An Update from the OAI: Presentation Session & Discussion Session

Herbert Van de Sompel (LANL), Carl Lagoze (Cornell U), Simeon Warner (Cornell U), Michael L. Nelson (ODU)

OAI-rights

In collaboration with Project RoMEO, we formed an OAI-rights working group in September 2003 (*http://www.openarchives.org /news/oairightspress030929.html*) to investigate and develop the means of expressing rights information about metadata and resources within the OAI-PMH. Initial discussions lead to a decision to split the work into two parts: rights expressions about metadata, and rights expressions about resources.

A beta specification (*http://www.openarchives.org/OAI/2.0/guidelines-rights.htm*) for conveying rights expressions about metadata was announced on November 3, 2004 and the final specification is expected by the end of 2004. This specification uses Creative Commons licenses as examples but provides a flexible way to associate any rights expression with an OAI-PMH metadata record, either by-value (included in the XML) or by-reference (at a specified URL). Manifests of applicable rights expressions may be expressed at the repository and set levels to inform harvesting decisions. However, only expressions associated with each record are authoritative.

The expression of rights information about resources will be addressed in the next phase of this work. This introduces a number of additional issues and potential semantic uncertainties. Not least is the accurate identification of digital resources within the metadata, an issue also encountered in work to use the OAI-PMH to facilitate resource exchange. Work is also underway to create a best practice document for the use of ODRL rights expressions in OAI-PMH.

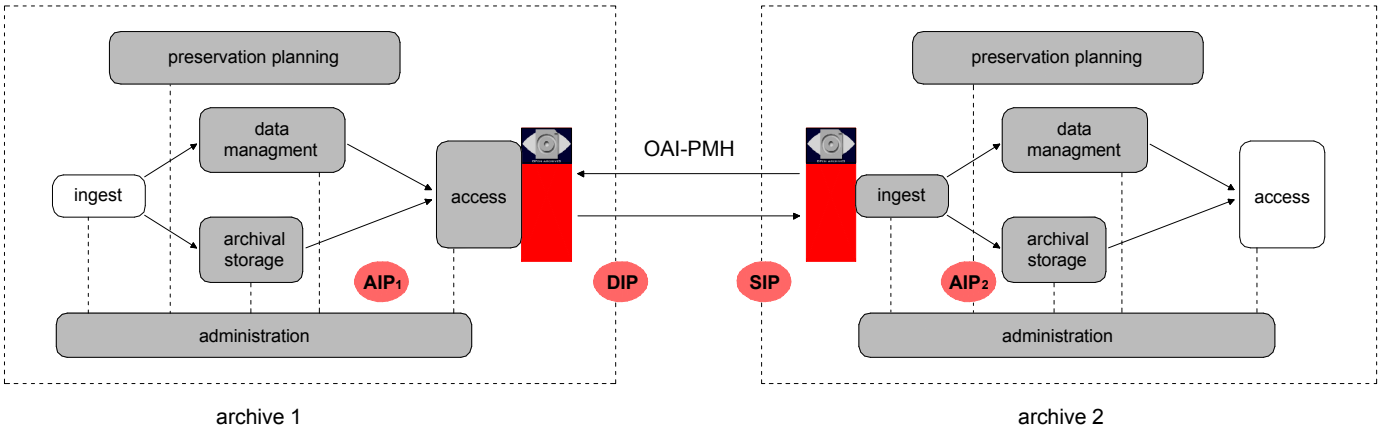
Resource Harvesting using the OAI-PMH

There is a growing need to make content, not only descriptive metadata, harvestable in an interoperable manner. So far, this need is manifested in two major use cases:

- Preservation: The need to recurrently transfer digital content from a data repository to one or more trusted digital repositories charged with storing and preserving safety copies of the content. The trusted digital repositories need a mechanism to permanently and automatically synchronize with the originating data repository.
- Discovery: The need to take discovery services beyond the level of descriptive metadata and to use the content itself in the creation of services. Examples frequently mentioned include search engines that make full-text from multiple data repositories searchable and citation indexing systems that extract references from the full-text content. Another compelling scenario is the inclusion of thumbnail versions of high-quality images from cultural heritage collections in services, and hence the need to be able to make these thumbnails harvestable in a protocol-based manner.

Current approaches to harvesting resources from OAI-PMH repositories use extensions or conventions outside of the OAI-PMH. These can work in localized applications, but not without a multitude of problems that may compromise the accuracy of the results of the content gathering process or the semantics of the metadata. While this lack of accuracy may be merely frustrating in some use cases, it clearly is unacceptable in others.

To bring resources into the OAI-PMH context, we use the concept of metadata as a *modeled representation* of a resource. In typical OAI-PMH implementations metadata is descriptive, with resources commonly being described by means of widely deployed metadata formats such as DC. Such metadata can be regarded a model of the resource itself according to the representational model underlying these metadata formats. Thus it is straightforward to accept the representation of a resource according to even more complex models – indeed, complex object models – to be metadata pertaining to the resource.



Ongoing projects have revealed the attractiveness of using complex object formats – such as MPEG-21 DIDL and METS – as metadata formats in the OAI-PMH. A resource can be modeled according to such complex object format, and the result is an XML document that is a modeled representation of the resource. These XML documents can be exposed by an OAI-PMH repository. Through the introduction of complex object formats into the OAI-PMH framework, the content gathering problem can be addressed through OAI-PMH harvesting of complex object XML documents. In such a solution all existing OAI-PMH concepts, such as sets and "about" containers, remain available. The notion of the OAI-PMH datestamp applied to complex objects yields an unambiguous technique to harvest resources. The solution can be fully specified within the boundaries of the OAI-PMH, and as such can be deployed on the basis of existing, widely deployed OAI-PMH tools. Deploying the solution boils down to specifying and implementing support for an additional metadata format.

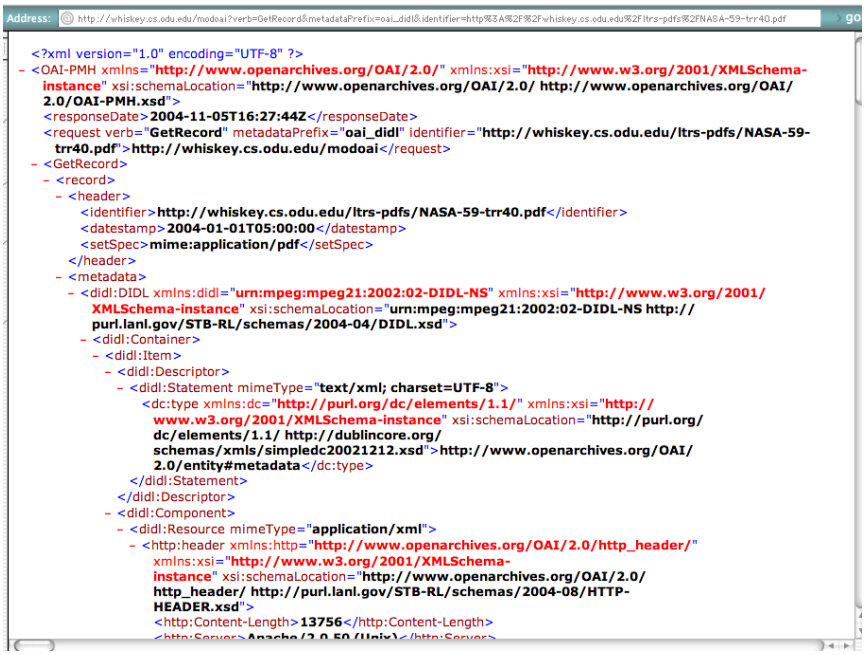
This approach introduces an archive export/ingest paradigm in which the concept of transferring content between archives is tackled in an application-independent and protocol-based manner, and at a more abstract level than is usually the case with mirroring solutions.

See forthcoming paper: Herbert Van de Sompel, Michael L. Nelson, Carl Lagoze, and Simeon Warner. 2004. "Resource Harvesting within the OAI-PMH Framework". D-Lib Magazine, Volume 10, Issue 12, December 2004. *http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html*

mod_oai

mod_oai is a joint project between Old Dominion University and the Los Alamos National Laboratory Research Laboratory that aims to bring OAI-PMH semantics to the web crawling community. An Apache module, mod_oai, is being developed that automatically responds to OAI-PMH requests on behalf of a web server. If Apache and mod_oai are installed at *http://www.foo.edu/*, then the baseURL for the server is *http://www.foo.edu/mod_oai*

While respecting the http access controls specified in httpd.conf, mod_oai provides 3 metadata formats in the OAI-PMH responses. Dublin Core is provided, but only administrative metadata such as file size and MIME type is included. A new metadata format is introduced, http_header, which contains all the http response headers that would have been returned if the resource had been obtained by a regular web crawler. In mod_oai, DC is a subset of the information returned in http_header. The third metadata format is oai_didl, which includes the metadata in the http_header format, as well as by-value, base64 encoded resource.



A mod_oai DIDL XML document

A number of subtle interpretations of the OAI-PMH data model are required. First, the URL of the resource serves as the OAI identifier. Second, the datestamp of the resource is the datestamp of all 3 metadata formats – there will never be a situation when a record in one metadata format has a different datestamp than another metadata format for the same identifier. Lastly, the set membership of item is based on the MIME type of resource.

Both the discovery and preservation use scenarios are applicable with mod_oai. For discovery, OAI-PMH offers incremental harvesting semantics with datestamp and MIME type as arguments. For preservation, mod_oai allows an entire website to be transformed into DIPs/SIPs and stored for later reconstitution. The http_header metadata, either by itself or included in the oai_didl metadata format, provides complete http header information about the resource as well; information that is otherwise not available in the standard OAI-PMH usage scenario.

mod_oai is funded by the Andrew Mellon Foundation. More information can be found at http://www.modoai.org/