

Joining Up ‘Discovery to Delivery’ Services

Ann Apps; Ross MacIntyre

Mimas, The University of Manchester
M13 9PL, UK

e-mail: ann.apps@manchester.ac.uk; ross.macintyre@manchester.ac.uk

Abstract

Zetoc is a bibliographic current awareness service that provides discovery of relevant literature within the British Library’s Electronic Table of Contents of journal articles and conference papers. A researcher having discovered an article of interest will wish to read it, preferring to locate an electronic copy of an article to be delivered directly to their desktop. However, until now, Zetoc was essentially the British Library’s document delivery catalogue, containing details of journals that are published traditionally. The lack of open access articles in Zetoc, because there would be no reason to order and pay for copies of these articles, implied a deficiency in Zetoc as a current awareness and general article discovery service. This paper describes the introduction of open access article records into Zetoc by OAI-PMH harvesting from UK PubMed Central. The prototype concentrates on biomedicine and initially BioMed Central journals. But the paper discusses future extension to other disciplines, as well as general requirements for sharing bibliographic article records.

Keywords: open access; bibliographic article record; OAI-PMH; harvest; biomedical research services.

1. Introduction

Zetoc [1] is a bibliographic current awareness service that provides discovery of relevant literature within the British Library’s Electronic Table of Contents of journal articles and conference papers. Hosted at Mimas, at The University of Manchester, Zetoc is available to researchers, learners and teachers in UK Higher and Further Education, as well as to members of several other organisations including Irish colleges and health care trusts. The Zetoc database holds details of around 20,000 current journal articles and 16,000 conference proceedings per year, covering all disciplines, data being available from 1993 and updated daily. Searches using the Zetoc Web or Z39.50 interfaces yield bibliographic citation details of the discovered articles. Zetoc Alert is an alerting service, via either email or RSS feeds, which provides tables of contents of new issues of journals. Each article in an alert is accompanied by a persistent URL enabling direct access to its full record within the Zetoc Web interface, and hence location of the article.

Once a researcher has discovered an article of interest, they will wish to read it, and therefore to locate, and request delivery of, its full text. Zetoc provides access to the British Library’s document delivery service, which requires payment, and also assistance with requesting articles from an institution’s library via traditional Inter-Library routes. But researchers prefer to locate an electronic copy of an article to be delivered directly to their desktop [2]. Thus the full record of an article in Zetoc also includes a link to the reader’s institution’s OpenURL resolver [3] if this service is available, or alternatively to free article location services including Google Scholar [4], Sirius [5] and Copac [6].

However, because Zetoc is essentially the British Library’s document delivery catalogue, it contains details of journals that are published traditionally, the majority still published in print as well as electronically. Zetoc does not include open access electronic journals because there would be no reason to order and pay for copies of these articles. Thus a natural extension to Zetoc, as a current awareness and general article discovery service, is to include details of open access articles where there is no overlap with existing content.

Mimas also hosts UK PubMed Central (UKPMC) [7], based on PubMed Central, the US National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature; providing this service jointly with the British Library. UKPMC provides an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interface [8], which includes timely details of new articles. This raised the possibility of incorporating details of open access journal articles in UKPMC into Zetoc, both into the searchable database and into Zetoc Alert. Because these articles are open access, it is possible to provide a direct link from the Zetoc full record to the full text of the article in UKPMC.

2. Methodology

2.1. Domain and Publisher for Prototype

It was decided to concentrate initially on importing details of biomedical open access journals into Zetoc, in particular those published by BioMed Central (BMC) [9]. BMC states the ‘all research articles published by BioMed Central are archived without delay in PubMed Central’.

2.2. Data Mapping

The first task was to map the UKPMC data within the OAI-PMH feed to the properties within the Zetoc namespace, to investigate any significant omissions and ascertain any requirements for new fields. The UKPMC OAI-PMH metadata format chosen is the article header (or front matter) details (‘pmc_fm’), the simple Dublin Core (‘oai_dc’) metadata format lacking precision. The disseminated UKPMC article header details are in XML according to the NLM DTD [10]. UKPMC also makes the full text articles available via a further OAI-PMH metadata format (‘pmc’), but that is beyond requirements because Zetoc contains only bibliographic citation details of articles. The data mapping to existing data fields in the Zetoc namespace is shown in Table 1.

Zetoc	UKPMC
Article Title	<article-title>
Author	<contrib contrib-type=“author”>
Journal Title	<journal-title>
ISSN	<issn pub-type=“ppub”>
Publication Year	<pub-date pub-type=“ppub”><year>
Date of Publication	<pub-date pub-type=“epub”><day>,<month>,<year>
Volume / Issue	<volume> / <issue>
Pagination	<fpage>,<lpage>
Publisher	<publisher><publisher-name>
Abstract	<abstract>
Keywords	<kwd-group><kwd>

Table 1: Data Mapping from UKPMC to Existing Zetoc Data Fields

It was necessary to add several new fields to the Zetoc namespace and database to accommodate significant details in the UKPMC data records. Articles are identified in UKPMC by their PubMed identifier (PMID), so its capture was essential for providing a persistent URL for direct access to the full article. The Digital Object Identifier (DOI) is also a persistent URI, enabling a link to the publisher’s copy of the article. The PMID is also used to construct a Zetoc identifier for the imported articles. The new fields are shown in Table 2.

Zetoc	UKPMC
PubMed Identifier	<article-id pub-id-type=“pmid”>
Digital Object Identifier (DOI)	<article-id pub-id-type=“doi”>
eISSN	<issn pub-type=“epub”>
Copyright	Constructed as agreed with BioMed Central

Table 2: New Zetoc Data Fields

The Zetoc Alert application requires a unique identification (within the Zetoc namespace) for a journal; for this it uses the British Library’s ‘shelfmark’. A similar identification for the new BMC journals was needed that does not conflict with the existing shelfmarks. This pseudo-shelfmark is constructed from the letters ‘PM’, the ISSN, and matching punctuation.

BMC open access articles are available according to a Creative Commons [11] ‘attribution required’ licence, with no restrictions on sharing or remixing or commercial use. Following discussion with BioMed Central, a copyright statement was agreed, for instance, where the first author’s family name is ‘Smith’: “© 2008 Smith et al; This article is distributed under the terms of the Creative Commons Attribution Licence (<http://creativecommons.org/licenses/by/2.0>)”.

There are a few data fields in the UKPMC record that are ignored on import to Zetoc. This is because they are irrelevant to Zetoc, for example publication history details. Or it is because they do not have a matching field in Zetoc, for instance author affiliations, which would need not only a new Zetoc field, but also cross-referencing to author names. However in the future author affiliations may be included to support improved author identification, which is becoming an increasing requirement, including: authenticating a researcher's work for assessment reporting, for which Zetoc has been used recently; or finding an author's papers in open access repositories. This is an area under investigation by various projects, including the Names Project [12], which has been informed by Zetoc as one of its sources of author data.

2.3. Data Enhancement

The noticeable omission in UKPMC article data is a subject classification, other than the author's keywords. Journals in Zetoc have Dewey Decimal Classification (DDC) [13] and Library of Congress Classification [14] terms, reproduced in each article record. The British Library provided DDC subject classifications for the BMC journals, which are added to the records during processing of the UKPMC import, via a look-up table using ISSN as its key. A process for the inclusion of DDC terms for new journals will be developed; Zetoc support staff will request these from the British Library and include them into the application via an administration interface. This introduces the possibility of a future extension to include improved subject classifications, which would enable possible future enhancements to Zetoc with more accurate subject-based end-user searching or alert selection, especially if coupled with a DDC-based terminology service such as the High-Level Thesaurus project (HILT) [15].

Another look-up table is needed for the publisher's country; UKPMC records the city. Currently this table contains only one entry for BioMed Central in Great Britain. Again Zetoc support staff will be able to update this table using an administration interface.

2.4. Data Load Implementation

Zetoc harvests open access OAI-PMH data (set = 'pmc-open') from UKPMC and selects from it articles published by BioMed Central. This harvest occurs nightly immediately following the upload of the British Library's Electronic Table of Contents data. Both sets of article records are aggregated to inform the Zetoc Alert service.

The implementation sends a RESTful URL over HTTP Get [16] according to the OAI-PMH protocol, which returns XML according to the article header part of the NLM DTD. The transformation of the UKPMC data is driven by a mapping template, defined in XML. This generic model should assist the future introduction of data harvesting and transformation from different suppliers and formats. The implementation is based on a similar application for the JISC Information Environment Service Registry (IESR) [17], which harvests details of ESDS International [18] macro and micro economic datasets from the UK Data Archive [19], transforming them from DDI metadata [20] into IESR format.

For the initial prototype only current data is imported into Zetoc. But it is planned to import all the BioMed Central back data from UKPMC, whose records date from 2001. This will enhance the corpus of biomedical articles in Zetoc for discovery.

3. Results

3.1. Data Harvest

Zetoc is able to find details of BMC open access journal articles in a timely fashion by selecting them from the daily UKPMC OAI-PMH feed, and thus to provide opportune alerts to researchers. Since its introduction, the harvest of BMC article details from UKPMC has been trouble free. The quantity of new articles introduced into Zetoc is relatively small: 1325 BMC articles in a month, as opposed to 162300 journal articles from the British Library.

It became apparent that some BMC articles do not have PubMed identifiers in the UKPMC data feed, and these are ignored in the Zetoc data load. On investigation it appears that the articles without PMIDs are supplementary or review articles. But PubMed Central have agreed to include PMIDs in the OAI-PMH data in the future wherever possible.

There was some concern during the initial design phase, that not all BMC content is open access, thus implying inaccessible links to the full article. However, on investigation, only open access BMC content is added to UKPMC; all research papers being open access.

3.2. Article Location

Figure 1 shows the full record for a BMC article in Zetoc; Figure 2 shows the document delivery links below the record. The result of following the UKPMC link is shown in Figure 3.

The screenshot shows a web browser window titled "Zetoc - Full Record Display - Windows Internet Explorer". The address bar contains a URL from zetoc.stmas.ac.uk. The page content is as follows:

Zetoc Full Records

Search terms: journal_1471-2105 (sorted on reverse date) [Reverse] [Reset]

Displaying record 4 of 91

4. Author(s): Yan, C.; Hu, J.; Wang, Y.
 Article Title: Discrimination of outer membrane proteins with improved performance

Journal Title: BMC Bioinformatics
 ISSN: 1471-2105
 eISSN: 1471-2105
 Year: 2008
 Volume/Issue: VOL 9
 Page(s): 47-47

Publisher: Great Britain : BioMed Central
 Date published: 20080124
 Dewey Class: 570.285
 PubMed ID: 18218108
 DOI: 10.1186/1471-2105-9-47
 BLDSC shelfmark: PM14.712105

Copyright: © 2008 Yan et al; This article is distributed under the terms of the Creative Commons Attribution Licence (<http://creativecommons.org/licenses/by/2.0>)

Abstract:

Background Outer membrane proteins (OMP) perform diverse functional roles in Gram-negative bacteria. Identification of outer membrane proteins is an important task. Results This paper presents a method for discriminating outer membrane proteins (OMP)

Figure 1: An Example Full Record for a BioMed Central Article in Zetoc

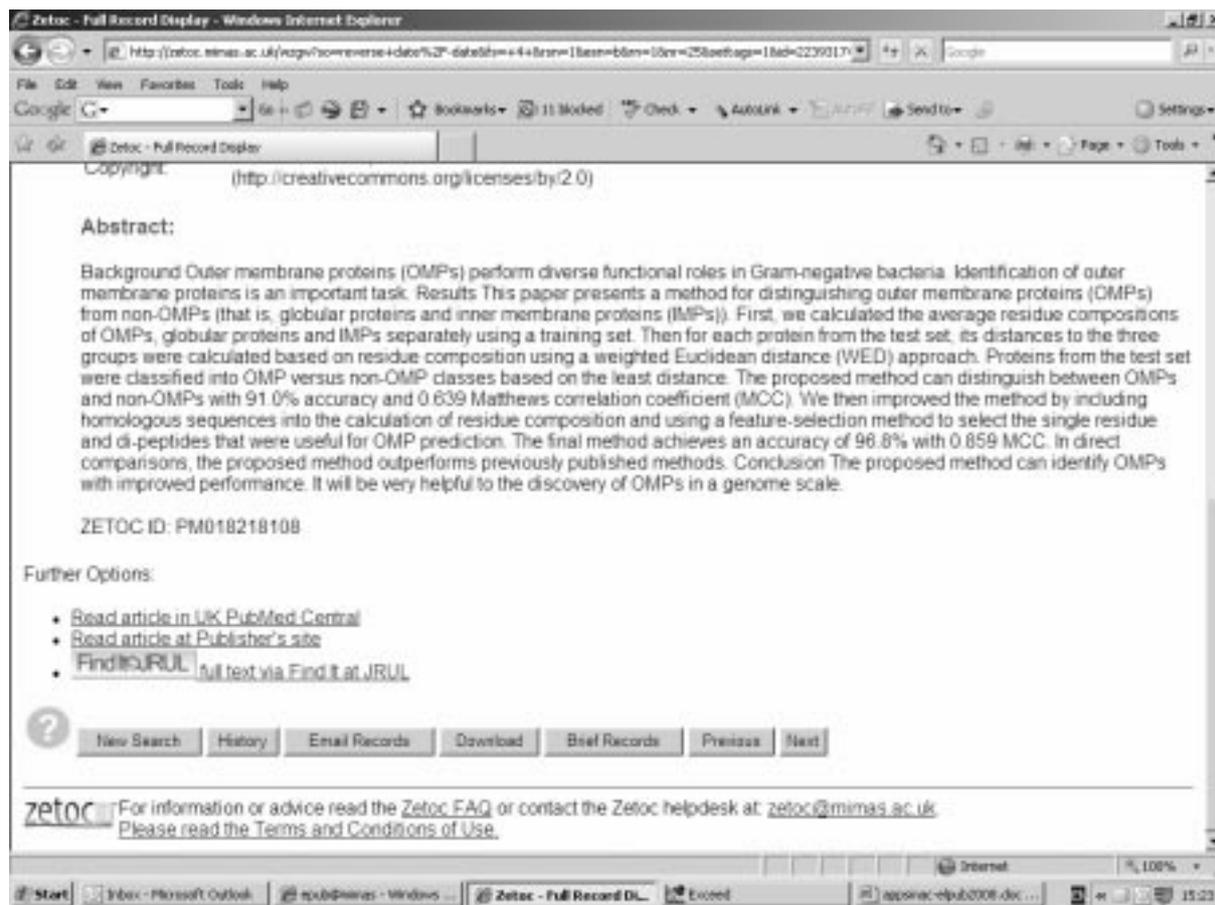


Figure 2: Links to the Full Article from Zetoc

A direct link is provided to the full text of the article in UKPMC, which, being open access, will be universally available. A second link is provided to the, again open access, article at BioMed Central, by their request. Because these links are based on global identifiers they are persistent, following good practice to minimise future loss of information [21].

The third link is to a user's institution's OpenURL resolver, in this example John Rylands University Library at The University of Manchester. This is provided because it is assumed to be some institutions' preference and will potentially lead the reader to additional services about the article. The 'hidden' COinS machine-parsable bibliographic citation [22] is also provided alongside the OpenURL Resolver link. The PubMed identifier and the Digital Object Identifier (DOI) are included in the OpenURL and COinS, information that is not available for other Zetoc articles. This more precise article identification will assist downstream consumers of the OpenURL or COinS.

These links differ from the document delivery options shown for other Zetoc articles, which are not open access as far as Zetoc is aware. For articles that are not open access, in addition to the OpenURL resolver link, there are the options of acquiring the article through Inter-Library routes or purchasing directly from the British Library. Obviously there is no point in providing links that require payment when a free version is available.

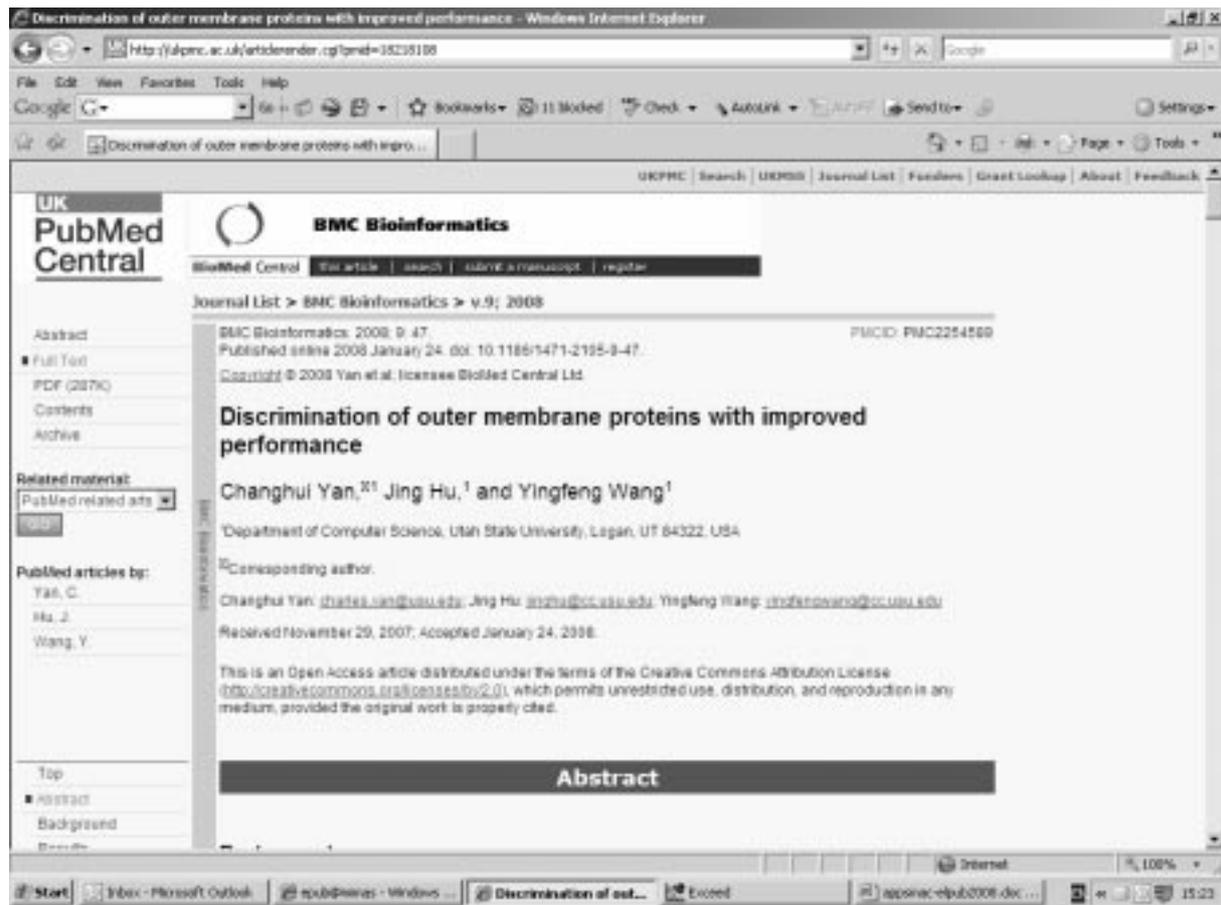


Figure 3: Full Article in UKPMC

3.3. Zetoc Alert

Inclusion of the new BMC journals into the Zetoc Alert system has been seamless because their article details are appended, in the Alert data feed, to those supplied by the British Library. Zetoc Alert maintains its journal list from this data feed, automatically adding new journals to the potential alert list. A user wishing to locate an article from a link in a received alert is taken via its full record page in the main Zetoc application, again making the introduction of these new journals seamless.

4. Discussion

4.1. Enhanced Functionality

The import from UKPMC of BioMed Central journal article details is a significant introduction of high profile and open access research literature in a particular discipline. Because the literature is open access, Zetoc is able to provide direct links to the full text of an article, at UKPMC and at BioMed Central. Although, at the time of writing, this is a recent introduction and has not yet been formally, publicly announced, links to full text articles are already showing use, with 38 accesses during April 2008.

Previous Zetoc evaluation studies have indicated some dissatisfaction about the availability of articles by users from institutions that do not have OpenURL resolvers, or that have a low level of journal subscriptions [23, 24]. Providing direct links to open access literature should address these concerns.

An additional data field in the UKPMC data is the article abstract. The majority of Zetoc article records do not contain abstracts. This inclusion of abstracts in Zetoc for the BMC articles is a significant enhancement, providing assistance to researchers when choosing full articles to read.

4.2. Joining Up BioMedical Services

The British Library, jointly with the Technical Computing Group at Microsoft, are developing the Research Information Centre (RIC) [25]. This Virtual Research Environment is a desktop application to support researchers through the complete research lifecycle. The primary focus of the prototype RIC is biomedicine, selected resources being those considered to be of importance to biomedical researchers. Zetoc, via its prototype Web Services SOAP interface [26], is one of the included resources. The introduction of BioMed Central journals into Zetoc will enhance the availability of biomedical literature within the RIC.

4.3. Extension to Other Open Access Journals

Clearly it should be possible to extend this model in the future to import into Zetoc details of all pure open access articles in UKPMC. But, as indicated above, journals will need manually assigned subject classifications and publisher countries. Thus it is probable that journals would be added to the import from UKPMC either by publisher or individually rather than developing an automatic mass ingest of all UKPMC open access articles.

It may be that some articles recorded in UKPMC are also published traditionally and so are already included in the Zetoc data feed from the British Library. Deduplication should be possible given the quality bibliographic citation data in both sources. However having two records for an article, would not really matter within a very large discovery service such as Zetoc, which already contains duplicate records for conference papers that are published in journals. It would rather give a reader the option of document delivery, by either open access or traditional routes.

Further, this solution could be used to import details of articles into Zetoc from other disciplines if a suitable open access archive, with a harvest interface and an acceptable data mapping, were available. BMC also publish a couple of Physics journals, according to their open access model, in PhysMath Central [27] and have suggested their inclusion in Zetoc. These journals are available within arXiv, the Physics repository [28], from which article details could be harvested. Implementation would require a similar process to that described above for ingest from UKPMC; the first step being the definition of a data mapping from arXiv to Zetoc. If this were successful it would open up the possibility of introducing more open access physics literature into Zetoc.

4.4. Single Article Publishing

The article publishing model exhibited by UKPMC differs from the traditional journal issue model. BioMed Central articles are received by UKPMC, and hence imported into Zetoc and users alerted, as soon as they are available, because they are not tied to the printed paper paradigm, in which a complete journal issue is assembled before publication. However they do include the traditional bibliographic citation details within a journal such as volume and pseudo-pagination, which is actually an article number within the volume. Although the Electronic Table of Contents data from the British Library consists of standalone records for articles, the complete table of contents for a journal issue is supplied consecutively in the data feed.

Thus users of Zetoc Alert who subscribe to a BMC journal will not receive a complete journal issue table of contents in a single email. Rather they will receive multiple emails, over a period of time, for a journal issue, each containing a subset of the eventual articles. However it was decided that the advantage to a researcher of receiving a very timely alert about the publication of an article of interest, which is also open access, should override any irritation at receiving multiple alerts rather than a single, complete table of contents. The introduction of the BMC journals into Zetoc is too recent to determine whether this situation is acceptable to researchers in practice.

4.5. Sharing Records of Open Access Literature

The data mapping, to Zetoc from the UKPMC records available for harvesting via OAI-PMH, has indicated a more general data requirement for sharing, between registries and repositories, bibliographic records of open access scholarly literature published in journals. The apparently necessary data fields are shown in Table 3, and further 'good to have' items in Table 4.

Data Field	Notes
Article Title	
Author Names	Family name and initials at a minimum
Journal Title	
ISSN	
Publication Year	
Volume and Issue	As per journal
Pagination	First page; last page optional
Subject Classification of journal	According to a standard scheme
Persistent URI	

Table 3: Data Fields for Sharing Journal Article Bibliographic Records

Data Field	Notes
eISSN	
Date of Publication	
Publisher	
Country of Publication	
Abstract	
Keywords	E.g author keywords
Global Identifiers	E.g. DOI, PubMed Identifier

Table 4: Additional Useful Data Fields

5. Conclusions

This development is ‘joining up’ two bibliographic services hosted by Mimas. It provides, through Zetoc, a single discovery and current awareness application, supplying bibliographic details of articles appropriate to a researcher’s request, with the addition of abstracts for BioMed Central articles. The full text of discovered open access biomedical articles is immediately available for reading from UKPMC. The objective is to enhance the usefulness of Zetoc to biomedical researchers by accelerating ‘discovery to delivery’, and in the future to researchers in other disciplines.

Acknowledgements

The Zetoc service, including its development, is supported by the Joint Information Systems Committee (JISC) [29] of the UK Higher and Further Education Funding Councils and the British Library [30].

Notes and References

- [1] Zetoc. Retrieved, April 25, 2008, from <http://zetoc.mimas.ac.uk/>
- [2] APPS, Ann; MACINTYRE, Ross. Customising Location of Knowledge. *DC2006: Proceedings of the International Conference on Dublin Core and Metadata Applications, Manzanillo, Colima, Mexico, 3-6 October 2006*. Universidad de Colima, 2006, p. 261-272. Retrieved, April 25, 2008, from <http://epub.mimas.ac.uk/papers/dc2006/appsmac-dc2006.html>
- [3] APPS, Ann; MACINTYRE, Ross. Why OpenURL? *D-Lib Magazine*. Vol 12 No 5, 2006, doi:10.1045/may2006-apps.
- [4] Google Scholar. Retrieved, April 25, 2008, from <http://scholar.google.co.uk/>
- [5] Sirus – for scientific information only. Retrieved, April 25, 2008, from <http://www.scirus.com/>
- [6] Copac Academic and National Library Catalogue. Retrieved, April 25, 2008, from <http://copac.ac.uk/>
- [7] UK PubMed Central Free Archive of Life Sciences Journals. Retrieved, April 25, 2008, from <http://ukpmc.ac.uk/>
- [8] LAGOZE, C.; VAN de SOMPEL, H.; NELSON, M.; WARNER, S. The Open Archives Protocol for Metadata Harvesting. 2004. Retrieved, April 25, 2008, from <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- [9] BioMed Central: The Open Access Publisher. Retrieved, April 25, 2008, from <http://www.biomedcentral.com/>

- [10] NLM Journal Archiving and Interchange Tag Suite. Retrieved, April 25, 2008, from <http://dtd.nlm.nih.gov/>
- [11] Creative Commons. Retrieved, April 25, 2008, from <http://creativecommons.org/>
- [12] HILL, Amanda. What's in a Name? Prototyping a Name Authority Service for UK Repositories. *ISKO2008 Conference, Montreal, Canada, August 2008*. 2008. (Accepted for publication). Retrieved, May 1, 2008, from http://names.mimas.ac.uk/documents/Names_ISKO2008_paper.pdf
- [13] OCLC Dewey Services: Dewey Decimal Classification. Retrieved, April 25, 2008, from <http://www.oclc.org/dewey/>
- [14] The Library of Congress Classification. Retrieved, April 25, 2008, from <http://www.loc.gov/aba/cataloging/classification/>
- [15] NICHOLSON, D; DAWSON, A; SHIRI, A. HILT: a Terminology Mapping Service with a DDC Spine. *Classification Quarterly*. Vol 42 No 3/4, 2006, p. 187-200. Retrieved, May 1, 2008, from <http://eprints.rclis.org/archive/00008767/>
- [16] WIKIPEDIA CONTRIBUTORS. Representational State Transfer. *Wikipedia, The Free Encyclopedia*. 2008. Retrieved, May 1, 2008, from http://en.wikipedia.org/wiki/Representational_State_Transfer
- [17] APPS, Ann. Using an Application Profile Based Service Registry. *DC2007: Proceedings of the International Conference on Dublin Core and Metadata Applications, Singapore, 27-31 August 2007*. Dublin Core Metadata Initiative and National Library Board Singapore, 2007, p. 63-73. Retrieved, May 1, 2008, from <http://epub.mimas.ac.uk/papers/2007/dc2007/apps-dc2007.html>
- [18] ESDS International. Retrieved, May 1, 2008, from <http://www.esds.ac.uk/international/>
- [19] UK Data Archive. Retrieved, May 1, 2008, from <http://www.data-archive.ac.uk/>
- [20] Data Documentation Initiative (DDI). Retrieved, May 1, 2008, from <http://www.icpsr.umich.edu/DDI/>
- [21] LAWRENCE, S; PENNOCK, DM; FLAKE, GW; KROVETZ, R; COETZEE, FM; GLOVER, E; NIELSEN, FA; KRUGER, A; GILES, CL. Persistence of Web References in Scientific Research. *Computer*. Vol 34 No 2, 2001, p. 26-31.
- [22] HELLMAN, E. OpenURL COinS: A Convention to Embed Bibliographic Metadata in HTML. 2006. Retrieved, April 25, 2008, from <http://ocoins.info/>
- [23] EASON, Ken; HARKER, Susan; APPS, Ann; MACINTYRE, Ross. Towards an Integrated Digital Library; Exploration of User Responses to a 'Joined Up' Service. *Lecture Notes in Computer Science (ECDL2004: Eighth European Conference on Research and Advanced Technology for Digital Libraries, University of Bath, UK, 13-15 September 2004)*. Vol 3232, 2004, p. 452-463. Retrieved, April 25, 2008, from <http://epub.mimas.ac.uk/papers/eham-ecdl2004.html>
- [24] EASON, Ken; MACINTYRE, Ross; APPS, Ann. A 'Joined Up' Electronic Journal Service: User Attitudes and Behaviour. *Libraries Without Walls 6: Evaluating the Distributed Delivery of Library Services*. London : Facet Publishing, 2005, p. 63-70. Retrieved, April 25, 2008, from <http://epub.mimas.ac.uk/papers/lww6/easonetal-lww6.html>
- [25] BARGA, RS; ANDREWS, S; PARASTATIDIS, S. The British Library Research Information Centre (RIC). *UK e-Science ALL HANDS MEETING 2007, Nottingham, UK, 10-13 September 2007*. 2007. Retrieved, April 25, 2008, from <http://www.allhands.org.uk/2007/proceedings/papers/800.pdf>
- [26] APPS, Ann. Zetoc SOAP: a Web Services Interface for a Digital Library Resource. *Lecture Notes in Computer Science (ECDL2004: Eighth European Conference on Research and Advanced Technology for Digital Libraries, University of Bath, UK, 13-15 September 2004)*. Vol 3232, 2004, p. 198-208. Retrieved, April 25, 2008, from <http://epub.mimas.ac.uk/papers/appsecdl2004.html>
- [27] PhysMath Central. Retrieved, April 25, 2008, from <http://www.physmathcentral.com/>
- [28] arXiv.org. Retrieved, April 25, 2008, from <http://arxiv.org/>
- [29] Joint Information Systems Committee (JISC). Retrieved, April 25, 2008, from <http://www.jisc.ac.uk/>
- [30] The British Library. Retrieved, April 25, 2008, from <http://www.bl.uk/>