

Exploiting Open Standards in Academic Web Services

Ross MACINTYRE,

Ann APPS,

and

Leigh MORRIS

MIMAS, University of Manchester

Manchester, M13 9PL, United Kingdom

ABSTRACT

In Digital Library-related technologies, there is a whole host of open standards and protocols that are at varying stages of definition or emergence and acceptance or agreement. Nevertheless, specifically in an academic context, these have led to some valuable improvements in the quality and value of services provided to teachers, learners and researchers alike. However, it often remains difficult for these information seekers to find relevant resources that are not immediately 'visible', they may be effectively hidden within database-driven web services or proprietary applications. The focus of this paper is upon a project based at the UK academic data centre, MIMAS, which provides web-based services to the education community in the UK, Ireland and beyond. The project's principle aim was to increase the visibility and accessibility of 'appropriate' resources by exploiting a number of relevant open standards and initiatives to ensure interoperability. This principally required focusing on machine-to-machine metadata interchange.

Keywords: Interoperability, Dublin Core, Metadata, Collection Level Descriptions, Z39.50, OAI and OpenURL.

1. INTRODUCTION

In the United Kingdom the government provides funds to the Higher and Further Education Funding Councils (FCs). They in turn provide funds to the many universities, colleges, etc. As well as this obvious funding stream, the FCs also allocate funding for products and services that benefit from a co-ordinated approach. In the area of technology, they do this via the Joint Information Systems Committee (JISC), which has, for example, invested in a shared academic network (SuperJANET). It also oversees the acquisition and provision of electronic resources for learning, teaching and research, some of which are provided from data centres. MIMAS[1], located at the University of Manchester, is the UK's largest national academic datacentre.

Work has taken place on the development of a technical architecture for the JISC's 'Information Environment' [2]

for resource discovery by researchers and learners. MIMAS services have to provide interfaces consistent with this architecture. However, many of the services pre-date this architecture and do not provide these interfaces. To complicate the situation, some of the services are products hosted and supported by MIMAS, but not developed in-house, making implementation of additional interfaces an issue. As the technical architecture was developing and certain key standards identified, MIMAS proposed implementing these key technologies within a real service environment. This was agreed, taken forward and is the subject of this paper.

MIMAS Services

The services are many and varied, by definition they are cross-domain and include:

- Bibliographic information such as citation databases offered by ISI Web of Science, the combined online catalogue of the UK research libraries in COPAC and a current awareness service called **zetoc**, based on the British Library's electronic table of contents database of journal articles and conference papers.
- Electronic journals via the JSTOR archive of scholarly journals.
- Archival information from the Archives Hub national gateway to descriptions of archives in UK universities and colleges.
- Statistical datasets including data from the UK censuses, international macro-economic data and UK government surveys.
- Satellite images and geographic datasets.
- Chemical database services including Beilstein Crossfire.

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.

2. PROJECT DESCRIPTION

The detail of the project was agreed with the technical architects. 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 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 and
 - 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 finish in August 2003.

3. THE METADATA REPOSITORY

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 [3] 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.

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)[4] and Dewey Decimal[5], UNESCO[6] 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.

Similar classification schemes are included for 'Type' to better classify the type of the resource for cross-domain searching. Each metadata record includes a 'Type' from the high-level DCMI Type Vocabulary [7] such as 'Service', 'Collection' or 'Dataset'. A further 'Type' field may include type indications, such as 'Bibliographical citations' and 'Online searching'. Again the MIMAS-specific resource type is included.

Countries covered by information within a MIMAS service are detailed according to their ISO3166 [8] 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[9] 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'.

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.

Application

The XML-encoded metadata is stored in a Cheshire II[10] database, which provides a World Wide Web and a

Z39.50 interface[11]. It is open source software, users include: US: NSF/NASA/ARPA and in EU: NESSTAR.

Using the Web interface to this metadatabase, 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.

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 [12], 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[13].

Metadata Maintenance

It is planned to develop a specific, 'wiki style'[14], web-form tool for metadata creation and updating. This tool will capture metadata by field and include links to standard schemes for subject keyword selection and classification, the required XML being created at its back end, effectively transparently. It will allow a metadata creator to view the eventual display of the record within the application before making a final 'commit' to the metadatabase. Such a tool will become essential when the metadata maintenance is performed by more than one person.

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[15]. This Collection database contains a record for each top-level collection at MIMAS, corresponding to the top-level descriptions of the services in the metadatabase.

Similar to the metadatabase, standard schemes are used to provide quality concepts for collection discovery. It is probable that the common subject classification used within the Information Environment will be Dewey Decimal, but LCSH and UNESCO concepts are also provided to allow searching by other sources.

MIMAS has extended the RSLP CLD schema to include administrative metadata needed for date stamping of records and quality control, including the record creation date, the name of the metadata record creator and the local identifier for the record.

In the web interface, the 'Describes' field is 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[16] 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[17]. 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.

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'[18] 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[19]. The resulting draft OpenURL[20] 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: J. Bloggs, "Reference Linking", **D-Lib**, Vol. 9, No. 1, 2002.

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

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

NISO Committee AX

NISO Committee AX [21] 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[22] and Open Archives Initiative. (An author of this paper, Ann Apps, is a member of the Committee.)

NISO committee AX has decided that the draft OpenURL Version 1.0 standard will be put out for 'trial use' during the second quarter 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 Fall 2003. The registration process will not be tested in the trial, for which the registry will be pre-defined and static.

OpenURL Implementation

At MIMAS, OpenURL support has been added to the **zetoc** service[23]. OpenURL links are provided from discovered citation records within **zetoc**, thus enabling it as an 'OpenURL Source'. It is also possible to (OpenURL) link 'in' to the record for a specific article in **zetoc**, enabling it as an 'OpenURL Target'. It provides a direct web link into a particular article's record using its citation metadata. **zetoc** has become a 'citation centre' providing discovery of an article by a definitive citation search and then location of an appropriate copy of that article along with other relevant services. Note that other hosted services are already OpenURL-enabled, including ISI Web of Science.

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 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. 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.

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 preprint 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.

Hosted SFX Instances: Significant effort has been devoted to the identification of appropriate targets for the universities that agreed to participate. 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.

Comparative Studies: As expected, UK institutions have also been licensing SFX and other products. This project has been used to provide a forum for comparisons to be made and experience shared amongst these early adopters of OpenURL resolution technology. A formal round table discussion was held, attended by ten UK universities and was the subject of a detailed report[24]. Other resolver services are now being licensed and considered by UK institutions, including Openly Informatics' iCate, Endeavour's LinkFinder+, Innovative Interface's WebBridge and Fretwell Downing Informatics' ZPORTAL software. It is intended to include these alternatives in a comparison report, though setting the scope for any formal comparison is not straightforward. A comparative study from the

Mississippi State University in the US was recently published[25].

7. CONCLUSIONS

At the time of writing this paper, the project has not completed, 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. 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 [26] for statistical datasets or a standard geographic scheme, such as ISO DIS 19115 Geographic Information – Metadata [27], for census and map datasets. Another possible future extension would be to include educational metadata, such as IMS[28], where appropriate datasets are learning resources. But the MIMAS metadata cross searching capability would of necessity still be based on the core metadata encoded in qualified Dublin Core.

Subject or concept keywords are included in the metadata according to several standard classification schemes, as are resource types and geographical names. The use of standard classification schemes will improve resource discovery, especially if faceted schemes 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[29].

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.

Overall, the project has made real strides towards an interoperable environment and as Miller[30] states: "Changing internal systems and practices to make them interoperable is a far from simple task. The benefits for the organisation and those making use of information it publishes are potentially incalculable."

- [1] Manchester Information and Associated Services (MIMAS). 2002. Including: Archives Hub, COPAC, ISI Web of Science, JSTOR, zetoc. Available from <http://www.mimas.ac.uk>.
- [2] A.Powell, and L. Lyon, "The JISC Information Environment and Web Services". **Ariadne**, 31, (2002) Available from <http://www.ariadne.ac.uk/issue31/information-environments>.
- [3] DCMI. 2002. "The Dublin Core Metadata Initiative". Available from <http://www.dublincore.org>.
- [4] Library of Congress. 2002. "Library of Congress Subject Headings". In: **Cataloguing Distribution Service**.
- [5] DDC. 2002. "Dewey Decimal Classification". OCLC Forest Press. Available from <http://www.oclo.org/dewey/>.
- [6] UNESCO. 2001. "UNESCO Thesaurus". Available from <http://www.ulcc.ac.uk/unesco/>.
- [7] DCMI. 2000. "DCMI Type Vocabulary". Available from <http://dublincore.org/documents/dcmi-type-vocabulary/>.
- [8] DIN. 2002. "ISO 3166-1: The Code List". Available from <http://www.din.de/gremien/nas/nabd/iso3166ma/codlstp1/>.
- [9] W3C. 1997 "Date and Time Formats". Available from <http://www.w3.org/TR/NOTE-datetime>.
- [10] R.R. Larson, "Cheshire II Project". 2002. Available from <http://cheshire.lib.berkeley.edu>.
- [11] NISO. 1995. "Information Retrieval (Z39.50)". Application Service Definition and Protocol Specification. Available from <http://www.niso.org/standards/resources/Z3950.pdf>.
- [12] Bath Group. 2001. "The Bath Profile: An International Z39.50 Specification for Library Applications and Resource Discovery". Available from <http://www.nlc-bnc.ca/bath/bp-current.htm>.
- [13] The Consortium for the Computer Interchange of Museum Information (CIMI). 2001. "Dublin Core Document Type Definition". Available from <http://www.nlc-bnc.ca/bath/bp-app-d.htm>.
- [14] B. Leuf and W. Cunningham, 2002. "The Wiki Way". Available from <http://www.wiki.org>.
- [15] A. Powell, M. Heaney and L. Dempsey. "RSLP Collection Description," **D-Lib Magazine** 6, No. 9,
- [16] S. Warner, "Exposing and Harvesting Metadata Using the OAI Metadata Harvesting Protocol: A Tutorial." **High Energy Physics Webzine**, No. 4, 2001. Available from <http://library.cern.ch/HEPLW/4/papers/3/>.
- [17] P. Cliff. "Building ResourceFinder", **Ariadne**, No. 30, Available from <http://www.ariadne.ac.uk/issue30/rdn-oai/>.
- [18] P. Caplan and W.Y. Arms, "Reference Linking for Journal Articles", **D-Lib Magazine** 5, No. 7/8, 1999.
- [19] H. Van de Sompel and O. Beit-Arie, "Open Linking in the Scholarly Information Environment Using the OpenURL Framework", **D-Lib Magazine** 7, No. 3, 2001.
- [20] OpenURL Syntax Description (v0.1). Available from <http://www.sfxit.com/OpenURL/openurl.html>
- [21] OpenURL, NISO Committee AX. Available from <http://library.caltech.edu/openurl/>
- [22] A. Powell and A. Apps, "Encoding OpenURLs in Dublin Core metadata", **Ariadne**, No. 27, 2001. Available from <http://ariadne.ac.uk/issue27/metadata>
- [23] A. Apps and R. MacIntyre, "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.
- [24] MIMAS. 2002. "SFX Talking Shop". Available from <http://www.mimas.ac.uk/metadata/ITAM/sfx.html>
- [25] M.D.D. Collins and C.L. Ferguson "Context-Sensitive Linking: It's a Small World After All", **Serials Review**, Vol. 28, No. 4, 2002, pp.267-282.
- [26] DDI. 2002. "Data Documentation Initiative Codebook DTD". Available from <http://www.icpsr.umich.edu/DDI/CODEBOOK/>.
- [27] ISO. 2002. "ISO/TC211: Geographic Information / Geomatics". Available from <http://www.isotc211.org>.
- [28] B. Olivier, O. Liber P. Lefrere, "Specifications and standards for learning technologies: the IMS project", **International Journal for Electrical Engineering Education**, Vol. 37, No. 1, pp. 26-37.
- [29] T. Koch and M. Day, "The role of classification schemes in Internet resource description and discovery", In: Work Package 3 of Telematics for Research project DESIRE, Available from <http://www.ukoln.ac.uk/metadata/desire/classification/>.
- [30] P. Miller, "Interoperability, what is it and why should I want it?" **Ariadne**, No. 24, 2000. Available from <http://www.ariadne.ac.uk/issue24/interoperability/>