Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 71-80

Exposing Cross-Domain Resources for Researchers and Learners

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

Abstract

MIMAS is a national UK data centre which provides networked access to resources to support learning and research across a wide range of disciplines. There was no consistent way of discovering information within this cross-domain, heterogeneous collection of resources, some of which are access restricted. Further these resources must provide the interoperable interfaces required within the UK higher and further education 'information environment'. To address both of these problems, consistent, high quality metadata records for the MIMAS services and collections have been created, based on Dublin Core, XML and standard classification schemes. The XML metadata repository, or 'metadatabase', provides World Wide Web, Z39.50 and Open Archives Initiative interfaces. In addition, a collection level database has been created with records based on the RSLP Collection Level Description schema. The MIMAS Metadatabase, which is freely available, provides a single point of access into the disparate, cross-domain MIMAS datasets and services.
Keywords. *Metadata, Dublin Core, cross-domain, collection level description, subject classification.*

1. Introduction

MIMAS [26] at the University of Manchester, UK, is a national data centre for higher and further education and the research community in the UK, providing networked access to key data and information resources to support teaching, learning and research across a wide range of disciplines. This cross-domain, heterogeneous collection of resources includes:

- Bibliographic information such as ISI Web of Science, COPAC providing access to the UK research libraries' online catalogue, and the *zetoc* current awareness service 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, and the UK National Electronic Site Licence Initiative (NESLI).
- Archival information from the Archives Hub national gateway to descriptions of archives in UK universities and colleges.
- Statistical datasets including data from several UK censuses, international macro-economic data and UK government surveys.
- Spatial, satellite and geographic datasets.
- Scientific, chemical data via Beilstein Crossfire.
- Software packages for manipulating some of these datasets.

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. Although most of these web pages contain high quality information relevant to their particular service, this information is not presented in a standard format and there is not a simple way to search for information across the services.

Some of the resources held at MIMAS are freely available globally, but access to many is restricted in some cases to members of UK academia, maybe requiring registration, in other cases by subscription. For resources where access is restricted, currently general resource discovery will find only shallow top-level information, and may not indicate to a prospective user the appropriateness of a resource to their interest.

MIMAS services are funded by the Joint Information Systems Committee (JISC) [17] of the UK Higher and Further Education Funding Councils. Thus they will be required to provide interfaces consistent with the architecture of the JISC 'Information Environment' [33] for resource discovery by researchers and learners. Currently many of the services do not provide these interfaces. Some of the MIMAS services are products hosted and supported by MIMAS, but not developed in-house, making implementation of additional interfaces unlikely.

https://doi.org/10.23106/dcmi.952106966

To overcome all of these problems consistent, high quality metadata records for the MIMAS services and collections have been created. These metadata records are standards-based, using Dublin Core [7], XML and standard encoding schemes for appropriate fields. Freely available access to this XML metadata repository, or 'metadatabase', is provided by an application which supports the interfaces required by the Information Environment, enabling information discovery across the cross-domain MIMAS resource collection and allowing users at all experience levels access to 'an interoperable information world' [3].

2. MIMAS Metadata Records

2.1. Cross-Domain Information Discovery

Because the MIMAS service consists of a heterogeneous collection of services and datasets across many disciplines, a common, cross-domain metadata schema is required for their description. The metadata created to describe them is based on qualified Dublin Core, which enables 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 [10] for statistical datasets or a standard geographic scheme, such as ISO DIS 19115 Geographic Information – Metadata [15], for census and map datasets. Another possible future extension would be to include educational metadata, such as IMS [30], 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.

2.2. An Example Metadata Record

The MIMAS metadata is encoded in XML and stored in a Cheshire II [19] database, described in more detail in section 5, which provides a World Wide Web and a Z39.50 interface. NISO Z39.50 [28] is a standard for information retrieval which defines a protocol for two computers to communicate and share information [25].

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.

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) [23] and a simple tagged reference format. In addition the MIMAS Metadatabase is compliant with the Bath Profile [2], 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 [5].

The MIMAS Metadatabase has the capability to expose simple Dublin Core metadata about the MIMAS resources for harvesting, conforming to the Open Archives Initiative (OAI) [29] Metadata Harvesting Protocol.

An example of a full record for one of the results retrieved by searching for a subject 'science', with web links underlined, but with an abbreviated description, is:

Title:	ISI Web of Science
Creator:	MIMAS; ISI
Subject: (LCSH)	Abstracts; Arts; Books Reviews; Humanities; Letters; Periodicals; Reviews; Science; Social sciences
Subject: (UNESCO)	Abstracts; Arts; Book reviews; Conference papers; Discussions (teaching method Periodicals; Science; Social sciences
Subject (Dewey):	300; 500; 505; 600; 605; 700; 705
Description:	ISI Citation Databases are multidisciplinary databases of bibliographic information gathered from thousands of scholarly journals
Publisher:	MIMAS, University of Manchester
Type (DC):	Service
Type (LCSH):	Bibliographical citations; Bibliographical services; Citation indexes; Information retrieval; Online bibliographic searching; Periodicals Bibliography; Web databases
Type: (UNESCO)	Bibliographic databases; Bibliographic services; Indexes; Information retrieval; Online searching
Type (Dewey):	005
Type (MIMAS):	bibliographic reference
Medium:	text/html
URL:	http://wos.mimas.ac.uk/
Language:	eng
isPartOf:	ISI Web of Science for UK Education
hasPart:	Science Citation Index Expanded
hasPart:	Social Sciences Citation Index
hasPart:	Arts & Humanities Citation Index
Access:	Available to UK FE, HE and research councils. Institutional subscription required
MIMAS ID:	wo000002

https://doi.org/10.23106/dcmi.952106966

2.3. Standard Classification and Encoding Schemes

To provide quality metadata for discovery, subject keywords within the metadata are encoded according to standard classification or encoding schemes. These subject keywords will enable discovery beyond simply the existence of a resource by allowing researchers to find resources which are relevant to their particular research field. In order to facilitate improved cross-domain searching by both humans and applications where choices of preferred subject 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) [22] and Dewey Decimal [9], UNESCO [38] 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 MIMAS support staff.

The use of standard classification schemes will improve resource discovery [40]. If schemes such as Dewey Decimal [37] were used, in the future, in a multi-faceted form they would lend themselves to use by developing search engines which create their indexes on faceted subject headings [11]. 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 [18].

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 [8], 'Service' in the example above, but for some MIMAS records this will be 'Collection' or 'Dataset'. In addition, the above example includes type indications, including 'Bibliographical citations' and 'Online searching', according to standard schemes. Again the MIMAS-specific resource type is included.

Countries covered by information within a MIMAS service are detailed according to their ISO3166 [12] names and also their UNESCO names, captured within the 'dcterms:spatial' element of the metadata record and shown on the web display as 'Country'. This is of particular relevance to the macro-economic datasets, such as the IMF databanks, which include data from many countries in the world. Temporal coverage, again of relevance to the macro-economic datasets, is captured within a 'dcterms:temporal' element and encoded according to the W3CDTF [41] 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'.

2.4. The MIMAS Application Profile

Where possible the metadata conforms to standard qualified Dublin Core. But this is extended for some Dublin Core elements to enable the capture of information which is MIMAS-specific or according to schemes which are not currently endorsed by Dublin Core. These local additions to qualified Dublin Core effectively make up the MIMAS application profile [14] for the metadatabase. The inclusion of UNESCO as a subject, type and spatial classification scheme described above is an example of local extensions, as is the capture of MIMAS-specific subjects and types. A possible future extension would be to capture the provenance of some metadata elements, such as subject keywords, where these were supplied by the content creator.

Some administrative metadata is included: the name of the person who created the metadata; the creation date; and the identifier of the record within the MIMAS Metadatabase. Capturing the name of the metadata creator will be of use for future quality checks and updating. The creation date, or 'date stamp', for the metadata, actually the date it is added into the database, is captured within a 'dcterms:created' element according to the W3CDTF scheme, for example "2002-05-27". The local MIMAS identifier, which is required to implement the functionality of the application as well as providing a unique identifier for each record within the database, is captured in a dc:identifier element with a MIMAS scheme.

2.5. The MIMAS Metadata Hierarchy

Although each of the records within the MIMAS Metadatabase is created, indexed and available for discovery individually, the records represent parts of the service within a hierarchy. In the example above, the record for 'ISI Web of Science' is a 'child' of the top-level record 'ISI Web of Science for UK Education', the umbrella term for the total service offered, and is a 'parent' of several records including 'Science Citation Index Expanded'.

During metadata creation only the 'isPartOf' relation is recorded, as the MIMAS identifier of the parent metadata record. The 'hasPart' fields and the displayed titles and links for parent and child metadata records are included by the MIMAS Metadatabase application as described in section 5.2. Hard coding 'hasPart' fields into a metadata record would necessitate the inefficient process of updating a parent record whenever a new child record were added. Dynamic generation of these links assists in simplifying the metadata creation and update process, and in maintaining the consistency of the metadata.

A further navigation hierarchy is provided by the application. If a parent and a child record, according to the 'isPartOf' hierarchy, also have a matching MIMAS subject keyword, the application includes a link from the parent's subject keyword to the particu-

https://doi.org/10.23106/dcmi.952106966

lar child record. For example a JSTOR fragment record could include:

Title:	JSTOR Ecology & Botany Collection
Subject (MIMAS):	Ecology / Journal of Applied Ecology
Subject (MIMAS):	Botany

where the text 'Ecology / Journal of Applied Ecology' is a web link to the record for that particular journal. Again this subject navigation hierarchy is provided dynamically by the application and does not depend on the accuracy of metadata creation beyond the 'isPartOf' identifier and the matching subject keyword.

The child, 'hasPart', links within the MIMAS metadata hierarchy are available in the web interface only. A metadata record retrieved through the Z39.50 or OAI interfaces will include a single 'isPartOf' relation at most, which will consist of the MIMAS identifier of the parent record. Any required linking between records would be provided by the application retrieving the records.

2.6. Metadata Creation

The initial MIMAS metadata covering all the MIMAS services has been created by one person as part of the set-up project, much of it being scraped from the existing MIMAS service web pages. The metadata records for each service have been checked manually by the particular support staff, thus ensuring quality metadata for each MIMAS service. It is envisaged that the metadata will be maintained by the service support staff in the future, as part of the standard support process for each MIMAS service. There are currently 57 records in the metadatabase, distributed unevenly across 14 services (maximum 14, minimum 1) but this will increase when the metadata is extended to lower levels in the hierarchy.

The metadata reaches appropriate levels of the hierarchy, differing for each service, but it may be extended to greater depth in the future, possibly to the data level in some cases. For instance, the individual journals and issues included in JSTOR could be listed in the metadatabase.

Lacking a suitable XML authoring tool, the MIMAS 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 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. The tool will be 'wiki style' [21] allowing a metadata creator to immediately 'publish' and 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.

3. The JISC Information Environment

All MIMAS resources are part of the JISC 'Information Environment' [33], which provides resources for learning, teaching and research to UK higher and further education, and thus must be consistent with its architecture. The Information Environment will enable resource discovery through the various portals in its 'presentation layer', including the discipline specific UK Resource Discovery Network (RDN) hubs [35], the RDN also being one of the participating gateways in the European Renardus service [36]. Content providers in the 'provision layer' are expected to disclose their metadata for searching, harvesting and by alerting. This means that all resources within the Information Environment should have a Web search interface and at least some of the following for machine-to-machine resource discovery: a Z39.50 (Bath Profile cross-domain compliant) search interface; an OAI (Open Archives Initiative) [29] interface for metadata harvesting; and an RDF Site Summary (RSS) [32] alert channel capability. In addition resources may support OpenURL [39] for article discovery and location, where appropriate. This initiative, based on standard metadata and methods, may be seen as moving the UK academic information environment 'from isolated digital collections to an interoperable digital library' [3].

The majority of MIMAS resources have a Web search interface to provide resource discovery within their particular service. A few MIMAS services, COPAC, *zetoc* and the Archives Hub, provide Z39.50 interfaces. Some services, being commercial products hosted by MIMAS, may never provide Z39.50 searching or OAI metadata. To overcome this lack of requisite interfaces for MIMAS content and access restrictions on some of the services, the MIMAS Metadatabase will act as an intermediate MIMAS service within the 'provision layer' of the Information Environment, functioning as the main resource discovery service for MIMAS content.

The MIMAS Metadatabase does not currently include an RSS alert facility. If thought necessary within the Information Environment, it would be possible to include an alerting service in the future where appropriate, which could inform researchers when new datasets or journals were added to the MIMAS collection.

OpenURL support is not included because the metadatabase is not primarily concerned with article discovery, although this is relevant to several of the MIMAS services. There is work underway to investigate the provision of OpenURL linking within *zetoc*, and ISI Web of Science provides OpenURL linking to users whose institution has an OpenURL resolver.

https://doi.org/10.23106/dcmi.952106966

4. MIMAS Collection Description

A further requirement of the Information Environment is a 'collection description service' [43], to allow portals within the 'presentation layer' to determine which content providers may have resources of interest to their users. This will maintain machine-readable information about the various resource collections available to researchers and learners within the Information Environment. A portal will ascertain from a collection description that a particular content provider may have resources of interest to an end-user, before pointing the end-user to the content service.

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 [34]. The MIMAS 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, with Web, Z39.50 and OAI interfaces.

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, there is a 'Describes' field which is a web link to the corresponding top-level service record in the MIMAS 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 MIMAS Metadatabase hierarchy. Including this link between the two applications, and so effectively between the two databases, removes the necessity to replicate within the MIMAS Collection Description all the MIMAS service descriptions at lower levels in the hierarchy. It is intended that the MIMAS Collection database will remain an exclusively top-level description.

4.1. An Example MIMAS Collection Record

An example collection description for a MIMAS service, *zetoc*, is as follows (with some abbreviation):

Collection Name:	zetoc
Concept (LCSH):	Arts; Business; Conference proceedings; Diseases; Economics; Engineering; Finance; Geography; History; Humanities; Language; Law; Library materials; Literature; Medical sciences; Medicine; Online library catalogs; Periodicals; Philosophy; Political science; Psychology; Religion; Science; Social sciences; Technology
(UNESCO):	Concept: Conference papers; Diseases Economics; Engineering; Finance; Law; Medical sciences; Periodicals; Science; Social sciences; Technology
Concept: (Dewey)	050; 100; 105; 200; 300; 320; 330; 340; 400; 405; 500; 505; 600; 605; 610; 620; 700; 705; 800; 805; 900; 905
Temporal Cover:	1993/
Description:	zetoc provides Z39.50-compliant access to the British Library's Electronic Table of Contents (ETOC)
Collection URL:	http://zetoc.mimas.ac.uk
Type (CLDT):	Catalogue.Library.Text
Accumulation:	2000/
Contents Date:	1993/
Accrual:	The database is updated nightly (active, deposit, periodic)
Legal Status:	Please see the Terms and Conditions of Use for further details
Access:	Available conditionally free to UK FE and HE. Available by institutional subscription to UK research councils, English NHS regions, CHEST associated and affiliated sites, and academic institutions in Ireland
Collector:	The British Library
Owner:	The British Library
Location:	Manchester Computing
Location URL:	http://zetoc.mimas.ac.uk
Administrator:	MIMAS
Admin Role:	Service provider
Admin Email:	info@mimas.ac.uk
Describes:	ze000001

4.2. Using the RSLP Collection Level Description Schema for Digital Collections

Because the RSLP schema was developed for the purpose of recording collections held by libraries and museums, some issues have arisen when using it to describe digital collections. Mostly these questions related to the irrelevance and apparent repetition of some of the fields, for instance the collection URL and the location URL in the above example. In many cases the distinction between 'collector' and 'owner' is not obvious. 'Physical location' is probably not of great importance for a digital collection which could easily be moved or distributed, and it is unlikely to be of interest to an end-user, whereas the physical location of a museum collection would be of significance. However, it is recognised that all the fields in the schema

https://doi.org/10.23106/dcmi.952106966

are optional. In general the RSLP CLD, although not an 'official' standard, seems to provide a suitable common format for interoperable collection descriptions.

The application's Z39.50 interface provides, amongst other formats, simple Dublin Core in XML, for Bath Profile cross-domain compliancy and for interoperability within the JISC Information Environment. Similarly the OAI interface provides metadata records in simple Dublin Core. The mapping from the RSLP CLD to simple Dublin Core inevitably 'dumbs down' the information provided and loses some of the richness of the RSLP CLD schema. The Z39.50 SUTRS full record results, which are similar to the web display, maintain the full RSLP CLD information, but may not be very easily parsable. Thus it appears that to use these collection description records for machine-to-machine data interoperability within the JISC Information Environment a further metadata schema based on RSLP CLD will be required for OAI harvesting. Similarly such a schema could be incorporated into the results returned according to the Z39.50 standard if a new profile were agreed.

5. The Cheshire II Information Retrieval System

The software platform used for the MIMAS Metadatabase is Cheshire II [20] which is a next generation online catalogue and full text information retrieval system, developed using advanced information retrieval techniques. It is open source software, free for non-commercial uses, and operates with open-standard formats such as XML and Z39.50, all reasons which influenced its choice for this project. Cheshire II was developed at the University of California-Berkeley School of Information Management and Systems, underwritten by a grant from the US Department of Education. Its continued development by the Universities of Berkeley and Liverpool receives funding from the Joint Information Systems Committee (JISC) of the UK Higher and Further Education Funding Councils and the US National Science Foundation (NSF). Experience and requirements from the development of the MIMAS Metadatabase have been fed back into the continuing Cheshire development. Although using evolving software has caused some technical problems, the Cheshire development team has been very responsive to providing new functionality, and this relationship has proved beneficial to both projects. Examples of new functionality are the sorting of result sets within the Cheshire Web interface and 'virtual' databases, described further in [1].

5.1. Z39.50 via Cheshire

Cheshire provides indexing and searching of XML (or SGML) data according to an XML Document Type Definition (DTD), and a Z39.50 interface. The underlying database for the MIMAS Metadatabase is a single XML data file containing all the metadata records, along with a set of indexes onto the data. The MIMAS metadata XML is mapped to the Z39.50 Bib-1 Attribute Set [4] for indexing and searching. The application's Z39.50 search results formats are detailed above in section 2.2. The mapping from the MIMAS metadata to the GRS-1 Tagset-G [23] elements is defined in the Cheshire configuration file for the database and is used by Cheshire to return data in GRS-1 format to a requesting client. The other Z39.50 result formats are implemented by bespoke filter programs which transform the raw XML records returned by Cheshire, the 'hooks' to trigger these filters being specified in the configuration file for the database. The mapping from the MIMAS metadata to simple Dublin Core, as required by the Bath Profile, is straightforward, the base data being qualified Dublin Core, albeit with some loss of information such as subject schemes. In order to obviate this information loss as much as possible, such details are included in parentheses in the supplied record. For example, a Z39.50 XML result for the example in section 2.2 may contain the element:

```
<subject>(LCSH) Abstracts</subject>
```

5.2. The Cheshire Web Interface

Cheshire also provides 'webcheshire' which is a basic, customisable World Wide Web interface. The web interface for the MIMAS Metadatabase is built on webcheshire as a bespoke program written in OmniMark (version 5.5) [31]. This web program provides a search interface which includes saving session information between web page accesses. It transforms retrieved records from XML to XHTML (version 1.0) for web display. OmniMark was chosen as the programming language for this interface because it is XML (or SGML) aware according to a DTD, a knowledge which is employed for the XML translations involved, and also because of existing expertise and availability on the MIMAS machine. Other suitable languages for the web interface implementation would have been Perl, or TCL which is the basic interface language to Cheshire.

The MIMAS Metadatabase web interface provides search results in discrete 'chunks', currently 25 at a time, with 'next' and 'previous' navigation buttons. This is implemented by using the Cheshire capability to request a fixed number of records in the result set, beginning at a particular number within that set. The application remembers the MIMAS identifiers of the results in the retrieved 'chunk', and extracts the record corresponding to a particular MIMAS identifier when an end-user selects a 'full record display'.

To implement the metadata hierarchy navigation functionality, described in section 2.5, an additional index, used internally by the application, is created

https://doi.org/10.23106/dcmi.952106966

on the 'isPartOf' fields of the records which denote the MIMAS identifiers of the parent records. When a record is displayed, this index is checked to find all metadata records which indicate the current record as parent, the titles of these children records also being determined from the database. For each child record found a 'hasPart' link is displayed. Similarly the title and link for the 'isPartOf' display are determined by a database look-up.

Within the MIMAS Collection database, when a record is displayed, the MIMAS Metadatabase is checked for a record with a matching identifier. If such a record is found the display includes a 'Describes' web link from the Collection database to the corresponding record in the metadatabase.

6. Exposing OAI Metadata

The Open Archives Initiative (OAI) has specified a Metadata Harvesting Protocol [42] 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 [6]. 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* which details the database and its OAI level of support; *ListMetadataFormats* to indicate the metadata formats available, currently only simple Dublin Core (oai_dc); *ListIdentifiers* to list the identifiers of all the available records; *GetRecord* to retrieve the metadata of a particular record; *ListRecords* to list the metadata of all records; and *ListSets* which returns an empty response, sets not being supported.

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.

6.1. An Example OAI Record

An example response to a GetRecord request would be as follows, abbreviated for conciseness:

```
<?xml version="1.0" encoding="UTF-8" ?>
<GetRecord
xmlns=
"http://www.openarchives.org/OAI/1.1/OAI_GetRecord"
xmlns:xsi=
"http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation=
"http://www.openarchives.org/OAI/1.1/OAI_GetRecord
http://www.openarchives.org/OAI/1.1/OAI_GetRecord.xsd">
<responseDate>
 2002-05-28T11:59:45+01:00
</responseDate>
<requestURL>
 http://irwell.mimas.ac.uk/cgi- bin/cgiwrap/zzmetadm/
 mimas_oai?
 verb=GetRecord&identifier=oai%3Amimas%3Aze000001
  &metadataPrefix=oai_dc
</requestURL>
<record>
 <header>
  <identifier>oai:mimas:ze000001</identifier>
  <datestamp>2002-04-24</datestamp>
 </header>
 <metadata>
  <dc xmlns="http://purl.org/dc/elements/1.1/"
      xmlns:xsi=
        "http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="http://purl.org/dc/elements/1.1/
        http://www.openarchives.org/OAI/1.1/dc.xsd">
   <title>zetoc</title>
   <creator>The British Library</creator>
   <creator>MIMAS</creator>
   <subject>(Dewey) 050</subject>
   <identifier>http://zetoc.mimas.ac.uk</identifier>
  </dc>
 </metadata>
</record>
</GetRecord>
```

6.2. Date Range

The OAI protocol allows harvesters to specify they want records 'from' a certain date and/or 'until' a certain date. Selecting records added to the Cheshire database before or after a certain date, in response to an OAI request, is implemented easily when a Cheshire index has been created for the 'date loaded' (dcterms:created) field. This field is also used to provide the date stamp on returned records.

6.3. Response Restriction

With no restrictions, OAI harvesting could result in effective 'denial of service' attacks because of the machine resources required, so there is generally a

https://doi.org/10.23106/dcmi.952106966

need for an 'acceptable use' policy to restrict how many records may be harvested at one time and how soon a repeat request may be made. This is probably not a serious consideration for the MIMAS Metadatabase, which currently does not contain a large number of records, but would be a significant issue if OAI interfaces were to be implemented on some of the very large datasets at MIMAS such as *zetoc* [1]. When supplying only part of a result set, the OAI protocol allows for the return of a 'resumptionToken' which the harvester uses to make repeat requests. The format of this 'resumptionToken' is not defined in the OAI protocol but by the source application.

The MIMAS Metadatabase returns a fixed number of records or identifiers in response to one request. If there are more records available a resumptionToken is returned. Because a repeat request will contain just the resumptionToken as an exclusive argument, details of the original request are included in the token to enable a repeat of the original search on the database.

The format of the resumptionToken for the MIMAS Metadatabase is:

<database>-<start>-<from>-<until>-<format>

where:

<database> is the database identifier
<start> is the number of the next record to be retrieved within the result set
<from> is the 'from' date specified in the original request (yyyymmdd) or zero
<until> is the 'until' date specified in the original request (yyyymmdd) or zero
<format> is the metadata format specified in the original request. This may be: 'dc' for Dublin Core; 'xx' for an unsupported metadata format; 'li' for a ListIdentifiers request where metadata format is irrelevant.

For example, a resumptionToken returned by a ListRecords request for Dublin Core records from 2002-04-01 until 2002-07-01 following the first 50 records would be:

mimas-51-20020401-20020701-dc

When an OAI request includes a resumptionToken, the cgi program parses the token, then performs the original search on the database, but requesting a result set beginning at the token's <start> number. For a large result set, this search may again result in a further resumptionToken. This implementation relies on Cheshire functionality which allows a search request to return a fixed number of results beginning at a stated point within the result set.

6.4. Subject Keywords in OAI Records

Because simple Dublin Core metadata format records are supplied to OAI harvesters, there is some loss of richness in the information from the base qualified Dublin Core data, similar to that described for Z39.50 XML results in section 5.1. In particular, the subject encoding scheme used is not included, unless in parentheses as part of a subject keyword text string. Knowledge of the encoding schemes used for subject keywords would be important to services which are providing search interfaces across metadata harvested from multiple repositories, both to ensure the quality of the metadata and for comparison between subject keywords from harvested sources [24]. If a qualified Dublin Core XML schema were available, and recognised by OAI and the JISC Information Environment, then more complete metadata, including relevant encoding schemes, could be supplied to metadata harvesters from the MIMAS Metadatabase.

7. Conclusion

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 MIMAS 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 MIMAS Metadatabase provides a single point of access into the disparate, cross-domain MIMAS datasets and services. It 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.

https://doi.org/10.23106/dcmi.952106966

The MIMAS Metadatabase may be searched at http://www.mimas.ac.uk/metadata/ and the MIMAS Collection Description at http://www.mimas.ac.uk/metadata/collection/ .

Acknowledgements

The authors wish to acknowledge the contribution to the development of the MIMAS Metadatabase by their colleagues at MIMAS who support the many services and datasets and provided quality metadata, and the Cheshire development team, Ray Larson at the University of California-Berkeley and Paul Watry and Robert Sanderson at the University of Liverpool. The development of the MIMAS Metadatabase and the MIMAS Collection database was funded as part of the 'Implementing the DNER Technical Architecture at MIMAS' (ITAM) project [16] and its predecessor 'MIMAS Metadata for the DNER' [27] by the Joint Information Systems Committee (JISC) for the UK Higher and Further Funding Councils within the Distributed National Electronic Resource (DNER) programme [13].

References

[1] Apps, A. and MacIntyre, R., "Prototyping Digital Library Technologies in zetoc", *Proceedings of ECDL2002: 6th European Conference on Research and Advanced Technology for Digital Libraries, Rome, September 16-18, 2002, 2002*, accepted for publication.

[2] The Bath Profile: An International Z39.50 Specification for Library Applications and Resource Discovery. http://www.nlc-bnc.ca/bath/bp-current.htm

[3] Besser, H., "The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries", *First Monday, 7 (6)*, 2002. http://firstmonday.org/issues/issue7_6/besser/

[4] Bib-1 Attribute Set. http://lcweb.loc.gov/z3950/agency/defns/bib1.htm

[5] 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

[6] Cliff, P., "Building ResourceFinder", *Ariadne*, 30, 2001. http://www.ariadne.ac.uk/issue30/rdn-oai/

[7] The Dublin Core Metadata Initiative. http://www.dublincore.org

[8] DCMI Type Vocabulary. http://dublincore.org/documents/dcmi-type-vocabulary/

[9] Dewey Decimal Classification. OCLC Forest Press. http://www.oclc.org/dewey/

[10] DDI, Data Documentation Initiative Codebook DTD. http://www.icpsr.umich.edu/DDI/CODEBOOK/

[11] Devadason, F., Intaraksa, N., Patamawongjariya, P. and Desai, K., "Search interface design using faceted indexing for Web resources", *Proceedings of the 64th ASIST Annual Meeting*. Medford: Information Today Inc, 38, 2001, pp. 224-238.

[12] ISO 3166-1: The Code List. http://www.din.de/gremien/nas/nabd/iso3166ma/codlstp1/

[13] The UK Distributed National Electronic Resource. http://www.jisc.ac.uk/dner/

[14] Heery, R. and Patel, M., "Application profiles: mixing and matching metadata schemas", *Ariadne*, 25, 2000. http://www.ariadne.ac.uk/issue25/app-profiles

[15] ISO/TC211: Geographic Information/Geomatics. http://www.isotc211.org

[16] Implementing the DNER Technical Architecture at MIMAS (ITAM) project. http://epub.mimas.ac.uk/itam.html

[17] JISC, The Joint Information Systems Committee. http://www.jisc.ac.uk

[18] Koch, T. and Day, M., "The role of classification schemes in Internet resource description and discovery", *Work Package 3 of Telematics for Research project DESIRE (RE 1004)*, 1999. http://www.ukoln.ac.uk/metadata/desire/classification/

[19] Larson, R.R., Cheshire II Project. http://cheshire.lib.berkeley.edu

[20] Larson, R.R., McDonough, J., O'Leary, P., Kuntz, L., Moon, R., "Cheshire II: Designing a Next-Generation Online Catalog", *Journal of the American Society for Information Science*, 47 (7), 1996, pp. 555-567.

[21] Leuf, B. and Cunningham, W., The Wiki Way. http://www.wiki.org

[22] Library of Congress Subject Headings. *Cataloguing Distribution Service*, Library of Congress. http://lcweb.loc.gov/cds/lcsh.htm

[23] The Z39.50 Generic Record Syntax (GRS-1) Tagsets. http://lcweb.loc.gov/z3950/agency/defns/tag-gm.html

https://doi.org/10.23106/dcmi.952106966

[24] Liu, X., Maly, K., Zubair, M., Hong, Q., Nelson, M.L., Knudson, F. and Holtkamp, I., "Federated Searching Interface Techniques for Heterogeneous OAI Repositories", *Journal of Digital Information*, 2 (4), 2002, http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/

[25] Miller, P., "Z39.50 for All", *Ariadne*, 21, 1999, http://www.ariadne.ac.uk/issue21/z3950

[26] MIMAS, Manchester Information and Associated Services, including Archives Hub, COPAC, ISI Web of Science, JSTOR, NESLI, zetoc. http://www.mimas.ac.uk

[27] MIMAS Metadata for the DNER project. http://epub.mimas.ac.uk/dner.html

[28] Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. http://www.niso.org/standards/resources/Z3950.pdf

[29] Open Archives Initiative. http://www.openarchives.org

[30] Olivier, B., Liber, O. and Lefrere, P., "Specifications and standards for learning technologies: the IMS project", *International Journal for Electrical Engineering Education*, 37 (1), 2000, pp. 26-37.

[31] OmniMark Technologies, http://www.omnimark.com

[32] Powell, A., "RSS FAQ, JISC Information Environment Architecture", 2002. http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/faq/rss/

[33] Powell, A. and Lyon, L., "The JISC Information Environment and Web Services", *Ariadne*, 31, 2002. http://www.ariadne.ac.uk/issue31/information-environments.

[34] Powell, A., Heaney, M. and Dempsey, L., "RSLP Collection Description", *D-Lib Magazine*, 6 (9), 2000. doi:/10/1045/september2000-powell.

[35] RDN, Resource Discovery Network. http://www.rdn.ac.uk

[36] Renardus. http://www.renardus.org

[37] Tinker, A.J., Pollitt, A.S., O'Brien, A. and Braekevelt, P.A., "The Dewey Decimal Classification and the transition from physical to electronic knowledge organisation", *Knowledge Organization*, 26 (2), 1999, pp. 80-96.

[38] UNESCO Thesaurus. http://www.ulcc.ac.uk/unesco/

[39] 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

[40] Vizine-Goetz, D., "Using Library Classification Schemes for Internet Resources", E. Jul, ed. *Proceedings of the OCLC Internet Cataloguing Colloquium, San Antonio, Texas, 19 January 1996*. OCLC, 1996. http://www.oclc.org/oclc/man/colloq/toc.htm

[41] W3C, Date and Time Formats. http://www.w3.org/TR/NOTE-datetime

[42] Warner, S., "Exposing and Harvesting Metadata Using the OAI Metadata Harvesting Protocol: A Tutorial", *High Energy Physics Webzine*, *4*, 2001, http://library.cern.ch/HEPLW/4/papers/3/

[43] Watry, P. and Hill, A., "Collection Description Service Scoping Study", JISC Report, 2002.

https://doi.org/10.23106/dcmi.952106966