

Open Archives Initiative Protocol for Metadata Harvesting Report to SAA Council and Standards Committee

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has been suggested as a relatively simple means by which metadata drawn from various sources and storage formats might be exchanged in a common format.¹ OAI's purpose has been fully summarized elsewhere,² but in general four main points can be made regarding OAI and its relationship to the archival profession.

First, OAI is an international initiative centered on increasing the interoperability of digital libraries, broadly construed. OAI originated in the scientific community's desire to exchange information regarding what they termed "archived" preprinted scientific papers. However, the protocol is content neutral and may be applied to any type of metadata, including metadata describing archives, manuscripts, photographs, or other cultural heritage materials, as well as journal articles, scientific papers, preprints, etc. As judged by the large amount of interest in OAI recently generated at the Joint Conference on Digital Libraries, the use of OAI is growing among many sectors of the digital library community. In particular the National Science Digital Library core architecture is built around OAI, meaning that the protocol will be well supported. The Andrew Mellon Foundation has also funded seven OAI projects. In the realm of cultural heritage materials, The University of Illinois, University of Michigan, Emory, and the University of Virginia are exploring the use of OAI with Mellon grants. The Digital Libraries Federation is also conducting an evaluation of OAI.³

Second, OAI is built upon a model of "metadata harvesting," (not distributed searching), and it is therefore much less technically complex than other digital library interoperability standards, such as Z39.50.⁴ In general, metadata is exchanged in a simple Dublin Core format, although interest groups may exchange metadata using any format which can be described using an XML Schema. (EAD could therefore be exchanged under OAI in its native format, if such a schema were developed to supplement the DTD which is maintained by the EAD Working Group.)

Third, OAI views the digital world into terms of data providers and service providers. Again, the distinction is described in detail elsewhere,⁵ but in general a data provider is any institution that agrees to provide its metadata in OAI format, using a simple Dublin Core/XML/RDF syntax, via

¹ <http://www.openarchives.org/> Accessed July 22, 2002.

² Lagoze, Carl, and Herbert Van de Sompel, The Open Archives Initiative: Building a Low-Barrier Interoperability Framework, *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries* (2001): 54-62. Available at <http://doi.acm.org/10.1145/379437.379449>;

³ <http://www.diglib.org/architectures/testbed.htm>

⁴ Michael L. Nelson, "Better Interoperability through the Open Archives Initiative," *The New Review of Information Networking* 7 (2001): 133.

⁵ Nelson, "Better Interoperability," 134.

an OAI "repository." If desired, OAI repositories may group their OAI records into sets to allow for selective harvesting based on topic, format, or other criteria. Establishing a repository means developing a relatively easy-to-implement software layer sitting on top of core services. Some "out of the box" applications and open source software to do this are available.⁶ Service providers, on the other hand, are those institutions or individuals which "harvest" this Dublin Core metadata from the data providers' repositories. Service providers often build value added services (such as search portals) which make use of the harvested metadata.

Fourth (and finally), OAI does not specify how, when, or under what conditions harvested metadata might be used. It simply specifies a common protocol under which the metadata might be exchanged. It is therefore application independent. Institutions may use the metadata in whatever way they wish (subject to service provider restrictions or copyright restrictions encoded in the rights field) to develop search portals or other services.

The remainder of this report provides a technical overview of the OAI-PMH, version 2.0 (stable).⁷ As noted in the specification, "OAI-PMH requests are expressed as HTTP requests [in a `get` or `post` method]. A typical implementation uses a standard Web server that is configured to dispatch OAI-PMH requests to the software handling these requests." The requests take the form of six "verbs" to which a service provider responds by sending OAI records. Appendix one lists the verbs and explains their usage.

OAI records may be static objects residing on an OAI server, or, more commonly, dynamically generated by scripts. Records sent via an OAI request consist of a unique identifier for the metadata record, a timestamp for when the metadata record was last modified, a list of sets to which the record belongs, and finally the metadata record itself. Additional metadata about the metadata record itself may be sent in an optional "about" container. Records may be grouped into sets, to allow for selective harvesting, and the timestamp also allows selective harvesting (i.e. so that only new or changed records might be harvested.)

All OAI responses must be in well-formed XML and must validate against the OAI schema. A sample OAI record is shown in appendix two. In order to allow for maximum interoperability, service providers must provide metadata in simple Dublin Core format, but providers may also use other metadata formats, provided that the metadata is described in an XML schema.

Chris Prom
Assistant University Archivist
University of Illinois
July 29, 2002

⁶ Several are provided by the University of Illinois OAI project. See. <http://oai.granger.uiuc.edu/ProviderTools/>

⁷ Full technical details are available at <http://www.openarchives.org/OAI/openarchivesprotocol.html>

Appendix 1: OAI Verbs

Verb	Purpose
GetRecord	used to retrieve an individual metadata record from a repository. Must specify the unique identifier and metadata format
Identify	used to retrieve information about a repository, such as its name, description, and technical details.
ListRecords	used to harvest records from a repository; may limit harvesting using sets and datestamp
ListIdentifiers	abbreviated form of ListRecords; returns only headers
ListMetadataFormats	used to retrieve a list of the metadata formats which the repository provides
ListSets	used to retrieve the set structure of a repository, useful for selective harvesting

Appendix 2: Sample OAI record

```
<record>
  <header>
    <identifier>oai:perseus:Perseus:text:1999.02.0084</identifier>
    <datestamp>2002-05-01T14:16:12Z</datestamp>
  </header>
  <metadata>
    <oai_dc:dc
      xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
      xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
        http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
      <dc:title>Opera Minora</dc:title>
      <dc:creator>Cornelius Tacitus</dc:creator>
      <dc:type>text</dc:type>
      <dc:source>Opera Minora. Cornelius Tacitus. Henry Furneaux.
        Clarendon Press. Oxford. 1900.</dc:source>
      <dc:language>latin</dc:language>
      <dc:identifier>http://www.perseus.tufts.edu/cgi-bin/ptext?
        doc=Perseus:text:1999.02.0084</dc:identifier>
    </oai_dc:dc>
  </metadata>
</record>
```