services is included as an example of using the default OAI schema of unqualified DC. OAlster makes use repository and harvester software developed by The University of Illinois at Urbana-Champaign Open Archives Initiative (UIUC OAI) Metadata Harvesting Project. The packages are largely Microsoft[®] centric in that they use Active Server Pages (ASP) or Visual Basic (VB); however, OAlster uses no longer supported software developed for a Linux/MySQL platform (Wilkin, J. et al, 2003). OAlster is included here as an example of harvesting an unqualified DC subset of much richer metadata and dynamically linking back to the latter through the DC source element

URL <http://www.aim25.ac.uk/cats/19/88.htm>.

Thus the OAI harvested data is here seen as a stage in resource discovery. In this case the link is to a rich content record held by the AIM25¹⁶ (Archives In London and the M25 area) project where metadata conforms to an archival standard called ISAD(G), the General International Standard on Archival Description produced by the International Council on Archives¹⁷. OAlster then acts as a pointer to rich content description. It reduces complexity in terms of implementation at a trade off in ease of resource discovery in being multi staged.

Implementation costs: low

Interoperability: high

Resource discovery: low

HEIRPORT/CIE/Arena

These three projects are grouped as they all use the same underlying technology as has already been described above. The technology can legitimately be described as supporting Web Services as it uses HTTP for communication and hence is part of the second generation of Z39.50 implementations. Like OAIster the records in these projects provide links to richer content metadata when available through the DC Source element .

CIE addresses some of the problems of data aggregation, be it federated or distributed, where thesauri used by specific communities can and do differ. CIE uses spatial location as a shared knowledge classification system in an attempt to overcome such problems. This clearly worked well with the wide range of datasets used but not all datasets will or can be spatially referenced. The Archaeological Records of Europe Network Access Project or ARENA¹⁸ project looks at problems of thesauri within multi-lingual environments.

Implementation costs: high
Interoperability: high
Resource discovery: medium

Archives Hub

The Archives Hub is one of an increasing number of organisations using the Cheshire II Project¹⁹ application software which supports many current information retrieval standards including Z39.50 and an SGML database. The Cheshire II software was specifically adapted by its developers in order to support ISAD(G) compliant Encoded Archival Description or EAD²⁰ metadata schema which is a standard for encoding archival finding aids using SGML. As already noted the recent CIE project was able to target the Archives Hub Z39.50 server with little difficulty because of Bath profile conformance with presumably a mapping between the required support for DC and EAD.

Implementation costs: medium (for EAD compliant archives)
Interoperability: high
Resource discovery: medium (if via Z39.50)

Cross-Domain Resource Discovery Project ("Cheshire")

The JISC/NSF funded "Cheshire" Project²¹ has led to the development of a number of recently released toolkits known collectively as Cheshire 3²² although there appear to be no implementations as yet. Toolkits so far developed are a CQL Parser (Common Query Language), SRW client and SRW server with support for a number metadata formats including DC, MarcXML, MODS, OAI MARC and EAD. They use SOAP or HTTP as communication or carrier protocols. A range of operating systems are supported suggesting platform independence. The final report of the "Cheshire" project suggests that Cheshire 3 is likely to become a core part of the JISC information architecture.

Because of its newness and lack of existing implementations it is difficult to judge if it will be easier to set up than existing systems such as the Java APIs underlying HEIRPORT. The documentation for Cheshire 3 notes that additional installs are required including ZSI and PyZ3950. It also notes non-trivial tasks dependent on existing set up. It is worth mentioning that other SRW toolkits are also becoming available, for example the YAZ toolkit from Index Data²³.

Implementation costs: medium to high?
Interoperability: high
Resource discovery: variable depending on metadata format

Discussion

Whilst there are other examples of the use of OAI and Z39.50 outside of the handling of bibliographic data take up is currently small; however, it is clear that implementations of both protocols can have a role in an information environment. OAI appears more suited to contexts where limited technical resources are available and datasets are relatively static whilst Z39.50 is seen as technically more challenging but has other advantages such as the provision of dynamic datasets. Developments in two of the above projects; CIE and Cheshire 3, appear to recognise these contextual differences and have implemented or made provision for the integration of data collected through both protocols. In supporting similar technologies there should be possibilities for interoperability these systems.

As already noted simplicity in terms of metadata schemas increases the possibility interoperability but reduces the ease of resource discovery; in short a trade off. In an effort to ease this contradiction HEIRPORT/CIE/ARENA and OAIster respectively present relatively simplistic qualified and unqualified DC; however, where records are a subset of richer metadata held elsewhere provide links back to the source record and thus staged resource discovery. At the other end of the spectrum the AHDS provide a subset mapping to DC of their relatively rich schema for use by other organisations.

What developments there have been then suggest a mix of technologies or a hybrid system will be needed for a substantial information environment and that resource discovery may need to be multi staged in order to facilitate interoperability between components of such an environment.

Bibliography

Austin, T., Pinto, F., Richards, J. and Ryan, N. 2002. 'Joined up writing: an Internet portal for research into the Historic Environment' in G. Burenhult (ed.) *Archaeological Informatics: Pushing the Envelope CAA2001*, BAR International Series 1016, 243-51.(also online at <http://www.cs.kent.ac.uk/pubs/2001/1261/content.pdf>)

Austin, T., Kilbride, W., Fernie, K. and Richards, J. forthcoming. 'An experiment in Space: Towards a Common Information Environment', *Computer Applications and Quantitative Methods in Archaeology 2004*

Pinto F, Ryan, N. Austin, T. and Richards, J. 2002. 'An Interoperable Portal for the Historic Environment', in Neuhold, E. & Kalinichenko, L. (eds). *Proceedings of the Third DELOS Workshop: Interoperability and Mediation in Heterogenous Digital Libraries*, DELOS Network of Excellence on Digital Libraries Workshop Series, Darmstadt (also online at <http://www.ercim.org/publication/ws-proceedings/DelNoe03/12.pdf>)

Polfreman, M. 2004. 'One way of looking at things (the AHDS Metadata Framework)', *AHDS Newsletter – Spring 2004* (also online at <http://ahds.ac.uk/news/newsletters/spring-2004/index.htm#4> - downloaded 11 July 2004)

Rust, G. (1998). 'Metadata: The Right Approach', *D-Lib Magazine*, July/August 1998 online at <http://www.dlib.org/dlib/july98/rust/07rust.html> (downloaded 13 May 2004).

Storey, T. 2003. 'University repositories: An extension of the library cooperative' *OCLC Newsletter 261* (also online at <http://www.oclc.org/news/publications/newsletters/oclc/2003/261/n261.pdf>)

Storey, T. 2004. 'Moving Z39.50 to the Web' *OCLC Newsletter 263* (also online at <http://www.oclc.org/news/publications/newsletters/oclc/2004/263/srw.html>)

Weibel, S., Iannella, R. & Cathro, W. 1997. 'The 4th Dublin Core Metadata Workshop Report', *D-Lib Magazine*, June 1997 online at <http://www.dlib.org/dlib/june97/metadata/06weibel.html> (downloaded 13 May 2004)

Wilkin, J., Hagedorn, K. & Burek, M. 2003. 'Creating an Academic Hotbot: Final Report of the University of Michigan OAI Harvesting Project', University of Michigan (online at <http://oaister.umdl.umich.edu/o/oaister/description.html> - download 11 June 2004)

URLs

- 1 OAI <http://www.openarchives.org/>
- 2 Z39.50 <http://www.loc.gov/z3950/agency/>
- 3 MARC <http://www.loc.gov/marc/>
- 4 DCMI <http://dublincore.org/>
- 5 Bath profile <http://www.collectionscanada.ca/bath/bp-current.htm>
- 6 CIMI profile http://www.cimi.org/public_docs/HarmonizedProfile/HarmonProfile1.htm
- 7 CIE http://www.jisc.ac.uk/index.cfm?name=wg_cie_home
- 8 HEIRPORT <http://ads.ahds.ac.uk/heirport/>
- 9 Archives Hub <http://www.archiveshub.ac.uk/>
- 10 Harvesting the FitzWilliam <http://www.fitzmuseum.cam.ac.uk/htf/about.html>
- 11 SRW <http://www.loc.gov/z3950/agency/zing/srw/>
- 12 AHDS catalogue <http://www.ahds.ac.uk/collections/>
- 13 OAICAT <http://www.oclc.org/research/software/oai/cat.htm>
- 14 RSLP collections descriptions <http://www.ukoln.ac.uk/metadata/rsrp/>
- 15 OAister <http://oaister.umdl.umich.edu/o/oaister/>
- 16 AIM25 <http://www.aim25.ac.uk/index.stm>
- 17 ISAD(G) <http://www.ica.org/biblio.php?pdocid=1>
- 18 ARENA <http://ads.ahds.ac.uk/arena/>
- 19 Cheshire II <http://sca.lib.liv.ac.uk/cheshire/index.html>
- 20 EAD <http://www.loc.gov/ead/>
- 21 "Cheshire" Project <http://cheshire.lib.berkeley.edu/FINALDLI.htm>
- 22 Cheshire 3 <http://srw.cheshire3.org%20/>
- 23 Index Data <http://www.indexdata.dk/>