

What Is OAI-PMH Good For?

Timothy W. Cole

University of Illinois at Urbana-Champaign, USA

The relatively high uptake of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) confirms the continuing interest on the part of librarians, curators, and other scholarly content providers in resource sharing and interoperability. As of December 2005, less than 5 years after the public introduction of OAI-PMH, there are almost 900 OAI data providers known to be up and running worldwide (source: *Experimental OAI Registry at UIUC* <<http://gita.grainger.uiuc.edu/registry/>>). Collectively, these repositories provide metadata describing several million digital objects. All manner of object types are represented, covering a broad spectrum of topic areas and temporal periods. Interest in other, overlapping and complementary protocols designed to facilitate interoperability and resource sharing (SRU/SRW, RSS, OpenSearch, ...) is also high. There's clearly a strong desire to get the word out about digital information resources in ways more proactive than passively waiting to be indexed by Google. High-level sharing of resources and resource descriptions is recognized as a prerequisite to enhanced and more ubiquitous end-user access to scholarly information.

In many respects, however, recent work with OAI-PMH, RSS, ZNG (singly and in combination) and with the many content management and aggregation tools that exploit these and other protocols -- ContentDM, EPRINTS.ORG, D-Space, Greenstone, DigiTool ... (the list goes on) -- has only served to whet the appetite. End-user services based on metadata and/or resource sharing are proving challenging to do well (one would expect no less). OAI-PMH provides an essential layer of standardization and community agreement, but on its own is not enough. Additional layers of community agreement on issues critical to the shared use of harvested metadata (and eventually the resources themselves) are also required.

Towards an Infrastructure supporting innovative DL Services

Developing the technical infrastructure for a large-scale digital library composed of disparate, physically scattered component collections and content repositories is a bit like reverse engineering an onion from the inside out with only a hazy vision of what the finished vegetable will look like. We understand that the technical infrastructure we want will be multi-layered, with each layer simultaneously dependent on those within it and a prerequisite for outer layers. We have an idea, albeit an imperfect one, of the skin of services that will overlie the layers of technical infrastructure. But we can't yet name or count all the layers of essential infrastructure individually. The truth is we don't yet know all the layers we need. As we design and add each new layer of technology or standardization we keep thinking (hoping) we'll be ready next for the outermost skin of robust and innovative digital library services, only to discover that there is more needed to enable the full suite of services and functions we want. Each additional layer tends to be larger and more complex. Each new layer makes clear the need for additional layers of technology and community consensus.

OAI-PMH serves as a layer somewhere midway between the innermost and outermost -- enveloping and reliant on core digitization technologies, collection building policies, and the essential transport and presentation protocols of the Web -- but still requiring to be surrounded

itself by layers of standard metadata schemas and community agreements on metadata authoring and processing. In retrospect it is clearly a useful step in the right direction; clearly it is also not the final step. It's difficult to visualize the complete vegetable when all you have in hand are the innermost layers of the flesh, but that's exactly what we must do if we are to succeed.

This workshop provides an excellent opportunity to look at what we've learned from work with OAI-PMH to date and to prioritize the near-term technical architecture issues which have surfaced from that work and most need to be addressed to build the next layers of digital library supporting infrastructure. In doing so we must re-examine critically and in some depth our shared understanding of the objectives of digital libraries -- the functions they are designed to support -- and begin to develop the additional community consensus and technical architecture agreements essential to support robust, high-quality, innovative digital library tools and services.

Lessons Learned from OAI-PMH: the UIUC-IMLS Digital Collections & Content Project

In 2001 IMLS convened a Digital Library Forum to "discuss the implementation and management of networked digital libraries, including issues of infrastructure, metadata, thesauri and other vocabularies, and content enrichment such as curriculum materials and teacher guides." In collaboration with several researchers involved in the (U.S.) National Science Digital Library (NSDL), this group promulgated the first edition of the *Framework of Guidance for Building Good Digital Collections*, now in its second edition and maintained by NISO <<http://www.niso.org/framework/Framework2.html>>. The Forum also recommended (among other things) that IMLS should create a registry of IMLS-funded digital collections and should encourage the use within the IMLS community of OAI-PMH (as a de facto emerging interoperability standard) by creating an item-level metadata repository for metadata describing items in IMLS-funded collections. This led in 2002 to a grant to the University of Illinois at Urbana-Champaign (UIUC) for a project to do just that. Three years on we have learned a great deal from this project (and from several other OAI-PMH based projects) about the use of OAI-PMH and about resource metadata sharing and aggregation more generally.

For instance, practical experience has confirmed the general utility of the OAI-PMH design which divides the digital library universe into content providers and service providers. Certainly this view of the digital library universe is simplistic. But just as the client-server model remains useful in some contexts as a view of many intrinsically complex network interactions, so too is the OAI content provider-service provider paradigm useful as a model for designing digital library architectures. Such a model encourages the creation of an architecture that attributes different roles and responsibilities to each participating partner. In such a model content providers can assert a measure of control over their resources and strive for a clearer, more explicit understanding of their obligations as content providers and of the expectations they should have of service providers harvesting resource metadata. Service providers can focus on the central aggregation services they want to build and leave the content creation and primary resource management responsibilities to the content providers.

Though arguably not as "light weight" technically as originally advertised, OAI-PMH has generally proven robust and relatively easy to implement, either directly, through shareware or vended software, or through special purpose gateways such as the OAI Static Repository gateways hosted at Los Alamos and UIUC (and elsewhere) and the FileMaker Pro - OAI

gateway developed at UIUC and being tested and further developed for ongoing UIUC projects and for the IMLS-funded *MOAC Toolbox* project being led by UC-Berkeley. These extensions and add-ons to canonical OAI-PMH have the potential to facilitate the development of technical infrastructure for the proposed Middle East Digital Library.

More problematically, work with OAI-PMH has shown that while sharing metadata is technically easy, sharing good and useful metadata remains hard. Efforts to aggregate metadata have brought to the fore a host of inconsistencies in metadata authoring practice, ranging from general metadata quality concerns [Shreeves et al. 2005], to problems that arise with loss of context when metadata is removed from local implementation and aggregated, to problems with descriptive granularity and definitions of exactly what level of resource abstraction should be described (depends on purpose, of course, and that's difficult to predict a priori). Add to these the additional problems associated with the cultural heritage institution spectrum of diverse descriptive traditions and practices (as apparent in the broad range of available metadata schemas and standards available), and the end result is a hodge-podge of metadata records that really don't function well collectively when aggregated. Experience with OAI-PMH the last few years has made shareable metadata issues far more visible and has demonstrated the shared responsibilities that both service providers (who harvest, normalize, and enrich metadata) and data providers have to ensure quality metadata aggregations. This recognition in turn, as a positive outcome and a further foundation, has stimulated work to begin the process of resolving these issues -- notably the *DLF-NSDL OAI and Shareable Metadata Best Practices* effort <http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?OAI_Best_Practices>, work funded in large part by IMLS and NSF. The proposed Middle East Digital Library will do well to build on the foundation this work offers with regard to descriptive metadata (in addition, of course, to the work being done by the BA on the unique problems of Arabic-language metadata authoring and dissemination).

Recently, we've also seen a growing interest in creating digital library services that go beyond simple search and discovery. While OAI-PMH anticipated this evolution, such efforts have served to show the limitations of OAI-PMH in the absence of additional community-based agreements and standards. The problem, of course, is less with OAI-PMH than it is with the lack of community agreement and consensus on metadata standards, crosswalks, thesauri, name and term authority, practices for normalization and enrichment of harvested metadata, ... (again, the list goes on). This has led to an emerging appreciation of the critical need for additional layers of technology (e.g., vocabulary and schema registries), standardization, and community agreement.

Priorities

Given all that needs to be done, the natural question is what first? Based on my experience I would suggest three critical needs in regard to digital library infrastructure:

1. *The need to more clearly define information resource granularity, to understand collection identity, and to appreciate the relationship between a collection and its constituent items.* The decision of IMLS to couple together both a collection registry and an item-level metadata repository in the grant to Illinois was prescient in that it has encouraged us to consider issues of collection identity and descriptive granularity as part of our work. Earlier work done at Illinois under funding from the Andrew W. Mellon Foundation, which included a focus on how EAD might fit with OAI-PMH, and ongoing work being done at Illinois in

collaboration with the U.S. Midwestern research universities that comprise the CIC, also have encouraged study of these issues. Results so far make clear (not surprisingly) that users do not always query indexes or use resources at a single level of granularity. Multi-word queries frequently contain keywords found at different levels of descriptive granularity [Foulonneau et al., 2005]. End-user information needs variously are best met by collections, by items, or by combinations of both. Data gathered in the UIUC-IMLS DCC project confirm that questions of collection definition represent not just a theoretical problem but one that is being contemplated by practitioners and both actively and passively responded to in the daily work of digital library development. We need to understand better how collection identity will be transferred, transformed, and created and how item granularity and interrelationships will be handled in development of the proposed Middle East Digital Library, and how we can exploit agreements on these issues to provide enhanced services to end-users.

2. *The need to understand archetypical behaviors (views) of specific classes of resource objects and to agree on means to obtain standard views of content.* The initial focus of OAI-PMH has been on descriptive metadata, but recent work demonstrates that the protocol has the potential to facilitate as well the sharing of complete resources [Bekaert et al. 2005]. The demand and utility of specific views of resources for use in specialized analytical tools and to facilitate specific DL functions and services is high. Current techniques to crawl or spider information resources directly from Web servers tend to be crude, and the objects gathered by such means are not always in the form or fidelity needed for desired purpose [Foulonneau et al. 2006]. To effectively exploit OAI-PMH to harvest complete resources, we need better and more explicit agreements on the ways resources are disseminated. Consider a complex textual resource made up of ordered page images and OCR'd text. For some purposes an end-user (or digital library tool) might require individual page images. For other purposes, the OCR'd text. For other purposes, scanned pages suitably arranged for page-turner software. There needs to be community agreement on how each of these views (and other potentially useful views of content class members) are labeled, requested, and disseminated.
3. *The need to define expectations.* In discussions about digital libraries we often make the mistake of assuming a common understanding of the functions and services that define a digital library. Indeed, there is some research about digital library defining functions. The IFLA Functional Requirements for Bibliographic Records (FRBR) model [IFLA 1998] defined four user tasks, *Find*, *Identify*, *Select* and *Obtain*, and suggested a possible fifth function: *Relate*. Others [e.g., Svenonius, 2000] have suggested additional functions that digital libraries should perform in support of end-users, e.g., *Navigate* and *Interpret*. However, current functional definitions of digital libraries are inconsistent and vary significantly from domain to domain. Relationships between digital libraries and applications such as institutional repositories, learning management systems, meta-search utilities, traditional OPACs remain hazy. There is not clear, explicit consensus on the essential functions and services that define a generic digital library. One can argue that this is because no complete exemplar yet exists, but nonetheless it behooves us at the start of a project like this one to reach explicit agreements on what we are trying to do. To realize essential technical infrastructure we need to understand how object properties and attributes relate to delivered functions and services. Given that many metadata elements can support multiple functions [Greenberg 201], we need especially to understand how normalization and enrichment done to support one function effects metadata usability for other functions.

Concluding Thoughts

Where does this leave us relative to the question I chose as a title for this brief? The answer I offer is that OAI-PMH is good for sharing descriptive metadata and possibly complete resources. But it is not sufficient onto itself. On its own, it has limited utility, or perhaps none at all. As a layer of a robust, multi-layered infrastructure underpinning a digital library, it has great value. Obviously OAI-PMH enables nothing in the absence of quality content held in well-managed and well-maintained content provider repositories. Perhaps less obviously, the full utility of OAI-PMH can't be realized without community agreement on a host of additional issues -- e.g., collection identity, descriptive granularity, common conventions for dissemination of full content, shared understanding of target digital library functions and services. Agreements, consensus, and technical conventions on such matters are essential to create an adequately robust technical infrastructure on which to build a useful Middle East Digital Library.

References

Bekaert, Jeroen, Xiaoming Liu, and Herbert Van de Sompel. (2005) aDORe, A Modular and Standards-Based Digital Object Repository at the Los Alamos National Laboratory. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital Libraries*. New York: The Association for Computing Machinery, Inc. (ACM): 367.

Foulonneau, Muriel, Timothy W. Cole, Thomas G. Habing, and Sarah L. Shreeves. (2005) Using collection descriptions to enhance an aggregation of harvested item-level metadata. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital Libraries*. New York: The Association for Computing Machinery, Inc. (ACM): 32-41.

Foulonneau, Muriel, Thomas G. Habing, and Timothy W. Cole. (2006) Automated capture of thumbnails and thumbshots for use by metadata aggregation services. *D-Lib Magazine* 12 (1).

Greenberg, Jane. (2001) A quantitative categorical analysis of metadata elements in image-applicable metadata schemas. *Journal of the American Society for Information Science and Technology*, 52(11): 917-924

IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998) *Functional Requirements for Bibliographic Records*. IFLA publications vol. 19. Available: <http://www.ifla.org/VII/s13/frbr/frbr.pdf>

Shreeves, S. L., E. M. Knutson, B. Stvilia, C. L. Palmer, M. B. Twidale, T. W. Cole. (2005) Is 'quality' metadata 'shareable' metadata? The implications of local metadata practices for federated collections. In H.A. Thompson (Ed.) *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, [April 7-10 2005, Minneapolis, MN]*. Chicago, IL: Association of College and Research Libraries, 223-237.

Svenonius, Elaine. (2000) *The intellectual foundation of information organization*. Cambridge: MIT Press.