*LIBER QUARTERLY, ISSN 1435-5205*

*K.G. Saur, Munich, printed in Germany*

# Open Archive Solutions to Traditional Archive/Library Cooperation

*By* DONATELLA CASTELLI

## INTRODUCTION

Searching and accessing in a seamless way a multitude of heterogeneous distributed information sources, like archives and libraries, is a requirement expressed by many categories of users. It can improve access to historical and cultural resources and create the condition for a better exploitation of them. The realization of this functionality requires the solution of a number of technical, organizational, sociological, and economical issues. In particular, the technological issues play a fundamental role and create a concrete basis for the discussion of the other issues. Few years ago a new technical approach to achieve cross-repository access has been proposed within the scientific communities that publish their pre-prints on electronic repositories (Eprints.org). This solution, named *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH), has rapidly become known worldwide, and many organisations are now experimenting with it. The results achieved have stimulated a diffusion of the OAI-PMH concepts and terminology also beyond the e-print application area. Today many people uses the expression *open archives* to mean repositories of digital information that provides a machine interface for making their content available to external services and view the OAI-PHM as a mechanism to achieve interoperability among different types of repositories. This paper briefly introduces the Open Archives Initiative (OAI) approach and surveys its application within the library and archival community. The paper concludes by presenting our view of the future role that OAI-PMH can play in supporting the collaboration between libraries and archives.

## TECHNICAL SOLUTIONS FOR SUPPORTING OPEN ARCHIVES

The Z39.50 protocol has been for many years the most widely knows mechanisms for supporting cross-repositories search. This is a standard protocol that specifies a client/server based communication for information retrieval. It specifies procedures and structures for a client to search a database provided by a server, retrieve database records identified by a search, scan a term list, and sort a result set. The protocol addresses communication between corresponding information retrieval applications, the client and the server, which may reside on different computers. Z39.50 is widely used in the libraries application framework, but it has also been used in other application areas like museums and geographic systems. Its application has been confined to institutions that

have strong motivations and enough resources to support an heavy commitment from both the client and the server side.

In order to archive a more wider applicability at a lower developmental and operational cost, few years ago, the Open Archives Initiative (Lagoze & Van de Sompel, 2001) proposed an alternative interoperability approach based on a very simple to implement metadata harvesting protocol, the OAI-PMH. This protocol comprises only six different service requests. It assumes a conceptual framework that distinguishes between data providers ('archives' in the OAI terminology) and service providers. In order to be OAI-PHM compliant, the data providers must implement the six service requests. Service providers can uniformly harvest metadata from the OAI-PHM compliant archives by invoking the protocol requests. A data provider decides which part of its information space is harvestable, i.e. it decides which items can be perceived outside and, for each of these items, which metadata are harvestable. An archive can decide to open multiple metadata formats for each visible items. A Dublin Core metadata record is mandatory for each item. This constraint ensures the conditions for the implementation of federated searching. The OAI-PHM has been initially proposed as a technical approach for solving the problems of interoperability among repositories of pre-print publications. By implementing this protocol the existing repositories make accessible the documentation that has been created by a specific community of interest also to other interdisciplinary communities. This creates the conditions for a wider exchange of information and a better exploitation of the existing resources. Currently there are more than one hundred and twenty worldwide distributed OAI-PHM compliant repositories and around twenty services built on them. It is expected that in a near future this number will grow very fast. The reason for the success of OAI-PHM is certainly its declared simplicity with respect to previous solutions and its capability to respond the requests of the most basic cross-repository services, like search and browse.

Very recently a new approach has been presented to further lower the level of complexity of the OAI-PMH protocol (Hochstenbach, Jerez & Van de Sompel, 2003). This approach allows a data provider to make a metadata collection available on a Web server as a XML file with a specific format. This file, named OAI Static Repository, is made OAI-PMH harvestable through the intermediation of a particular software, an OAI Static Repository Gateway, operated by a third party. This approach lowers significantly the barrier for sharing data via the OAI-PMH since the data provider only has to create and update the XML file of its metadata records, place it on its Web server and register it with a Static Repository Gateway. It is expected that this approach will be widely applied in the next future since it responds to concrete requirements raised during the experimentation and operation of the OAI-PMH in different application areas. Currently, the Open Languages Archive Community (OLAC) has indicated interest in migrating to the Static Repository approach, union catalogues in Belgium, Brazil and United States are considering adoption, and institutions collaborating with the Digital Library Federation (DLF) and the National Science Digital Library (NSDL) projects are

exploring the use of this approach as a means to increase the amount of metadata records they can make harvestable at limited expenses.

## USING OAI-PMH FOR LIBRARIES AND ARCHIVES

Many different user communities are now evaluating the applicability of the OAI-PMH technical solution to other application domains. A lot of experimentation has been done in the library domain and some initial attempts to explore its potentiality is also been conducted in the archival one. The next subsections briefly survey the work carried out in these two domains.

### *Open Access to Libraries*

Libraries had always been concerned to pool the records they have developed in order to document their respective holdings (Lynch, 2001). The mechanisms used to achieve this shared aim have been union catalogues and distributed search services. Union catalogues are designed around a single type of metadata, which limit the catalogue to materials that can be described by such metadata. The search across libraries represents an alternative approach that overcomes this problem. This approach has been based mainly on the use of the Z39.50 protocol. While this approach in itself has the advantage of offering virtual union search capability across repositories with differing underlying data formats, the application of the particular search and retrieve protocol presents problems of its own. As outlined previously, the data providers must support complex Z39.50 server software, and considerable coordination is required to set up workable profiles. Z39.50 search also works best across a limited number of services; it does not scale to the thousands of potential sources of digital content. The OAI-PMH approach has been perceived by the library community, especially by the academic and scholarly library communities, as a promising approach that is capable of combining the best of library and Internet techniques into a wholly new model for accessing library resources. In particular, the metadata harvesting technique proposed by the OAI-PMH has the potential of supporting the construction of library services that organise access to a rich variety of information resources to meet very specific user needs.

In the last few years many initiatives have been planned all around the world to experiment the OAI-PMH technical solution in the library domain. The Digital Library Federation (DLF) has been the first institution that has sponsored the OAI and a number of projects for exposing collections through the OAI-PMH and for implementing a number of harvesting-based services on them. The Library of Congress has also been an early tester of the OAI-PHM protocol. In particular, it has implemented the protocol to make harvestable a subset of collections from the American Memory and Prints & Photographs (P&P) Online Catalog. The result of this experimentation has shown that

the protocol was straightforward to implement and the harvesting traffic has no perceptive effect on the primary users of the American Memory project.

On the European side, a number of projects funded by the European Commission under the V Framework Programme, have decided, during the course of their activity, to experiment the OAI- PHM approach to the interoperability of library resources. The European Library Project (TEL) is one of these projects. TEL brings together ten major European National libraries and library organisations to investigate the technical and policy issues involved in the sharing of digital resources. The objective of TEL is to set up a co-operative framework, which will lead to a system for access to the major national and deposit collections in European National libraries. This project aims to create a new library organisation able to offer access to different services like multilingual services, name authority services and links to local services. One of the keys to meet the above challenges is integration: metadata generated by each service should be usable when accessing other services. This requires a common understanding of metadata, an easy way to carry metadata from one service to other services and an easy way to associate related metadata. The major challenges in the technical work of this project are related to the diversity of collections, languages and local services. Two alternative approaches have been used to support this facility: a Z39.50 approach and, a more recently introduced searching on a central index of metadata harvested from other collections via the OAI-PHM.

The MALVINE project offers another example of how the Open Archives Initiative approach could be exploited to create new sources of information from existing ones. MALVINE provides access to distributed holdings of modern manuscripts kept in European libraries and archives. MALVINE has to be intended as the beginning of a new phase of activities in the sector of modern manuscripts. At the time of launching MALVINE the OAI-PMH was not yet known, then no initial planning was done regarding its use. The MALVINE consortium is currently evaluating the implications of using this protocol in a landscape of European institutions, be these great or small, which provide data to a joint service. LEAF is another well known project that aims at developing a model architecture for establishing links between distributed authority records (personal names only) and providing access to them. The system allows uploads of the distributed authorities to a central system and automatically links those authorities concerning the same entity. In this framework, the OAI-PMH protocol plays a vital role in keeping the central repository up-to-date at any time.

*Open Access to archives*

In the last few years many archive institutions have begun to aggregate their finding aids - e.g. the Online Archive of California (OAC) and the Texas Archival Resources Online (TARO); the Access to Archives (A2A) project in the UK, etc. The development of aggregated services of this kind indicates that there is currently a considerable interest in

the archive world in interoperability. However, this world is somewhat behind libraries in understanding the potential and developing practical implementations. The main reason of this delay has to be understood in the diversity of the material held in archives and in their uniqueness. Until fairly recently, archives described their holdings in individual ways, using locally determined rules. This did not matter, since users of archives had to make physical visits to see the records, and could have the catalogues explained by the archivists. However, as electronic communications have spread, archivists have seen the desirability of exchanging and disseminating data, and have recognized the need to have standardized tools to do so. For users consulting catalogues over the Internet, there is no archivist available to mediate and explain the local rules.

A number of archives has recently adopted common portals or network systems, often on a geographic, thematic or institutional basis (e.g. the Archives Hub, the Scottish Archive Network, and the Swedish Nationell Arkivdatabas). These are all examples of an aggregated service, which adds value by bringing diverse descriptions together. They are broadly similar in principle, though not in application, to an OAI service provider. Currently, few conventional archives know about OAI-PMH and even fewer are using it. It is important to stress that these systems provide access to catalogue information only and not to digital materials. The most significant experimentation of OAI-PMH in the archival area has been carried out at the University of Illinois at Urbana-Champaign (UIUC). This is one of seven institutions funded by the Mellon Foundation to carry out research into interoperability and the OAI-PMH (Cole [et al.], 2002). UIUC is an OAI service provider, taking descriptive metadata from around forty institutions, mainly American university and research libraries. Its contributing institutions are also, in some cases, aggregating descriptive data from other agencies, for example the Online Archive of California, which brings together over sixty archival and other bodies holding archives. The metadata contributed ranges in size from a few tens of records to nearly a million, and the materials described range from museum artefacts to archival documents.

UIUC is particularly investigating conversion from Encoded Archival Description (EAD) to Dublin Core for exposure to OAI harvesters. Their conclusion is that there are difficulties, but not insurmountable barriers in doing so. The barriers lie mainly in the inconsistency with which archivists have employed EAD, which in turn is a result of its permissive model and its relatively relaxed rules. UIUC researchers found that many of the descriptions encoded in EAD were not well enough structured to allow them to be searched properly in an OAI environment. One of the conclusions of this experience is that paradoxically, the OAI-PHM approach may provide the incentive to overcome the problems of inconsistency. If it is used as a front end to EAD metadata, OAI-PMH records could possibly mitigate the encoding differences found between institutions and between the finding aids of different cataloguers. Another barrier in exposing EAD description into Dublin Core was found when attempting to transfer the different hierarchies in archival descriptions from EAD to OAI. To do so many OAI records had to be created from one EAD, with links between them to indicate the hierarchies. This resulted in ballooning file sizes and many empty records. Their solution adopted in the

project was to eliminate the listing of most levels of the EAD hierarchy in the OAI record, which means that the searcher will need to rely on a link to the full finding aid in order to view the all important context.

Another experimentation of OAI-PMH in the archival framework is carried out in UK by the AIM25 project. This project covers a range of archive repositories in London that provide OAI compliant descriptions to UIUC. These descriptions have been produced in a simpler fashion than in most of the UIUC examples. AIM25 holds its descriptions in a database and can export them in EAD or other format as required. The AIM25 experimenters did not find OAI-PMH compliance difficult to achieve, but found it difficult to reproduce the linkages between archive material from the same collection, or between material created by the same person or organisation held at different repositories.

The Bright Sparcs project in Australia is another tester of the OAI-PMH in the library and archival world. This project, that provides biographical and name authority information on Australian scientists, is involved in work with the National Library of Australia (NLA). Bright Sparcs brings together information on over 4,000 people involved in the development of science, technology and medicine in Australia, from archives and libraries. It is therefore not a typical archive site, though it does illustrate the potential of using OAI for name authorities. Bright Sparcs provides Dublin Core compliant descriptions of its pages, and these are harvested by the NLA and mounted on their site.

The Access to Archives (A2A) project in UK is planning to use OAI-PHM. It brings together catalogue descriptions at collection and item level from a range of participating archive organizations in England, with central editorial control provided by a team based at the Public Record Office (PRO). They have around four million records captured or planned for capture. They intend to produce OAI compliant descriptions at collection level only, and use these almost as a virtual ‘signpost’ to the fuller, item level description, which will be available through the A2A site. They looked at the UIUC method of producing many OAI records from one EAD record to represent different levels, and linking them together, but felt this would be impractical in their application. Instead they intend to convert their EAD/XML descriptions to OAI via XSL. They expect this to require minimum effort, which is an important consideration in view of the size of their application. They recognize the importance of judging whether users find their approach of having harvested metadata as a signpost to fuller descriptions is helpful.

## A VISION OF THE FUTURE

There are no agreed business models for how archives and libraries will make their resources available in the future, but there is growing realization that remote access to finding aids and bibliographic information, coupled with some form of access to digitised information, be it in native form or an image of the original, will offer a significant way forward. National initiatives already has been set up to achieve this aim in the UK, Sweden, Canada, Australia, etc. If interest in OAI-PHM is confirmed and varied organizations and domains use it, there will be immense quantities of information available across different fields. In these circumstances smart systems to let the user navigate and evaluate resources will be needed. The challenge will be if OAI-PHM can provide sufficient support for enabling the user-required functionality.

The recent call for proposal of the EU VI Framework in the Culture Heritage domain will offer the possibility of making advances on across-archive services. In this domain the OAI-PHM protocol will cover one of the many technical aspects related to the provision of services capable of satisfying the needs of potential users. Most likely, the access to archives and libraries will be provided thought new digital systems capable of supporting a multitude of services. Different user communities will have the possibility to select their preferred information space and the services that operate on this space. Technical infrastructures will be developed that will allow to dynamically include in the digital library other content providers and other services. By exploiting this heterogeneous information space, it will be possible to construct authoring services that will support the creation of new types of digital objects built by composing parts extracted from different and possibly distributed sources. It will thus be possible, for example, to create documents that mix information extracted not only from different media archives and libraries, but also from geographical databases and Web pages. These new objects will permit new forms of communications and will create the basis for a new kind of culture exchange.

*Acknowledgements*

Much of the considerations presented in this paper are the result of discussions among the participants of the 2nd Open Archives Forum Workshop "Open Access to Hidden Resources" held in Lisbon on December 2002. Many thanks to all the participants for their fruitful inputs. A particular thank to George MacKenzie and Goran Christiansson who have written a report, commissioned by the Open Archives Forum project, on how the OAI-PMH protocol can been exploited to open conventional archives. This report, titled "How Real Archivists can learn to love the OAI" has been another great source of inspiration for this paper.

REFERENCES

1. Cole, Timothy W. [et al.]. "Now That We've Found the 'Hidden Web,' What Can We Do With It? The Illinois Open Archives Initiative Metadata Harvesting Experience". Papers *Museums and the Web 2002*. http://www.archimuse.com/mw2002/papers/cole/cole.html

2. Hochstenbach, Patrick, Henry Jerez & Herbert Van de Sompel, "The OAI-PMH Static Repository and Static Repository Gateway", *Proceedings JCDL 2003*, pp. 210-217, Houston, May 2003.

3. Lagoze, Carl, Herbert Van de Sompel. "The Open Archives Initiative: Building a low barrier interoperability framework", *Proceedings JCDL 2001*, Roanode, VA, USA, June 2001.

4. Lynch, Clifford A. "Metadata Harvesting and the Open Archives Initiative", *ARL*, issue 217, August 2001, http://www.arl.org/newsltr/217/mhp.html

5. MacKenzie, George and Göran Christiansson. "How Real Archivists can learn to love the OAI". *Open Archives Forum*, Expert Report 2, 13 March 2003. http://www.oaforum.org/otherfiles/oaf_d44_cser2_kenzie-krist.pdf

6. "Open Access to Hidden Resources". *Open Archives Forum*, Workshop Report 2: 6-7th December 2002, Lisbon. Portugal. http://www.oaforum.org/otherfiles/oaf_d43_workshop2.pdf

WEB SITES REFERRED TO IN THE TEXT

A2A - Access to Archives. http://www.a2a.pro.gov.uk/

AIM25 - Archives in London and the M25. http://www.aim25.ac.uk/

Archives Hub - A national gateway to descriptions of archives in UK universities and colleges. http://www.archiveshub.ac.uk/

Bright Sparcs. http://www.asap.unimelb.edu.au/bsparcs/bsparcshome.htm

DLF - Digital Library Federation. http://www.diglib.org/dlfhomepage.htm

DLF evaluation of the Open Archives Initiative. http://www.diglib.org/architectures/testbed.htm

Dublin Core Metadata Initiative. http://www.dublincore.org/

Encoded Archival Description (EAD). http://www.loc.gov/ead/

EPrints.org. http://www.eprints.org/

The European Library (TEL) - The Gate to Europe's knowledge. http://www.europeanlibrary.org

LEAF - Linking and Exploring Authority Files. http://www.leaf-eu.org

Library of Congress. http://www.loc.gov/

Library of Congress collections for which records are available for harvesting through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). http://memory.loc.gov/ammem/oamh/lcoa1_content.html

Library of Congress Prints & Photographs (P&P) Online Catalog. http://lcweb.loc.gov/rr/print/catalog.html

MALVINE - - Manuscripts and Letters via Integrated Networks in Europe. http://www.malvine.org

National Library of Australia (NLA). http://www.nla.gov.au/

Nationell Arkivdatabas. http://www.nad.ra.se

NSDL - The National Science Digital Library. http://nsdl.org/render.userLayoutRootNode.uP

OAC - Online Archive of California. http://www.oac.cdlib.org/

Open Archives Forum. http://www.oaforum.org/

Open Archives Initiative. http://www.openarchives.org/

The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html

OLAC - Open Languages Archive Community. http://www.language-archives.org/

PRO - Public Record Office. The National Archives. http://www.pro.gov.uk/

Scottish Archive Network. http://www.scan.org.uk

TARO - Texas Archival Resources Online. http://taro.lib.utexas.edu/

UIUC - University of Illinois at Urbana-Champaign. http://www.uiuc.edu/index.html

Z39.50 International Standard Maintenance Agency. http://lcweb.loc.gov/z3950/agency