



Detailed Report : Pilot to explore the creation of an OAI-compliant metadata repository for a specialist publisher of e-journals

<i>Author(s)</i>	Linda Kerr, Santy Chumbe
<i>Last Updated</i>	23 st September 2003
<i>Version</i>	1.0
<i>Document Name</i>	OAI_finalreport

Document Summary

This document is the final report of the project: Pilot to explore the creation of an OAI-compliant metadata repository for a specialist publisher of e-journals, funded by JISC as part of the PALS Metadata and Interoperability Projects Programme.

Contents

1. Background to the Project	2
2. Rationale for Work	2
3. Project Details	2
4. Aims and Objectives	2
5. Methodology and Activities	2
6. Outputs	4
7. Impacts and Conclusions.....	4

1. Background to the Project

EEVL: the Internet Guide to Engineering, Mathematics and Computing, based at Heriot Watt University, was awarded funding under strand A of the JISC Metadata & Interoperability Projects Programme (05/03) to develop a pilot to explore the creation of an OAI-compliant metadata repository for a specialist publisher of ejournals, Inderscience Publishers Ltd. Inderscience currently publishes more than 80 scientific journals. Its publication, the "International Journal of Technology Management", played a major role in developing the term "technology management".

2. Rationale for Work

Bibliographic and full text databases are extremely important information resources for academics in engineering, mathematics and computing. The existing landscape for these subjects is, however, quite complex. Two large publishers, Elsevier and ISI, through their products ScienceDirect, Ei Village and Web of Science, currently dominate that landscape. There are also, however, numerous other publishers providing content and access tools. Many of these publishers have only a basic understanding of the JISC Information Environment, and the benefits of releasing their metadata to allow for a more seamless discovery of their resources, either within the JISC IE or as part of commercial aggregation/presentation services. Normally such publishers make their content available only from their own web sites. As a result, access to their content is under-exploited, and users will increasingly by-pass their content to that more easily accessible.

As part of its work in the JISC-funded Subject Portal Project (SPP), EEVL identified a number of content providers, mainly small commercial or academic publishers, who expressed an interest in being a target in the cross-searching "portlet", but who did not have the technical expertise or the time to reach the required technical level. The interoperability barrier was too high. EEVL saw this project as a way of procuring content for the portal, and also as a way of developing guidelines to encourage other publishers to take a similar path.

3. Project Details

Project Length: 4 months

Partners: EEVL: The Internet Gateway to Engineering, Mathematics and Computing (based at Heriot Watt University) and Inderscience Publishers Ltd.

4. Aims and Objectives

The aim of the project was to test the feasibility of creating an OAI journal metadata repository at Inderscience Publishers Ltd, to allow future interoperability with JISC Information Environment services and projects, and to encourage publishers of bibliographic, image and other databases to make their metadata available for searching and/orharvesting.

The project had the following objectives:

- Creation of an OAI-compliant metadata repository
- Implementation of an OAI-PMH tool for harvesting the OAI repository
- Production of publisher-friendly guidelines, including practical examples, recommendations and details of useful tools and services.
- Dissemination of results
- The feasibility of implementing the OpenURL standard in this context will be explored

5. Methodology and Activities

WP1

Investigation of the publisher's data

An initial survey was made of the data held by Inderscience, how the data was structured, and what was required to make the data ready for OAI harvesting.

WP2

Definition of the methodology for implementing an OAI repository for Inderscience.

A methodology for implementing the OAI repository was developed, and the repository created accordingly. Inderscience publishes more than 80 scientific journals, and most of their articles are relevant to EEVL. Almost 70% of the articles are available online from the publisher's web site. The articles are stored in a SQL database, and managed from a web-based content management system (CMS). It was found that the CMS is mainly orientated to support the printed production of complete journals and to allow full-text searching of their contents, without taking into account interoperability aspects, nor leaving the possibility to give open access to their database to potential aggregators and harvesters. However, because the RDBMS and CMS were developed in-house at Inderscience, it was envisaged that they could be adapted to support OAI technology. Next, the metadata was examined. It was decided that the metadata format of the repository would be based on the Dublin Core Metadata Element Set. Analysis of the database content revealed that all the required metadata was available. This was good news, as this is not always the case with publisher's metadata.

WP3

Creation of an OAI repository at Inderscience.

A methodology for implementing the OAI repository was developed, and the repository created accordingly. It involved the execution of the following tasks:

- Creation of a development area on the publisher web server and copy relevant databases for testing purposes.
- Development or adaptation and installation of a PHP/MySQL based OAI V2 data-provider software on the publisher web server.
- Integration of the data-provider software with the RDMS through the CMS, to form the OAI repository
- Helping the publisher to enforce the means to keep the relevant databases up-to-date. For instance, generating guidelines to make the DC metadata elements mandatory in their database.

Dublin Core Elements used in the project:

- DC 1 = DC.Title
- DC 2 = DC.Creator (author)
- DC 3 = DC.Subject (engineering, mathematics, computing)
- DC 4 = DC.Description (abstract)
- DC 5 = DC.Publisher (Inderscience))
- DC 6 = DC.Date (Last Updated)
- DC 7 = DC.Identifier (DOI based)
- DC 8 = DC.Date Stamp (Creation date)
- DC 9 = DC.Type (single article)
- DC 10 = DC.Format (text/plain)
- DC 11 = DC.Source (Journal Code)
- DC 12 = DC.Language (English)
- DC 13 = DC.Relation (bibliographic reference: Volume, Issue No. and publication year)
- DC 14 = DC.Rights (Inderscience)

WP4

Development of an OAI-PMH harvester at EEVL

An OAI-PMH harvester was developed on the EEVL server.

WP5

Testing of the OAI repository within a implementation of a cross-search prototype

It was originally intended, in the project proposal, that the Inderscience OAI repository would be added to the list of targets searched from within the cross-search portlet, developed as part of the JISC-funded Subject Portal Project (SPP), and an evaluation of the data structure carried out using that interface. However, the Subject Portal Project portlet does not, at this point, support searching via OAI. A more limited evaluation was carried out using a more basic interface, and the results collated. The main findings were that the search interface was clear, but basic. There was a good spread of

entry options, and the options of having summary or detail on the browse and search results list. It was commented that the detailed record was fine, but the source field only gave the journal acronym, which doesn't mean much to the user. The details of volume, month, year occasionally contained an extra comma, and this was caused by the DC Identifier: Relation being used for details which may have originally been in more than one field, a failing of unqualified Dublin Core. It was commented that there was no help available. It was suggested that providing direct links to the full text would be useful.

WP6

Dissemination

A case study was written, giving guidelines for publishers intending to develop similar repositories. A paper disseminating the findings of the project has been accepted for the IADIS International Conference, WWW/Internet 2003, Algarve, Portugal, 5-8 November 2003.

The outputs of the project are listed below, and links to the relevant documents are available from the project website: http://www.eevl.ac.uk/projects_503.htm

6. Outputs

The outputs of the project were:

- Brief report on the database structure and management at Inderscience, including the IT technology. (WP1)
- Brief report on the methodology and architecture for Inderscience's OAI repository. (WP2)
- OAI Repository at Inderscience (WP3)
- OAI-PMH harvester for harvesting the Inderscience OAI repository (WP4)
- User testing report (WP5)
- Guidelines/case study (WP6)
- Journal article (conference paper) (WP6)
- Project report submitted to the PALS website (WP6)

7. Impacts and Conclusions

The project has demonstrated how content which has been "hidden" or had a relatively limited visibility in the user community can be made more accessible. It has also shown that to satisfy users needs, access to the full-text article needs to be as seamless as possible. It has also demonstrated how, in a relatively short period of time data held in a flat file format can be restructured, and made interoperable. Initial doubts by the technical staff at Inderscience that this was not part of their core business were overcome by the simplicity of the process.

The user testing raised the question of whether Dublin Core metadata was sufficiently rich to cope with the detailed bibliographic information present in journal articles. The issue of whether it is better to mandate a low level interoperability standard, which will be used but which supplies limited information, is under current discussion on Open Archive forums. The conclusion the project came to is that it may be useful, in future, to apply a more detailed metadata standard in the repository to ensure maximum interoperability, providing Dublin Core is also implemented. More detailed metadata would be an overhead for the publisher as repository creator, and its usefulness would need to be practically demonstrated.

EEVL wishes to test of the data within the SPP project interface, however, the issue of unqualified Dublin Core presenting less helpful bibliographic information may need to be explored with the publisher. However, although the presentation of bibliographic information in the record is not as clear as in a MARC record, it provides basic reference.

Having created fairly readily an OAI repository at Inderscience, the next logical stage is to explore making the full-text easily available, as at present the user has to use a little ingenuity to find papers. This raises the issue, as far as JISC subject/institutional portals are concerned, of authentication. Would it be financially feasible for a smaller/medium sized publisher to join Athens, or will some other kind of authentication or rights evaluation system be supported within a JISC portal? There was not

time in the project to explore the feasibility for implementing the OpenURL standard, and this is recommended for further work by JISC.

Opportunities have subsequently arisen since the start of this project to work with other publishers to make their metadata interoperable, using OAI. Interest has been expressed by The Institution of Civil Engineers in creating an OAI repository of their Virtual Library, the largest repository of full text civil engineering papers in the world.

It is hoped that the Case Study resulting from this project will provide an introduction for publishers to the process of creating a repository. However, what is now required are guidelines on how to exploit this resource, and work within the JISC Information Environment (and perhaps commercial aggregators), with particular reference to issues such as authentication. The information on the PALS web site (http://www.jisc.ac.uk/index.cfm?name=wg_palsinter_faq) for publishers is extremely clear, but more guidance is required, either documentation or activity, to allow publishers to participate in the JISC IE. At a basic level, list of service providers would be of use.

Commercial aggregators can come to agreements to both access the metadata and deal with downloads of full-text documents. For the JISC Information Environment, and subject portals, the situation is less clear how the full-text can be delivered, as smaller publishers are less likely to join Athens, and will have their own authentication, subscription and "pay-per-view" schemes. A project like this only starts to deal with the practicalities.

In conclusion, this was a successful project technically, and raised some interesting issues surrounding the priorities of JISC and of smaller publishers, both of whom have similar aims in serving their users, but different cultures. Now that the repository has been set up, the next stage is to turn this into a service, and to use the lessons learned to work with other, similar publishers to make their content interoperable, and work towards seamless access via subject portals and eprint service providers.