11-22-2006

Providing an OAI-PMH Interface to the Storage Resource Broker With OAISRB

Michael Witt

Purdue University, mwitt@purdue.edu

Jigar Kadakia

Follow this and additional works at: http://docs.lib.purdue.edu/lib_research

Witt, Michael and Kadakia, Jigar, "Providing an OAI-PMH Interface to the Storage Resource Broker With OAISRB" (2006). *Libraries Research Publications*. Paper 78.

http://docs.lib.purdue.edu/lib_research/78

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

ABSTRACT

The OAISRB software was developed by Michael Witt and Jigar Kadakia of the Purdue University Libraries to provide an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interface to expose metadata from digital objects residing on the Storage Resource Broker (SRB) to OAI service providers. By harvesting metadata from the SRB, grid resources such as large research datasets from computational sciences can be represented alongside more conventional digital library objects such as e-prints and digitized archival collections.

OAI-PMH

The OAI-PMH defines a protocol that allows a service provider to harvest metadata from a compliant repository, also known as a data provider. OAI-PMH requests are constructed using six different verbs:

- Identify
- ListSets
- ListMetadataFormats
- ListIdentifiers
- ListRecords
- GetRecord

STORAGE RESOURCE BROKER

The SRB functions as a data grid, allowing multiple physical storage units to be aggregated and represented as a single logical storage unit. An assortment of client programs facilitate access to files on the SRB and perform functions such as filesystem virtualization. A metadata catalog (MCAT) service keeps track of the relationships among users, files, and storage units as well as supporting different levels of description that can be queried by users to perform various actions to access and use objects.

OAISRB

OAISRB essentially serves as middleware, converting OAI-PMH requests into SRB queries and returning the results from the MCAT in an OAI-compliant manner. It is implemented as a Java servlet that runs as a web service, for example, on an Apache Tomcat server.

Figure 1. OAISRB Flow Diagram

Its external interface listens to TCP port 80 for an OAI-PMH request, which arrives via HTTP. The response to an Identify request provides basic information about the repository such as its name, administrator email address, and protocol version from a configuration file, `oaihandler.properties`. An OAI set is analogous to a collection in SRB, and OAISRB responds to the ListSets verb with the list of collections that are statically defined in `oaihandler.properties`. Because the current version of OAISRB only supports the `oai_dc` (Dublin Core) schema, the response to the ListMetadataFormats request is hard-coded. For all other requests, OAISRB extracts the verbs and other parameters from the URL, constructs a query, connects to SRB using the internal interface, and passes the query to MCAT. The results are received from MCAT as an array that is formatted into XML and validated before being returned to external interface and the harvester. The GetRecord verb retrieves an individual metadata record. The ListRecords verb harvests a group of metadata records from a collection that meet the parameters of a query. The ListIdentifiers verb returns a list of abbreviated metadata records. Both OAISRB interfaces use open source software; the external OAI-PMH interface is based on OAI-Cat, developed by Jeff Young at the Online Computer Library Center (OCLC), and the internal SRB interface uses the Jargon Java client API that was developed by the San Diego Supercomputer Center (SDSC).

The `oaihandler.properties` file also defines the base URL and port for the OAI-PMH interface, the SRB connection parameters (e.g., username, password, zone), and information about the server and collections. A crosswalk feature converts from one metadata format to `oai_dc` on-the-fly after the recordset is returned from MCAT. Custom field mappings can be defined for each collection. Lastly, a logging function records information related to performance and errors for debugging purposes.

OAISRB (con't)

OAISRB was approved for open source distribution by Purdue's Office of Technology Commercialization (request #64660) on August 30, 2006. It was tested in collaboration with the Rosen Center for Advanced Computing, a division of Information Technology at Purdue, on the TeraGrid. Feedback from developers and users will help identify bugs to fix and features to add in new releases.

FUTURE WORK

Future work on OAISRB includes adding support for metadata formats other than Dublin Core and the ability to reference an XML schema on the web to perform dynamic crosswalks. More flexibility and granularity in defining objects in collections could be provided by adding pattern-matching capability. For example, a service provider may only be interested in graphic files in a collection and apply a regex filter (`\.tif$`) in its collection definition. Once standards are defined, OAISRB could be extended to allow harvesting of objects as well as metadata. In the case of large datasets, it is impractical to ingest them into a single repository. A new set of middleware tools can be developed to use existing protocols to provide interoperability to access data wherever it resides (e.g., on the grid, on a researcher's desktop, etc.).

LINKS

Purdue University Libraries, OAISRB:
<http://www.lib.purdue.edu/research/oaisrb>

The OAI-PMH Version 2.0:
<http://www.openarchives.org/OAI/openarchives/protocol.htm>

San Diego Supercomputer Center, SRB:
<http://www.sdsc.edu/srb>

Online Computer Library Center, OAI-Cat:
<http://oclc.org/research/software/oai/cat.htm>

San Diego Supercomputer Center, Jargon:
<http://www.sdsc.edu/srb/index.php/Jargon>

Providing an OAI-PMH Interface to the Storage Resource Broker With OAISRB

Michael Witt & Jigar Kadakia, Purdue University Libraries, West Lafayette, Indiana, USA.