METADATA HARVESTING AND THE OPEN ARCHIVES INITIATIVE

Purushothama Gowda M M K Bhandi

Abstract

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a collaborative effort that provides an application- independent interoperability framework based on Metadata Harvesting. Though the OAI-PMH is a very recent development it is being regarded as an important step towards information discovery in the digital library arena. This paper looks into the issues leading to its development as well as gives an inside view of the proposed model.

Keywords: Metadata Harvesting Protocol, Open Archives Initiative, 239.50, Repositories

1 Introduction

There has been considerable confusion about the Open Archives Initiative (OAI)- Metadata Harvesting Protocol (MHP), mostly beginning with and stemming from its name. The protocol no longer has much to do with archiving or archives, other than in terms of its heritage. The OAI-PMH is a means of making machine-readable metadata widely available for use. The Open Archives Initiative was originally proposed to enhance access to e-print/pre-print archives. Gradually, however, the scope of the initiative has broadened to cover any kind of digital content including images and videos. It is available to all regardless of economic mechanism surrounding the content. The fundamental idea here is that authors would deposit preprints and/or copies of published versions of their articles into such servers, thus providing readers worldwide with a free way of obtaining access to these papers, without needing paid subscription access to the source electronic journals. The proponents of this movement argue that the refereed scholarly journal literature really belongs to the scholarly community and by extension to the world at large, and that such free access is better aligned with the interests of both authors and readers. The deposit of preprints would also speed up and democratize the frontiers of research and access to new knowledge; instead of a privileged circle of members of “invisible colleges” sharing preprints, these preprints would be available to everyone immediately, without the delays introduced by the journal refereeing and publication cycle. Proposals such as PubMed Central and the Public Library of Science build upon these ideas [9].

The Open Archives Metadata Harvesting Protocol grew out of an effort to solve some of the problems that were emerging as e-print servers became more widely deployed; it originated in the community concerned with advancing the development of e-print archives. The protocol was widely known as the Open Archives Protocol, and the program to develop it was widely known as the Open Archives Initiative, so the decision was made to maintain the popularly known terminology.

The protocol is now often referred to as the Open Archives Metadata Harvesting Protocol in an attempt to reintroduce a bit more clarity. This Metadata Harvesting Protocol can employ to make metadata describing objects housed at that server available to external applications that wish to collect this metadata. A server does not need to be part of an e-print program to use the protocol; indeed, it does not need to house journal papers at all. The server does not need to offer free access to the digital objects that it stores.

2. History of OAI

The origin of OAI can be traced back to the efforts to increase interoperability among the e- print/pre- print servers that hosted scientific and technical papers [3]. A number of factors led to the development of the pre-print archives most important of which was the rising cost of journals. Scholars and researchers would deposit their articles and papers into these servers, which allow for the dissemination of information among the scholarly community much more rapidly than through traditional print journals. The number of e-print/pre-print repositories was growing steadily in the nineties. This growth created an information overload and some other problems, which can be summarized as:

1. The end-users/scholars may not be able to know the existence of a repository.
2. Overlapping of coverage in terms of subjects.
3. Multi-disciplinary nature of subjects needed the documents to be kept at a number of repositories.
4. Discipline-specific and institution-specific archives created duplication efforts.
5. The end-users/scholars had to search individual repositories to get documents of his interest.
6. Also, it was undesirable to require scholars to deposit their work in multiple repositories.

Need was felt to build a framework to bring about a kind of integration of these e-print/pre-print archives to solve these problems. A meeting was convened in late 1999 at Santa Fe, New Mexico to address problems of the e-print world. The major work was to define an interface to permit e- print servers to expose their metadata for the papers it held, so that search services or other similar repositories could then harvest its metadata. These archives would then act as a federation of repositories by giving a single search platform for multiple collections.After the meeting, the agreed principles were launched in January 2000 as the Open archives Initiative specification by Herbert Van de Sompel, Rick Luce, and Paul Gisparg among others.

The Digital Library Federation, the Coalition for Networked Information, and the National Science Foundation sponsored it.The OAI Steering Committee was formed in August 2000 to give the strategic direction to the protocol. The protocol version 1.1 was launched in July 2001. The Open Archives Initiative Technical Committee (OAI-TC) was formed to develop and write version 2 of the Open Archives Protocol for metadata Harvesting based on feedback from implementers. The OAI-PMH version 2.0 was eventually released in June 2002 (

3 OAI vs. Z39.50

There was a debate as to why not use the existing Z39.50 protocol, which is also used for the search and transfer of metadata. The OAI's metadata - harvesting approach might look operationally much different to the Z39.50, but both achieve what's often called "federated searching." The federated searches allow users to gather information from multiple related resources through a single interface. The basic difference between the two protocols is in the search approach. The Z39.50 allows clients to search multiple information servers in a single search interface in *real time,* whereas the OAI-PMH allows bulk transfer of metadata from the repositories to the Service Providers' database. Hence the clients do not need search multiple data providers in real time rather they search the metadata database of the Service Provider who collect and aggregate the metadata from different data providers.

There were many reasons to have a completely new protocol rather than implementing the Z39.50 as it stands. Some of the reasons are:

1. Z39.50 is a mature, sophisticated, but unfortunately very complex protocol. It can be used as a tool to build federated search systems; in such a system, a client sends a search in parallel to a number of information servers that comprise the federation, and then gathers the results, eliminates or clusters duplicates, sorts the resulting records and presents them to the user.

2. It has been proven that it is very difficult to create high-quality federated search services across large numbers of autonomous information servers through Z39.50 for several reasons.

3. Retrieval accuracy is a problem: different servers interpret Z39.50 queries differently, in part due to lack of specificity in the standard, leading to semantic inconsistencies as a search is processed at different servers.

4. There are scaling problems in the management of searches that are run at large numbers of servers; one has to worry about servers that are unavailable (and with enough servers, at least one always will be unavailable), and performance tends to be constrained by the performance of the slowest individual server participating in the federation of servers.

5. Compromising speed of access since the user has to wait for a lot of record transfer and post-processing before seeing a result, making Z39.50-based federated search performance sensitive to participating server response time, result size, and network bandwidth.

The open archives committee adopted a model that rejected distributed search in favor of simply having servers provide metadata in bulk for harvesting services, subject only to some very simple scoping criteria, such as providing all metadata added or changed since a specified date, or all metadata pertaining to papers meeting matching gross subject partitions within an archive [4]

Implementing PMH is very simple since one does not need a different port like Z39.50 (which uses port 210). It works over the HTTP, which any web server listens, and any web browser or web-downloader talks. It means one can use common Linux programs such as wget or curl to harvest the metadata from repositories. One does not need a special toolkit (like Yaz for Z39.50).According to Lynch “These two protocols are really meant for different purposes, with very different design parameters, although they can both be used as building blocks in the construction of similar services, such as federated searching. Neither is a substitute for the other […] and we should not think about the world becoming partitioned between Z39.50-based resources and MHP-speaking resources, but rather about bridges and gateways.”

4 Metadata Standards and OAI-PMH

For the purpose of interoperability, the OAI Protocol for metadata Harvesting specifies unqualified Dublin Core, encoded in XML, a mandatory metadata schema as the lowest common denominator. It is certainly clear that almost any metadata scheme can be “downgraded” into unqualified Dublin Core. However, each server is also free to offer metadata in one or more other schemas, and a harvester can request that metadata in any format in addition to the unqualified Dublin Core.

The ListMetadataFormats request will return the metadataPrefix, schema, and optionally a metadataNamespace, for either a particular record or for the whole repository (if no identifier is specified). In the case of the whole repository, all metadata formats supported by the repository are returned. It is not implied that all records are available in all formats.

5 The Metadata Harvesting Interface

Harvesting Protocol uses a very simple HTTP-based request-response transaction framework for communication between a harvester and a repository. A harvester can ask for metadata to be returned with optional restrictions based on when the metadata has been added or modified (in other words, it can obtain new or changed metadata since its last harvest interaction with a repository); it can also restrict metadata by server-defined “partitions” .The server returns a series of sets of metadata elements (in XML) plus identifiers (i.e., URLs) for the objects that the metadata describes.

Multiple metadata schemes are supported in the Open Archives Metadata Harvesting Protocol—this is really the key architectural change from the Santa Fe Convention. The protocol requires that all servers offer unqualified Dublin Core metadata (encoded in XML) as a lowest common denominator; however, each server is also free to offer metadata in one or more other schemes, and a harvester can request that metadata be provided in a scheme other than Dublin Core as part of the harvest request. There is also another auxiliary transaction that permits a harvester to obtain a list of the names of the metadata schemes that a given repository supports. The underlying idea here is that we will see communities of practice evolve that define metadata schemes that are richer and more precise than unqualified Dublin Core; for example, the e-print archives community is already working on one that encodes various important data elements for e-prints, such as author affiliations, bibliographic information if the paper has been published in a journal, and even the paper’s cited references in a structured form. These community-specific schemes could be handled as qualified Dublin Core, or as de novo schemes; the only requirement is that they be transportable in XML.

The protocol does not address the very real issue of how harvesters will identify repositories that they wish to harvest, nor does it provide information to help determine when harvesting should occur, or how frequently. Questions about acceptable use of harvested metadata are not addressed by the protocol; these might be agreed upon explicitly as part of establishing a harvesting relationship with a server that is access-controlled, or they might be simply advertised as terms and conditions that any harvester automatically agrees to in the case of a publicly-accessible server, but in any case this is outside the scope of the harvesting protocol.

6 Applications Enabled by the Metadata Harvesting Protocol

The most obvious applications that are enabled by the Metadata Harvesting Protocol are those that helped to motivate the work at the initial Santa Fe meeting: repository synchronization and federated search. For repository synchronization, one compares metadata from two or more repositories and decides what objects should be copied from one repository to another (along with the necessary metadata). The hard part here is in the application: deciding what repositories to examine, and determining the criteria for identifying what to copy. There is also a problem with the propagation of metadata from one repository to another; it’s not clear (other than by using community standards) how to determine the most comprehensive metadata set describing an object so that all of the relevant metadata can be copied over.

Similarly, federated search using MHP is not hard in principle; one collects metadata from a number of sites, normalizes it, clusters it in some fashion to deal with duplicates as appropriate, and offers search services against the resulting database. In practice, all of the details are complex: what sites to harvest, how often to harvest them, how to normalize metadata, how to handle duplicate objects—these are all key design issues that need to be addressed. MHP provides a very powerful framework for building union-catalog-type databases for collections of resources by automating and standardizing the collection of contributions from the participating sites, which has traditionally been an operational headache in building and managing union catalogs. But there are many complex specifics that need to be coded into any actual implementation.

A set of applications closely related to federated search deal with the potential enhancement of web search engines in at least two distinct dimensions. One is providing a more efficient way for web search engines to crawl static HTML pages, and also to obtain metadata associated with these pages. The second is being able to integrate various parts of what is sometimes called the “deep web” or the “invisible web” with the indexing of static web pages, including repositories of digital objects and databases that do not exist as retrievable and indexable static web pages, and also proprietary content, where the content owner may be willing to make metadata about the content available to facilitate finding it, but may be unwilling to permit arbitrary web-indexing programs to have direct access to the content in order to index it.

7. Open Questions and Future Directions for Open Archives Metadata Harvesting

While the Open Archives Metadata Harvesting Protocol solves one very important set of problems, it also focuses attention on a number of other issues that will have to be addressed as applications proliferate. Some of them will require progress in standards and/or other networked information infrastructure components; others are simply not well understood at this point and will require considerable research and experimentation to allow the development of a body of design knowledge and community practice. In this final section I will briefly sketch some of these issues.

8. Selective Harvesting

Harvesters can also limit the metadata to be returned by applying restrictions based on two relatively simple criteria:

Date-based: Harvesters may use datestamps to harvest only those records that were created, deleted, or modified within a specified date range. To specify datestamp-based selective harvesting, datestamps are included as values of the optional arguments, *from* and *until*, in the *ListRecords* and *ListIdentifiers* requests.

***Ex*ample:**

http://arxiv.org/oai2?verb=ListRecords&from=20021112&untill=20030212&metadataPrefix=oai_dc

Set-based: Harvesters may specify set membership as a criterion for selective harvesting. To specify set-based selective harvesting, a setSpec is included as the value of the optional set argument to the ListRecords and ListIdentifiers requests, thereby specifying selective harvesting of records from items within the respective set.

Example:

http://rocky.dlib.vt.edu/jcdlpix/cgi-bin/OAI/jcdlpix.pl?verb=ListRecords &set=200105dle&metadataPrefix =oai_dc

9. Flow Control and the Resumption Token

One of the concerns with the PMH model involves how a service provider can obtain large numbers of metadata records from a data provider without overburdening the system. The way that metadata records are transferred remains under the control of the data provider.Flow control is supported with the HTTP retry-after status code 503. This allows a server (data-provider) to tell the harvesting agent (service-provider) to try the request again after some interval. It is left entirely up to the server implementer to determine the conditions under which such a response will be given. The server could base the response on current machine load or limit the frequency at which requests will be serviced from any given IP

address. The retry-after response may also be used to handle temporary outages without simply taking the server off-line. In an environment where one of a set of servers may handle a request, the server may dynamically redirect a request using the HTTP 302 response.The PMH takes into consideration that the data provider will have preferences regarding when it will want to respond to harvester and how many records it will deliver in a given time. PMH includes a control mechanism called a *Resumption Token*. At any time, a data provider's server can return an incomplete set or records in response to a request, issuing a *resumptionToken.* To retrieve the next portion of the complete list the next request must use the value of that *resumptionToken* element as the value of the *resumptionToken* argument of the request. Optionally, this token may be valid for a certain period of time only mentioned as *expiration Date*.

9.1 Exception Condition and Error Handling

The OAIMH protocol has very simple exception handling: syntax errors result in HTTP status code 400 replies, and parameters that are invalid or have values that do not match records in the repository result in empty replies. For example, a *ListRecords* request for a date range when there were no changes, or for a metadata format not supported, will result in a reply with header information but no *<record>* elements [5].

10 Some Existing Data Providers

As discussed earlier the Data Providers are repositories or archive of a digital content with some kind of metadata describing the content. The Data Providers expose their metadata, by installing a piece of software, in such a manner that harvesters can harvest their metadata to build value added services.

10.1 ArXiv E-Print Archive

Description : ArXiv is an e-print service in the fields of physics, mathematics, non-linear science and computer science.

Homepage: http://arxiv.org/

Base URL: http://arXiv.org/oai2

10.2 E-Prints in Library and Information Science (E-LIS)

Description: E-LIS is an electronic open access archive for scientific or technical documents, published or unpublished, in Librarianship, Information Science and Technology, and related application activities.

Homepage: http://eprints.rclis.org/

Base URL: http://eprints.rclis.org/perl/oai2

10.3 CogPrints

Description: Cognitive Sciences E-print Archive. An electronic archive for self-archive papers in any area of Psychology, neuroscience, and Linguistics, and many areas of Computer Science , Philosophy, Biology, Medicine, Anthropology, as well as any other portions of the physical, social and mathematical sciences that are pertinent to the study of cognition.

Homepage: http://cogprints.ecs.soton.ac.uk/

Base URL: http://cogprints.ecs.soton.ac.uk/perl/oai2

10.4. Open Video Project

Description: The Open Video Project is a shared digital video repository and test collection intended to meet the needs of researchers in a wide variety of areas related to digital video.

Homepage: http://www. Open -video.org/

Base URL: http://www.pen -video.org/oai2.0/

11. Some Existing Service Providers

The Service Providers harvest the metadata exposed by the Data Providers. Their job is similar to the web-crawlers of the Internet search engines. Theygo to the individual repositories to harvest their entire metadata, collects in its database in the XML format. The collected metadata is then parsed to provide an integrated search interface and browsing indices to the collections of all the participating data providers/ repositories.

11.1. OAIster

Description: OAIster is a project of the University of Michigan Digital Library Production Services, originally funded through a Mellon grant.

Homepage: http://oaister.umdl.umich.edu/o/oaister/

11.2. Networked Computer Science Technical Reference Library

Description: The Networked Computer Science Technical Reference Library (NCSTRL -pronounced as “ancestral”) is an international collection of computer science research reports made available for non-commercial use from over 100 participating organizations worldwide. The organizations that participate in NCSTRL include Ph.D. granting computer science departments, research laboratories, ePrint repositories, and electronic journals.

Homepage: http://www.ncstrl.org

11.3. iCite: CITATION INDEXING

Description: iCite is a citation indexing service based on OAI-PMH by Scuola Internazionale Superiore di Studi Avanzati (SISSA, International School for Advanced Studies), Italy.

Homepage: http://icite.sissa.it:8888/icite/

11.4. Electronic Thesis/Dissertation OAI Union Catalog

Description: This is a service built by harvesting metadata from Open Archives of electronic theses and dissertations. The underlying technology is based on layered Open archives with data being harvested from source archives and then stored in a Union Catalog.

Homepage: http://rocky.dlib.vt.edu/etdunion/cgi-bin/index.pl

12 Conclusions

The Open Archives Metadata Harvesting Protocol 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 important to be ignored.

For content suppliers, the way forward seems clear. They should prepare to offer metadata through the MHP interface. Yet they will need to think very carefully about what they are doing, both in terms of what metadata they want to expose and at what level of granularity, and in terms of the potential reuse of this metadata. This is particularly true for operators of online catalogs, though it is also a question for organizations mounting special collections of all kinds. Any organization offering access to a sophisticated networked information resource may find the MHP is a new way to make content available to a variety of innovative service providers.

For data-intensive scholarly communities in which data is widely distributed rather than centralized into a few key community databases, this interface may offer a new way to translate rather abstract investments in metadata standardization into tangible opportunities to contribute to operational systems for locating information resources. And it may have other far-reaching implications; for example, in communities where the resources to underwrite centralized databases haven't been available, or where the community practices emphasize local control of datasets by individual research groups, the base of available information may become much more visible to the community.

Finally, OAI metadata harvesting may offer a new bridge to bring innovation in networked information services and applications out of the research community more rapidly than has been the case in the past. Organizations that manage large databases and production information services are generally slow to innovate because their first priorities appropriately reflect the needs to exercise stewardship over the data and to provide reliable service to their user communities; most of their resources tend to be tied up in operations and maintenance. Researchers who want to explore new ways of organizing, presenting, or using these large data resources will now have a standardized way of extracting content without much disruption or cost to existing operational systems. This may be a powerful mechanism for enabling the development of new applications and services that have never before been possible.

13. References

1. Breeding, M. (2002, April). The Emergence of the Open Archives Initiative: This Protocol could become a key part of the digital library infrastructure. Information Today from http://www.findarticles.com/cf_0/m3336/4_19/85251474/p1/article.jhtml
2. Breeding, M. (2002). Understanding the Protocol for metadata Harvesting of the Open Archives Initiative. Computers in Libraries, 22(8).
3. Lagoze, C., & Sompel, H. V. d. (2001, January). The Open Archives Initiative Protocol for metadata Harvesting, from http://www.openarchives.org/OAI/openarchivesprotocol.htm
4. Lynch, C. A. (2001, August). metadata Harvesting and the Open Archives Initiative. ARL Bimonthly Report 217. from http://www.arl.org/newsltr/217/mhp.html
5. Shearer, K. (2002, March). The Open Archives Initiative: Developing an Interoperability Framework for Scholarly Publishing. CARL/ABRC Background Series, No. 5. from http://www.carl-abrc.ca/projects/scholarly/open_archives.PDF
6. Suleman, H., & Fox, E. A. (2001, December). A Framework for Building Open Digital Libraries. D-Lib Magazine, 7(12). from http://www.dlib.org/dlib/december01/suleman/12suleman.html

7. Sompel, H. V. d., & Lagoze, C. (2000, February). The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, 6(2). from http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

8. Warner, S. (2001, June). Exposing and Harvesting metadata Using the OAI Metadata Harvesting Protocol: A Tutorial. HEP Libraries Webzine Issue 4. from http://library.cern.ch/HEPLW/4/papers/3/

9. <http://www.pubmedcentral.nih.gov/> and <http://www.publiclibraryofscience.org/> for more information.

10. <http://web.mit.edu/dspace/>.

11. <http://www.openarchives.org/meetings/SantaFe1999/sfc_entry.htm>.

12. <http://www-cse.ucsd.edu/~rik/others/lynch-trust-jasis00.pdf>.

About Authors

Shri. Purushothama Gowda M is a Senior Assistant Librarian at Mangalore University Library
Email: purushotham@mangaloreuniversity.ac.in

Dr M K Bhandi is a University Librarian, at Mangalore University, Mangalore.
E-mail : mkb @mangaloreuniversity.ac.in