Interoperability and Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

Martha Latika Alexander J N Gautam

Abstract

Interoperability refers to the ability of a Digital Library to work cooperatively with other Digital Libraries in an attempt to provide higher quality services to users. There are many approaches to achieve some degree of interoperability and one such approach involves the creation and use of Open Archives. The OAI is an initiative to develop and promote interoperability standards that aim to facilitate the efficient dissemination of content. The OAI Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework based on metadata harvesting. OAI-PMH enables automated distribution of any kind of metadata, which may be aggregated into searchable databases by "harvesting" systems. It has reached version 2.0, intended for stable, production services.

Keywords : Open Archive Initiative(OAI), Metadata Harvesting, Interoperability, Digital Library

0. Introduction

Interoperability is the ability of information systems to operate in conjunction with each other encompassing communication protocols, hardware, software application and data compatibility layers. (Interoperability Clearinghouse Glossary of Terms). Interoperability is a broad term, touching many diverse aspects of archive initiatives, including their metadata formats, their underlying architecture, their openness to the creation of third-party digital library services, their integration with the established mechanism of scholarly communication, their usability in a cross-disciplinary context, their ability to contribute to a collective metrics system for usage and citation, etc.

Mechanisms for interoperability offer the potential for discovery tools and virtual collections that extend across the contents of multiple archives. Author also benefit from such archive spanning tools since their works will be accessible by a wider audience.

The Mechanisms for establishing this interoperability are :

- The definition of a common protocol to enable extraction of metadata from participating archives.
- The definition of a set of simple metadata elements for the sole purpose of enabling coarse granularity document discovery among archives.
- The agreement to use a common syntax, XML, for representing and transporting both metadata sets and archives-specific metadata.

1. The Mission of Open Archives Initiative

The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. It is dedicated to solving problems of digital library interoperability. Its focus has been on defining simple protocol, most recently for the exchange of metadata from archives.

Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. Continued support of this work remains a cornerstone of the Open Archives program. The fundamental technological framework and standards that are developing to support this work are, however, independent of both the type of content, and promise to have much broader relevance in opening up access to a range of digital materials. As a result, the OAI is currently an organization and an effort explicitly in transition, and is committed to exploring and enabling this new and broader range of applications.

The OAI evolved out of a need to increase access to scholarly publication by supporting the creation of interoperable digital libraries. As a first step towards such interoperability, a metadata harvesting protocol was developed to support the streaming of metadata from one repository to another, ultimately to a provider of user services such as browsing, searching or annotation. The name OAI means :

Open means the protocol is openly documented and metadata is ‘exposed’ to at least some peer group.

Archives means ‘collection of stuff”. The OAI uses the term ‘Archive’ in broader sense : as a repository for the stored information.

Initiative means that OAI is happening at break-neck speed.

2. Brief History of OAI-PMH

The OAI-PMH has its technical roots in the Universal Preprint Service (UPS) and Dienst protocol. In turn, Dienst is based on the Kahn-Wilensky Framework (KWF). Thus, KWF led to Dienst, Dienst to UPS and UPS to OAI-PMH.

In the late 1999, a meeting was convened in Santa Fe, New Mexico to identify the key issues preventing the implementation of services such as linking and searching across large, diverse, distributed E-print archives. Attendees of Santa Fe Convention developed a consensus to adopt a UPS Prototype-based metadata harvesting model as a workable technical and organizational framework for delivering digital archive content and services to end users. The harvesting model allowed “E-print (content) providers to expose their metadata via an open interface, with the intent that this metadata be used as the basis for value-added service development”.

Meeting participants also agreed upon the basic definitions, concepts, technical components and organizational aspects of interoperable E-print archives. Theses agreements became known as the “Santa Fe Convention”. Shortly after the meeting in Santa Fe, members of UPS changed the name to the Open Archives Initiative (OAI) to refer to the overall group of people and its philosophy, and named the protocol itself, the “OAI-PMH”.

Herbert Van de Sompel and Lagoze (2001), along with the members of the OAI-Technical Committee, released version 1.0 of the OAI-PMH in January 2001. The authors did not plan to make changes to the protocol version 1.0 for a period of 12-18 months after the initial release, but they adopted the newly released World Wide Web Consortium (W3C) XML Standards, and upgraded the OAI-PMH in July 2001. The authors considered the version 1.1 of the protocol to be experimental, and the 12 to 18 month observation phase provided a static time-period during which problems with the protocol were identified and evaluated.

In June 2002 Lagoze, Van de Sompel, Nelson and Warner (2002), along with the members of the new OAI Technical Committee, released version 2.0 of the OAI-PMH. This was considered to be stable, non-experimental version. Changes from 1.1 to 2.0 included referring , to “resources” rather than “document like objects”. Table 1 lists the changes made from Santa Fe Convention to OAI-PMH Ver.2.0.

	Santa Fe Convention	OAI-PMHv.1.0/1.1	OAI-PMHv.2
Nature	Experimental	Experimental	Stable
Verbs	Dienst	OAI-PMH	OAI-PMH
Request	HTTP GET/POST	HTTP GET/POST	HTTP GET/POST
Responses	XML	XML	XML
Transport	HTTP	HTTP	HTTP
Metadata	OAMS	Unqualified Dublin Core	Unqualified Dublin Core
About	E-print	Document like objects	Resources
Model	Metadata harvesting	Metadata harvesting	Metadata harvesting

Table 1 : Changes made from Santa Fe Convention to OAI-PMH v.2.0.

The OAI-PMH Version History

OAI-PMH Ver.1.0 21 Jan. 2001 OAI-PMH Ver.1.1 02 July 2001

OAI-PMH Ver.2.0 14 June 2002

Lagoze, Van de Sompel, Nelson and Warner (2002) defined an OAI-PMH record and the process for an Service Provider to obtain it from Data Provider as “metadata expressed in a single format. A record is returned in an XML encoded byte stream in response to an OAI-PMH request for metadata from an item”.

3. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework based on metadata harvesting. Metadata harvesting is a formalized framework for the co-ordinated exchange of metadata in distributed and decentralized electronic information environments. It enables automated distribution of any kind of metadata, which may be aggregated into searchable databases by “harvesting” systems. It is independent of both the types of content offered and the economic mechanisms surrounding that content, and promise to have much broader relevance in opening up access to a range of digital materials.

The OAI technical framework is not intended to replace other interoperability standards (for example Z39.50) but to provide an easy-to-implement and easy-to-deploy alternative for different constituencies or different purposes than those addressed by existing interoperability solutions. Table 2 lists the points of comparison between OAI-PMH and Z39.50.

	Z39.50	OAI-PMH
Content (objects)	Distributed	Distributed
World View	Bibliographic	Bibliographic
Object Presentation	Data Provider	Data Provider
Searching is	Distributed	Centralized
Search done by	Data Provider	Service Provider
Metadata searched is	Up-to-date	Stale
Semantic Mapping	When searching	Metadata Delivery

Table 2: OAI-PMH as Compared to Z39.50

This article is not intended to be a definitive technical summary of the protocol ; documents providing such a discussion can be found at http://www.openarchives.org/OAI/2.0/openarchivesprotocol. htm. Rather, the focus here is on the uses of the protocol and its strategic significance as an enabling technology.

4. Dublin Core Metadata Element Set (DCMES)

The DCMES is a standard for cross-domain information resource description. Here an information resource is defined to be "anything that has identity". (Dublin Core Metadata Initiative, 2003).

The OAI Community has defined a common denominator for interoperability among multiple communities while satisfying community specificity. As mapping among multiple metadata formats would place a considerable burden on service providers, who harvest the metadata and use it to build higher level services. So the protocol mandates a common metadata format : DCMES. It has been adopted as a lowest common-denominator metadata format which all data providers should support. The fifteen elements Dublin Core has over the past several years evolved as a defacto standard for simple cross-discipline metadata and is thus the appropriate choice for a common metadata set. Table 3 lists the fifteen elements of DCMES.

Title	Contributor	Source
Creator	Date	Language
Subject	Type	Relation
Description	Format	Coverage
Publisher	Identifier	Rights

Table 3: Elements of DCMES

It is not intended that the requirement to export Dc metadata should preclude the use of other metadata set that may be more appropriate within particular communities. In other words the common metadata set is DCMES but the OAI-PMH can be extended to use other metadata sets (AGLS,MARC etc).

The OAI encourages the development of community-specific standards that provide the functionalities required by specific communities. Cooperation between the OAI and the Dublin Core metadata Initiative has led to a common XML Schema for unqualified DC.

5. Extensible Markup Language (XML)

To form a record, the DCMES has to be encoded with Extensible Markup Language (XML). In other words, the DCMES has formed the building block for resource description, while XML has provided the framework for resource discovery across multiple networked systems.

The W3C ISO Standard for SGML (Standard Generalized Markup Language) has defined it as a system for creating a document markup language or tag set. W3C developed XML from SGML, and it "is a pared-down version of SGML, designed especially for web documents. It allows designers to create their own customized tags, enabling the definition, transmission, validation and interpretation of data between applications and between organizations" (Webopedia,2002).

The OAI technical framework defines a record, which is an XML encoded-byte-stream that serves as a packaging mechanism for harvested metadata. The data provider support the protocol definition if there are XML Schemas to validate all responses. Thus OAI-PMH can be extended to any metadata format that can be encoded in XML.

6. The OAI : general assumptions

OAI-PMH is not about direct interoperability between archives. It is based on a model which puts a very clean divide between data providers and service providers, the two classes of participants in the OAI-PMH framework. It is a light weight protocol which allows data providers to expose metadata records for retrieval by service providers. Figure 1 shows OAI : general assumptions.

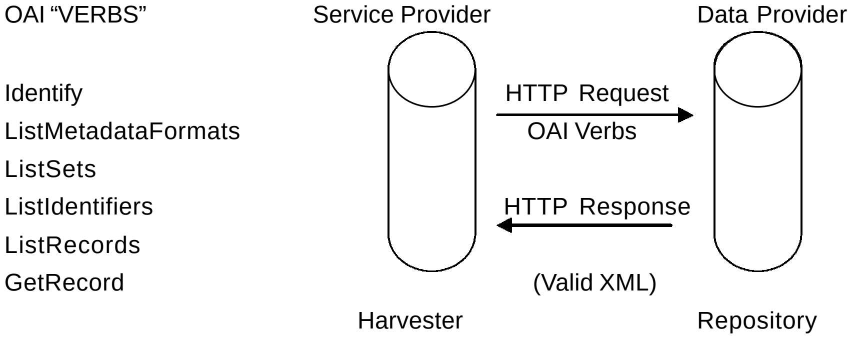

Figure 1: OAI General Assumptions.

There are two groups of 'participants': Data Providers and Service Providers.

Data Providers (open archives, repositories) provide free access to metadata, and may, but do not necessarily, offer free access to full texts or other resources. OAI-PMH provides an easy to implement, low barrier solution for Data Providers.

Service Providers use the OAI interfaces of the Data Providers to Harvest and store metadata. This means that there are no live search requests to the Data Providers; rather, services are based on the harvested data via OAI-PMH. Service Providers offer (value-added) services on the basis of the metadata harvested and they may enrich the harvested metadata in order to do so.

The Figure 2 shows OAI-PMH Data Model.

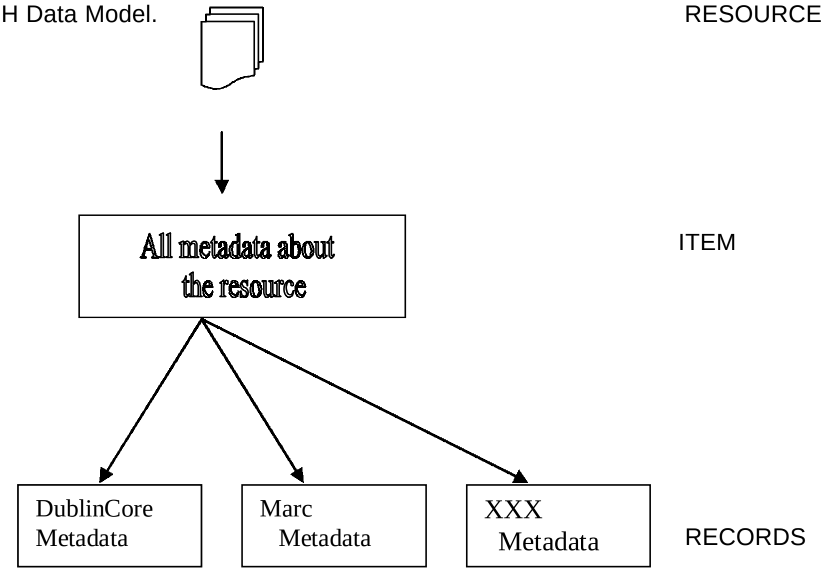

Figure 2: OAI-PMH Data Model.

7. OAI-PMH Concepts

- Harvester : a client application that make a OAI-PMH request.
- Repository : network accessible server able to process a OAI-PMH request.
- Resource : the stuff the metadata is about.
- Item : a constituent of a repository, conceptually, it is the container of the metadata.
- Identifier : Unique identifier that unambiguously identifies an item within a repository.
- Record : an XML-encoded set of metadata expressed in a specific format.
- Datestamp : date of creation / modification / deletion of a record.
- Set : optional construction for grouping items in the purpose of selective harvesting.

8. OAI-PMH Request

The OAI-PMH is based on HTTP (HyperText Transfer Protocol). Request arguments are issued as GET or POST Parameters. OAI-PMH support six request types (known as “verbs”).

- Identify: retrieve repository information.
- List Metadata Formats : what metadata formats in repository.
- List Sets : retrieve repository set structure.
- Get Record : retrieves a single metadata record.
- List Records : harvest records from a repository.
- List Identifiers : harvest record headers only.

9. OAI-PMH Responses

Responses are encoded in XML syntax. OAI-PMH supports any metadata format encoded in XML. Dublin Core in the minimal format specified for basic interoperability.

- General Information
- Metadata formats
- Set structure
- Record identifier
- Metadata

Example OAI-PMH Transaction

Request :

http://arXiv.org/oai2?
verb=GetRecord&identifier=oai:arXiv:cs/0112017&metadataPrefix=oai_dc

Response :

<?xml version=”1.0" encoding=”UTF-8"?> <OAI-PMH xmlns=”http://www.openarchives.org/OAI/2.0/” xmlns:xsi=”http://www.w3.org/2001/XMLSchema-instance” xsi:schemaLocation=”http://www.openarchives.org/OAI/2.0/http://www.openarchives…… >

10. Some Usage Cases of OAI-PMH

A number of projects use OAI-PMH in conjunction with Open URL, a specification for reference linking that is currently being standardized by NISO. The following are a few examples:

10.1 Andrew W. Mellon Foundation

In 2000 the Andrew W. Mellon Foundation sought to explore, how libraries and other repositories of scholarly information make metadata about scholarly collections more visible/useful to Internet users through the use of OAI-PMH.

Mellon hosted a series of planning meetings and eventually awarded seven grants to fund test projects. These grants are provided to the following Institutions:

The Research Libraries Group, Emory University (MetaArchive.Org), SOLINET/ASERL (AmericanSouth.Org), The University of Michigan (OAIster), University of Illinois at Urbana-Champaign, University of Virginia, Woodrow Wilson International Center for Scholars.

10.2 The MetaScholar Initiative

The MetaArchive and AmericanSouth projects merged to form the MetaScholar Initiative. This initiative is creating an extended metadata aggregation network encompassing some two dozen academic libraries, archives and museums across the United States.

10.3 The University of Illinois Urbana-Champaign

The University of Illinois Urbana-Champaign, through an Andrew W. Mellon Foundation grant, is exploring the feasibility of using OAI-PMH to build services to reveal and make more accessible collections of cultural heritage material. By Feb. 2002, the Illinois OAI-PMH project has harvested metadata from 25 different institutions or consortium.

Some other projects based on OAI-PMH Usage are :

The Project of Open Archives Initiative Virginia Tech DLRL Project, NSDL (National Science Digital Library), BOAI (Budapest Open Access Initiative), NDLTD (Networked Digital Library of These and Dissertation), Internet Archives, eprints.org , rclis (Research in Computing Library and Information Science), IMS (International Metadata Standard), LAOAP (Latin American Open Archives Portal), PhysDoc, MathDoc (Germany), the California Digital Library eScholarship program (USA), the MIT Dspace project (USA), Theses Electronique and Hyper Articles at CNRS project (France), projects at Lund University (Sweden) and Caltech (USA).

All the above projects show the recent developments in the OAI-PMH Usage.

Some Open Source softwares which support OAI-PMH are : Greenstone Digital Library Software from New Zealand Digital Library Projects, DSpace, eprints archiving software, DLESE OAI software, CERN’s CDSware document Server.

11. Conclusion

The Open Archives Initiative Protocol for Metadata Harvesting “has emerged as a practical foundation for digital library interoperability”. It can be used by a variety of communities who are engaged in publishing content on the web, as the OAI protocol has great potential for increasing access to exposure of hidden resources via the web.

Many features make the protocol special in the world of digital libraries and one such feature is its ability to allow co-existence of multiple domain specific metadata vocabularies, collection description and resource organization schemes more like a cross platform situation. Within two-three years, OAI-PMH is likely to be the primary means of making research information known to search services.

The OAI-PMH opens many new possibilities, which are yet to be explored. This means that it is difficult and speculative, to establish strategies to exploit the new technology. But these opportunities are too import to be ignored.

12. References

1. Andrew W. Mellon Foundation. 2002. The Mellon Metadata Harvesting Initiative http://www.diglib.org/forums/fall2002/mellon%20metascholar.htm (Accessed on 02.09.03)
2. Suleman, Hussein. 2002. Introduction to the OAI-PMH (v.2.0) http://www.dlib.vt.edu/projects/OAI/reports/solinet_2002_oai2_overview.pdf (Accessed on 05.11.03)
3. Cole, Timothy W. et.al. 2002. Now that We’ve found the ‘Hidden Web’, What Can We Do With it? http://www.archimuse.com/mw2002/papers/cole/cole.html (Accessed on 05.12.2003)
4. Halbert, Martin. 2002. The Open Archives Initiative and MetaScholar Projects http://www.rlg.org/events/metadata2002/halbert/tsld001.htm (Accessed on 06.11.2003)
5. Carpenter, Leona. 2003 OAI for Beginners : Overview http://www.oaforum.org/tutorial/page1.htm (Accessed on 02.09.2003)
6. Sompel, Herbert Van de. et.al. 2002. 2nd workshop on the Open Archives Initiative (OAI), CERN, 17-19. October 2002, Report. http://library.cern.ch/sompeletal.htm (Accessed on 19.11.03)

About Authors

Ms. Martha Latika Alexander is Research Scholar and part time lecturer in SOSLIS at Jivaji University Gwalior 474011, India and holds B.Sc. and MLISc. He is Life Member of ILA and contributed two research papers.
E-mail: marthalatika@hotmail.com

Dr. J. N. Gautam is Reader and Head of S.O.S in Library and Information Science at Jiwaji University, Gwalior 474 011, India and holds M.Sc., MLISc. and Ph.D. He contributed 50 papers and 6 books in LIS and his fields of specialization are User Education, Bibliometrics, Information and Communication Systems, Library Organisation and Processing and University Library System.
E-mail : jngautam@yahoo.com