Open Archives Metadata Harvesting: An Overview

Shipra Awasthi Babita Jaiswal

Abstract

The goal of the Open Archives Initiative Protocol for Metadata Harvesting is to supply and promote an application-independent interoperability framework that can be used by a variety of communities who are engaged in publishing content on the Web. This paper discusses about OAI and a protocols which harvests metadata i.e. OAI-PMH.

Keywords : Open Access Initiative, Metadata Harvesting

1. Introduction

1.1 Metadata: Metadata can be defined as "data about data" describe the content, quality, condition, and other characteristics of data. Metadata is vital in helping potential users to find needed data and determine whether a data set will meet their needs before they spend the time and money to obtain and process it.

1.2 Harvesting: In the Open Archives Initiative context, harvesting refers specifically to the gathering together of metadata from a number of distributed repositories into a combined data store.

1.3 Open Archives Initiative (OAI): Open access scientific outputs are scattered across many disciplinary archives, institutional e-print archives, institutional repositories and open access journals. Therefore, it is difficult for scholars to locate all needed works on a particular subject.

Open Archives Initiative, is one of the international movement to solve this problem. It aims to develop and promote the use of a standard protocol, known as the Open Archives Metadata Harvesting Protocol (OAMHP). It is designed for better sharing and retrieval of e-prints residing in distributed archives. It also promotes interoperability standards that aim to facilitate the efficient dissemination of content.

Open Archives Initiative is an attempt to build a "low-barrier interoperability framework" for archives containing a digital content. It allows people to harvest metadata from Data Providers.

The mission of the Open Archives Initiative is to "develop and promote interoperability standards that aim to facilitate the efficient dissemination of content". The Protocol for Metadata Harvesting, a tool developed through the Open Archives Initiative, facilitates interoperability between disparate and diverse collections of metadata through a relatively simple protocol based on common standards such as XML, HTTP, and Dublin Core. The Open Archives Initiative world is divided into data providers or repositories, which traditionally make their metadata available through the protocol, and service

pr in
th

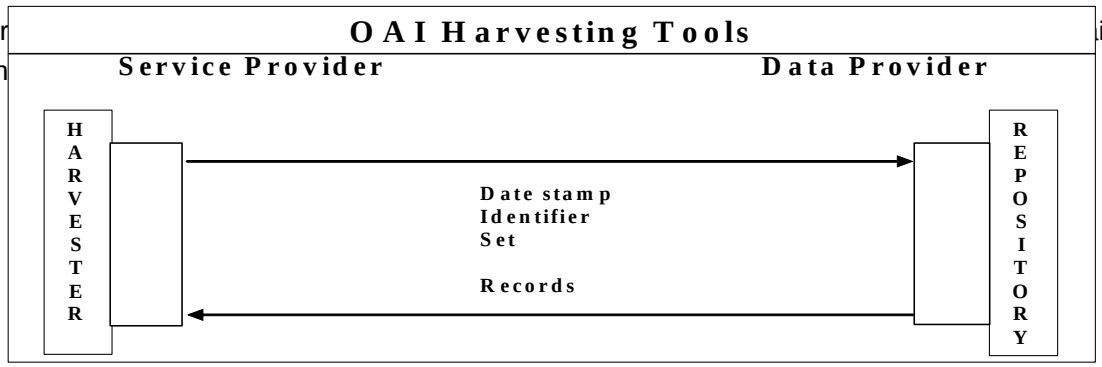


Figure 1: OAI Harvesting Tools

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

OAI-PMH is a protocol developed by the Open Archives Initiative. It is used to harvest (or collect) the metadata descriptions of the records in an archive so that services can be built using metadata from many archives. The protocol is usually just referred to as the OAI Protocol. OAI-PMH uses XML over HTTP. The current version is 2.0, updated in 2002. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has been widely adopted since its initial release in 2001.Initially developed as a means to federate access to diverse e-print archives through metadata harvesting, the protocol has demonstrated its potential usefulness to a broad range of communities. According to the experimental OAI Registry at the University of Illinois Library at Urbana – Champaign there are currently over 300 active data providers using the production version (2.0) of the protocol from a wide variety of domains and institution types.

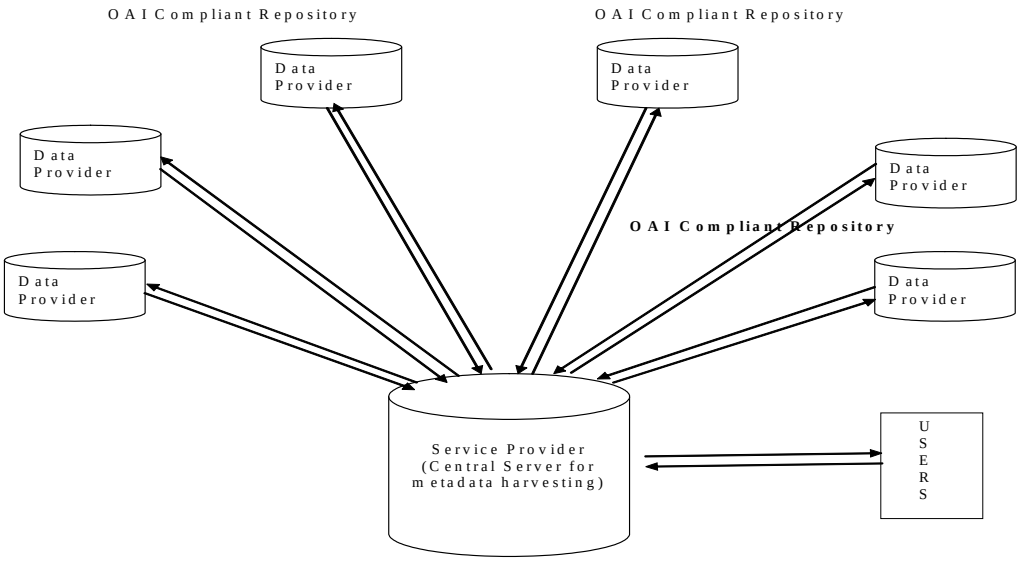


Figure : 2 A Schematic of OAI-PMH compliant Institutional Repositories

The OAI-Protocol for Metadata Harvesting (OAI-PMH) defines a mechanism for harvesting records containing metadata from repositories. The OAI-PMH gives a simple technical option for data providers to make their metadata available to services, based on the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). Thus, metadata from many sources can be gathered together in one database, and services can be provided based on this centrally harvested or "aggregated" data.

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a low-barrier mechanism for repository interoperability. Data Providers are repositories that expose structured metadata via OAI-PMH. Service Providers then make OAI-PMH service requests to harvest that metadata. OAI-PMH is a set of six verbs or services that are invoked within HTTP. OAIPMH is not limited to Dublin Core (DC) metadata. However, since Open Archives Initiative aims to promote interoperability, DC metadata has been adopted as a lowest common-denominator metadata format which all data-providers should support.

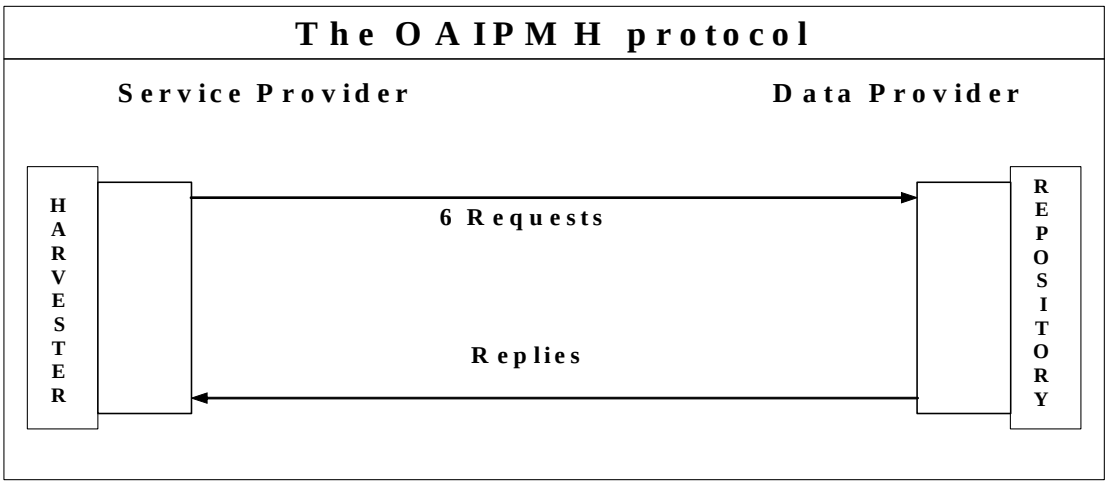


Figure : 3 The OAIPMH Protocol

Interoperability is the ability of two or more systems to exchange information and to use the information that has been exchanged. National Information Standards Organization (NISO) defines interoperability as "the ability of multiple systems, using different hardware and software platforms, data structures and interfaces, to exchange and share data"

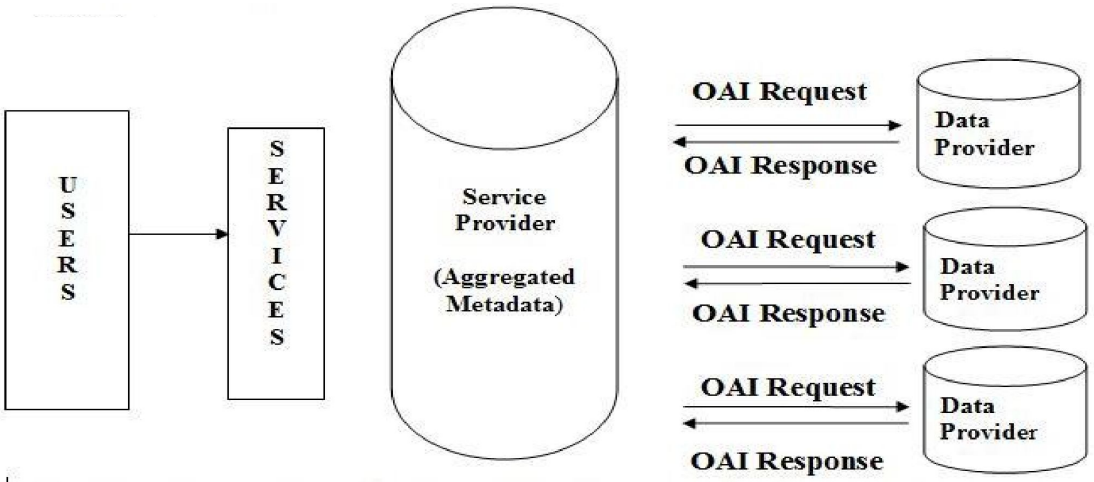
3. Key players in OAI-PMH

There are two classes of participants in the OAI-PMH work:

3.1 Data providers: A data provider maintains one or more repositories (web servers) that support the OAI-PMH as a means of exposing metadata. These are the repositories which process

the request and respond to service providers with appropriate OAI-PMH responses. They are creators and keepers of the metadata for objects (repositories) and archives of resources.

3.2 Service Providers: A service provider issues OAI-PMH requests to data providers and uses the metadata as a basis for building value-added services. They are harvesters of the metadata for the purpose of providing a service such as a search interface, peer-review system, etc.



Interaction between Service Provider & Data Provider & the process of Metadata Harvesting

Figure -4

4. Software

OAI-PMH is based on client-server architecture, in which "harvesters" request information on updated records from "repositories". Requests for data can be based on a date stamp range, and can be restricted to named sets defined by the provider. Data providers are required to provide XML metadata in Dublin Core format, and may also provide it in other XML formats. A number of software systems support the OAI-PMH, including GNU EPrints from the University of Southampton and DSpace from MIT.

4.1 Software to run Open Archives Initiative Repository

- Eprints.org – University of Southampton
- Open source metadata server – OCLC
- NT OAI Server - University of Illinois
- Aleph 500 – Ex Libris
- Z39.50 – OAI gateway – Virginia Tech (ongoing)
- MARC to DC converter – OCLC

4.2 Existing OAI-PMH based Approaches

Typical scenario:

- An OAI-PMH harvester harvests Dublin Core records from the OAI-PMH repository.
- The harvester analyzes each Dublin Core record, extracting dc.identifier information in order to determine the network location of the described resource.
- A separate process, out-of-band from the OAI-PMH, collects the described resource from its network location.

4.3 Benefits of OAI-PMH

The benefits are as follows:

- Simple, easy and built on existing technology
- Web friendly
- Multiple service providers can harvest from multiple data providers ensuring a wider spread of metadata.
- Low cost
- An open standard

4.4 Major software supporting OAI-PMH

- Arc (http://arc.cs.odu.edu/)
- Citebase (http://citebase.eprints.org/cgi-bin/search)
- CYCLADES (http://www.ercim.org/cyclades)
- DP9 (http://arc.cs.odu.edu:8080/dp9/index.jsp)
- MeIND (http://www.meind.de/)
- METALIS (http://metalis.cilea.it/)
- My.OAI (http://www.myoai.com)
- NCSTRL (http://www.ncstrl.org/)
- Perseus (http://www.perseus.tufts.edu/cgi-bin/vor)
- Public Knowledge Project – Open Archives Harvester (http://pkp.ubc.ca/harvester/)
- OAICAT (http://www.oclc.org/research/software/oai/cat.htm)
- OAI Repository Explorer (http://re.cs.uct.ac.za)
- OAIster (http://oaister.umdl.umich.edu/o/oaister/)
- OASIC (http://www.oclc.org/research/software/oai/harvester.htm)
- OAIHarvster (http://www.oclc.org/research/software/oai/harvester.htm)
- DLESE OAI Software (http://dlese.org/oai/index.jsp)

4.5 Major Metadata Harvesting Services in India

A metadata harvesting service harvests or indexes metadata from OAI-compliant archives or repositories through harvesting software that supports a protocol known as OAI-PMH (Open Access

Initiative Protocol for Metadata Harvesting). Some Indian institutions have been experimenting with metadata harvesting services and installed metadata harvesters. Major Metadata harvesting services in India are

- Search digital Libraries (SDL)
- SJPI (Scientific Journal Publishing in India)Cross Journal Search Service
- Open J-Gate
- Knowledge Harvester@INSA

Sl.No	Name	URL	Host	Software
1	Search Digital Libraries (SDL)	http://drtc.isibang.ac.in/sdl	DRTC, Bangalore	Public Knowledge Project system
2	SJPI Cross Journal Search Service	http://144.16.72.144/harvester/	NCSI, IISC	Public Knowledge Project system
3	SEED	http://eprint.iitd.ac.in/seed/	IIT, Delhi	Public Knowledge Project system
4	Open J-Gate	www.openj-gate.com/	Informatics India Ltd.	----------
5	Knowledge Harvester@INSA	http://61.16.154.195/harvester/	INSA	Public Knowledge Project system

Table : 1 Metadata Harvesting Services in India

5. Conclusion

Metadata is a key part of the information infrastructure necessary to help create order in the chaos of the Web, infusing description, classification, and organization to help create more useful stores of information. OAI metadata harvesting offers a new bridge to bring new innovation in networked information services and applications, out of the research community more rapidly. The spreading of the OAI-PMH may allow such experiments to be launched and tools and procedures to be exchanged.

This is a challenge for funders and institutional stakeholders of digital heritage, to include the Open Archives Initiative model within the digital content creation framework, to anticipate the major initiatives for standardization within the cultural heritage sector and ensure the involvement of major industrial heritage partners to implement OAI repositories for document management applications. The OAI-PMH provides an easy and effective solution for scholarly data circulation and academic communication.

References

1. Hirwade, Mangala and Hirwade, Anil. Metadata Harvesting Services in India.
2. Prasad, A. R. D. Interoperability and the OAI-PMH. Proceedings of the National Conference on Information Management in Digital Libraries (NCIMDIL) 2-4Aug, 2006.
3. Shreeves, Sarah L. et.al. Current Developments and future trends for the OAI Protocol for Metadata harvesting .Library Trends, 53(4).
4. Sompel, Herbert Van de. OAI-PMH for Resource Harvesting. OAI-PMH for Resource Harvesting Tutorial OAI4, October 20th 2005, CERN, Geneva, Switzerland
5. Sompel, Herbert Van de. OAI metadata harvesting specifications. Workshop on OAI and peer review journals in Europe Geneva, Switzerland – March 22nd to 24th 2001.
6. Sompel, Herbert Van de and Lagoze ,Carl. Version of the OAI-PMH and some other stuff. 2nd Workshop on the OAI, CERN Geneva, October 17th 2002
7. Warner, Simeon. Exposing and Harvesting Metadata using the OAI Metadata Harvesting Protocol: A tutorial. High Energy Physics Libraries Webzine, Issue 4, June 2001.

About Authors

Ms. Shipra Awasthi, Assistant Librarian, National Institute of Technology, Rourkela, Orissa.
Dr. Babita Jaiswal, Lecturer, Department of Library and Information Science, University of Lucknow, Uttar Pradesh.