mod_oai: An Apache Module for Metadata Harvesting

Michael L. Nelson [1], Herbert Van de Sompel [2], Xiaoming Liu[2],
Terry L. Harrison[1], and Nathan McFarland[2]

[1] Old Dominion University, Department of Computer Science, Norfolk VA 23508 USA
{mln, tharriso}@cs.odu.edu
[2] Los Alamos National Laboratory, Research Library, Los Alamos NM 87545 USA
{herbertv, liu_x, nmcfarl}@lanl.gov

Abstract. We describe mod_oai, an Apache 2.0 module that implements the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH is the de facto standard for metadata exchange in digital libraries and allows repositories to expose their contents in a structured, application-neutral format with semantics optimized for accurate incremental harvesting. mod_oai differs from other OAI-PMH implementations in that it optimizes harvesting web content by building OAI-PMH capability into the Apache server.

1 Introduction

There has been considerable attention given to increasing the efficiency of web crawlers through more accurate estimation of updates [1]. This problem arises from the fact that http does not support semantics of the form "what resources have changed since 2004-12-27?" Although syndication formats such as RSS are widely implemented, these formats are either in the process of standardization or optimized for syndicating web ephemera and not for accurate incremental harvesting. Within the digital library community, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is the de facto standard for metadata interchange.

We have developed an Apache module, mod_oai, 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 OAI-PMH requests 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 technical metadata such as file size and MIME type is included. We introduce a new metadata format, http_header, which contains all the http response headers that would have been returned if the resource had been obtained directly. The third metadata format, oai_didl, encodes the web resource with the MPEG-21 Digital Item Declaration Language (DIDL) [2]. This representation includes the metadata in the http_header format, as well as the web resource itself, either base64 encoded ("by-value"), as a URL ("by-reference"), or both. 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 a standard OAI-PMH usage scenario. The introduction of the oai_didl metadata format allows for the incremental harvesting of resources while remaining within the boundaries of the OAI-PMH [3].

A. Rauber et al. (Eds.): ECDL 2005, LNCS 3652, pp. 509 – 510, 2005.

A number of subtle interpretations of the OAI-PMH data model are made to achieve optimal functionality of mod_oai. First, the URL of the resource serves as the OAI-PMH identifier. Second, the last modified date of the resource is used as the OAI-PMH datestamp of all 3 metadata formats. As a result, all metadata for a given OAI-PMH identifier will share an OAI-PMH datestamp. Lastly, the set membership of item is based on the MIME type of resource.

There are two general classes of mod_oai use. The first is to issue only ListIdentifiers as a way of identifying new URLs to be added to a regular web crawler. In the ListIdentifiers scenario, mod_oai offers incremental harvesting semantics with datestamp and sets (i.e. MIME types) as arguments. The second scenario, is to issue ListRecords, which causes an entire website to be transformed into OAIS Archival Information Packages (AIPs) and stored for later reconstitution.

2 Conclusions

mod_oai currently works for static files only; we are adding support for dynamic pages in a future release. mod_oai is not intended to replace existing OAI-PMH repositories, but rather to bring the OAI-PMH semantics of incremental harvesting based on datestamps and sets to general web servers. A full architectural discussion and performance evaluation of mod_oai can be found in [4] and more information can be found at http://www.modoai.org/.

Acknowledgements

mod_oai is supported by the Andrew Mellon Foundation.

References

[1] Cho, J., Garcia-Molina, H. Estimating Frequency of Change. *ACM Transactions on Internet Technology, 3*, 3, 2003, 256-290.

[2] Bekaert, J. Hochstenbach, P., Van de Sompel, H. Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library. *D-Lib Magazine, 9*, 11, 2003.

[3] Van de Sompel, H., Nelson, M. L., Lagoze, C. L., Warner, S. Resource Harvesting within the OAI-PMH Framework, *D-Lib Magazine, 10*, 12, 2004.

[4] Nelson, M. L., Van de Sompel, H., Liu, X., Harrison, T. L., McFarland, N. mod_oai: An Apache Module for Metadata Harvesting, arXiv Technical Report cs.DL/0503069, 2005.