OAI-PMH Architecture for the NASA Langley Research Center Atmospheric Science Data Center

Churngwei Chu[1], Walter E. Baskin[1], Juliet Z. Pao[1], and Michael L. Nelson[2]

[1]NASA Langley Research Center, Hampton VA, USA
[2]Old Dominion University, Norfolk VA, USA
{c.chu, w.e.baskin, j.z.pao}@larc.nasa.gov, mln@cs.odu.edu

Abstract. We present the architectural decisions involved in adding an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) interface to the NASA Langley Research Center Atmospheric Science Data Center (ASDC). We review four possible implementation strategies and discuss the implications of our choice. The ASDC differs from most OAI-PMH implementations because of its complex data model, large size (1.3 petabytes) of its Earth Science data holdings and its rate of data acquisition (>20 terabytes / month).

1 Introduction

The National Aeronautics and Space Administration (NASA) Langley Research Center Atmospheric Science Data Center (ASDC) [1] supports 42 science projects with over 1700 data sets and 2M data granules in a combination of 1.3 petabytes of online and nearline storage. ASDC is one of 8 NASA Distributed, Active Archive Centers (DAACs) in the U.S. that provide curation of federally-funded Earth Science data. The DAACs are arranged by discipline; ASDC's data sets involve radiation budget, clouds, aerosols and tropospheric chemistry. These data sets were produced to increase academic understanding of the natural and anthropogenic perturbations that influence the global climate change. In addition to archiving, distributing and processing data, ASDC also distributes metadata to other trading partners. To increase visibility of its holdings and facilitate more automated interchange with data partners, a pilot project was implemented for providing an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [2] interface for the ASDC.

2 OAI-PMH Architecture Options

The largest difficulty faced was mapping the ASDC data model into the relatively simple OAI-PMH data model of resource/item/record. Although the data model is fully discussed in [2], we present a highly summarized review here. Resources are the objects of interest and exist outside of the OAI-PMH data model; they are the "stuff" the metadata is "about". Items are in the data model and have a unique OAI identifier; they represent all the metadata records that describe a resource. Set level information is attached to the item. Items have 1 or more records. Records are

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 524 – 527, 2006.

metadata in a particular format ("metadataPrefix") and inherit their identifier from the item. Records are the unit of discourse in OAI-PMH transactions.

Conventional bibliographic applications have simple data models. For example, the resource is a book, the item uses an ISBN for its unique identifier, and the metadata is available natively in MARC as well as a Dublin Core (DC) [3] subset intended for general resource discovery. ASDC's resources are not that uniform; they have a hierarchy of "project/collection/granule" (in descending order of magnitude and rate of change) to specify provenance and versioning. For example, the International Satellite Cloud Climatology Project (ISCCP) has 9 collections, one of which (ISCCP_B3_NAT) has more than 0.5M granules. Project level metadata describes the project's purpose, instruments used and spatial/temporal coverage. Collection level metadata includes descriptions for scientific keywords and platforms such as satellites and airplanes. Granule level metadata describes only spatial/temporal coverage. In the ASDC, there are approximately 1000 record/collection metadata records and 2 million granule metadata records. Four main options were considered for mapping projects/collections (P/C) and granules into OAI-PMH:

1. Three separate repositories and corresponding baseURLs are created. The repositories are separated by how they define items: projects, collections or granules. This option maintains uniformity within each repository, but each repository must be harvested separately to acquire all the data.
2. Granules are defined as items and a single repository/baseURL is used. Separate metadataPrefixes are used to convey project and collection level metadata. This option promotes granules to the primary focus of the repository, but would result in significant duplication of project and collection metadata records because of the imbalance between granules and projects/collections: a single project metadata record is likely to be associated with thousands of granules.
3. Similar to #2, granules are defined as items and a single repository/baseURL is used. But in this case, project and collection metadata exists outside of the OAI-PMH framework. For example, since there are so few project and collection records and are likely to change very infrequently, they could be simply referenced as external URLs from the granule metadata records. This option has the advantage of uniformity (all items are granules), but not all information is directly OAI-PMH accessible.
4. The final option considered also uses a single repository/baseURL, but items are projects, collections and granules. The nature of a metadata record can be inferred from its metadataPrefix, identifier and set information. This option does not have a uniform concept of items, but the items are highly interrelated and all ASDC metadata records are accessible from a single baseURL.

After careful consideration, we decided to implement option #4. The records are differentiated by their metadataPrefix: project and collection records are available in DC and Directory Interchange Format (DIF) [4], whereas granules are only available in DIF. Although this is not compliant with the OAI-PMH since there is not a DC representation for every record, this approach was chosen because we considered the metadataPrefix as an indicator of intention. DIF is a highly specialized science data format harvested and likely to be understood only by known trading partners active in Earth Science research. We interpret a request for DC as an indication of non-expert usage, and thus harvesters requesting DC receive only the project and collection based records, which are suitable for cross-domain service providers.

The project/collection/granule hierarchy is incorporated in both identifiers and sets. This identifier specifies a granule 445025 in the collection ISCCP_B3_NAT (124th revision), and the set value describes the membership of the granule.

Identifier: oai:asdc.larc.nasa.gov:ISCCP:ISCCP_B3_NAT_124:445025
Set: info:asdc.larc.nasa.gov:ISCCP:ISCCP_B3_NAT

Similarly, the corresponding collection and project identifiers and set values would be, respectively:

Identifier: oai:asdc.larc.nasa.gov:ISCCP:ISCCP_B3_NAT
Set: info:asdc.larc.nasa.gov:ISCCP:ISCCP_B3_NAT

Identifier: oai:asdc.larc.nasa.gov:ISCCP
Set: info:asdc.larc.nasa.gov:ISCCP

Notice that for projects and collections, the set values match the identifiers. It also worth noting that in OAI-PMH, the ":" character is recursive. So requests of the form:

?verb=ListRecords&metadataPrefix=DIF&set= info:asdc.larc.nasa.gov:ISCCP

Would return all project, collection and granules records in the DIF metadata format. This request specifies DC as the metadataPrefix, so it would return metadata for projects and collections, but not metadata:

?verb=ListRecords&metadataPrefix=oai_dc&set= info:asdc.larc.nasa.gov:ISCCP

This request would return only the project metadata in DIF:

?verb=GetRecord&metadataPrefix=DIF&identifier=oai:asdc.larc.nasa.gov:ISCCP

3 Future Work and Conclusions

There are several implications for adding an OAI-PMH interface to the ASDC. First, it will result in much greater exposure of ASDC collection. We plan to expose the project and collection records to the NASA Technical Report Server (NTRS) [5]. NTRS is a "one-stop shop" for NASA authored publications, and since NTRS already uses OAI-PMH to harvest from other NASA institutional repositories, adding the ASDC collection will be easy. Furthermore, since Google supports OAI-PMH, it will be easy to expose project and collection metadata records to Google as well. Increased coverage in additional services such as NTRS and Google is an example of the "inverted repository" model in which the data objects themselves are exposed and harvested by many services and point back to their main home page. This is contrast to the more conventional model where home pages (such as [1]) are the only resource indexed and the data objects are discovered only through the services at the home page.

Our future plans also include moving from metadata harvesting to actual resource harvesting. This involves bringing the resource into the realm of the OAI-PMH data model by encoding in a complex object format (e.g., MPEG-21 Digital Item Declaration Language (DIDL) or Metadata Encoding and Transmission Standard (METS)) and treating the resulting object as a metadata format. This approach has been shown to allow for accurate repository synchronization using off the shelf harvester software [6].

ASDC is in the process of testing the repository implementation, currently based on OAICat [7], and will make the URL initially available to selected trading partners. We are working with potential partners to establish OAI-PMH use in the Earth Science community. Our implementation currently violates the letter of the OAI-PMH specification by not returning DC records for all items. Since the conversion of DIF-to-DC for granules is very lossy, we adopted this approach to prevent the likelihood of unwary harvesters acquiring millions of DC records with little value. If we choose to become technically compliant in the future, we could not expose granule identifiers for harvesters that do not provide some well-known user id and password (e.g., "earth" and "science").

We plan to use OAI-PMH to facilitate interchange between ASDC and other partners, to encourage the development of new, specialized services based on Earth Science data, and to increase the exposure of our Earth Science data holdings through increased search engine and service provider coverage.

References

1. Atmospheric Science Data Center. http://eosweb.larc.nasa.gov/
2. C. Lagoze, H. Van de Sompel, M. L. Nelson and S. Warner. *The Open Archives Initiative Protocol for Metadata Harvesting Version 2.0.* http://www.openarchives.org/OAI/openarchivesprotocol.html
3. S. Weibel, J. Kunze, C. Lagoze and M. Wolf. *Dublin Core Metadata for Resource Discovery.* Internet RFC-2413, 1998.
4. *Directory Interchange Format (DIF) Writer's Guide. Version 9.4.* 2005. http://gcmd.nasa.gov/User/difguide/
5. M. L. Nelson, J. R. Calhoun, and C. E. Mackey. "The OAI-PMH NASA Technical Report Server." *Proceedings of JCDL 2004*, (Tucson, Arizona; June 2004): 400.
6. J. Bekaert and H. Van de Sompel. "A Standards Based Solution for the Accurate Transfer of Digital Assets." *D-Lib Magazine* **11**(6) (June 2005).
7. OAICat.. http://www.oclc.org/research/software/oai/cat.htm