zetoc SOAP: a Web Services Interface for a Digital Library Resource

Ann Apps

MIMAS, University of Manchester, M13 9PL, UK
ann.apps@man.ac.uk

Abstract. This paper describes the provision of a Web Services interface that will extend the possibility of machine-to-machine access to the **zetoc** current awareness service, within the JISC Information Environment and eScience applications. This bespoke interface includes open standard XML metadata for searches and responses where possible. Elements from the OpenURL XML metadata formats for journals and books are used to transmit the bibliographic citation information that is an integral part of a **zetoc** record for a journal article or conference paper.
Keywords. Web Services, bibliographic citation, metadata, OpenURL, Dublin Core, SRW.

1 Introduction

zetoc [1] [2] is a current awareness and document delivery service based on the British Library's [3] Electronic Table of Contents of journal articles and conference papers. Hosted at MIMAS [4], **zetoc** is available to researchers, teachers and learners in UK Higher and Further Education under a 'strategic alliance' [5], and to practitioners within the UK National Health Service. The **zetoc** database, updated daily, contains details of articles from approximately 20,000 current journals and 16,000 conference proceedings published per year. With over 20 million article and conference paper records from 1993 to date, the database covers every imaginable subject in science, technology, medicine, business, law, finance and the humanities. Human users can search **zetoc** through its Web interface to retrieve articles by one of the document delivery options. They may also use the popular email alert service to maintain current awareness of new articles of possible interest. Machine-to-machine searching is available by Z39.50 [6], the NISO (North American National Information Standards Organization) standard for information retrieval that provides a protocol for two computers to communicate and share information. It is also enabled by OpenURL [7], another standard way of passing information between machines, **zetoc** being enabled as an OpenURL 'link-to' resolver.

A new Web Services SOAP [8] (the World Wide Web Consortium's server-to-server protocol for object retrieval) interface to **zetoc** has been developed as part of the A2Z (Akenti acces to **zetoc**) project [9], the main purpose of which was to investigate digital certificate authentication, in particular using Akenti

[10], within eScience applications as well as within the Information Environment [11] under development by the JISC (the Joint Information Services Committee of the UK Higher and Further Education Funding Councils). Because eScience projects are generally outside the digital library domain, experimenting via the **zetoc** Z39.50 interface did not seem appropriate. Whereas a Web Services interface would be suitable for use in areas such as workflow modelling of composite services within eScience projects such as myGrid [12]. Within the digital library based JISC Information Environment portals and virtual learning development projects are beginning to use Web Services for machine-to-machine communication.

2 A SOAP Interface for zetoc

A Web Services interface deals with messages that are sets of XML elements wrapped within SOAP envelopes. The requesting and responding servers, both understanding the SOAP protocol, are able to extract the XML data from the messages sent to them and to package their responding XML results accordingly. SOAP messages are passed between machines by RPC (Remote Procedure Call). The **zetoc** SOAP interface is implemented by RPC over the Web Common Gateway Interface (CGI), and thus its address appears like a URL.

Provision of a Web Services interface is in two parts. Firstly a 'search request' is needed to submit search terms for discovery to the application. Secondly a 'search response' will return details of the results. To design the **zetoc** SOAP interface there appeared to be two options: use a standard or generally accepted schema; or develop a bespoke interface, that is an interface designed specifically for the particular service.

SRW (Search - Retrieve - Web) [13] is a specification for general search request and response developed under the auspices of the Z39.50 community, with the possibility of becoming a NISO standard for meta-searching [14]. SRW emulates Z39.50 by including various fields to return the number of search hits and to request the start position within the result set of the returned records. For the actual search SRW provides a Common Query Language (CQL) to enable Z39.50-like and interoperable search requests. The expected returned response for each record within the result set is simple Dublin Core [15].

In fact the **zetoc** SOAP interface developed as part of the A2Z project is bespoke, although based on open standard metadata schemes where possible. It seems that SRW is ideal for distributed searching within a wide domain such as the JISC Information Environment because it allows common search requests to be sent to a range of services. However SRW seems less appropriate for making a connection to a single service whose capability and specific domain is well understood. Similarly returning simple Dublin Core records provides clear interoperability for distributed searching. But a simple Dublin Core description for a result that is a bibliographic record would lose richness and significant detail, in particular the bibliographic citation information for a journal article or conference paper.

The XML elements that make up the various requests and responses of the **zetoc** SOAP interface are defined formally [16] as a Dublin Core Application Profile [17]. An application profile was a useful way to document all of these properties and their corresponding namespaces. However strictly a Dublin Core application profile is a flat structure, so some distortion of the application profile has been made to specify the hierarchical structure of the returned result set. This application profile effectively defines the **zetoc**-specific terms within a **zetoc** namespace.

3 Metadata for zetoc SOAP

Having decided to implement a bespoke SOAP interface for **zetoc**, it was desirable to use metadata properties from open standards wherever possible.

3.1 zetoc Search Request

The available **zetoc** SOAP search requests replicate the searches available on the **zetoc** Web interface. Thus three search requests are provided: general that searches over all the data; journal article; and conference paper. The search fields include the obvious possibilities such as 'all fields', article title, author, publication year, and ISSN, to specify a journal, or ISBN, a book identifier to specify a conference proceedings. The journal and conference searches include more specific fields related to those genre. To support the retrieval of large result sets in manageable chunks the **zetoc** SOAP search requests also need to indicate the position within the result set of the first record to be returned. Currently no Boolean operators are available. As in the **zetoc** Web interface, when several search terms are provided the implicit Boolean operator is 'and'.

There is an additional, fourth, 'identifier' request that returns a single **zetoc** full record corresponding to a specific **zetoc** identifier.

3.2 zetoc Response

The three search requests result in a search response that is a list of brief descriptions of **zetoc** records matching the search. The 'identifier' request results in a single, full **zetoc** record.

To avoid returning unmanageably large result sets, the **zetoc** search response is a list of a fixed number (25) of brief records. Thus the response must include the total number of hits and the number of the next record in the result set following those returned. Along with the 'first record position' requested, this data enables repeated requests to obtain the full result set. An indication of the search performed is also returned.

The brief records returned correspond to the **zetoc** brief records available from a Web search, including the position of the record within the entire result set, but with the addition of the **zetoc** identifier. Returning the **zetoc** identifier with the brief record enables a subsequent 'identifier' request to retrieve the full details of an item of particular interest.

An 'identifier' response results in a single, full **zetoc** record that corresponds to a full **zetoc** record available from the Web interface. It includes all details about the article or paper available from the database.

3.3 Dublin Core Properties

For maximum interoperability properties are taken from Dublin Core where possible. Thus from the simple Dublin Core namespace ('dc') the following terms represent: dc:title, the article or paper title; dc:creator, the authors; and dc:identifier, the identifier of the resource within the **zetoc** database, currently an identifier local to the **zetoc** service. In a search request 'title' and 'creator' could contain keywords from the field rather than the entire value. In addition dc:subject (for conference keywords), dc:contributor (for editors), dc:publisher, dc:language, dc:format and dc:type are used to return some detailed information in an 'identifier' full record response.

From the wider Dublin Core namespace ('dcterms') the following properties represent: dcterms:issued, the publication year of an article; and dcterms:bibliographicCitation, the citation details in a brief record of a search response.

3.4 SRW and Z39.50 Bath Profile Properties

The SRW namespace includes obvious properties to implement the retrieval of large result sets in manageable pieces. Thus from the SRW namespace ('srw') are taken: srw:numberOfRecords, the total number of search hits; srw:startRecord, the requested start position; srw:nextRecordPosition, the number of the record following those returned; and srw:recordPosition, the number within the result set of each brief record returned.

The Bath Profile [18] is a derivation of Z39.50 for digital library applications, defining search request attributes. From this namespace ('bath') these search request terms are taken: bath:any, an 'all fields' search; and bath:conferenceName, the conference details in a conference paper search.

3.5 OpenURL Properties

Because **zetoc** is a citation database providing bibliographic information to enable article requests, it is essential that the **zetoc** SOAP interface includes bibliographic details in its search requests and responses. It was preferred that this bibliographic information be passed using open standard properties where possible. Dublin Core provides 'dcterms:bibliographicCitation', which is used to return the information as a string value within a brief record response. But Dublin Core does not provide bibliographic properties at any finer granularity.

OpenURL was developed as a standard way of passing information about a resource between a source application and an OpenURL-aware resolver [19]. Its original and primary purpose is to enable a researcher to link from a referenced article to a full text copy of that article where the researcher's institution has a valid subscription to read the article. During the process of proposing the OpenURL Framework as a NISO standard, Z39.88-2004 [20], other possible uses of OpenURL were envisaged including server-to-server communication. Thus an XML schema for the OpenURL 'payload' (the ContextObect) was developed. The OpenURL Framework is extensible by means of a Registry [21]. The initial content of the OpenURL Registry, and hence the standard, includes metadata formats as XML schema for journals and books.

The OpenURL journal [22] ('oujnl') and book [23] ('oubook') metadata formats are used to capture the bibliographic citation properties within **zetoc** SOAP. Thus from the OpenURL journal metadata format are taken: oujnl:jtitle, the journal title; oujnl:issn, the journal ISSN; oujnl:volume, oujnl:issue and oujnl:spage, the volume and issue number and start page of an article within a journal search request; and oujnl:pages, the page range of an article or paper in a full record response. Similarly from the OpenURL book metadata format are taken: oubook:isbn, the ISBN of a conference proceedings; and oubook:spage, the start page within a conference paper search.

3.6 zetoc Properties

Although open standards are used as far as possible it was necessary to include several **zetoc**-specific properties within a **zetoc** namespace. This namespace includes all the containing XML elements of **zetoc** SOAP comprising the search and 'identifier' requests and responses, and the brief record and its containing array. The only **zetoc** term within the search requests is a field 'ISSN or ISBN' included in the general search. The only **zetoc** term in a search response holds a string value representing the search performed on the **zetoc** database.

Inevitably the full record 'identifier' response includes several **zetoc**-specific terms, for example the British Library's 'shelfmark' and 'location' information, and the frequency of publication for some journals. A 'zetoc:type' property indicates whether a returned record is for a journal article or a conference paper.

Subject terms are available in **zetoc** as Dewey and Library of Congress Classification. These are returned as properties 'dewey' and 'lccn' in the **zetoc** namespace. Ideally they would be returned as 'dc:subject' with an XML attribute 'xsi:type' of 'dcterms:DDC' or 'dcterms:LCC', something that may be implemented in future versions of **zetoc** SOAP.

Within the **zetoc** database journal volume and issue information is run together into a single field, reflecting the data supply from the British Library, thus necessitating another **zetoc**-specific property 'volissue'. A **zetoc**-specific term is used to return any journal issue title, for example the name of a special issue, recorded in **zetoc**, this being outside the scope of general journal metadata formats.

There did not appear to be an existing open standard metadata scheme to describe conference details. It would be possible to record the proceedings as a book title using the OpenURL book metadata format, but this would not necessarily capture all the data about the conference such as its venue and date. Thus a term within the **zetoc** namespace is used to return all the conference details concatenated into a string value. Another **zetoc** field returns the conference sponsors.

3.7 Examples

Journal Search. Some possible fields in a journal search request may be as in Table 1.

Table 1. Example journal search terms

Property	Value
dc:creator	apps
oujnl:jtitle	materialia
oujnl:issn	1359-6462
oujnl:volume	48
oujnl:issue	5
oujnl:spage	475
dcterms:issued	2003

Search Response. A search response for the above example would return a list of brief records containing the single record shown in Table 2.

Table 2. Example brief record response

Property	Value
srw:recordPosition	1
dc:title	Phase compositions in magnesium-rare earth alloys
dc:creator	Apps, P. J.; et-al
dcterms:bibliographicCitation	SCRIPTA MATERIALIA - 2003; VOL 48; NUMBER 5; Pages: 475-481
dc:identifier	RN125218404

'Identifier' Response. The 'identifier' full record response (omitting conference paper properties that are irrelevant to this article) would be as in Table 3.

Table 3. Example full record response

Property	Value
srw:numberOfRecords	1
dc:identifier	RN125218404
zetoc:type	J (ie. journal)
dc:title	Phase compositions in magnesium-rare earth alloys...
dc:creator	Apps, P. J.; Karimzadeh, H; King, J. F.; Lorimer, G. W.
zetoc:dewey	669
zetoc:lccn	TT273
oujnl:jtitle	SCRIPTA MATERIALIA
oujnl:issn	1359-6462
zetoc:volissue	VOL 48; NUMBER 5
oujnl:pages	475-481
dcterms:issued	2003
dc:publisher	Great Britain : Elsevier Science B.V., Amsterdam.
zetoc:frequency	Fortnightly: 15-30 issues per year
dc:language	English
zetoc:shelfmark	8212.970000

3.8 An Alternative 'Identifier' Response

An alternative approach to implementing a full record 'identifier' response would be to return a simple Dublin Core record for the discovered article, including salient information such as its title and authors. This simple Dublin Core record would contain a 'by-reference' link, a pointer as the value of a 'dc:relation' property, to a full **zetoc** XML record. This pointer could be an OpenURL that would return an XML record for the item in **zetoc** as in the following example. Note that in this example: a hypothetical resolver address is used, and an actual OpenURL would be 'URL escape encoded', with special characters in hexadecimal format for safe transmission, but this encoding has been omitted, and line-breaks have been added to the OpenURL, for readability.

```
http://zetoc.mimas.ac.uk/openurl/linkto?
    url_ver=Z39.88-2004
    &url_ctx_fmt=info:ofi/fmt:kev:mtx:ctx
    &rft_val_fmt=info:ofi/fmt:kev:mtx:dc
    &rft.identifier=RN125218404
    &svc_val_fmt=info:ofi/fmt:kev:mtx:dc
    &svc.format=text/xml
```

This OpenURL uses the Dublin Core metadata format to describe the **zetoc** record recquired, the referent, as a **zetoc** identifier. That identifier being local to **zetoc** makes an OpenURL 'referent-identifier' key inappropriate. The Dublin Core metadata format is also used to request a service type that returns an XML record.

An alternative OpenURL could use private data to pass the **zetoc** identifier, in which case the fourth and fifth lines of the above example would be replaced by:

```
&rft_dat=RN125218404
```

If the **zetoc** identifier were to become a URI, possibly by registering it within the new 'info' URI scheme [24], then the fourth and fifth lines of the above example could be replaced by the preferable:

```
&rft_id=info:zetoc/RN125218404
```

The returned record could possibly be an XML Dublin Core description with related metadata using the OpenURL journal or book metadata format as suggested in [25].

The advantage of this approach is that the 'identifier' response would return an interoperable simple Dublin Core record. The disadvantage is that any service retrieving this record would have to make a further retrieval to obtain the full **zetoc** reord, including its bibliographic citation information that could not be captured in the simple Dublin Core record. This approach would be suitable for returning a simple Dublin Core record from a **zetoc** SRW implementation.

4 Authentication

zetoc is available to members of institutions in UK Higher and Further Education and the UK National Health Service. It is also provided by modest subscription to various other bodies in UK academia, including the research councils, and to institutions in Ireland. Authentication is firstly by a machine domain name (DNS) or IP address check, and failing that by Athens [26], the access authorisation system used within UK Higher and Further Education. **zetoc** SOAP allows access using the same machine address checks. Access via Athens is not supported, human intervention not being possible. The same terms and conditions for the use of **zetoc** apply. This means that any portal must first check that a user has a right to use **zetoc** before providing a search through the **zetoc** SOAP interface.

The A2Z project has investigated and succcessfully demonstrated the use of Akenti digital certificate authenticated access to the **zetoc** Web interface, as reported eslewhere [9]. The original intention of the 'Web Services' part of the A2Z project was to investigate the use of digital certificate authenticated access to a **zetoc** SOAP interface within an eScience application such as myGrid. However it became apparent that this was not viable within the time frame of the project because digital certificates are not yet in use by the potential user base. Thus providing digital certificate authentication has not been taken forward. It would simply involve replacing the current machine address authentication module with the A2Z digital certificate 'black box' module and installing the access point to **zetoc** SOAP on the A2Z secure server.

5 Implementation

zetoc SOAP is implemented in C++ using gSOAP. gSOAP is a set of compiler tools that provide a SOAP/XML-to-C++ language binding to ease the development of SOAP/XML Web services and client applications in C++. Developed by Prof Robert van Engelen and his team in the Department of Computer Science and School of Computational Science and Information Technology at Florida State University, USA [27], it is available under a GNU licence from SourceForge [28]. gSOAP is used to implement several major applications, including Adobe Version Cue, an innovative file-management feature of Adobe Creation Suite.

gSOAP takes care of all the details of the XML to support the SOAP protocol and also the serialisation of the XML elements of the **zetoc** requests and responses to and from the C++ public data of the **zetoc** SOAP server implementation. gSOAP also generates the requisite WSDL (Web Services Description Language) file that provides a machine readable definition of the interface. Requests are translated into searches in the format of the underlying Livelink Discovery Server (previously known as BRS/Search) [29] database, the searches being performed by existing C++ code modules. A **zetoc** SOAP client was implemented with gSOAP alongside the server for testing.

6 Conclusion

The use of Web Services is becoming increasingly important for machine-to-machine communication. It is already used within eScience Grid applications and projects. It is mandated for machine-to-machine applications within the UK government's interoperability framework (eGif) [30]. Within the JISC Information Environment portals are starting to use Web Services, the Resource Discovery network (RDN) [31] has a SOAP/SRW interface, and collections with Web Services access are recorded in the Information Environment Service Registry [32].

As discussed above, the **zetoc** SOAP interface is bespoke rather than using SRW. The short timescale available for the development of the **zetoc** SOAP interface within the funding of the A2Z project did not allow for any investigation into the provision of an SRW interface. A future SRW interface for **zetoc**, given the availability of funding, will be developed to allow its inclusion in Web Services distributed search requests within the JISC Information Environment, although distributed searching is already enabled via Z39.50. But it seems appropriate to provide an interface to an application that is specific to its data and purpose. The Common Query Language of SRW will provide interoperability but it seems to be too general when specific requests and results of an application are required. Also SRW will allow search requests inappropriate to an application resulting in null or distorted responses. For example a search for 'dc:description' in **zetoc** would return results from around 1994 only, later records not having abstracts.

Similarly the requirement of SRW to return simple Dublin Core records provides an interoperable result set for a distributed search but it does not cater for the return of richer application-specific details. **zetoc** SOAP provides a bespoke result format to include the important bibliographic citation details necessary to make use of any record from **zetoc**. There is no recommended way to include bibliographic citation information about a resource within a simple Dublin Core record. The alternative approach given above in section 3.8 would resolve this problem if **zetoc** were to provide an interoperable SRW impementation to support distributed searching. But it seems that a retrieval by a server that understands the **zetoc** application would be simpler using the bespoke interface.

Resembling the **zetoc** Web search interface, the **zetoc** SOAP interface does not allow the inclusion of Boolean operators in search requests, all search terms being implicitly 'anded'. Future developments to the **zetoc** SOAP interface would investigate the provision of Boolean operators between search terms when assembling searches on the underlying database. This functionality would be enabled if the Common Query Language of SRW were supported.

Developing **zetoc** SOAP has been a useful experience in exploring the design and specification of such an interface and the issues involved. Investigation into alternative implementations, such as SRW, was limited by the short timescale of this part of the A2Z project. But this development will provide a prototype for Web Services implementations for further information collections.

Acknowledgements. The A2Z project [33], including the **zetoc** SOAP interface development, was funded by the Joint Information Systems Committee (JISC) [34] of the UK Higher and Further Education Councils as part of the 'AAA' (Authentication, Authorisation, and Accounting) programme [35]. The **zetoc** data is provided by the British Library. The author wishes to acknowledge the assistance of colleagues in the development of **zetoc** SOAP, Ashley Sanders for the C++ search interface to **zetoc**, Ross MacIntyre the project manager and Mike Jones who implemented the Akenti authentication 'black box'.

References

1. Apps, A., MacIntyre, R.: Prototyping Digital Library Technologies in zetoc. Lecture Notes in Computer Science. **2458** (2002) 309-323
2. zetoc, Electronic Table of Contents from the British Library. `http://zetoc.mimas.ac.uk`
3. The British Library. `http://www.bl.uk`
4. MIMAS, a UK Higher and Further Education Data Centre. `http://www.mimas.ac.uk`
5. Strategic alliance emphasises British Library's central role in support of higher education. Press release, 19 March 2003. `http://www.bl.uk/cgi-bin/press.cgi?story=1231`
6. Z39.50, the North American National Information Standards Organisation (NISO) standard for information retrieval. `http://www.niso.org/standards/resources/z3950.pdf`
7. NISO Committee AX, Apps, A. Z39.88-2004: The Key/Encoded Value Format Implementation Guidelines. (2004). `http://www.openurl.info/registry/docs/implementation_guidelines`

8. W3C Web Services SOAP. http://www.w3.org/2000/xp/Group/
9. Jones, M.A.S., Apps, A., Hewitt, W.T., MacIntyre, R., Sanders, A., Weeks, A.: Akenti Access to zetoc. Paper and Poster at AHM2003 eScience 'All Hands' Meeting, Nottingham, 2-4 September 2003. (2003)
10. Akenti Distributed Access Control. http://www.itg.lbl.gov/akenti
11. JISC.: Investing in the Future: Developing an Online Information Environment. (2003). http://www.jisc.ac.uk/index.cfm?name=ie_home
12. myGrid project. http://www.ebi.ac.uk/mygrid/
13. SRW: Search - Retrieve - Web. http://www.loc.gov/z3950/agency/zing/srw/
14. NISO MetaSearch Initiative. http://www.niso.org/committees/MS_initiative.html
15. The Dublin Core Metadata Initiative. http://www.dublincore.org
16. Apps, A.: zetoc SOAP Interface Application Profile. (2004). http://zetoc.mimas.ac.uk/soap/
17. CEN/ISSS CWA 14855, Dublin Core Application Profile Guidelines. (2003). ftp://ftp.cenorm.be/PUBLIC/CWAs/e-Europe/MMI-DC/cwa14855-00-2003-Nov.pdf
18. Z39.50 Bath Profile Indexes. http://www.loc.gov/z3950/agency/zing/cql/bath-indexes/v1.0/
19. 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
20. ANSI/NISO.: Z39.88-2004, The OpenURL Framework for Context-Sensitive Services. Available via http://library.caltech.edu/openurl/Standard.htm
21. Registry for the OpenURL Framework. http://www.openurl.info/registry/
22. OpenURL Journal XML Metadata Format. http://www.openurl.info/registry/docs/info:ofi/fmt:xml:xsd:journal
23. OpenURL Book XML Metadata Format. http://www.openurl.info/registry/docs/info:ofi/fmt:xml:xsd:book
24. 'info' URI Scheme. http://info-uri.info
25. Apps, A., MacIntrye, R.: Using the OpenURL Framework to Locate Bibliographic Resources. In: Proceedings of the 2003 Dublin Core Conference (DC2003 - Supporting Communities of Discourse and Practice - Metadata Research and Application), Seattle, Washington, USA, 28 September - 2 October 2003. ISBN 0-9745303-0-1. (2003) 143-152
26. Athens Access Management System. http://www.athens.ac.uk
27. Van Engelen, R.: gSOAP. http://www.cs.fsu.edu/~engelen/soap.html
28. gSOAP from SourceForge. http://sourceforge.net/projects/gsoap2
29. Livelink Discovery Server. http://www.opentext.com/brs/
30. The UK e-Government Interoperability Framework (eGif). http://www.govtalk.gov.uk/egif/contents.asp
31. RDN, the Resource Discovery Network. http://www.rdn.ac.uk
32. The JISC Information Environment Service Registry (IESR). http://www.mimas.ac.uk/iesr/
33. The A2Z Project. http://a2z.mimas.ac.uk
34. JISC, the Joint Information Systems Committee of the UK Higher and Further Education Funding Councils. http://www.jisc.ac.uk
35. The JISC 'AAA' (Authentication, Authorisation and Accounting) programme. http://www.jisc.ac.uk/index.cfm?name=aaa_docs