# DP9: An OAI Gateway Service for Web Crawlers

**Xiaoming Liu, Kurt Maly and Mohammad Zubair**
Computer Science Department
Old Dominion University
Norfolk, Virginia USA
{liu_x,maly,zubair}@cs.odu.edu

**Michael L. Nelson**
NASA Langley Research Center
Hampton, Virginia USA
m.l.nelson@larc.nasa.gov

### ABSTRACT

Many libraries and databases are closed to general-purpose Web crawlers, and they expose their content only through their own search engines. At the same time many researchers attempt to locate technical papers through general-purpose Web search engines. DP9 is an open source gateway service that allows general search engines, (e.g. Google, Inktomi) to index OAI-compliant archives. DP9 does this by providing consistent URLs for repository records, and converting them to OAI queries against the appropriate repository when the URL is requested. This allows search engines that do not support the OAI protocol to index the "deep Web" contained within OAI compliant repositories.

## Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries - *Systems issues.*

## General Terms

Design, Experimentation

### Keywords

Open Archives Initiative, deep Web, gateway service

### INTRODUCTION

Many Digital Libraries (DLs) and databases are closed to general-purpose Web crawlers. The "deep Web" or "invisible Web" refers to vast repositories of content, such as documents in online databases, that general-purpose Web crawlers cannot reach. The deep Web content is estimated at 500 times that of the surface Web, yet has remained mostly untapped due to the limitations of traditional search engines [1]. On the other hand, many researchers use general-purpose search engines to locate research papers more frequently than they use specific DLs. A study about ResearchIndex query logs showed that only about 6% of the total number of sessions started with a search query from ResearchIndex itself, the majority of the sessions have been initiated by linking through a search engine such as Altavista or Google [2]. We are not claiming that an Internet search engine is the best way to discover information from DLs. The precision of search engines not aware of metadata is generally not as good as that of metadata-aware search engines. But Internet search engines do provide an efficient way to discover research information. The gateway service makes the contents of previously closed DLs visible to search engines, which can provide a search interface across DLs and hide multiple query languages and search interfaces to end users.

The Open Archives Initiative (OAI) is an effort to address interoperability issues among many existing and independent DLs [3]. An OAI-compliant Data Provider (see [3] for OAI notation and terms) exposes metadata through an HTTP/XML based protocol. The OAI has already attracted many adopters, and more than 70 archives with well over 1 million records have been harvested by Arc, a cross-archive search service [4].

DP9 (http://dlib.cs.odu.edu/dp9) is an open source gateway service that allows general search engines, (e.g. Google, Inktomi) to index OAI-compliant archives. DP9 does this by providing consistent URLs for repository records, and converting them to OAI queries against the appropriate repository when the URL is requested. This allows search engines that do not support the OAI protocol to index the "deep Web" contained within OAI-compliant repositories.

### ARCHITECTURE

Internet search engines cannot index OAI collections since they are not aware of the OAI protocol. We introduce an OAI gateway architecture to address this problem. Typically, a Web crawler indexes a Web site by starting with a base HTML page and by following the links on this page to retrieve deeper pages on the Web site. To support this for an OAI collection, which only responds to OAI requests (and only in XML), we begin by dynamically creating a starting HTML page for an OAI collection. Although an individual data provider may have its own mechanism of creating this page, DP9 provides a general solution that fits all OAI compliant data providers. In DP9, the starting page for a data provider would be constructed by issuing an OAI ListIdentifier request and translating the response into a HTML format containing a series of links. A link on this HTML page, when invoked, would result in

*JCDL'02*, July 13-17, 2002, Portland, Oregon, USA.

another OAI GetRecord request for a specific identifier. Again, the response for such a request would be translated into an HTML page with appropriate links. In other words, an HTML page presented to a Web crawler is a result of an OAI request, and the links on the Web page lead to other OAI requests. DP9 supports the resumption token and HTTP 503 status code "retry-after" and thus provides a basic flow control for large data providers. Note that the flexibility in the OAI protocol allows different ways of constructing the HTML pages to expose an OAI collection. For example, the starting page could have been constructed using the OAI request ListRecords. The sequence of OAI requests we have used in our design was driven by what would be useful for crawlers.

DP9 uses links on Web pages that have the following format: http://{hostname}/dp9/getrecord/{MetadataFormat}/{OAI_ID}. An example is:

http://arc.cs.odu.edu:8080/dp9/getrecord/oai_dc/oai:NACA:1917:naca-report-10

DP9 creates a series of ListIdentifiers pages for each archive with links to all individual records. These URLs are static and will be only activated when a HTTP request is received. DP9 provides an entry page and if a Web crawler finds this entry page, it may follow the links on this page and send requests to DP9. DP9 will then forward the request to corresponding OAI data providers and process the returned XML records. Depending on the depth a crawler follows, it can index all records in an OAI data provider.

DP9 consists of three main components (Figure 1), an URL wrapper, an OAI handler and an XSLT processor. The URL wrapper accepts the persistent URL and calls internal JSP/Servlet applications. The OAI handler issues OAI requests on behalf of a Web crawler. The XSLT processor transforms the XML content returned by the OAI archive to an HTML format suitable for a Web crawler. XSLT allows DP9 to support any XML metadata format simply by adding an XSL file. DP9 is based on Tomcat/Xalan/Xtag technology from Apache.

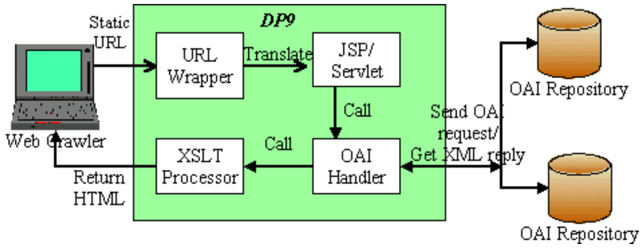

**Figure 1. DP9 Architecture**

Some crawlers use the HTML meta tags to index a Web pages; so in addition to creating the user friendly HTML page, DP9 also maps Dublin Core metadata to corresponding HTML meta tags. For pages that are designed exclusively for robots navigation, a noindex robots meta tag is used.

## RESULTS AND FUTURE WORK

We have collected 70 OAI repositories with well over one million records. Considering Parallel Metadata Sets [3] are supported by OAI leading to more references, potentially several millions of pages could be indexed by Web search engines. With DP9 now being deployed, thousands of documents in OAI collections have been indexed by search engine such as Inktomi and Google. Web logs show that more than 1000 queries are issued from popular Web search engines each day.

DP9 is a gateway service, it does not cache the OAI records and only forwards requests to corresponding OAI data providers. This insures DP9's records are always up-to-date; however, its quality of service is highly dependent on the availability of OAI data providers. On the other hand, an aggressive crawler using DP9 can rapidly send requests without regard for the load they are placing on the data providers. The robot exclusion protocol [5] at the data provider site will not be observed because the requests come from DP9, an OAI service provider. We are studying the possibility of using an OAI mirror/caching mechanism such as OAI Aggregator [6] and HTTP throttle software to relieve the overhead on data providers.

DP9 also provides an easy way to build services for OAI-compliant repositories. Indexing tools such as htdig and GreenStone are designed to index websites, they could be used to build searching services for OAI collections with DP9 support.

## ACKNOWLEDGEMENTS

We thank Rick Luce, Tim Brody, Herbert Van de Sompel, Andy Powell, Thorsten Schwander, Larry Stone, Walter Underwood, for valuable feedback and recommendations. This work was supported by the ODU DL group and the LANL Research Library.

## REFERENCES

1. M. K. Bergman. The Deep Web: Surfacing Hidden Value. Journal of Electronic Publishing, 7(1), 2001.
2. M. Mahoui and S. J. Cunningham. Search Behavior in a Research-Oriented Digital Library. Proceedings of ECDL2001, Darmstadt, Germany, September 4-9, 2001, LNCS 2163, pp. 13-24.
3. C. Lagoze and H. Van de Sompel. The Open Archives Initiative: Building a low-barrier interoperability framework. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, Roanoke VA, June 24-28, 2001, pp. 54-62.
4. X. Liu, K. Maly, M. Zubair, and M. L. Nelson. Arc - An OAI Service Provider for Digital Library Federation, D-Lib Magazine 7(4), April 2001.
5. M. Koster. The Web Robots Page. Available at http://info.webcrawler.com/mak/projects/robots/robots.html
6. OAI Perl. Available at http://oai-perl.sourceforge.net/