July, 2004

Integration of Non-OAI Resources for Federated Searching in DLIST

Anita S. Coleman, *University of Arizona*
Paul Bracke, *University of Arizona*
S. Karthik, *University of Arizona*

ARTICLES

D-Lib Magazine
July/August 2004

Volume 10 Number 7/8

ISSN 1082-9873

Integration of Non-OAI Resources for Federated Searching in DLIST, an Eprints Repository

Anita Coleman
<asc@u.arizona.edu>

Paul Bracke
<paul@ahsl.arizona.edu>

S. Karthik
<skarthik@cs.arizona.edu>

Abstract

Federated, distributed, and broadcast searches on the Internet depend on an underlying common metadata framework by which the information resources to be searched are organized. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is designed to facilitate searches across OAI-compliant databases. Software such as Arc allow service providers to offer federated searching of multiple, OAI-compliant resources. The majority of web-accessible information resources, however, are not OAI-compliant. This article describes a process whereby readily available open source tools and customized scripts were developed for integrating metadata from non-OAI compliant repositories for a federated search. The work described is being carried out as part of the development of the Digital Library of Information Science and Technology (DLIST), an Eprints repository.

1. Introduction

Many institutions and individuals have taken advantage of the Web as a medium for disseminating their work. Disciplinary and institutional repositories, agency websites, museums, and other organizations have made large bodies of scholarly and cultural materials available online. Individuals have published their materials on personal web pages and other distributed venues. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) was designed to facilitate the technical interoperability among distributed archives. The objective of OAI-PMH is to develop a low-barrier, lightweight framework to facilitate the information discovery of content in distributed archives [1].

OAI-PMH has been a success to a great extent, and it has speeded the development of federated service providers such as Arc and OAIster [2, 3]. At the same time, OAI-PMH has not provided a complete solution to the issue of interoperability with non-OAI archives; information discovery of quality non-OAI, discipline specific resources continues to be elusive

as much information remains hidden in web sites that do not adhere to any standards. A related problem is that not all resource providers wish to build structured digital collections or participate in a digital repository, and therefore users continue to express the need for subject gateways and portals.

The work reported in this article is an attempt to address these problems in the context of our Eprints repository, the Digital Library of Information Science and Technology (DLIST). We provide a description of DLIST, our guiding principles for a solution to the challenge of integrating non-OAI resources, the DLIST systems architecture and finally the integration procedure with which we experimented in fall 2003. In the conclusion, we highlight some of the limitations of our proposed solution as well as the added benefits for digital repository development that non-OAI integration offers.

2. DLIST

DLIST is the Digital Library of Information Science and Technology, a repository of electronic resources for Library and Information Science (LIS) and Information Technology (IT) [4]. DLIST is an initiative of the School of Information Resources and Library Science[1] and the Arizona Health Sciences Library[2] at the University of Arizona[3]. It is an ambitious project that is seeking to build an international consortium for scholarly communication in the LIS discipline [5]. The primary objective of DLIST is to create a web-accessible, open, digital repository—an intellectual commons for LIS and IT.

DLIST is based on free software. At its core is the Eprints software developed at the University of Southampton, UK [6]. Eprints requires the Linux operating system; the Apache web server with mod_perl, the Perl programming language with a handful of extra modules, and the MySQL database system [7, 8, 9]. An open source software, Webalizer, is used to analyze and prepare DLIST usage reports [10]. Paracite, software developed with Eprints for reference parsing and location on the WWW, is also implemented in DLIST [11].

Eprints software is generally used to build institutional or discipline repositories of scholarship, the outputs of research such as preprints, journal articles, technical reports, conference proceedings, theses, dissertations, and books. Whole journals, books, and conference proceedings or their components such as individual chapters, and articles can be deposited into a web-accessible digital storage system and described using a database form. Examples of well-known Eprint archives that exist for other disciplines include Cog Prints for Cognitive Science and ArXiv for Physics [12, 13]. Figure 1 shows a screenshot of the DLIST home page.

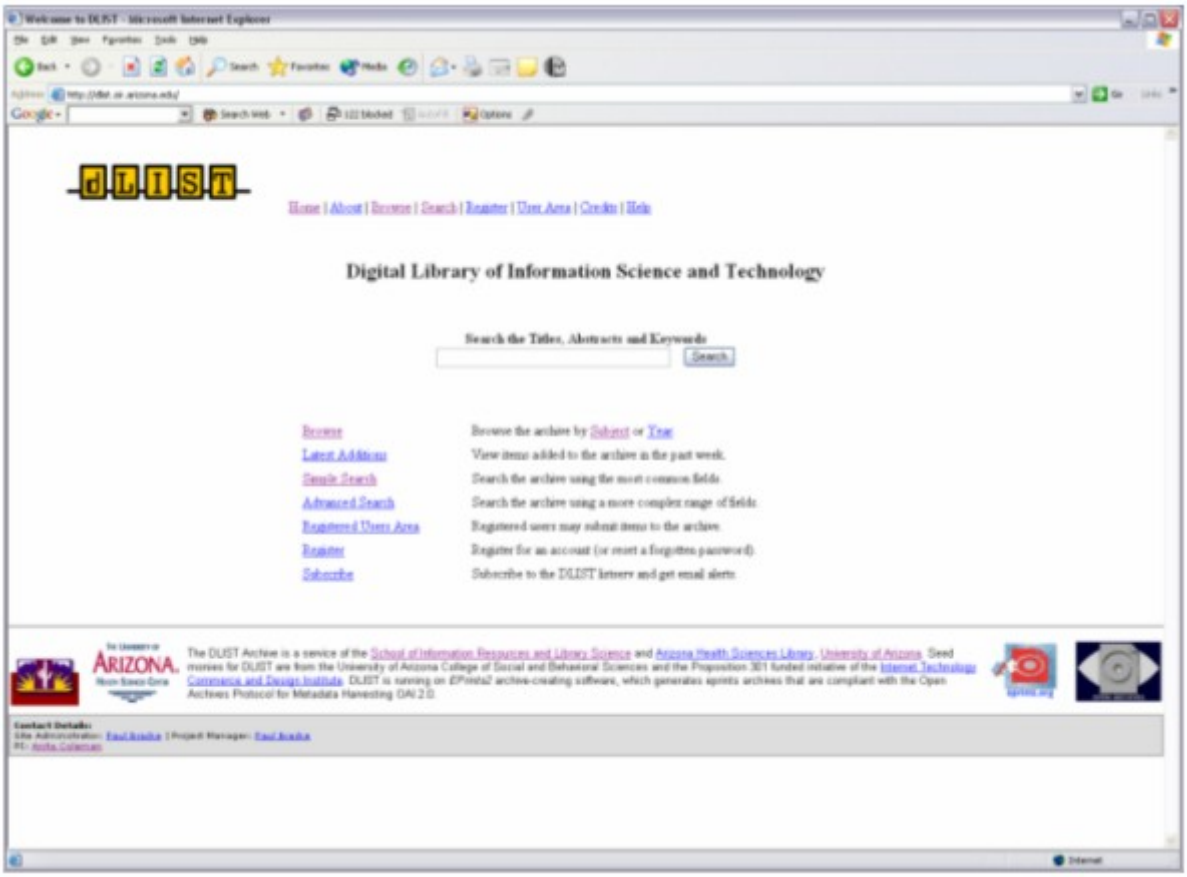

Figure 1: Screenshot of the DLIST home page.

(For a larger view of the DList home page, click here.)

Despite the easy availability of software for developing digital repositories, self-deposit of research articles by authors and the accompanying processes such as metadata creation have proven to be significant road-blocks in the growth and use of digital repositories. For example, the growth of the DLIST collection has been far slower than has its user base. Since its inception in June 2002, DLIST has amassed only 105 items through self-deposit. However, it has three times that number of registered users, and this number continues to increase (we currently have 330 registered users). In our first year of operation we learned that in LIS and affiliated disciplines such as Information Systems, neither the argument that open access improves research visibility nor the economic crisis overtaking scholarly journals are powerful enough motivators to encourage participation and author self-deposit in a repository [14, 15]. It became clear that another model, distinct from the Eprints model of author self-deposit, for growing the repository collection was necessary. Several academic units indicated that they would participate in a discipline-based repository (as opposed to an institutionally based one) but these academic units also stated that author deposit and metadata creation were barriers to such participation. A quick review of the literature showed that this was by no means unusual; in fact, the CiteSeer model, whereby scientific literature on the Web is crawled, harvested, and indexed, appeared to be a more successful one than even the institutional repositories initiative for growing collections and services quickly [16, 17, 18, 19]. Focused crawling has also been reported as part of the strategy to build NSDL collections [20]. However, we did not want to just crawl, harvest, and index relevant materials from the Web; rather, we wanted to see how focused mining of selected partners' websites could be automated so that DLIST could do two things efficiently:

1. build a rich collection of subject-specific LIS and IT research materials and

2. build communities and collections in symbiosis without requiring partners to conform to standards.

Very simply, the problem we wished to solve was this: how can we integrate non-OAI resources into DLIST automatically?

For automation purposes, non-OAI resources were characterized as follows:

1. Unstructured personal web pages of individuals/organizations, and
2. Structured web sites that use a web content management system, database, or other system.

The problem of automating the integration of non-OAI resources from large structured sites is the one described in the next sections.

3. Guiding Principles

The technical approach to integrate non-OAI resources was guided by three considerations:

1. The tools used for automatic metadata extraction must be available as open source.
2. It must be possible to schedule harvesting of non-OAI resources regularly and automatically.
3. The process developed must be transferable.

4. Systems Architecture

In order avoid needless code modification and enhance long-term stability, OAI-PMH was used as an integration layer. Metadata from non-compliant sites was harvested through the locally developed process, exposed through OAI-PMH, and then harvested by Arc, which provided the federated search service and capabilities [21]. Since then, the PMC has become OAI-compliant, but fortunately the scripts we developed for PMC can be used for other non-OAI resources as well, such as the ones on the Association of Research Libraries (ARL) website [22]. PMC is the National Library of Medicine (NLM) digital archive of life sciences journal literature, and it contains a wealth of valuable research of interest to a number of service providers including DLIST.

We also chose PMC for our experiments because it contained a large volume of valuable content (specifically, the backfiles of the *Journal of the Medical Informatics Association* and the *Bulletin* and *Journal of the Medical Library Association*) and the site was well structured, with consistent directory structures, citation information, and HTML coding conventions.

In order to avoid modifying Arc code and enhance long-term stability, OAI-PMH was used as an integration layer. Metadata from non-compliant sites was harvested through the locally developed process, exposed through OAI-PMH, and then harvested by Arc, which provided the search capabilities.

5. Integration Procedure

The process of integrating non-OAI compliant resources with OAI-compliant resources through Arc was accomplished in four steps:

1. Data capture
2. Metadata extraction
3. Metadata mapping

4. Metadata importation for federated search

These are further described below.

Data capture

The first step is the automated data capture phase. WGET is used to harvest pages from resources of interest [23]. WGET has been configured to crawl websites of interest, and save their contents to local files. In our configuration, only unchanged web pages are harvested to reduce load on remote servers and simplify local processing.

Metadata extraction

Once the new pages of interest are available locally, the next phase is to parse the pages and extract all necessary metadata. A parser, written in awk, must be developed for each non-compliant site to be integrated. The parser uses the HTML structure of a document to extract metadata, removes all extraneous markup tags, and saves the results to a local file. Sometimes a parser must also be created for subsections (e.g., journals) within a site.

Data capture and metadata extraction are all neatly contained in a script file that can be setup as a cron job. The script takes several parameters applicable to all potential sites: the base URL from which to harvest and the output directory to which the metadata needs to be stored. Additionally, the script can be extended to take advantage of site-specific features that can help in the harvesting process. For example, articles within PMC are sequentially numbered, with a page of metadata for each article. The script could check the identifier of the last article harvested and start crawling from the next article. Once set up as a cron job, this script can be executed periodically to harvest new articles and extract the metadata from them.

Metadata mapping

The next step in the integration process is the mapping of metadata from our custom format into Dublin Core elements so that it can present an OAI interface. For this we use the Rapid Visual OAI Tool (RVOT) developed at the Old Dominion University [24]. RVOT can be used to make collections OAI compliant in a quick and convenient way. It can be used to graphically construct an OAI-PMH repository from a collection of files, and it comes inbuilt with a metadata translation tool, a lightweight HTTP server, and an OAI-PMH request handler.

We chose RVOT for its easy to use interface, flexibility in handling custom metadata formats and built-in HTTP server providing a data provider interface. RVOT requires the development of a parser for custom metadata format in Java so that it can be mapped to Dublin Core (DC) format. Once the metadata is imported, fields in the local format are mapped to DC elements through a graphical, drag and drop procedure. RVOT then uses the mapping to convert local metadata files to a new file for its use. Once the conversion is complete, the metadata is immediately available for OAI-PMH requests through the data provider interface.

Metadata importation for federated search

Once the non-OAI resource has been made available through an OAI server, the Arc harvester component is used to import metadata into its federated search service, just as it would import the data from any OAI-compliant repository. Arc is also used to provide federated search of OAI-compliant repositories, as well as the non-OAI resources harvested through our procedure.

Like any screen-scraping approach to information system integration, there are disadvantages to this approach. First, the parsers are site-specific and must be developed for each resource to be integrated. This requires considerable labor and in-depth knowledge of regular expressions or other text-matching techniques. It is a reasonable approach for large, consistently structured sites such as PMC, in which a single parser can extract the metadata for hundreds of articles. For smaller sites, or sites with very inconsistent HTML coding, this approach will likely require too much labor to justify the expenditure of the required development time. Second, such an approach is very sensitive to any design change at the remote resource. Even small changes to the HTML structure of a site can require extensive modification or complete refactoring of a parser.

6. Conclusion

We have described a simple and efficient architecture for integrating non-OAI compliant resources found on the Web with OAI-compliant repositories in a federated search system. Initial tests have shown us that the system is flexible and can be extended to other web sites or resources on the Internet.

At present the parser for extracting metadata is hard-coded to support a single non-compliant repository. We plan to make this process more data-driven by using XML configuration files to specify details about resources to be harvested, including the site's metadata organization. The current system also requires human intervention for setting up the mapping between custom metadata format and DC elements, and for the conversion. We plan to make this process as automatic as possible in the future [25].

Ideally, more resources will become available through OAI-compliant archives and an architectural process of the sort we have described to integrate non-OAI resources might become unnecessary. Unfortunately, this may not be the case. There will always be a number of valuable non-compliant resources that could significantly enhance existing digital repositories and information discovery search services. We hope more such sites will become OAI-compliant over time, so this system will be needed less and less. Until then, our suite of open source tools, scripts and processes provide a relatively low-barrier approach to integrating non-compliant resources into federated search. Another benefit of our approach is that it is proving to be an attractive method for data providers who do not want to re-build their own repositories but continue to make their data available through structured, non-standardized websites. In these cases, we can set up automated processes for resource harvesting, metadata creation, and integration into the disciplinary repository. This is a win-win situation for both data providers and repositories, as the repository collections and users are growing in symbiosis. Data providers do not have to engage in the provision of costly web upkeep and searching mechanisms and yet they are able to ensure wider access to their materials and increased visibility and publicity for the constituencies and organizations they represent.

7. Notes

1. School of Information Resources and Library Science. <http://www.sir.arizona.edu/>.

2. Arizona Health Sciences Library. <http://www.ahsl.arizona.edu/>.

3. University of Arizona. <http://www.arizona.edu>.

8. References

[1] Lagoze, Carl, Van de Sompel, Herbert, Nelson, Michael, and Warner, Simeon. *The Open Archives Initiative Protocol for Metadata Harvesting: Protocol Version 2.0 of 2002-06-14*. <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

[2] Liu, X., et al. Arc: an OAI service provider for cross archive searching. In *Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries* (Roanoke VA, June 24-28, 2001), ACM Press, New York, NY, 2001, 65-66.

[3] OAIster. <http://oaister.umdl.umich.edu/o/oaister/>.

[4] DLIST Home. <http://dlist.sir.arizona.edu/>.

[5] Coleman, A. S. and Bracke, P. Building an international scholarly communication consortium for Library and Information Science. In *Mapping Libraries and Technology Across People: Proceedings of the First International Conference on Library Automation in Educational and Research Institutions (CALIBER 2003)*, Edited by T.A.V. Murthy et al. Ahmedabad, India, INFLIBNET, 2003. <http://dlist.sir.arizona.edu/pubs/6.htm>.

[6] Eprints. <http://www.eprints.org/>.

[7] Apache. http://www.apache.org/>.

[8] mod-perl. <http://perl.apache.org/>.

[9] MySQL. <http://www.mysql.com/>.

[10] Webalizer. <http://www.mrunix.net/webalizer/>.

[11] Paracite. <http://paracite.eprints.org/developers/>.

[12] CogPrints. <http://cogprints.ecs.soton.ac.uk/>.

[13] ArXiv. <http://arxiv.org/>.

[14] Lawrence, S. Free online availability substantially increases a paper's impact. *Nature Web Debates*. <http://www.nature.com/nature/debates/e-access/Articles/lawrence.html>.

[15] Harnad, S. 1995. The PostGutenberg Galaxy: How to get there from here. <http://www.ecs.soton.ac.uk/~harnad/THES/thes.html>.

[16] Suber, P. 2003. Removing the barriers to research: An Introduction to open access for librarians. <http://www.earlham.edu/~peters/writing/acrl.htm>.

[17] Hitchcock, S. 2003. Metalist of open access eprint archives: The Genesis of institutional archives and independent services. *ARL Bimonthly Report*, 223, April. <http://www.arl.org/newsltr/227/metalist.html>.

[18] Lynch, C. 2001. Metadata harvesting and the Open Archives Initiative. *ARL Bimonthly Report*, 217, August. <http://www.arl.org/newsltr/217/mhp.html>.

[19] Hitchcock, S. Developing services for open eprint archives: globalisation, integration, and the impact of links. <http://opcit.eprints.org/dl00/dl00.html>.

[20] Bergmark, D., Lagoze, C., and Sbityakov, A. Focused crawls, tunneling and digital libraries. In *European Conference on Digital Libraries 2002*. <http://mercator.comm.nsdlib.org/CollectionBuilding/ECDLpaper2.pdf>.

[21] Liu, X, Maly, K, Xubair, M, and Nelson, M. 2001. Arc - An OAI Service Provider for Digital Library Fedeeration. *D-Lib Magazine* 7 (4): April. <doi:10.1045/april2001-liu>.

[22] Association of Research Libraries. <http://www.arl.org/>.

[23] GNU WGET. <http://www.gnu.org/software/wget/wget.html>.

[24] Sathish, K. Maly, K., Zubair, M. and Liu, X. 2003. RVOT: a tool for making collections OAI-PMH compliant. *Russian Conference on Digital Libraries, 2003*. <http://rvot.sourceforge.net/documents/RapidVisualOAITool.pdf>.

[25] DLIST Non-OAI Resources Integration Suite. <http://dlist.sir.arizona.edu/pubs/>.

doi:10.1045/july2004-coleman