The Dienst-OAI Gateway

Terry L. Harrison, Michael L. Nelson, Mohammad Zubair
Old Dominion University
Department of Computer Science
Norfolk VA, 23529 USA
+1 757 683 4817
{tharriso,mln,zubair}@cs.odu.edu

ABSTRACT

Though the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) is becoming the defacto standard for digital libraries, some of its predecessors are still in use. Although a limited number of Dienst repositories continue to be populated, others are precariously unsupported. The Dienst Open Archive Gateway (DOG) is a gateway between the OAI-PMH and the Dienst (version 4.1) protocol. DOG allows OAI-PMH harvesters to extract metadata records (in RFC-1807 or Dublin Core) from the Dienst servers.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: Systems issues, Standards.

General Terms

Design, Reliability, Standardization.

Keywords

OAI-PMH, Dienst, Gateways, Metadata, Preservation

1. INTRODUCTION

Dienst was once a popular, rich digital library (DL) protocol, comprised of over 30 verbs [1]. The Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) [2] represents some of the lessons learned during the deployment of Dienst, including a significantly decreased scope. As such, the OAI-PMH has only 6 verbs and introduces the division of responsibility for harvesters and repositories.

While many of the 100+ institutions that once used Dienst to participate in the Networked Computer Science Technical Reference Library (NCSTRL) have transitioned to the OAI-PMH [3], not all have. Some Dienst repositories continue to be populated and supported, while others are no longer being updated, and in the worst case, no longer being maintained. The Dienst OAI-PMH Gateway (DOG) was created to allow OAI-PMH harvesters to extract data from existing and at-risk Dienst repositories (Dienst versions 4.1.x). DOG was initially created to allow for the harvesting of the ICASE Dienst repository for the OAI-PMH enabled NASA Technical Report Server [4], but the general nature of DOG allows it to be used for any Dienst repository, including those current DL projects based on Dienst, such as OpenDLib [5].

JCDL03, May 27-31, 2003, Houston, TX.

2. OAI-PMH VERB IMPLEMENTATION

DOG is a Java servlet and is available both as a demonstration service at ODU and as a tar file for local implementation (dlib.cs.odu.edu). DOG allows a specified Dienst repository to appear as a normal baseURL to a harvester:

http://128.82.7.113:5187/dog/servlet/dataprovider/ (DOG)
dienst.iei.pi.cnr.itSLASHDienstSLASH/ (Dienst)
?verb=ListRecords&metadataPrefix=oai_dc (OAI-PMH)

The URL of the Dienst server is imbedded the URL. DOG parses out this URL and issues the appropriate Dienst verbs to the Dienst repository (Table 1 has the OAI-PMH to Dienst mapping), parsing the RFC-1807 plain text results and returning an OAI-PMH formatted response. Metadata can be returned as either RFC-1807 or Dublin Core (DC) The mapping from RFC-1807 to DC is described in [6] (as well as more detailed implementation notes), and the mapping is modifiable if a local installation is chosen.

Table 1. OAI-PMH to Dienst Mapping

OAI-PMH Verb	Dienst 4.1 Service/Verb
Identify	Repository/2.0/List-Contents
ListIdentifiers	Index/2.0/List-Contents
ListRecords	Index/2.0/List-Contents
GetRecord	Index/2.0/Bibliography/*handle*
ListSets	no calls made (response hardcoded)
ListMetadataFormats	no calls made (response hardcoded) or: Index/2.0/Bibliography/handle if an identifier argument is supplied

2.1 Identify

The Identify verb extracts much of its required content from the hard-coded variables or parsed from the requestURL, with the exceptions of earliestDatestamp and sampleIdentifier. Worthy of note is the earliestDatestamp element, which guarantees to harvesters a lower bound on record datestamps. To discover the earliest date is an expensive operation and requires the parsing of all records for a given Dienst repository. To reduce this overhead, the resulting value is cached so that the operation need only be performed once per repository. As with all datestamps, RFC-1807 format (May 31, 2003) is converted to ISO-8061 format (2003-05-31). Values for the sampleIdentifier element are intentionally not cached, as the Dienst call needed to extract this from the repository also provides a status check on it, returning an http exception if the archive is no longer available. Some escaped character sequences had to be created to accommodate characters in some Dienst repository identifiers that did not conform with

OAI-PMH character restrictions for the sampleIdentifier element content. One such case is the underscore in repository identifiers such as ncstrl.mit_ai. To resolve this, the "_" char is replaced with "UNDERSCORE".

2.2 ListSets

Since the Dienst collection service was never widely adopted, most Dienst implementations have no equivalent to the OAI-PMH concept of sets. Thus, the response to a ListSets request is hardcoded to be the "noSetHierarchy" error.

2.3 ListMetadataFormats

Since Dienst metadata is in RFC-1807, and OAI-PMH requires support of DC, DOG supports both formats. ListMetadataFormats without an identifier argument will show support for both formats. Requests with an identifier argument require issuing a Dienst verb to verify the metadata for the given identifier exists.

2.4 GetRecord

Getting a single record from the Dienst archive is relatively simple, as the request maps well into Dienst protocol. The OAI-PMH identifier is formatted as a Dienst handle and appended into a Dienst record request. The record is returned in RFC-1807 as plain text, which DOG parses and returns as an XML formatted RFC-1807 or DC record.

2.5 ListRecords & ListIdentifiers

ListIdentifiers and ListRecords utilize the same Dienst calls and only differ in the amount of metadata returned for each record. DOG issues an Index/2.0/List-Contents verb that gives it an RFC-1807 metadata dump of all the records. Initially the parameter "file-after" was utilized to implement an OAI-PMH "from" argument, but since many Dienst archives had not implemented this argument, its use threatened to corrupt result sets. Once the metadata dump is received, DOG parses it, extracting any records that meet any given "from" or "until" arguments. Interestingly, by its ability to handle "until" parameters, DOG adds utility that did not previously exist within the Dienst protocol. DOG does not cache the results of the metadata dump, which permits it to provide realtime harvesting.

3. ISSUES & ERROR HANDLING

OAI-PMH 2.0 adds many requirements for handling OAI-PMH errors, which are distinct from http errors. DOG also is careful to distinguish http errors and errors to be handled by the OAI-PMH. For example, an http 404 (File Not Found) error could be the result of the archive being down, a bad identifier in the request, or the file requested by a valid identifier does not exist. Invalid syntax issues like a malformed identifier are easily resolved by pattern matching to OAI-PMH regular expressions provided in its schema. Should the syntax be deemed valid, then it must be determined if the problem is a communications fault or a non-existent file. To resolve these uncertainties, a subsequent call is made ("Repository/2.0/List-Contents/"). If a connection is made, an OAI-PMH error "idDoesNotExist" is returned instead of an http 404 error.

A different development issue was the hiding of the Dienst archive URL into the request URL to DOG. Traditionally this might be accomplished through the use of standard escape character '"%2F" for the forward slashes, but a documented feature in Tomcat 4.0.4 which does not handle these escape characters before the method call, necessitated the use of the custom "SLASH" chars to escape instances of "/" in the Dienst URL. In anticipation of the resolution of this issue, DOG handles escape characters "%2f", "%2F" and "SLASH".

Another issue was the metadata inconsistency in the use of the RFC-1807 fields "ID::" and "HANDLE::". Since handles are unique permanent identifiers of the form HANDLE:: <repository>/<identifier> they seem the most logical choice for <identifier>. However, because they are optional, they may be missing altogether from an archive (i.e.: dienst.iei.pi.cnr.it). In this case, DOG uses data from the mandatory ID:: field, which is typically the same data. This also works when searching for a record. For example, when a GetRecord request is made and DOG is comparing Dienst record metadata for a match, should there not be a HANDLE:: field, then the ID:: field is used instead. Due to RFC-1807 syntax, the ID:: field data will contain two forward slashes (ID:: <publisher-id>//<free-text>) instead of one as in the HANDLE:: field. To avoid confusion the DOG response represents the <identifier> using only one slash. DOG can resolve this, should it need to match the ID:: field later. Additionally, DOG will handle requests that mistakenly use two slashes in the identifier.

4. CONCLUSIONS

While the previous NCSTRL transition project provided a specific strategy for OAI-PMH conversion, DOG provides on-demand, general OAI-PMH compatibiltiy for any Dienst archive. DOG has been exhaustively tested with the Repository Explorer and has been used to harvest ICASE Dienst repositories. Sadly, ICASE no longer exists at NASA and it's Dienst repositories have been shut down. This underscores the importance of providing protocol gateways with an eye toward long-term preservation.

5. REFERENCES

[1] Davis, J., and Lagoze, C. NCSTRL: design and deployment of a globally distributed digital library. Journal of the American Society for Information Science, 51(3), 2000, 273-280.

[2] Van de Sompel, H., and Lagoze, C. Notes from the interoperability front: A progress report on the Open Archives Initiative. in Proceedings of ECDL 2002 (Rome, Italy, September 2002), 144-157.

[3] Anan, H., Liu, X., Maly, K., Nelson, M., Zubair, M., French, J., Fox, E., and Shivakumar, P. Preservation and transition of NCSTRL using an OAI-based architecture. in Proceedings of JCDL 2002 (Portland OR, July 2002), 181-182.

[4] Nelson, M., Rocker, J., and Harrison, T. OAI and NASA scientific and technical information. Library Hi-Tech, 21(2), 2003.

[5] Castelli, D., and Pagano, P. OpenDLib: A digital library service system. in Proceedings of ECDL 2002 (Rome, Italy, September 2002), 292-308

[6] Harrison, T., Nelson, M., and Zubair, M. The Dienst-OAI gateway: A preservation gateway for a legacy protocol. ODU CS TR 2003-01, Feb. 2003.