Utility of an OAI Service Provider Search Portal

Sarah L. Shreeves, Christine Kirkham, Joanne Kaczmarek, Timothy W. Cole
University of Illinois Library at Urbana-Champaign
{sshreeve, ckirkham, jkaczmar, t-cole3}@uiuc.edu

Abstract

The Open Archives Initiative (OAI) Protocol for Metadata Harvesting (PMH) facilitates efficient interoperability between digital collections, in particular by enabling service providers to construct, with relatively modest effort, search portals that present aggregated metadata to specific communities. This paper describes the experiences of the University of Illinois at Urbana-Champaign Library as an OAI service provider. We discuss the creation of a search portal to an aggregation of metadata describing cultural heritage resources. We examine several key challenges posed by the aggregated metadata and present preliminary findings of a pilot study of the utility of the portal for a specific community (student teachers). We also comment briefly on the potential for using text analysis tools to uncover themes and relationships within the aggregated metadata.

1. Background

The Open Archives Initiative (OAI) Protocol for Metadata Harvesting (PMH) is now established as an important tool for interoperability between digital collections. It was designed as a technically low-barrier means to share metadata, particularly metadata describing XML documents, digital images, content in other non-HTML formats, or resources contained in databases i.e. formats and locations not readily available to current Web search engines. [1] Based on a harvesting model, the OAI-PMH relies on both *data providers*, who expose their metadata through the protocol, and *service providers*, who harvest and aggregate metadata from one or more providers. [3]

OAI-PMH service providers can facilitate efficient interoperability among data providers by constructing, for example, search portals that present aggregated metadata to specific communities. The OAI-PMH project based at the University of Illinois at Urbana-Champaign tested the efficacy of the OAI-PMH model for search and discovery of metadata describing content in the domain of cultural heritage. The Illinois OAI-PMH project began in June 2001 and ends in May 2003.

0-7695-1939-3/03 $17.00 © 2003 IEEE

2. Building a portal to aggregated metadata

The Illinois project built a repository which can be accessed through a search portal called the UIUC Digital Gateway to Cultural Heritage Materials (http://oai.grainger.uiuc.edu/search). The repository contains approximately 1.1 million original metadata records. The web portal uses the XPAT indexing and search tools developed by the Digital Library Extension Service (DLXS) at the University of Michigan. As of December 2002, we had collected metadata from 39 providers, including museums, archives, academic and public libraries, historical societies, consortiums, and digital libraries. The aggregated metadata describes an array of cultural heritage resources held by more that 500 institutions worldwide. Some resources exist in digital formats, such as .JPG images. Other resources exist only in analog format and are represented digitally through the metadata.

The common schema used for metadata stored in the repository is Dublin Core (DC). Approximately half of the participating institutions are registered OAI data providers, whose records are harvested directly from their own servers. The non–OAI-registered providers delivered "data dumps" of metadata, which are used as sources for surrogate metadata provider sites implemented at Illinois (only for harvest by this project). Included in the repository are item-level metadata records derived from more than 8,000 Encoded Archival Description (EAD) finding aids. Using an algorithm developed as part of this project [4], these 8,000 EAD files generated more than 1.5 item level DC records (describing mostly analog resources) bringing the total number of item-level DC records to approximately 2.5 million.

Analysis of a subset of approximately 600,000 records provided natively in DC revealed wide variations in the interpretation and application of DC elements by different communities. [5] For example, 93% of records from museums used the *subject* element versus only 15% of records from academic libraries. Such disparities, coupled with the variety of controlled vocabularies in use, present specific problems for anyone attempting to build an

effective search service for aggregated metadata. We developed a variety of strategies to minimize these disparities, including indexing and organizing metadata by type of material (image, text, physical object, etc.) and applying a normalization vocabulary to the *date*, *coverage*, and *type* elements. [2]

3. Testing utility of search portal

A goal of the Illinois project was to evaluate the utility of a search portal to aggregated metadata. We examined how one group of users interacted with the portal. Our user population was comprised of 23 college students training to become K-12 social studies teachers in an honors-level curriculum and instruction course. They were asked to use the site to find primary sources for use in preparing a lesson plan on a specific social sciences topic, write short papers about their experience, and participate in one focus group interview. Prior to their searches, users were introduced to the concept of *metadata aggregation* and were informed that the search portal would provide pointers to digital content held elsewhere and would include analog resources. Preliminary results from this pilot study highlight three key issues.

First, despite their prior introduction to the nature of the portal, in practice the test group expected all records to point directly to corresponding digital objects. They reported feelings of frustration in finding analog resources when they expected digital resources. This was exacerbated by the large number of item-level records derived from EAD files that described analog resources. Thus, a user who selected a result for "letters from a WWI soldier" might find that the record referred to the holding institution's finding aid instead of to the letters themselves. Likewise, they reported a significant slowing of their efforts when the pointers (the URLs within the record) went to a top-level or intermediate page, where they might have to resubmit their request using the institution's own search engine.

Second and not unexpectedly, the lack of a ranking facility in our portal resulted in the test group feeling overwhelmed by the quantity of unsorted results. Because of the lack of consistent metadata caused by variations in controlled vocabularies and disparities in the use of DC, the Illinois team had decided to enable greater recall by designing the default search screen as a keyword search on all elements. This exacerbated the lack of a ranking facility. In an attempt to address these known limitations we provided an advanced search screen, which included standard methods for refining a search, such as restricting searches to specific groups of fields and setting limits. However, the test group seldom used the advanced search tools, and the few users who did attempt to refine their searches were unfamiliar with the types of entries required by metadata fields like "Format." This suggests that a robust ranking facility is of great importance.

Third, users accorded equal credibility to all contributing collections. They reported that they made no decisions about which items to examine based on the name of the holding institutions. Feelings of frustration around failed searches were directed at the search portal rather than at individual institutions. Thus, users held the portal responsible for the usability of its aggregated metadata, even when that metadata originated elsewhere and remained outside the control of the Illinois project.

4. Conclusions and future work

A clear and perhaps obvious finding of our work is that, while the OAI-PMH itself is readily implemented, the challenges posed by large amounts of heterogeneous metadata are significant. Certainly the application of more sophisticated pre-processing tools as well as robust, scalable search tools would aid in making the search portal a more effective tool for users. Other options include the development of thematic exhibits (based on human and/or machine analysis of metadata) that would offer glimpses into the range and type of materials available through the search portal, as well as the ability for users to annotate individual records, thus highlighting particularly useful resources.

While normalizing scripts for elements such as *type* and *date* are feasible because of the limited range of variations in those elements, manual normalizing of more complex free text elements such as *subject* and *description* would require time-consuming and cost-prohibitive efforts. We are exploring the use of an automated text analysis tool to learn whether such a tool can ferret out shared concepts or themes hidden in many thousands of *subject/description* fields. ThemeWeaver is a data mining tool developed by the automated learning group at the National Center for Supercomputer Applications (NCSA). Although this tool was designed to analyze large sets of documents with large amounts of text per document, we are testing whether it can provide natural groupings of metadata within the search portal. Thus far, we have found that the content of the metadata fields tested was too sparse and/or inconsistent to enable this text-analysis application to uncover useful clusters. The Illinois team continues to work with the developers of ThemeWeaver to test upcoming versions.

Although the Illinois OAI-PMH project ends in May 2003, members from the project team are continuing investigations into ways OAI-PMH–based services can be built and sustained. These include (1) an IMLS-funded project to create an item-level metadata repository of

digital content created under the auspices of the IMLS National Leadership Grant program; (2) a project to create a state-wide repository of Illinois government and library information; and (3) a NSF-funded project to allow harvesting of a mathematics digital library. With the support of Grainger Engineering Library, the UIUC Gateway to Cultural Heritage Materials will be continued as a search portal to cultural heritage resources available from registered OAI data providers.

5. Acknowledgments

This material is based upon work supported by a grant from the Andrew W. Mellon Foundation.

6. References

[1] M.K. Bergman. "The Deep Web: Surfacing Hidden Value," *Journal of Electronic Publishing*, vol. 17, Aug. 2001, http://www.press.umich.edu/jep/07-01/bergman.html.

[2] T.W. Cole, et al, "Now That We've Found the 'Hidden Web' What Can We Do With It? The Illinois Open Archives Initiative Metadata Harvesting Experience", *Museums and the Web 2002: Selected Papers from an Int'l Conf*. (MW 2002*),* Archives & Museum Informatics, Toronto, 2002*,* pp. 63-72.

[3] C. Lagoze and H.V. de Sompel, "The Open Archives Initiative: Building a Low-Barrier Interoperability Framework", *Proc. 1st ACM-IEEE Joint Conf. on Digital Libraries*. (JCDL 2001), ACM Press, New York, NY, 2001, pp 54-62.

[4] C.J. Prom, "Reengineering Archival Access Through the OAI Protocols", to be published in *Library Hi Tech*. vol. 21, no.2, Jun.2003.

[5] S. L. Shreeves, J. Kaczmarek, and T.W. Cole. "Harvesting Cultural Heritage Metadata Using the OAI Protocol," to be published in *Library Hi Tech.* vol. 21, no.2, Jun. 2003.