

Old Dominion University ODU Digital Commons

Computer Science Faculty Publications

Computer Science

2005


Lessons Learned with Arc, an OAI-PMH Service Provider

Xiaoming Liu

Kurt Maly
Old Dominion University

Michael L. Nelson
Old Dominion University

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_fac_pubs

 Part of the [Computer Sciences Commons](#), and the [Digital Communications and Networking Commons](#)

Repository Citation

Liu, Xiaoming; Maly, Kurt; and Nelson, Michael L., "Lessons Learned with Arc, an OAI-PMH Service Provider" (2005). *Computer Science Faculty Publications*. 25.
https://digitalcommons.odu.edu/computerscience_fac_pubs/25

Original Publication Citation

Liu, X., Maly, K., & Nelson, M.L. (2005). Lessons learned with arc, an OAI-PMH service provider. *Library Trends*, 53(4), 590-603.

This Article is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Faculty Publications by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

Lessons Learned with Arc, an OAI-PMH Service Provider

XIAOMING LIU, KURT MALY, MICHAEL L. NELSON, AND
MOHAMMAD ZUBAIR

ABSTRACT

Web-based digital libraries have historically been built in isolation utilizing different technologies, protocols, and metadata. These differences hindered the development of digital library services that enable users to discover information from multiple libraries through a single unified interface. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a major, international effort to address technical interoperability among distributed repositories. Arc debuted in 2000 as the first end-user OAI-PMH service provider. Since that time, Arc has grown to include nearly 7,000,000 metadata records. Arc has been deployed in a number of environments and has served as the basis for many other OAI-PMH projects, including Archon, Kepler, NCSTRL, and DP9. In this article we review the history of OAI-PMH and Arc, as well as some of the lessons learned while developing Arc and related OAI-PMH services.

Interoperability is one of the significant research problems in the field of digital libraries (DLs) (Lynch & Garcia-Molina, 1995). The inability to federate, filter, and provide value-added services on remote content limits DLs to covering only local holdings. The Open Archive Initiative (OAI) is a major, international effort to address technical interoperability and facilitate discovery of content among distributed repositories. OAI differs from other interoperability approaches, such as Z39.50 (Lynch, 1997) or SDLIP (Paepcke et al., 2000), through its emphasis on a limited, simple, and easy to implement protocol that layers over an existing repository. The

Xiaoming Liu, Los Alamos National Laboratory, Research Library, Los Alamos, NM 87545, and Kurt Maly, Michael L. Nelson, and Mohammad Zubair, Old Dominion University, Department of Computer Science, Norfolk, VA 23529

LIBRARY TRENDS, Vol. 53, No. 4, Spring 2005 ("The Commercialized Web: Challenges for Libraries and Democracy," edited by Bettina Fabos), pp. 590–603

© 2005 The Board of Trustees, University of Illinois

OAI framework defines two functional roles: data providers (also “repositories”) and service providers (also “harvesters”). Service providers develop value-added services that are based on the metadata collected from data providers. These value-added services could take the form of cross-archive search engines, linking systems, and peer-review systems.

The roots of the OAI lie in a vision to stimulate the growth of open e-print repositories. This concept began to be developed with the Universal Preprint Service (UPS) prototype (Van de Sompel et al., 2000), and was further advanced with the Santa Fe Convention (Van de Sompel & Lagoze, 2000). The UPS prototype was the discussion piece during an invitation-only workshop in Santa Fe, New Mexico, in the fall of 1999. This workshop brought together many of the leaders in the e-print community for the purpose of fostering interoperability between the various author-contributed e-print servers and institutional repositories in use at the time. Contemporary approaches toward interoperability were ad hoc at best. One of the distinguishing factors for the Santa Fe Workshop was the collective experience in building DLs and the associated interoperability problems; earlier interoperability workshops (Scherlis, 1996) were comparatively premature. The immediate result of this workshop was the Santa Fe Convention, an intermediate step toward the metadata harvesting model that would become the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

Realizing that the simple metadata harvesting idea had appeal to a broader reach of communities than that engaged in e-print publishing, version 1.0 of the OAI-PMH was released in January 2001. Following an extended period of evaluation and alpha and beta testing, version 2.0 of the OAI-PMH was released as a stable specification in June 2002 (Lagoze, Van de Sompel, Nelson, & Warner, 2002a). The development, history, impact, and secondary effects of OAI-PMH have been discussed in several publications, including Lynch (2001), Nelson (2001), Lagoze and Van de Sompel (2001), Van de Sompel and Lagoze (2002) and Lagoze and Van de Sompel (2003).

ARC

Arc (<http://arc.cs.odu.edu>) is the first end-user federated search service based on the OAI-PMH (Liu, Maly, Zubair, & Nelson, 2001). The Repository Explorer (Suleman, 2001) was released prior to Arc, but its targeted audience is mainly repository developers and maintainers, not end-users. Arc was initially released as an experimental service to investigate issues in metadata harvesting in October 2000. The software developed for the Arc service (<http://oaiarc.sourceforge.net/>) was released as an open source system under NCSA-style license in September 2002. It has been used in several production and research projects (see Table 1).

Arc was first developed as a proof-of-concept service for OAI-PMH;

Table 1. Other Projects Using the Arc Software

Digital Library	URL	Description
MetaArchive	www.metaarchive.org	Sharing resources related to politics and religion (Halbert, 2003)
NCSTRL	www.ncstrl.org	A collection of computer science technical reports and e-prints (Anan, Liu, & Maly et al. 2002)
RDN	www.rdn.ac.uk/	Backend harvester of RDN ResourceFinder http://walrus.rdn.ac.uk/docs/oai_z3950
snelonline	www.snelonline.net	Cross Archive Search Engine

however, the development of Arc revealed interesting problems and inspired further research in these domains. In this article we introduce the development and architecture of the Arc system and follow-up research that attempted to improve or optimize the metadata harvesting system and search performance. We will discuss the Archon project for building value-added services to take advantage of rich metadata beyond Dublin Core (DC) (Weibel & Lagoze, 1997); the DP9 service to allow general search engines (Google, Yahoo, etc.) to index OAI-PMH compliant collections; and the recently funded Andrew Mellon Foundation DL Grid project for building a high-performance federated search service. When possible, interesting and general features resulting from these research projects are incorporated back into the publicly available Arc source code distribution.

DEVELOPMENT OF ARC

Arc was initially released as an experimental service to investigate issues in metadata harvesting. It immediately attracted interest because it was the only vehicle to demonstrate the potential and promise of OAI-PMH at that time. As new data providers appeared, they often requested to be added to the Arc system for demonstration purposes; by continuously integrating various new data providers, the software was made stable and fault tolerant. Originally conceived as more of a tour de force, Arc has become a useful tool for helping new data providers to make their collections truly OAI-PMH-compliant by giving them feedback on errors during harvesting.

When applying the Arc software in various environments, we encountered a number of problems such as inconsistent metadata, lack of controlled vocabulary, and XML errors. Based on feedback from other adopters, we have been able to address these problems and have consequently added many new features for customization and installation. The architecture of the Arc system has been refined to easily add or extend new functionalities.

Arc is available for download (<http://sourceforge.net/projects/>

oaiarc/) and can be used with either Oracle or MySQL. OAI-PMH uses unqualified Dublin Core as the default metadata set, and most Arc end-user services are implemented on the data provided in the DC metadata. The current supported end-user services include simple search, advanced search, interactive search, annotation service, and browse/navigation over search result. Arc has a Web-based administration interface, which allows users to configure various parameters for harvesting and to check harvester logs to handle various error situations such as erroneous XML replies from data providers.

ARCHITECTURE OF ARC

The basic structure of OAI-PMH supports two basic components: the service provider and the data provider. Data providers administer systems that support the OAI-PMH as a means of exposing metadata, and service

Figure 1.

The screenshot shows the ARC web interface in a Mozilla browser window. The address bar displays <http://128.82.7.99:8080/oai/results.jsp>. The page header identifies the service as "ARC - A Cross Archive Search Service" from the "Old Dominion University Digital Library Research Group". Navigation links include Home, Simple Search, Advanced Search, Browse, Administration, OAI, and Help. A sidebar on the left lists "Archive Groups" such as ACL(2149), AIM25(11039), ALADIN_OAI2(370), ATILF(1), Archive Lyon_2(21), ArchiveEnsls...(83), ArchivesEIAH(11), BDBComp Archive(2469), BieSon(1), CAV2001(111), CBOLD(89), CCSDJeanNicod(77), CCSDarchiveSIC(160), CCSDthesis(1641), CDLCIAS(37), CDLDERM(5), CDLTC(1), CPS(769), CSTC(105), CULeuclid(14718), and CULeuclid-test(882). The main content area shows search results grouped by archive, currently showing 'arXiv'. The query is "Get All Records>". The results are paginated, showing page 1 of 20. The first result is "A note on canonical functions" by Jech, Thomas J. and Shelah, Saharon, dated 1989-04-15, with OAI ID oai:arXiv:math/9201239. The second result is "Convex bodies with few faces" by Ball, Keith J. and Pajor, Alain, dated 1989-10-26, with OAI ID oai:arXiv:math/9201203. The third result is "Shadows of convex bodies" by Ball, Keith, dated 1989-10-26, with OAI ID oai:arXiv:math/9201204. The fourth result is "Volume ratios and a reverse isoperimetric inequality" by Ball, Keith, dated 1989-10-26, with OAI ID oai:arXiv:math/9201205. The fifth result is "On the volume of the intersection of two ℓ_p balls" by Schechtman, Gideon J. and Zinn, Joel, dated 1989-11-09, with OAI ID oai:arXiv:math/9201206. The bottom status bar indicates "Applet DetectPluginApplet notinited".

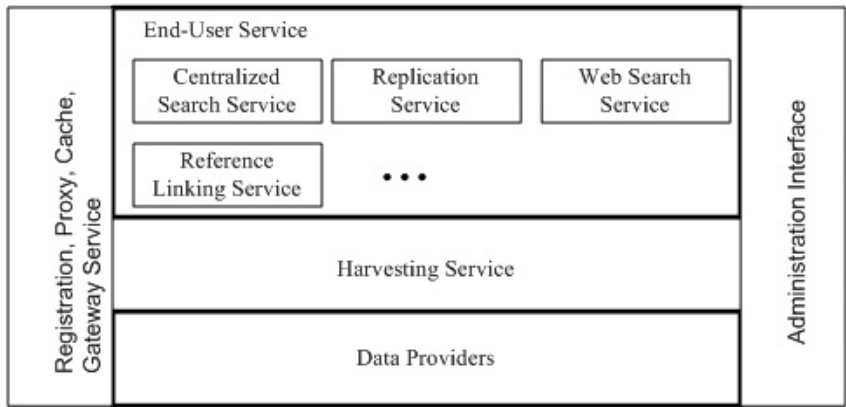
providers use metadata harvested via the OAI-PMH as a basis for building value-added services.

The OAI-PMH focuses on the clear interface between data providers and service providers. In Figure 2 we define the Arc model for metadata harvesting that addresses many of these issues. The data provider maintains one repository for digital records. Then a number of service providers work together to conduct metadata harvesting. The harvester is the key service that uses OAI-PMH to maintain the synchronization between data providers and other services, such as centralized federation services, replication services, and citation linking services. In addition, the Arc system includes OAI-PMH proxy, cache, and gateway services to optimize the functioning of the model underlying the OAI-PMH technology (Liu, Brody, Harnad et al., 2002). These services provide an infrastructure that can be used by all other components to achieve interoperability, scalability, and reliability.

Data Providers

A data provider supports the OAI-PMH as a means of exposing metadata. The design of a good data provider presents many challenges, including metadata quality, server availability, and service quality. The quality of data providers has been a significant problem since the genesis of OAI-PMH at the Santa Fe Convention in 2000. Data providers are frequently unreachable on the network or have errors in the data they return in response to OAI-PMH harvesting requests. Details of this phenomenon can be found at Celestial (<http://celestial.eprints.org>), a service that tracks the stability of data providers over time as well as provides a cache of the data provider's contents. During the testing of harvesting from OAI-PMH data providers, we were able to overcome particular problems regarding compatibility and adaptability.

Figure 2.



Initially, the use of unqualified DC as a common metadata format in OAI-PMH proved to be very helpful for building a quick prototype, and thus it is continuously used by several service providers. However, richer metadata formats are essential for building a richer service. Dublin Core is a “lingua franca” metadata format only—it is best suited for resource discovery rather than rich semantic description. The OAI encourages the simultaneous exposure of richer, community-specific metadata formats as well. Although a number of data providers have supported richer metadata formats, it is difficult to implement richer services over these metadata without individually studying each format. Consequently, we developed a series of interfaces relying on interactive user refinement to navigate a large corpus of metadata with varying quality, structure, and semantics (Liu, Maly, & Zubair et al., 2002). We found that, while it is possible to search the corpus in this manner, it places a high cognitive load on the user. Further research in this area is required to achieve the “deep semantic interoperability” identified by Lynch and Garcia-Molina (1995).

Harvesters

Similar to a Web crawler, an OAI-PMH harvester traverses the data providers automatically and extracts metadata. However, there are two significant differences between the OAI-PMH harvester and a Web crawler: the harvester recognizes metadata formats, thus allowing use of structured data, and the harvester exploits the incremental harvesting defined by the OAI-PMH, allowing more efficient extraction of information than a regular Web crawler.

For example, consider a Web site of 100 pages that is available through regular Web crawling and an OAI-PMH repository interface. Assume 5 of the pages are updated weekly and that the Web site is harvested weekly both by a Google robot and by an OAI-PMH harvester. Assume that the OAI-PMH interface is configured to distribute 10 documents, batched together, per connection. Assuming that smart Web robots perform conditional HTTP GETs (http status code 304) based on last modified dates, the Google robot will not download more files than it needs to, but it will have to query every individual Web page of the Web site to determine its date of last modification. Figure 3 illustrates this model.

But the OAI-PMH model saves the considerable overhead of establishing TCP/IP and HTTP connections for documents that have not changed. Instead of having to ask each of the 100 files if their modification date has changed, the harvester asks the OAI-PMH interface which files have changed, and the OAI-PMH interface only responds with the files that meet the criteria (see Figure 4).

Given the parameters stated above, Table 2 shows the relative load placed by each method. If the Web site is larger, say 1,000 or 10,000 files, the unnecessary network traffic avoided with OAI-PMH would be even

Figure 3.

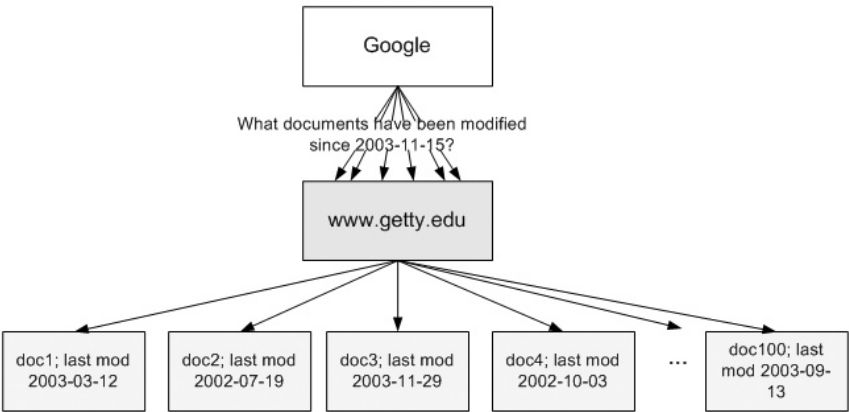
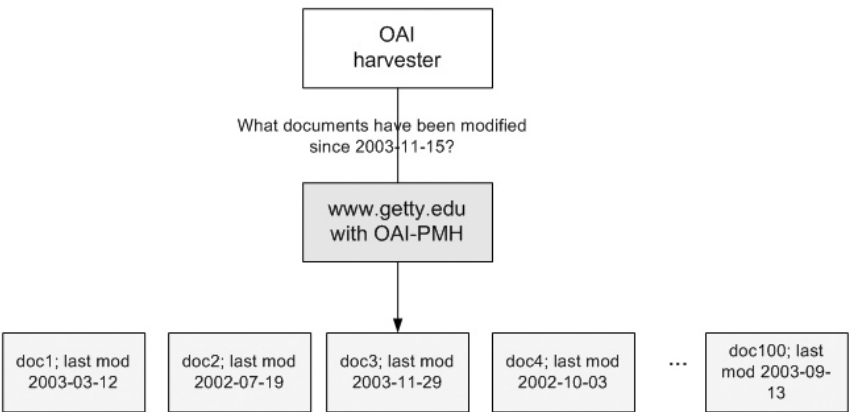


Figure 4.



greater. Even if Web sites updated their content (or added new content) more rapidly, the OAI-PMH approach would still reduce the number of connections by a factor of the number of files batched together in the response (in this example, by a factor of 10). Not only would this reduce the network load for the robots and Web sites, it would also allow for much quicker harvesting of updates and thus more up-to-date Web indices.

Because data providers are different in data volume, partition definition, service implementation quality, and network connection quality, all these factors influence the harvesting procedure. Historical and newly published data harvesting have different requirements. When a service

Table 2. Network and Server Load by Harvesting Type

	Initial Harvest	Weekly Updates
Google Robot	100 connections	100 connections
OAI-PMH Harvester	10 connections	1 connection

provider harvests a data provider for the first time, all past data (historical data) needs to be harvested, followed by periodic harvesting to keep the data current. Historical data harvests are high volume and more stable. The harvesting process can run once or, as is usually preferred by large archives, as a sequence of chunk-based harvests to reduce data provider overhead. To harvest newly published data, data size is not a major problem, but the scheduler must be able to harvest new data as soon as possible and guarantee completeness. The OAI-PMH provides flexibility in choosing the harvesting strategy, although optimizing it remains an open question.¹

Scalability Through Hierarchical Harvesting and “Aggregators”

A service provider can also act as a data provider, disseminating meta-data harvested from other data providers. This allows for the hierarchical harvesting of content and removes a limitation of having all data providers be at the same “level.” This structure has a great deal of flexibility in how information is filtered and interconnected between data providers and service providers. While hierarchical harvesting was not originally part of the OAI-PMH, there was nothing in the protocol that prohibited it. Arc was the first service provider to introduce hierarchical harvesting, and services that provide hierarchical harvesting are now known by the name of “aggregators.”

Aggregators may normalize, correct, transform, or otherwise change the harvested metadata. Thus, the re-exposed data might not be the same data harvested from the original data providers. Unless the metadata exposed by the aggregator is completely unchanged, the aggregator must issue new identifiers for the OAI-PMH records it makes available for harvesting. The OAI-PMH defines provenance containers to assist in the de-duplication of metadata harvesting from various sources. Guidelines have since been written to assist in the development of OAI-PMH proxies, caches, and aggregators (Lagoze, Van de Sompel, Nelson, & Warner, 2002b).

An illustrative example of an aggregator in action is the NASA Technical Report Server (NTRS) (Nelson, Rucker, & Harrison, 2003). NTRS (<http://ntrs.nasa.gov>) is a public digital library that provides users access to NASA-authored reports, reprints, and other aerospace related materials. NTRS uses the OAI-PMH to harvest metadata from twelve different data providers. If the developers of other digital libraries wish to include NASA material into their DL, they could harvest directly from the twelve data providers from which NTRS harvests. Or, because NTRS is also an

OAI-PMH aggregator, they could simply harvest the metadata via OAI-PMH directly from NTRS. NTRS effectively becomes a one-stop shop for NASA and aerospace materials. This hierarchical construct could be repeated many times, allowing complex harvesting systems to be constructed. In the example above, harvesting directly from NTRS not only removes the burden of harvesting from the twelve sites directly, but it also addresses the problem of discovering new repositories. If NTRS adds a new aerospace repository to its collection, the sites harvesting NTRS will automatically gain this new collection on their next harvest.

Registration and "Friends"

Discovery of new and existing data providers is an important and difficult topic for the OAI community. The OAI-PMH does not define an explicit registration service: repository awareness either happens out of band from the OAI-PMH (for example, a repository administrator manually submits the repository base URL to a service provider), or a harvester learns of other repositories through an optional "friends" container in the "Identify" response. This allows a repository to "link" other repositories of which the administrator is aware. This is similar to a Web crawler learning of new Web servers by crawling and analyzing Web pages.

It might seem tempting at first to try to include a registration service as part of OAI-PMH. But, ultimately, this would not allow large-scale deployment of OAI-PMH data providers. In the same way that you can install a Web server and not have to register it with any authority, data providers can be set up in the same way. There are Web sites that do try to keep track of the publicly available data providers, but these are definitely not complete, just as any list of public Web servers is not complete. Discovery of data providers remains a difficult issue within the OAI community, but we feel that "friends" and aggregators will provide the necessary mechanisms for addressing this issue.

Proxy, Cache, and Gateway

The current and emerging applications based on metadata harvesting require a scalable and reliable infrastructure to support them. OAI-PMH proxies, caches, and gateways are tools to optimize the functioning of the data provider/service provider model underlying the OAI-PMH. An OAI-PMH proxy dynamically forwards OAI requests to data providers. For example, it can dynamically fix common XML encoding errors and translate between different OAI-PMH versions. An OAI-PMH cache caches metadata and can filter and refine them before exposing them to service providers. It also serves as a simple cache that reduces the load on source data providers and improves server availability. An OAI-PMH gateway can convert the OAI-PMH to other protocols and applications. For example, the gateway could convert between different protocols (for example, SOAP) and OAI-

PMH. The goal is to achieve interoperability, scalability, and reliability of OAI-PMH services.

End-User Services

Applying the OAI-PMH harvester to a compatible data provider can lead to many harvesting possibilities—the most obvious being federated search services and repository synchronization. Based on OAI-PMH, there are two approaches to building a federated digital library that allow users to access content in all the libraries through a single interface: centralized and replicated. In the centralized approach, a federation service harvests metadata from the OAI-PMH-enabled libraries and provides a unified interface to search all the collections. This approach has been adopted by Arc and other OAI-PMH service providers.

However, a centralized search service is not a suitable approach if the primary objective is to use native library interfaces. A centralized approach also suffers from the organizational logistics of maintaining a centralized federation service and having a single point of failure. The replicated approach addresses these problems. This approach can be viewed as mirrored OAI-PMH-compliant repositories, where every participant has its own federation service. The consistency between these services is maintained using OAI-PMH. As a federation service is locally available, it becomes easy to push other participants' metadata into the native library. In addition, this approach supports several levels of redundancy, thereby improving the availability of the whole system. For example, a failure of a system at one repository would not severely impact users at other repositories. In fact, users at the affected repositories would continue to search and discover reports from other repositories, though they may not be able to see reports that are added to the system at other repositories during the down time.

The OAI-PMH also provides an interface that exposes the “hidden” information to general Web search engines. Other services such as cross-archive citation linking are also emerging.

RELATED PROJECTS AND IMPROVEMENTS

The construction of the Arc system motivates studies on registration service, repository synchronization problems, metadata quality, and scalability. The architecture of Arc is generally designed so that different modules can be plugged in and tested, and, when appropriate, they are incorporated into the Arc system and open-source software.

As of 2004 we have been involved in several projects toward improving Arc in various aspects. Archon is a project we launched in response to the problem of how to build rich service based on complicated metadata; DP9 is an effort to make OAI-PMH repositories open to general Web crawlers; and the DL Grid project addresses the scalability problem of a large meta-

data harvesting system by using DL Grid technology. We will discuss each of these in more detail below.

Archon

Funded by the National Science Foundation (NSF), Archon is geared toward building a federated digital library focused on physics with support of rich metadata (Anan et al., 2003). By design, OAI-PMH does not provide support for services beyond basic harvesting of metadata records. Due to the different quality of services and metadata implemented by various providers, the challenge exists to provide rich services besides simple keyword searches. With the Archon project, we have tried to address providing value-added services like citation and reference processing, equation searching, and data normalization to the process of dynamic harvesting. In order to improve several aspects of the Arc system, Archon focuses on

- harvesting and parsing richer metadata formats and full text besides DC
- automatic citation linking across heterogeneous digital archives (APS, arXiv, and other physics collections)
- normalizing and automating the generation of missing metadata fields (for example, subject)
- equation searching—displaying and searching equations that are embedded in metadata fields such as title or abstract in formats such as LaTeX or MathML. Equations in these formats are not easy for users to browse or view.
- post-processing after search—the varying quality of metadata makes it difficult to build a unified search interface. Post-processing provides an alternative way to take advantage of richer metadata sets without complex search interface.

The development of the Archon system makes it clear that a rich search environment can be developed in a metadata harvesting system. However, it also reveals a lack of standards in this environment, such as metadata formats, controlled vocabulary, citation and reference information, and standard equation expression, which places the burden on service providers to understand all proprietary formats.

DP9

DP9 is an open-source gateway service that allows general search engines (for example, Google, Inktomi, etc.) to index OAI-PMH-compliant archives (Liu, Maly, Zubair, & Nelson, 2002). DP9 does this by providing a persistent URL for repository records and converting this to an OAI-PMH request to the appropriate repository when the URL is requested. Search engines that do not support the OAI-PMH can thus index the “deep Web” contained within OAI-PMH-compliant repositories.

Currently, indexing OAI collections via an Internet search engine is

difficult because Web crawlers cannot access the full content of an archive, are unaware of OAI-PMH, and cannot handle XML content very well. DP9 solves these problems by defining persistent URLs for all OAI-PMH records and dynamically creating a series of HTML pages according to a crawler's requests. DP9 provides an entry page and, if a Web crawler finds this entry page, the crawler can follow the links on this page and index all records in a data provider. DP9 also supports a simple name resolution service: once an OAI Identifier is given, DP9 responds with an HTML page, a raw XML file, or forwards the request to the appropriate data provider.

Various caching mechanisms are also implemented in the DP9 service to make it more crawler friendly. However, due to the limitation of Web crawling technology, there is no guarantee that a document can be indexed by any Web crawler. Further research, such as the *mod_oai* project (<http://www.modoai.org/>), is underway to ensure better integration between OAI-PMH and Web-crawling technology.

Digital Library Grid

When dealing with a large number of data providers and documents, we discovered that it is necessary to parallelize each individual module of the Arc system, such as metadata harvesting, indexing, and searching. This leads to the Mellon-funded Digital Library Grid project (http://saturn.seven.research.odu.edu/grid/index_new). The objective of the Digital Library Grid project is to develop a high-performance federated search service that exploits the resources of a grid. It will make available a large amount of information that is distributed amongst heterogeneous digital libraries. In this project, we are developing the software tools to

- adapt Arc and *Lucene* indexing software (<http://jakarta.apache.org/lucene/>) to the grid
- deploy a cluster for parallel, high-performance searching based on *Lucene*
- develop software support to move indices and metadata between low- and high-latency nodes

In this project we propose to distribute the cost of publishing to collection builders (data providers), distribute the cost of harvesting and indexing to existing grid nodes, and only leave the cost of maintaining the federated search service to one institution (service provider), thus making it more sustainable. Since grid nodes by definition have unused capacity, no new hardware needs to be acquired, and we can, in essence, piggyback the onus of maintaining the infrastructure on the efforts to maintain the grid. The second advantage of this approach is availability of the service. The current Arc is running on a single processor without any redundancy. In the new approach, we plan to use hardware redundancy by exploiting the grid technology. For searching, we plan to exploit parallelism by partitioning

the indices amongst a cluster of PCs. A user query will be executed in parallel across these partitions, resulting in high performance. For supporting parallel indexing and searching, we will extend the open source Apache Jakarta Lucene search engine.

CONCLUSION

The Open Archives Initiative Protocol for Metadata Harvesting has transformed the way Web-based digital libraries are built. Originally conceived as a way of federating e-print repositories, OAI-PMH is now used by a variety of academic, government, and even commercial publishing interests. It has proved to be useful even beyond "document-like objects." For example, Van de Sompel, Young, and Hickey (2003) discuss how OAI-PMH is used for thesauri, Web-usage logs, and an OpenURL registry.

The Arc service provider traces its roots to the original UPS prototype system that was featured at the initial workshop in Santa Fe. Arc debuted as a public service in 2001 and has been in continuous operation since then. It now indexes nearly 7,000,000 records harvested from several hundred repositories. Arc is available as an open source and has been deployed by a number of different institutions. It has also served as a platform for many other OAI-PMH projects, including Archon, Kepler, and DP9. Arc was the first service provider to introduce "hierarchical harvesting," which formed the basis for what are known as OAI-PMH aggregators. Because of Arc's immense scale, it has informed the community on a number of issues related to synchronization, scheduling, caching, and replication. Arc is currently being used in an ongoing project to merge OAI-PMH digital libraries with grid computing.

NOTE

1. Further discussion regarding the synchronization of harvesters and repositories can be found in Liu, Maly, Zubair, & Nelson, 2003.

REFERENCES

- Anan, H., Liu, X., Maly, K., Nelson, M. L., Zubair, M., & French, J. C., et al. (2002). Preservation and transition of NCSTRL using an OAI-based architecture. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 181–182). New York: ACM Press.
- Anan, H., Maly, K., Nelson, M. L., Zubair, M., Liu, X., & Gao, J., et al. (2003). *ARCHON: Building and learning environments through extended digital library services*. Paper presented at the 5th International Conference on New Educational Environments (ICNEE), May 26–28, 2003, Lucerne, Switzerland.
- Halbert, M. (2003). The Metascholar Initiative: AmericanSouth.Org and MetaArchive.Org. *Library Hi-Tech*, 21(2), 182–198.
- Lagoze, C., & Van de Sompel, H. (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 54–62). New York: ACM Press.
- Lagoze, C., & Van de Sompel, H. (2003). The making of the Open Archives Initiative protocol for metadata harvesting. *Library Hi Tech*, 21(2), 118–128.
- Lagoze, C., Van de Sompel, H., Nelson, M. L., & Warner, S. (2002a). *Open Archives Initiative Protocol for Metadata Harvesting—Version 2.0*. Retrieved November 28, 2004, from <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

- Lagoze, C., Van de Sompel, H., Nelson, M. L., & Warner, S. (2002b). *Implementation guidelines for the Open Archives Initiative Protocol for Metadata Harvesting*. Retrieved November 28, 2004, from <http://www.openarchives.org/OAI/2.0/guidelines.htm>.
- Liu, X., Brody, T., Harnad, S., Carr, L., Maly, K., & Zubair, M., et al. (2002). A scalable architecture for harvest-based digital libraries: The ODU/Southampton experiments. *D-Lib Magazine*, 8(11). Retrieved November 28, 2004, from <http://www.dlib.org/dlib/november02/liu/11liu.html>.
- Liu, X., Maly, K., Zubair, M., Hong, Q., Nelson, M. L., & Knudson, F., et al. (2002). Federated searching interface techniques for heterogeneous OAI repositories. *Journal of Digital Information*, 2(4). Retrieved November 28, 2004, from <http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/>.
- Liu, X., Maly, K., Zubair, M., & Nelson, M. L. (2001). Arc: An OAI service provider for digital library federation. *D-Lib Magazine*, 7(4). Retrieved November 28, 2004, from <http://www.dlib.org/dlib/april01/liu/04liu.html>.
- Liu, X., Maly, K., Zubair, M., & Nelson, M. L. (2002). DP9: An OAI gateway service for Web crawlers. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 283–284). New York: ACM Press.
- Liu, X., Maly, K., Zubair, M., & Nelson, M. L. (2003). Repository synchronization in the OAI framework. In *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 191–198). New York: ACM Press.
- Lynch, C. A. (1997). The Z39.50 information retrieval standard. Part 1: A strategic view of its past, present and future. *D-Lib Magazine*, 3(4). Retrieved November 28, 2004, from <http://www.dlib.org/dlib/april97/04lynch.html>.
- Lynch, C. A. (2001). Metadata harvesting and the Open Archives Initiative. *ARL Bimonthly Report* 217, 1–9.
- Lynch, C. A., & Garcia-Molina, H. (1995). *Interoperability, scaling, and the digital libraries research agenda: A report on the May 18–19, 1995 IITA digital libraries workshop*. Retrieved November 28, 2004, from <http://www.diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html>.
- Nelson, M. L. (2001). Better interoperability through the Open Archives Initiative. *New Review of Information Networking*, 7, 133–145.
- Nelson, M. L., Rocker, J., & Harrison, T. L. (2003). OAI and NASA scientific and technical information. *Library Hi-Tech*, 21(2), 140–150.
- Paepcke, A., Brandiff, R., Janee, G., Larson, R., Ludascher, B., & Melnik, S., et al. (2000). Search Middleware and the Simple Digital Library interoperability protocol. *D-Lib Magazine* 6(3). Retrieved November 29, 2004, from <http://www.dlib.org/dlib/march00/paepcke/03paepcke.html>.
- Scherlis, W. L. (1996). Repository interoperability workshop: Towards a repository reference model. *D-Lib Magazine*, 2(10). Retrieved November 29, 2004, from <http://www.dlib.org/dlib/october96/workshop/10scherlis.html>.
- Suleman, H. (2001). Enforcing interoperability with the Open Archives Initiative repository explorer. In *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 63–64). New York: ACM Press.
- Van de Sompel, H., Krichel, T., Nelson, M. L., Hochstenbach, P., Lyapunov, V. M., & Maly, K., et al. (2000). The UPS prototype: An experimental end-user service across e-print archives. *D-Lib Magazine*, 6(2). Retrieved November 29, 2004, from <http://www.dlib.org/dlib/february00/vandesompel-ups/02vandesompel-ups.html>.
- Van de Sompel, H., & Lagoze, C. (2000). The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, 6(2). Retrieved November 29, 2004, from <http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html>.
- Van de Sompel, H., & Lagoze, C. (2002). Notes from the interoperability front: A progress report on the Open Archives Initiative. *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries* (pp. 144–157). London: Springer-Verlag.
- Van de Sompel, H., Young, J. A., & Hickey, T. B. (2003). Using the OAI-PMH . . . differently. *D-Lib Magazine*, 9(7/8). Retrieved November 29, 2004, from <http://www.dlib.org/dlib/july03/young/07young.html>.
- Weibel, S., & Lagoze, C. (1997). An element set to support resource discovery—The state of the Dublin Core. *International Journal on Digital Libraries*, 1(2), 176–186.