

DEVELOPING SEAMLESS DISCOVERY OF SCHOLARLY AND TRADE JOURNAL RESOURCES VIA OAI AND RSS

Santiago Chumbe and Roddy MacLeod
ICBL (Institute for Computer Based Learning)
School of Mathematics and Computer Science
Heriot Watt University
Edinburgh EH14 4AS
Tel: 0131 451 3280
Email: santiago@macs.hw.ac.uk

ABSTRACT

The usefulness of online information such as e-publishing and timely notification on the latest scientific or professional news has been widely accepted. However, access to such valuable information is often limited by lack of mechanisms for interoperability and distributed harvesting of the source databases. Following recent experiments at EEVL (Internet Guide to Engineering, Mathematics and Computing [1]), we have shown that emergent web-based technology can effectively and positively impact on access to this online information. The Open Archives Initiative (OAI) [2] is one major effort to address technical interoperability among distributed sources of information. The objective of OAI is to provide a framework to facilitate the discovery of content in distributed archives. RSS [3] is another XML-based format for sharing data on the web. This paper discusses two areas of current development at EEVL. The first is a project to set up a seamless access mechanism based on OAI for an important publisher of scientific journals; the second is a project to release metadata from a consortium of engineering trade information publishers using RSS. Outcomes will include metadata ready to be embedded in various subject-based or institutional portal services, including a bibliographic cross-search service, and 'one-step' aggregated engineering trade news, jobs and conference services.

KEYWORDS

OAI, RSS, XML, e-publishing, interoperability, distributed harvesting

1. INTRODUCTION

A large number of online journals and scientific digital publications are available today on the web. However, very few of them are accessible from a federated service that can inter-operate with these e-publications through open access protocols and which can harvest their metadata to provide a unified interface to them. In this paper we present a suitable implementation of software which provides a framework for interoperability with e-publishing providers, by enabling metadata from their databases to be harvested and aggregated into one searchable database/interface. We will focus our attention on e-journals, taking into account that the journal is the primary publication channel for communicating research results. The software makes intensive use of XML standard formats. Specifically, we use the Open Archives Initiative (OAI) [2] specification, considering that OAI is becoming widely accepted, and provides the normalization needed for overcoming the problems that arise when harvesting different e-journal providers with different format/naming conventions. Thus, this software can inter-operate with any OAI-compliant e-journal repository, because they share the same metadata specification, making their contents interoperable with one another. The web user can then harvest their metadata into a global "virtual" database that is seamlessly accessible and navigable from a unified interface.

A complementary contribution of this communication will be a mechanism to syndicate and aggregate metadata from a consortium of engineering trade information publishers using RSS [3]. The fact that scholars value e-notifications [19], makes important RSS news-feeds services, especially in fields with fast paces of innovation where awareness of new discoveries is critical for scholars.

The work described is being funded by JISC [10]. Publishers taking part include Inderscience [11], Centaur Communications [12], ProTalk [13] and GoJobSite [14].

2. ARCHITECTURE

The architecture of the software framework, currently in development, includes an OAI-compliant repository (Data Provider) for managing the e-journal metadata, a Service Provider or harvester based on the OAI Protocol for Metadata Harvesting (OAI-PMH) [4], and a back-end facilitator to make cross-searchable the harvested e-journals. Figure 1 shows the envisaged software architecture.

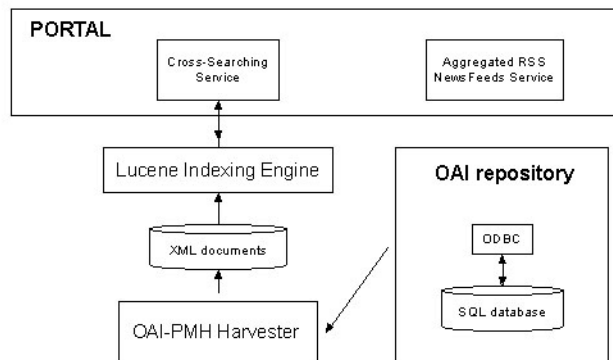


Figure 1. OAI Software Framework Architecture vision

The OAI software is being written at Heriot Watt University, UK, by EEVL [1] and it will be freely available to download from the EEVL web site. Once developed, the software will be relatively straightforward to install, although knowledge of XSLT transformations, Java servlets, PHP and MySQL will be required for adequate technical support.

2.1 The OAI Data Provider

The first task is to make the existing collection of e-journals stored in the publisher site OAI-compliant. Thus, we have extracted both the metadata and the data from the existing structured and unstructured databases. The extracted metadata can be stored in an archive of well-formed XML files, or in any indexed SQL database. In this implementation, we use the RDBMS open source MySQL for storing the actual metadata and for keeping track of harvested records. The repository supports an on-the-fly compression, to reduce the amount of data being transferred. It will make the OAI repository accessible as a compressed XML file for the OAI harvesters. The implementation will comply with the OAI-PMH 2.0 specification [4], and it was inspired by the work done by U. Müller [5] at the Humboldt University of Berlin.

The metadata format of the OAI repository is based on the Dublin Core Metadata Element Set [6] and includes such information as title, author, subject and abstract. Once installed, the OAI repository will be automatically ready to expose this metadata in a form, which can be picked up by OAI harvesters.

2.2 The OAI-PMH harvester

Similar to a web crawler, the OAI-PHM will extract metadata from the OAI repository and put it into a searchable database located on the OAI service provider, in this case EEVL. It will make use of enhanced features defined by the OAI protocol, such as the possibility to make incremental and selective harvesting.

The OAI harvester is being developed as a Java thread, which will fetch periodically the metadata exposed by the OAI data provider, and using XSLT [7] transformations, will produce a searchable database.

2.3 Cross-searching service

The XML documents generated by the OAI-PMH harvester will then be fed into an indexing engine. This engine will be developed around the open source software Lucene [8]. Lucene also will provide the searching algorithms required for supporting search by query, as well as results ranking.

The user interface for searching the harvested metadata will be embedded in a bibliographic cross-searching facilitator (portlet) or channel of a portal service. This facilitator will provide the unified interface, from where any user can then search and seamlessly access the harvested metadata from different distributed e-journal archives. Currently, EEVL is actively involved in the development of the Subjects Portal Project [9], which includes the implementation of this cross-searching facilitator.

3. RSS SYNDICATION AND AGGREGATION SERVICE

The second XML-related project carried out at EEVL, aims to release the latest scientific or professional news from publishers using RSS. The core component of this project is a web-based syndicated content reader, which also will be offered, in time, embedded as cross-searchable aggregated RSS newsfeeds services in a portal. This will also act as a showcase for other publishers to follow.

One advantage of having the service incorporated into a portal is that its usefulness will be enhanced by the user profile (personalization) and alerting facilities offered to all registered users of the portal.

4. CONCLUSION

Serious studies have found that a majority of scholars read e-alerts regularly, and a minority read full-text onscreen [19,20,21]. These studies have shown that real time cross-referencing and e-notifications are the two key components of any achievement in interoperability among e-publishing providers. By implementing the XML-based de facto standards OAI and RSS, we are working towards an environment that will provide both real time cross-referencing and e-notifications within the context of a subject-based portal.

XML and XML Schema have proven to be effective for structuring data for cross-referencing. In this context, the use of an XML-based repository such as OAI, opens the way for cross-referencing or inter-linking of distributed electronic resources, which is one of the hallmarks of the Web. Current technologies addressing this issue, from different perspectives, involve CrossRef [15], DOI [16], openURL [17] and semantic web [18].

We notice that cross-searchable access to e-publishing will have a strong influence on the open dissemination of research literature and its implementation is a new challenge for publishers. These efforts are justified by the multiple benefits that seamless access to scholarly and e-journal resources will provide for the society. Scholars tend to use e-journal articles as nodes in a network of linked, shareable content. Here there is another benefit of an OAI source of e-journal articles: they are set to work with technologies such as CrossRef to support cross-search, cross-reference and cross-hyperlink. An OAI repository can become the

heart of cross-reference and interoperability, giving the user the ability to create and follow networks of linked knowledge on the web. Seamless interoperation gives the possibility of non-linear reading, and also new levels of scholarly activities such as personalized desegregation of content and peer-to-peer communications.

From a commercial perspective the efforts towards interoperability are vital if businesses are to be agile and for content to effectively reach its intended audience. Providing interoperable, standards based, services can have a positive impact on the way a business operates. In fact, technologies such as OAI and RSS are bridging the digital divide between consumers and providers of relevant information.

A direct application of the results produced by this work will be the possibility of embedding the OAI software framework and the aggregated RSS services within institutional portals. Our development is based on Open Source software and we would like to make it freely available for general use. Both implementations are projected to be complete by the end of August 2003 and we expect to be ready to deliver further results during the conference.

ACKNOWLEDGEMENT

The research work and the two projects presented in this contribution were funded in part by the JISC Metadata & Interoperability Projects (5/03). Thanks go to all colleagues at EEVL for their very useful and important suggestions and to the four organizations that have been working in partnership with EEVL to encourage the wide access and exposure of their metadata on the web.

REFERENCES

- [1] The Internet Guide to Engineering, Mathematics and Computing, EEVL. Available at <http://www.eevl.ac.uk>.
- [2] Lagoze, C. and Van de Sompel, H. 2001. *The Open Archives Initiative: building a low barrier interoperability framework*. Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries (pp 54-62), Roanoke, VA.
- [3] Ben Hammersley, 2003. *Content Syndication with RSS*. O'Reilly & Associates Publishers, San Francisco, USA
- [4] The Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMH. Available at <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- [5] Müller, U. et al, 2003. *Example of a Data Provider Implementation*. In Open Archives Forum 2003. Humboldt University of Berlin, Germany.
- [6] Dublin Core Metadata Elements. Available: <http://dublincore.org/documents/dces/> (02 June 2002)
- [7] Tidwell, D. 2001. *XSLT*. O'Reilly & Associates Publishers, San Francisco, USA
- [8] Lucene Search Engine. Available at <http://jakarta.apache.org/lucene/docs/index.html>
- [9] Subject Portals Project. Available at <http://www.portal.ac.uk/spp>
- [10] The Joint Information Systems Committee, JISC. Available at <http://www.jisc.ac.uk>
- [11] Inderscience Publishers Ltd. Available at <http://www.inderscience.com>
- [12] Centaur Communications Ltd. Available at <http://www.centaur.co.uk>
- [13] Pro-Talk Internet Publishing. Available at <http://www.pro-talk.com>
- [14] GoJobSite Worldwide Ltd. Available at <http://www.gojobsite.com>
- [15] Pentz, E., Winter 2001. *CrossRef: A Collaborative Linking Network*. Science and Technology Librarianship Journal.
- [16] Sidman, D., 2002. *Digital Object Identifiers: Not just for publishers*. CMS Watch.
- [17] Walker, J. 2001. *Open linking for libraries: the OpenURL framework*. New Library World Journal, Vol. 102 - No. 1163/1164 (pp. 127-133) MCB University Press.
- [18] Hender, J., Berners-Lee, T. and Miller, E. 2002. *Integrating applications on the Semantic Web*. Journal of the Institute of Electrical Engineers of Japan, Vol 122(10) (pp. 676-680).
- [19] Online Features Survey. 2002. Available at <http://ejust.stanford.edu/findings3/Q9-11.pdf>
- [20] Jeon-Slaughter, H., 2002. *Designing Electronic Pages: User's behaviour and their needs*. Stanford University Libraries. USA.
- [21] *Information for Implementing the Library Strategic Plan. Results of User Focus Groups*. 2002. UBC Library. University of British Columbia. Canada.