"Open wide: using open standards in academic web services"

Ross MacIntyre, Ann Apps and Leigh Morris
MIMAS, University of Manchester, Oxford Road, Manchester, M13 9PL, UK.
ross.macintyre@man.ac.uk, ann.apps@man.ac.uk, leigh.morris@man.ac.uk

Abstract

Despite enormous strides recently, it remains difficult for researchers to find relevant resources that are not 'flat' html files and may be hidden within web services. This paper will draw upon recent experiences at the UK's largest academic data centre providing web services to the education community. Since May 2001 a project has been underway attempting to make the resources more visible by exploiting a number of relevant open standards and initiatives to ensure interoperability, including: Dublin Core, XML, Z39.50, Open Archives Initiative, OpenURL and Collection Descriptions. The aim of the project being to increase the visibility and accessibility of 'appropriate' resources. This principally requires focusing on machine-to-machine metadata interchange. This paper documents some of the realities faced when implementing these potentially valuable, though sometimes 'over-hyped' technologies.

1 Introduction

MIMAS[1], located at the University of Manchester, is the UK's largest national academic datacentre and receives funding from the Joint Information Systems Committee(JISC)[2] , of the Higher and Further Education Funding Councils, to provide services to the academic and research communities in the UK and beyond.
These services are many and varied. By definition they are cross-domain and include: bibliographic reference, electronic journals, archive information, statistical datasets, satellite images and geographic datasets, chemical database services and software packages for users to manipulate their own data.

Until now there was no consistent way of discovering information within these MIMAS collections and associated services, except by reading the web pages specific to each service. It was clear that some work could usefully be done making the resources more visible and accessible.

At the same time, work was underway within JISC to develop a technical architecture for their 'Information Environment' [3] for resource discovery by researchers and learners and certain key standards were being identified. MIMAS proposed implementing these key technologies within a real service environment. This was agreed and taken forward.

2 Project Description

The detail of the project was agreed with UKOLN[4], who were developing the technical architecture on behalf of JISC. The project itself consists of six strands:

1) Creating a repository of metadata describing the web services provided by the data centre. This metadata was to be available for searching directly via a web interface and remotely via the Z39.50 search protocol.

2) Creating sharable Collection Level Descriptions of the datasets offered. This would allow discovery of data collections within the emerging UK academic 'Information Environment'.

3) Exposing resource metadata for harvesting via the Open Archives Initiative Metadata Harvesting Protocol.

4) Introducing support for OpenURLs within web services, as both source and target, thus providing the ability to generate and receive transportable bibliographic metadata.

5) Trialing an OpenURL resolver (Ex Libris' SFX) in a service environment. The objectives being:

- to examine in detail the issues associated with supporting access to the web services, covering: hosted services, interfacing (fusion) services and locally-developed services
- to evaluate a potential default (UK) national resolver service
- to explore what would be involved in hosting institution-specific resolvers
- to compare any/all of the above with other OpenURL-resolving mechanisms implemented elsewhere.

6) Independently evaluating the results of the project. How much difference do these technologies actually make to the end-user? What benefits will or might they bring to learners and researchers?

The project commenced in May 2001 and is due to complete during 2003.

3 The Metadata Repository

3.1 Approach to Metadata Creation

One person was given the task of creating the initial set of metadata for the MIMAS services. This was really a 'bootstrap' approach and feasible due to the relatively small number of resources being described. The person concerned was a service support officer, had a good working knowledge of the services and was known and respected by the other support staff. Using one person ensured a consistency of approach to metadata creation. Subsequently, all metadata was quality assured by the relevant support staff. This was essential, as they are to undertake the metadata maintenance activity.

The metadata is currently created as XML files using an XML template and a text editor. The created XML is validated by parsing against an XML Document Type Definition before the record is indexed in the metadatabase. It is planned to develop a specific, 'wiki style'[5], web-form tool for metadata creation and updating. Such a tool will become essential when the metadata maintenance is performed by more than one person.

3.2 Metadata Standards Employed

Because the MIMAS service consists of a heterogeneous collection of services and datasets across many disciplines, a common, cross-domain metadata schema was required for their description. The metadata created to describe them is based on qualified Dublin Core [6] encoded in XML, enabling cross-searching using a common core of metadata. This allows someone searching for information about for example 'economic' to discover results of possible interest across many of the MIMAS services beyond the obvious macro-economic datasets, including JSTOR, census data, satellite images and bibliographic resources.

It is possible that in the future the metadata will be extended to include records according to domain-specific standards, such as the Data Documentation Initiative (DDI) Codebook [7] for statistical datasets or a standard geographic scheme, such as ISO DIS 19115 Geographic Information – Metadata [8], for census and map datasets.

3.3 Classification Schemes

To provide quality metadata for discovery, subject keywords within the metadata are encoded according to standard classification schemes. In order to facilitate improved cross-domain searching by both humans and applications where choices of preferred classification scheme might vary, MIMAS Metadata provides subjects encoded according to several schemes. As well as the encoding schemes currently recognised within qualified Dublin Core, Library of Congress Subject Headings (LCSH)[9] and Dewey Decimal[10], UNESCO[11] subject keywords are also available. In addition, MIMAS-specific subjects are included to capture existing subject keywords on the MIMAS web site service information pages supplied by the content or application creators as well as the support staff.

The use of standard classification schemes will improve resource discovery[12], especially if faceted schemes such as Dewey Decimal[13] are used. The development of more sophisticated ontology-based search engines will make the use of standard schemes even more important. Employing standard schemes will also assist in the provision of browsing structures for subject-based information gateways[14].

Similar classification schemes are included for 'Type' to better classify the type of the resource for cross-domain searching. Countries covered by information within a MIMAS service are detailed according to their ISO3166 [15] names and also their UNESCO names captured within the 'dcterms:spatial' element of the metadata record and shown on the web display as 'Country'. Temporal coverage is captured within a 'dcterms:temporal' element and encoded according to the W3CDTF[16] scheme. This is displayed as 'Time' and may consist of several temporal ranges. Information about access requirements to a particular MIMAS service is recorded as free-text within a 'dc:rights' element and displayed as 'Access'.

3.4 Application

The XML-encoded metadata is stored in a Cheshire II[17] database, which provides a World Wide Web and a Z39.50 interface[18].

Using the Web interface to this metadatabase[19], searches may be made by fields *title*, *subject* or '*all*', initially retrieving a list of brief results with links to individual full records. An example of a full record for one of the results retrieved by searching for a subject 'IMF', with web links underlined, is:

Title: IMF Databanks
Creator: MIMAS; International Monetary Fund
Subject (LCSH): Finance; International trade
Subject (UNESCO): Finance; Trade
Subject (Dewey): 330; 332; 339
Description: MIMAS hosts four major databanks from the International Monetary Fund:
The IMF Direction of Trade Statistics provides data on imports and exports for 184 countries and their trading partners.
The IMF Balance of Payments Statistics contains the standard Balance of Payments components and aggregates for over 160 countries.
The IMF Government Finance Statistics provides detailed figures for central, state and local government revenues and expenditures for 149 countries.
The IMF International Finance Statistics covers banking, national accounts and other financial indicators for 196 countries.
Publisher: MIMAS, Manchester Computing, University of Manchester
Contributor: Russell, Celia (editor)
Type (DC): collection
Type (LCSH): Economic statistics; Information retrieval; Online databases
Type (UNESCO): Databases; Economic statistics; Information retrieval; Online searching; Statistical data
Type (Dewey): 005
Type (MIMAS): socio-economic data
Medium: text/html
URL: http://www.mimas.ac.uk/macro_econ/imf/
Language: eng
isPartOf: Macro-Economic Time Series Datasets
hasPart: IMF Balance of Payments Statistics
hasPart: IMF Direction of Trade Statistics
hasPart: IMF Government Finance Statistics
hasPart: IMF International Finance Statistics
Access: Available to UK HE. Conditionally free. Registration required.
MIMAS ID: me000002

Following a Z39.50 search, records may be retrieved as Simple Unstructured Text Record Syntax (SUTRS), both brief and full records, full records being similar to the above example, GRS-1 (Generic Record Syntax) and a simple tagged reference format. In addition the metadatabase is compliant with the Bath Profile [20], an international Z39.50 specification for library applications and resource discovery, providing records as simple Dublin Core in XML according to the CIMI Document Type Definition[21].

4 Sharable Collection Descriptions

In line with the requirement of the JISC's 'Information Environment', MIMAS has developed a further metadata application, implemented using the same architecture as the metadatabase, to provide collection description metadata for its resources, based on the Research Support Libraries Programme (RSLP) Collection Level Description (CLD) Schema[22]. This Collection database contains a record for each top-level collection at MIMAS, corresponding to the top-level descriptions of the MIMAS services in the metadatabase.

In the web interface[23], the 'Describes' field has been implemented as a web link to the corresponding top-level service record in the metadatabase application. This link is inserted automatically by the application, based on the local MIMAS identifier within the collection record, rather than being hard-coded by the metadata creator, thus avoiding maintenance problems. Following this link enables navigation to lower level records within the metadatabase hierarchy. Including this link between the two applications, and so effectively between the two databases, removes the necessity to replicate all the lower level data. It is intended that the MIMAS Collection Description will remain an exclusively top-level metadata.

5. Preparing for Harvesting

The Open Archives Initiative (OAI) has specified a Metadata Harvesting Protocol[24] which enables a data repository to expose metadata about its content in an interoperable way. The architecture of the JISC 'Information Environment' includes the implementation of OAI harvesters which will gather metadata from the various collections within the 'Information Environment' to provide searchable metadata for portals and hence for end-users[25]. Portals will select metadata from particular subject areas of relevance to their user community. Thus there is a requirement for collections and services within the 'Information Environment' to make their metadata available according to the OAI protocol, including a minimum of OAI 'common metadata format', i.e. simple Dublin Core, records.

An OAI metadata harvesting interface has been added to both the MIMAS Metadatabase and the Collection database, as a 'cgi' program, written in TCL which is the native language of Cheshire. This program responds appropriately to OAI requests, implementing the OAI 'verbs': *Identify*; *ListMetadataFormats*; *ListIdentifiers*; *GetRecord*; *ListRecords* and *ListSets*.

In order to implement the OAI interface, three new search result formats have been defined for the databases, which return in XML, respectively, according to the required OAI format: the identifier of a record; the metadata of the record in Dublin Core; an identifier and date stamp for a record, where an unavailable metadata format is requested. The OAI cgi program performs the search on the Cheshire database according to the appropriate result format for the OAI verb and arguments, then passes the result to the harvester wrapped by the required OAI response format.

6. OpenURL

The capability to provide links to full text articles from MIMAS bibliographic services would be highly desirable for researchers. However, to ensure such a link is not a 'dead-end', it is necessary firstly to translate the citation information for the article into a URL link, and secondly to link, if possible, to an 'appropriate copy'[26] of the article which is available to the researcher, ideally free, say via a valid institutional subscription. Development of the OpenURL framework for open reference linking began with research conducted by Herbert Van de Sompel and his colleagues at Ghent University, Belgium[27]. The resulting draft OpenURL[28] has been 'pinned down' as version 0.1 to enable its use by early implementers of context sensitive linking technology. This draft OpenURL provides a syntax for transmitting the metadata of a citation of a scholarly paper (or the referent) to a baseURL (or resolver) using the Web 'HTTP Get' protocol. For example for a citation to a paper:

- Bloggs, J. (2002) Reference Linking. D-Lib 9(1)

A version 0.1 OpenURL (the query/referent part) would be:

- ?genre=article&title=D-Lib&jtitle=Reference%20Linking&aulast=Bloggs&auinit=J &issn=1082-9873&date=2002&volume=9&issue=1

At MIMAS, OpenURL links have been added to a prototype enhanced version of the zetoc service[29].

6.1 NISO Committee AX

NISO Committee AX [30] is now developing the OpenURL framework to become a standard. The committee consists of representatives of implementers of both link resolution and link source applications, including major publishers of scholarly works and abstracting and indexing services, as well as librarians and academics. It includes members of other citation metadata initiatives who have an interest in liaison with or utilising OpenURL, such as DOI, CrossRef, Dublin Core[31] and Open Archives Initiative. (An author of this paper, Ann Apps, is a member of the Committee.)

The OpenURL framework for Version 1.0 of OpenURL is a more general abstraction, the 'Bison-Futé' model[32], which describes the context of a citation reference (referent). NISO committee AX has decided that the draft OpenURL Version 1.0 standard will be put out for 'trial use' during the first months of 2003. The trial will involve OpenURL source and resolution services and end users. Feedback from this trial will be taken into account in the final standard expected to go to NISO vote in Summer 2003. The registration process will not be tested in the trial, for which the registry will be pre-defined and static.6.2 OpenURL Resolver

In order to more fully explore the extensive linking enabled via OpenURL, it was proposed that a resolver be implemented at MIMAS. Ex Libris agreed to participate in the project, offering their SFX[33] software. An 'off-the-shelf' solution was suggested principally to substantially reduce

development overhead, but would be a useful counterpoint to other solutions existing at the time, such as Openly Jake[34]. As OpenURL-related technologies potentially have significant implications for service providers and libraries themselves, it would be useful to gain a full understanding in a 'real' service environment, hence the different perspectives set out in the project description.

MIMAS wanted to install SFX within a shared server environment. This led to some minor installation issues, as the software is more normally run on a dedicated server and typically installed by Ex Libris themselves. However, all installation issues were resolved promptly.

Updates to SFX and its underlying 'KnowledgeBase', though initially 'hand-to-mouth' and time consuming, are now proceeding more smoothly, with improved documentation. Note though that they require operating system command knowledge.

6.2.1 National Default Resolver Service

The intention for this service was to offer additional services to institutions who did not have their own resolver service. Consequently, the emphasis was on offering SFX target services that provided:

- unrestricted material, such as arXiv, the physics pre-print archive, various free online journals, such as the British Medical Journal and database services such as PubMed
- full-text providers who offer free access to certain levels of material, such as article abstracts and journal table of contents
- widely used services appropriate to the UK academic community, such as ISI Web of Science, JSTOR and COPAC
- support facilities such as feedback, FAQ etc.

Two sites have agreed to trial the use of this service for the next academic year. They will identify which SFX sources they wish to enable.

6.2.2 Hosted SFX Instances

Significant effort has been devoted to the identification of appropriate targets for the initial university that has agreed to participate. The university is a very large research institution with a substantial number of licensed electronic resources and services. Unsurprisingly, the SFX targets are large aggregated services, including: Elsevier Science Direct, IEEE, JSTOR, Kluwer Academic, ProQuest, Synergy, Wiley Interscience and ISI Web of Science.

The effort required to activate targets, following initial discussion and discovery, has been low. The major effort required, in the case of this university's instance, was (and will be) to maintain an accurate list of journals by target and to test the targets.

It is planned to add a further university's hosted instance at MIMAS and have both trial throughout the next academic year.

7. Conclusions

At the time of writing this paper, the project is only partially complete, but the following conclusions can be drawn based on the experience so far.

MIMAS has aimed to describe its collection of datasets and services using quality metadata. Quality assurance has been achieved by checking of the metadata records for a particular service by the relevant support staff. Continued metadata quality will be ensured by maintenance of the metadata by these support staff.

Subject or concept keywords are included in the metadata according to several standard classification schemes, as are resource types and geographical names. Use of standard schemes enhances the quality of the metadata and enables effective resource discovery.

Another objective of the project was to develop an interoperable solution based on open standards and using leading-edge, open source technology. This has been successfully achieved using a Cheshire II software platform to index Dublin Core records encoded in XML. A spin-off has been improvements to Cheshire following feedback from MIMAS.

Use of other standard or experimental technologies such as the Z39.50 and OAI metadata harvesting interfaces in addition to the web interface will enable the metadatabase and Collection database to be integrated into the JISC 'Information Environment', thus providing a valuable resource discovery tool to the stakeholders within that environment.

The metadatabase provides a single point of access into the disparate, cross-domain MIMAS datasets and services. Note that in the case of hosted services, i.e. not created by MIMAS, no application or data provider development work has been required.

The metadatabase provides a means for researchers to find and access material to aid in the furtherance of their work, thus assisting in the advancement of knowledge. Learners and their teachers will be able to discover appropriate learning resources across the MIMAS portfolio, improving the educational value of these datasets.

References

1. MIMAS, 2002. Manchester Information and Associated Services, including Archives Hub, COPAC, ISI Web of Science, JSTOR, NESLI, zetoc. http://www.mimas.ac.uk .
2. JISC, 2002. The Joint Information Systems Committee. http://www.jisc.ac.uk .
3. Powell, A. and Lyon, L., 2002. The JISC Information Environment and Web Services. *Ariadne*, 31, http://www.ariadne.ac.uk/issue31/information-environments.
4. UKOLN http://www.ukoln.ac.uk
5. Leuf, B. and Cunningham, W., 2002. The Wiki Way. http://www.wiki.org .
6. DCMI, 2002. The Dublin Core Metadata Initiative. http://www.dublincore.org .
7. DDI, 2002. Data Documentation Initiative Codebook DTD. http://www.icpsr.umich.edu/DDI/CODEBOOK/ .
8. ISO, 2002. ISO/TC211: Geographic Information / Geomatics. http://www.isotc211.org .
9. Library of Congress, 2002. Library of Congress Subject Headings. In: Cataloguing Distribution Service.
10. DDC, 2002. Dewey Decimal Classification. OCLC Forest Press. http://www.oclo.org/dewey/ .
11. UNESCO, 2001. UNESCO Thesaurus. http://www.ulcc.ac.uk/unesco/ .
12. Vizine-Goetz, D., 1996. Using Library Classification Schemes for Internet Resources. *In:* E. Jul, ed. *Proceedings of the OCLC Internet Cataloguing Colloquium, San Antonio, Texas, 19 January 1996*. OCLC, http://www.oclc.org/oclc/man/colloq/toc.htm .
13. Tinker, A.J., Pollitt, A.S., O'Brien, A. and Braekevelt, P.A., 1999. The Dewey Decimal Classification and the transition from physical to electronic knowledge organisation. *Knowledge Organization*, 26 (2), 80-96.
14. Koch, T. and Day, M., 1999. The role of classification schemes in Internet resource description and discovery. *In:* Work Package 3 of Telematics for Research project DESIRE (RE 1004), http://www.ukoln.ac.uk/metadata/desire/classification/ .
15. DIN, 2002. ISO 3166-1: The Code List. http://www.din.de/gremien/nas/nabd/iso3166ma/codlstp1/ .
16. W3C, 1997. Date and Time Formats. http://www.w3.org/TR/NOTE-datetime .
17. Larson, R.R., 2002. Cheshire II Project. http://cheshire.lib.berkeley.edu .
18. NISO, 1995. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. http://www.niso.org/standards/resources/Z3950.pdf .
19. The MIMAS Metadatabase http://www.mimas.ac.uk/metadata/
20. Bath Group, 2001. The Bath Profile: An International Z39.50 Specification for Library Applications and Resource Discovery. http://www.nlc-bnc.ca/bath/bp-current.htm .
21. CIMI, 2001. The Consortium for the Computer Interchange of Museum Information (CIMI) Dublin Core Document Type Definition. http://www.nlc-bnc.ca/bath/bp-app-d.htm .
22. Powell, A., Heaney, M. and Dempsey, L., 2000. RSLP Collection Description. *D-Lib Magazine*, 6 (9), doi:/10/1045/september2000-powell .
23. The MIMAS Collection Metadata http://www.mimas.ac.uk/metadata/collection/
24. Warner, S., 2001. Exposing and Harvesting Metadata Using the OAI Metadata Harvesting Protocol: A Tutorial. *High Energy Physics Webzine*, 4, http://library.cern.ch/HEPLW/4/papers/3/ .
Watry, P. and Hill, A., 2002. Collection Description Service Scoping Study. (To be published).
25. Cliff, P., 2001. Building ResourceFinder. *Ariadne*, 30, http://www.ariadne.ac.uk/issue30/rdn-oai/ .
26. Caplan, P., Arms, W.Y.: Reference Linking for Journal Articles. D-Lib Magazine **5**(7/8) (1999). doi://10.1045/july99-caplan
27. Van de Sompel, H., Beit-Arie, O.: Open Linking in the Scholarly Information Environment Using the OpenURL Framework. D-Lib Magazine **7**(3) (2001). doi://10.1045/march2001-vandesompel
28. OpenURL Syntax Description (v0.1). http://www.sfxit.com/OpenURL/openurl.html
29. Apps, A., MacIntyre, R.: Prototyping Digital Library Technologies in zetoc. Lecture Notes in Computer Science (Springer-Verlag): Proceedings of Sixth European Conference on Research and Advanced Technology for Digital Libraries (ECDL2002), Rome Italy, 16-18 September 2002 (accepted for publication). (2002)
30. OpenURL, NISO Committee AX. http://library.caltech.edu/openurl/
31. Powell, A., Apps, A.: Encoding OpenURLs in Dublin Core metadata. Ariadne 27 (2001). http://ariadne.ac.uk/issue27/metadata
32. Van de Sompel, H., Beit-Arie, O.: Generalizing the OpenURL Framework beyond References to Scholarly Works. D-Lib Magazine 7(7/8) (2001). doi://10.1045/july2001-vandesompel
33. SFX, Ex Libris. http://www.sfxit.com
34. Openly Jake, Openly Informatics Inc. http://www.openly.com/jake/