Digital Library Federation (DLF) Aquifer Project

By: Katherine Kott, Jon Dunn, Martin Halbert, Leslie Johnston, Liz Milewicz, Sarah Shreeves

"Digital Library Federation (DLF) Aquifer Project," (with Katherine Kott, Jon Dunn, Leslie Johnston, Liz Milewicz, and Sarah Shreeves) D-Lib Magazine, May 2006, Volume 12 Number 5. http://www.dlib.org/dlib/may06/kott/05kott.html

Abstract:

In January 2005, the Digital Library Federation (DLF) renewed its commitment to developing a distributed open digital library by assigning a full time project director to lead the initiative. Called DLF Aquifer to symbolize the pooling of content into a community resource, and the piping or siphoning of content to meet specific needs, the collaboration amongst a subset of DLF member libraries has produced standards, reports, and prototypes over the past year. The work was accomplished through four working groups focused on services, technology/architecture, metadata, and collections, with additional participation by library staff at DLF Aquifer institutions. One prototype was created through a collaboration that extended beyond DLF membership. This project update highlights "Phase I" DLF Aquifer deliverables and provides a brief overview of future plans. Additional details, including the DLF Aquifer business plan, working group rosters, and mode of operation can be found on the DLF Aquifer web site. A list of participating institutions is included here, as an appendix.

Keywords: Digital Library Federation | Libraries

Article:

Introduction

In January 2005, the Digital Library Federation (DLF) renewed its commitment to developing a distributed open digital library by assigning a full time project director to lead the initiative. Called DLF Aquifer to symbolize the pooling of content into a community resource, and the piping or siphoning of content to meet specific needs, the collaboration amongst a subset of DLF member libraries has produced standards, reports, and prototypes over the past year. The work was accomplished through four working groups focused on services, technology/architecture, metadata, and collections, with additional participation by library staff at DLF Aquifer institutions. One prototype was created through a collaboration that extended beyond DLF membership. This project update highlights "Phase I" DLF Aquifer deliverables and provides a brief overview of future plans. Additional details, including the DLF Aquifer business plan,

working group rosters, and mode of operation can be found on the DLF Aquifer web site. A list of participating institutions is included here, as an appendix.

External Deliverables

In support of DLF Aquifer goals, the working groups created an array of reports, studies and guidelines that are useful both within the project and for the broader digital library community. In addition, one participant library, the University of Michigan, is hosting a DLF Aquifer "administrative portal" that enables a new level of Open Archives Initiative (OAI) metadata harvesting, and the DLF Aquifer Technology/Architecture Working Group embarked on an innovative experiment to enable a more consistent user experience across diverse image collections. Called "asset action packages," the test included collaborators from outside the DLF: Bill Parod of Northwestern University, and Rob Chavez of Tufts University.

To enable project progress, the Aquifer working groups also produced internal guidelines and policies, including target audience definition, collection scope and architectural principles. Project deliverable descriptions follow, organized by working group.

Services Working Group

The DLF Aquifer Services Working Group (SWG) faced the challenging task of systematically identifying what DLF institutions need and desire in a digital library, defining meaningful and unique services for such a collaborative venture, and in the process articulating what this collaboration would really involve. In short, what would DLF Aquifer be, and how would this differ from other, ubiquitous digital library projects? The DLF Aquifer project should draw on the strengths of the participating institutions and demonstrate the nature of collaboration. Merely producing a "digital library," as evidenced most typically by online search portals, would fail to realize the potential for capturing the essence of libraries in digital form: interoperability, shared standards, and collaboration.

Additionally, DLF institutions already possess considerable resources geared towards meeting the digital research requirements of their own user populations. What would the DLF Aquifer project offer these institutions that they could not already produce or procure for themselves? If the DLF Aquifer project was truly to be a collaborative enterprise, decisions about the services that define DLF Aquifer would have to be made with a clear understanding of DLF institutions' own perceptions of the digital-services landscape, the needs and wants of their users, and the types of services they would value and to which they would contribute in order to further their mission to these users.

The SWG quickly realized that its first task was to gather information on what digital resources and services DLF institutions currently provide, what obstacles hinder their ability to meet users' needs, and what services are needed or desired that they are unable to offer. Simultaneously, this group would consider how the collaborative nature of the DLF Aquifer project could be brought to bear on the development and distribution of these services.

Using interviews with DLF Aquifer participant library user services staff, the Services Working Group designed a survey to gather information about use of digital collections in DLF libraries. Among the 23 DLF institutions that responded to this survey, cross-resource discovery tools emerged as a needed and desired service. Several institutions named this as a gap in their current service offerings, along with helping users personalize their service options. As one respondent put it, "Users want (and expect) much more in the way of 'My Library' features, massively cross-collectional searching." Asked what impedes their ability to provide these and other useful services to their users, institutions resoundingly named budgetary, time, and personnel constraints. Asked to identify ways in which their digital collections were currently used, institutions consistently named searching as a common activity. Currently, most institutions are supporting their digital resources through metadata standardization and other metadata-management activities. The complete report, including anonymized qualitative responses, is available through the Digital Library Federation web site <http://www.diglib.org/aquifer/SWGisrfinal.pdf>.

As part of the interviews and surveys, SWG members asked participants if they could provide studies they had conducted of their own services and resources. Among those studies and reports uncovered during the interview phase of this process, metadata harvesting or finding services were more frequently examined, suggesting general interest in the development of these services. Other services tested or investigated among different user populations were semantic clustering, browsing, and metadata remediation and enhancement, and to a lesser extent, exporting, annotation, and integration with course-management systems.

Studies elicited by the survey were clustered by activity and abstracted in the survey report. A number of studies considered challenges and approaches to navigating and using digital object collections. Faceted browsing and topic maps emerged as desirable services for locating digital objects. These studies also uncovered a range of uses for digital objects, from research to online exhibitions, and the desire among users for more personalized organization and use of digital objects. Several studies identified cross-resource discovery services as desirable tools for locating resources.

Proposed services and next steps

Findings from the survey and from the user studies suggested that, in order to uniquely, usefully, and collaboratively serve the needs of DLF institutions, the services developed for DLF Aquifer should be geared towards the needs and desires of users to search across resources, to locate resources at point of need, and to personalize the tools at their disposal. At the same time, DLF Aquifer services should not simply support the needs of end users but should contribute to the construction of a larger infrastructure, one that would support the collective sharing that Aquifer embodies and specifically would assist institutions with meeting the digital resource demands of their users. Based on the collaborative aims of the Aquifer project, user studies and reports on digital collections and services, and feedback from DLF institutions, the SWG recommended that the DLF Aquifer project concentrate its services in three main areas.

- Developing tools and services that support meta-searching.
- Developing middleware tools that support metadata management.

- Developing tools and services that enable the integration of digital content into course management systems.

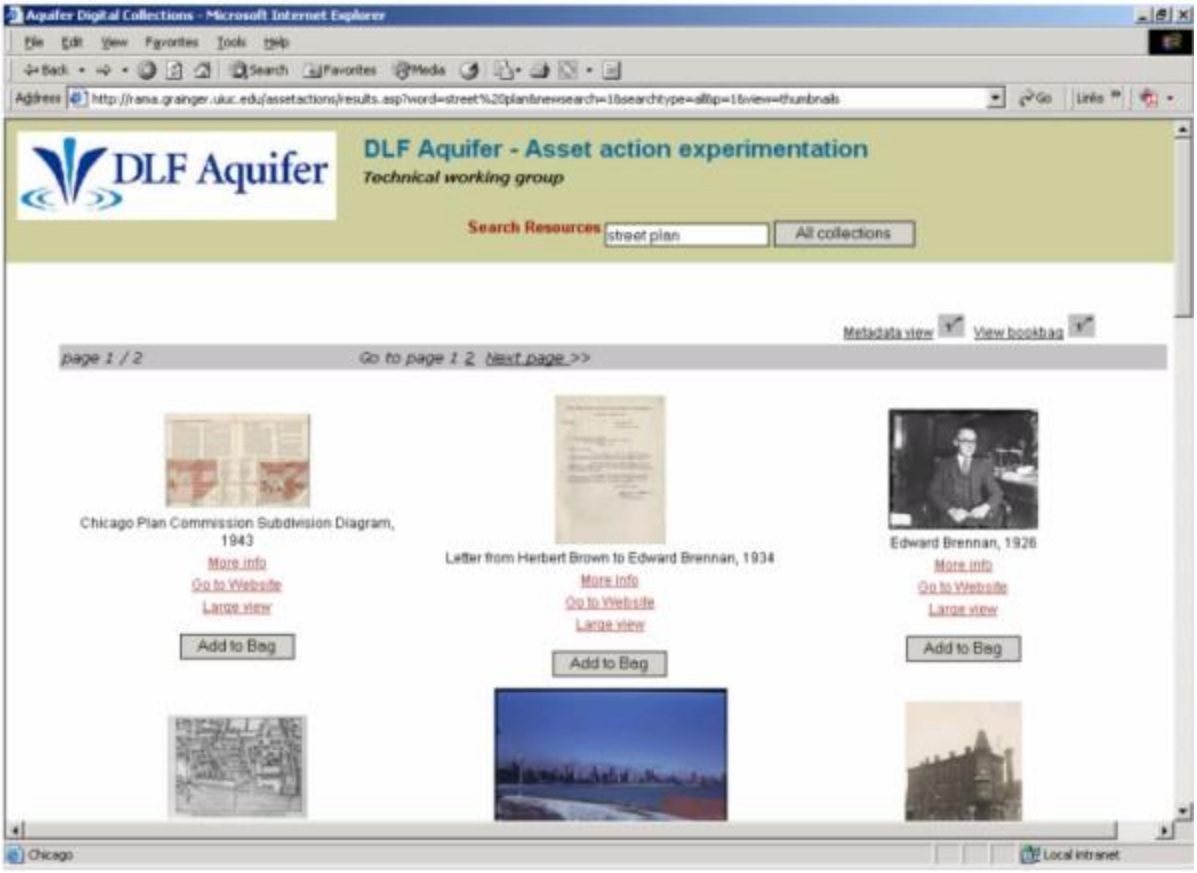
The SWG has currently defined twelve services that address these focal areas and, with the advisement of the Aquifer director and members of the other Aquifer working groups, are prioritizing and creating functional requirements for their development.

Technology/Architecture Working Group

The Technology/Architecture Working Group (TWG) created architectural policies and principles for the initiative, assuming harvested metadata at the core for Phase I, with additional services built outward. Services should be loosely coupled and lightweight, based on open standards, and repository-agnostic. Developed along a spectrum of compliance, services implemented with higher levels of conformity will allow greater interoperability. An early decision to begin project implementation by building on the past success of OAI harvesting, enabled the TWG to evaluate proposals and help select the University of Michigan as the DLF Aquifer metadata harvesting host for Phase I.

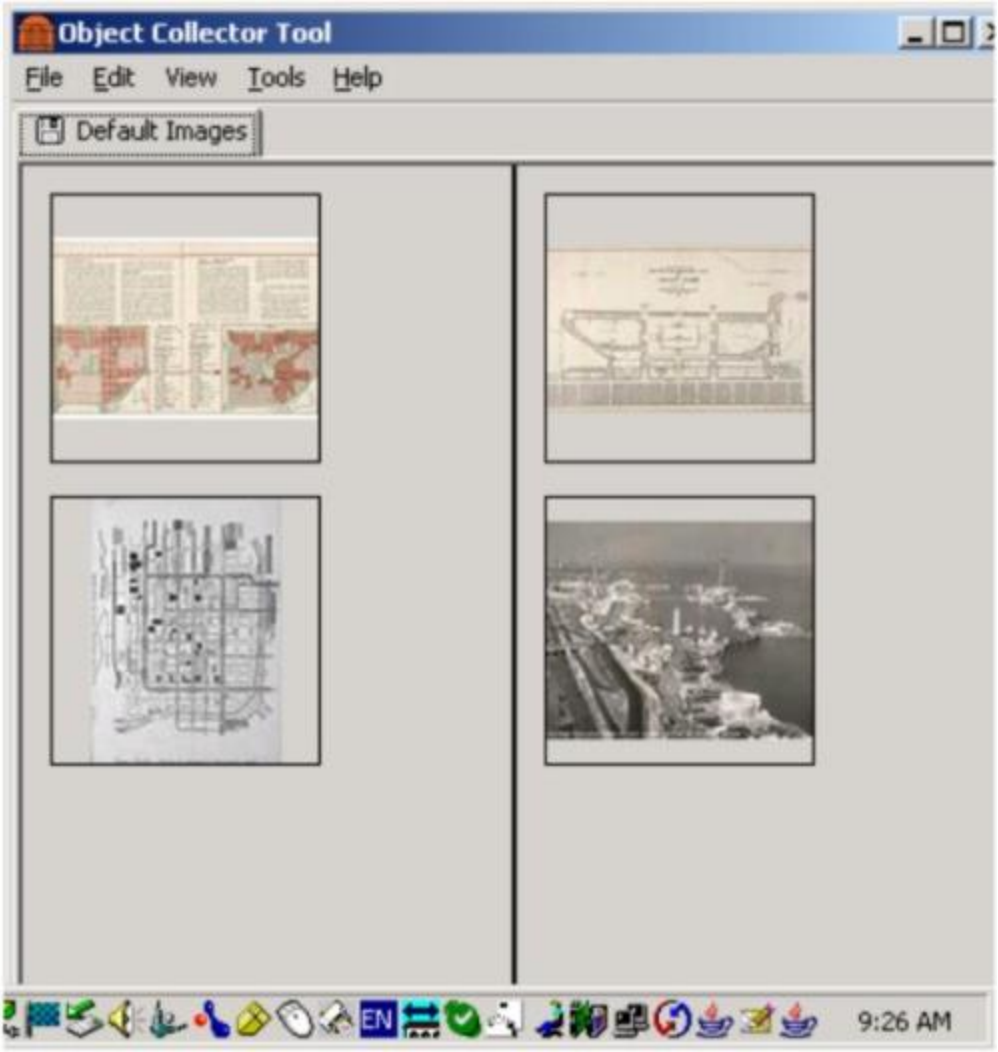
While the TWG awaited direction from the SWG on functional requirements for specific service development to insure project priorities would meet identified need, the clear DLF Aquifer goal to enable "deep sharing" across collections prompted an experiment the TWG named asset action packages. Built on an existing collaboration between the University of Virginia, Northwestern University, and Tufts University, asset action packages are designed to provide a consistent user experience across diverse collections.

An asset action package is an XML-defined set of actionable URIs for a digital resource that delivers named, typed actions for that resource. Packages are made up of action groups, which define behaviors that can be generalized. Every asset action package contains at least a "default" action group, which provides a basic high-level set of behaviors. An asset action experiment hosted by the University of Illinois, Urbana-Champaign, using image collections from Indiana University, Northwestern University, and Tufts University, demonstrates the utility of this lightweight protocol by showing an implementation that enables image thumbnails from multiple distributed collections to be displayed along with brief metadata, in one view, although the images may not exist with these properties in their native environment (see Example 1).



Example 1: Thumbnail view of images from collections at Indiana University, Northwestern University, and Tufts University

The application to the experiment of a Collector tool developed by the University of Virginia effectively demonstrates one of the underlying principles of the DLF Aquifer initiative: that lightweight, low-overhead agreements can be implemented within diverse local technical environments (see Example 2). Further asset action package testing and development will include comparison with unAPI and experiments with textual material, deployed in a variety of technical environments.



Example 2: Images displayed in parallel slide carousels within Collector tool

Although DLF Aquifer is designed to be a loosely coupled set of tools and services to support taking the digital library to the user by enabling local integration, the project requires a portal that can be used for internal evaluation of harvested metadata upon which services are being based. As noted above, the University of Michigan, Digital Library Production Service is hosting the DLF Aquifer administrative portal, which harvests MODS metadata.

Metadata Working Group

Background

In June 2005, the Metadata Working Group (MWG) along with the other working groups proposed a range of general activities for each of the three phases of the DLF Aquifer Initiative. Activities in the first phase were focused specifically on building on the best practices already established for 'shareable' metadata through a related Digital Library Federation activity and, in

particular, establishing a set of best practices and implementation guidelines for use of the Metadata Object Description Standard (MODS) in a shared environment. The proposed activities for the second phase are largely in the realm of post-aggregation activities including metadata remediation and enhancement, as well as investigations that look forward to the sharing of not only metadata but also content. What additional technical, structural, and rights metadata will be necessary when content is shared for re-use and re-purposing? The proposed activities in the third phase investigate the question of user-annotated metadata.

MWG work in progress

In July 2005, the MWG began work on a set of implementation guidelines for MODS records meant to be shared for the DLF Aquifer Initiative. The DLF Aquifer Initiative had determined at the June 2005 meeting that participants in DLF Aquifer would be expected to share MODS records via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). This was not an easy decision; many present acknowledged that their own institutions would not be able to provide MODS records at the start because of software or resource limitations. There was a sense from the participants, however, that DLF Aquifer should be working from already established best practice expressed in the Best Practices for Shareable Metadata, a collaboratively built set of guidelines for OAI data providers (DLF/NSDL 2005); in this case, use of metadata schemas in addition to unqualified Dublin Core. Requiring MODS records would start the DLF Aquifer Initiative with rich, semantically complex records and would avoid the already well-documented challenges of working with unqualified DC (Arms et al. 2003; Hagedorn 2003; Halbert 2003; Shreeves, Kaczmarek, & Cole 2003).

The starting point for the DLF MODS Implementation Guidelines for Cultural Heritage Materials were the MODS User Guidelines maintained by the Library of Congress Network Development and MARC Standards Office and the Best Practices for Shareable Metadata mentioned above (DLF/NSDL 2005; Library of Congress 2005). The MWG purposefully developed the document as a set of guidelines for creating rich, shareable metadata that is useful to aggregators and end users. While the MWG did not intend the guidelines to dictate local metadata practices, the group does hope that these guidelines will provide a best practice on which to build as institutions both within and outside of the Digital Library Federation implement MODS in their digital environments and share these records with others.

The Implementation Guidelines were released for public comment and review in late November 2005 for approximately two months (DLF Aquifer Metadata Working Group 2005). Thirteen comments were received back from a range of institutions from within as well as external to the DLF. Using a wiki hosted by Indiana University, the MWG reviewed and categorized the comments. In many cases a small change was needed. The MWG is still resolving several larger issues raised by reviewers and from members of the MWG. These issues were recently discussed at a Birds of a Feather session held on April 11, 2006, at the DLF Spring Forum in Austin, TX. The two major issues discussed were:

1. The requirement of one and only one <location><url> pair for the main portion of the record (i.e. multiple <location><url> pairs could be used within the <relatedItem> element). This issue in particular reflects a tension between the needs of a service provider or aggregator who needs

to determine both which URL to point a user to and what that URL resolves to, and the needs of a data provider who often has multiple URLs for a digital object for a multitude of reasons (multi-volumes, access to the jpg and a tiff file, et cetera). This issue may very well be resolved by a combination of factors including the proposed revisions to the MODS schema, particularly the introduction of an access attribute to the <url> subelement that would allow the data provider to specify what the URL was pointing to. The asset action package work conducted by the Technical Architecture Working Group has the potential to resolve the ambiguity inherent in the current state of the <url> subelement.

2. How and where to describe the digital surrogate and the analog original, or how to present information about the content of a resource and its carrier(s). In general, the draft guidelines attempted to distinguish between description of the intellectual content and genre of a resource and its digital format and location and made use of the <relatedItem> element to further describe the analog original of a digital object. The MWG acknowledges that the specific recommendations for some elements deviated from the original intention. Many of the reviewer comments included suggestions for how to change the approach taken; there were in total no fewer than four different approaches suggested! The MWG worked through multiple approaches, presented these at the Birds of a Feather session, and solicited comments and feedback from the participants there.

MWG next steps

The obvious next step for the MWG is to complete the work on the DLF MODS Implementation Guidelines for Cultural Heritage Materials using the comments received during the formal comment period, as well as those received during the Birds of a Feather session. The group hopes to complete this work by June 2006.

After the Guidelines are complete, the MWG will develop tiers of adoption levels to these guidelines and will work with the Services Working Group to explore tying the tiers of adoption levels to services that can be made available (i.e., the more tightly the metadata conforms to the guidelines, the more services can be made available with those records). In conjunction with these activities, the MWG will be developing modified crosswalks – particularly the MARC to MODS crosswalk – to reflect the implementation guidelines. Finally the MWG will be looking forward to the next phase of activities as described above.

Collections Working Group

In May 2005, the Collections Working Group began development of a collections development policy for the DLF Aquifer project. Approved in July 2005, the DLF Aquifer Collections Development Policy identifies the following criteria for inclusion:

In its prototype Phase 1, the materials aggregated into the DLF Aquifer Collection should:

- Fall into the subject scope of American history, society and culture;
- Serve research and instructional needs of libraries and of scholars in the humanities and social sciences;

- Represent collections of all media types (e.g., images, text, video, audio, etc.);
- Come from a selected subset of Digital Library Federation member institutions' local or aggregated collections – content quality is assumed given prior selection and digitization by the participating institutions;
- Be available for OAI harvesting, complying with both DLF OAI best practices and DLF Aquifer metadata standards;
- Include a persistent URL that leads to a publicly available digital object; and
- Include assurances from the hosting institutions that the content will remain available for the foreseeable future.

The policy also outlines criteria for later phases of the project, along with describing the acquisitions process and documenting the policy for the reconsideration of collections.

In July 2005, the DLF Aquifer Collections Working Group surveyed collections that were currently available via the OAI protocol from participating DLF Aquifer institutions. The goal was to identify sets that contained items that fit the scope of American history, society and culture. The group reviewed sets, limited in this phase to DLF member institutions, that were included in a number of aggregating resources:

- The American West and American South subject-based projects;
- The DLF Collections Registry;
- The prototype DLF OAI Portal;
- OAIster;
- The Experimental OAI Registry at the University of Illinois, Urbana-Champaign; and
- Sets that were proposed by DLF Aquifer or DLF member institutions.

Over 350,000 records in more than 100 sets were identified from fifteen institutions, potentially representing many more individual items. Not all will be included, as many records may lack the required links to digital objects.

The range of materials include scholarly articles, historical journals and newspapers, census data, geological survey data, historic photographs, photographic cultural documentation, aerial photography, art objects, natural history collections, posters and broadsides, photographic architectural documentation, historic maps, personal papers and institutional archives, sheet music, literature, historical texts, and video recordings of scholarly presentations.

In the fall of 2005, it was decided to narrow the focus of Phase 1 of DLF Aquifer to materials related to the American Civil War and Abraham Lincoln. A task group comprised of members from the Collections and Metadata working groups reviewed the full, proposed collections list and narrowed down the list to sets that contained materials that fit the topics. Seven institutions were invited to contribute eleven sets:

- Over 12,000 works of American literature, history, religion and science from three sets.
- 2,240 Civil War maps and charts and 76 atlases and sketchbooks.
- Approximately 24,000 pieces of sheet music, songbooks, and folios.

- Over 400 Civil War era broadsides, photographs, printed works, maps, Confederate currency, and manuscript letters and diaries from two sets.
- Political satires, caricatures, and allegories printed in the U.S., ca. 1766-1876.
- The complete run of an anti-slavery journal from the 1820s.
- Over 6,000 Civil War photographs and 4,000 panoramic photographs from two sets.

While the work of identifying collections was progressing, so was the work to identify the metadata requirements for participation, eventually resulting in the MODS Guidelines for Cultural Heritage Materials noted in the Metadata Working Group activity report, above. The majority of the collections identified for participation were not yet available via OAI in the MODS format, so a subset comprised of five sets from two institutions were included in the spring 2006 Phase 1 proof-of-concept.

A number of issues were raised during these focusing and collecting efforts:

- An institution's OAI sets often correspond to specific physical collections or discrete projects. When harvesting efforts are aimed at collecting around a specific subject, how are records filtered out of larger sets?
- The DLF Aquifer Project has set a high bar for its shareable metadata standard. If an institution cannot fully meet that standard, can harvested metadata for an OAI set be programmatically remediated after harvesting? Will there be more than one level of "compliance" for the metadata standard?
- There is a very strong focus on the development of tools on top of the DLF Aquifer collections in a distributed environment, where many institutions might be creating tools that take advantage of the collections. What does it mean to contribute materials to DLF Aquifer? What sorts of participation agreements might be needed?

The collection development work is continuing to move forward, with the development of participation agreements, an analysis of the potential barriers to participation, documentation of the collection review process, and the review of the Collection Development Policy in light of the work accomplished since it was written.

Aquifer Future Plans

In recent months, two more DLF libraries have joined the DLF Aquifer initiative, bringing the number of participant libraries from 12 to 14. The initiative is ambitious and will continue to make significant progress on issues of interest to the digital library community, with DLF support and DLF Aquifer participant effort. An infusion of outside funding would enable more rapid development in the areas outlined in the working group reports. Examples of specific sub-projects on the DLF Aquifer to-do list are:

Metadata

- Develop, adopt or adapt a rights expression syntax
- Generalize the date normalization tool developed by the California Digital Library

- Develop additional metadata enhancement tools and workflow, based on analysis of harvesting experiments

Service development

- Implement a "collecting" service that can be implemented in a variety of local technology environments
- Extend asset action packages to text

The Digital Library Federation and DLF participant libraries will also devote attention to fostering and brokering collaboration within the DLF and beyond, continuous innovation, identifying which experiments to discard and which to sustain, and creating models for sustainability. As the initiative develops, the business plan that was written more than a year ago will need to be adjusted and participants will continue to refine answers to strategic and tactical questions, such as:

- "How much of the initiative is about the collections?"
- "What level of resources should be devoted to building on OAI?"
- "What other experiments in federation are important to try?"
- "What is the relationship of DLF Aquifer to other digital library initiatives?"

Based on community response to requests for comments on the MODS Implementation Guidelines, broad engagement in the institutional survey results, and the interest asset action packages, the DLF Aquifer initiative is making a contribution to the community. The DLF Aquifer director as well as working group chairs and participants welcome questions and comments to insure that the initiative continues to complement and leverage other work in the field.

Acknowledgements

DLF Aquifer working groups gratefully acknowledge the contributions of Muriel Foulonneau and Tom Habing, University of Illinois, Urbana-Champaign; and Kat Hagedorn and Qian Liao, University of Michigan.

Appendix: Participating Institutions

California Digital Library
Cornell University (joined in January, 2006)
Emory University
Indiana University
Johns Hopkins University
Library of Congress
New York University
Stanford University
University of Illinois, Urbana-Champaign
University of Michigan

University of Minnesota
University of Southern California (joined in March, 2006)
University of Tennessee, Knoxville
University of Virginia

References

Arms, W.Y., Dushay N., Fulker, D. & Lagoze, C. (2003). A case study in metadata harvesting: the NSDL. *Library Hi Tech* 21(2), 228-237. <doi:10.1108/07378830310479866>.

Digital Library Federation and National Science Digital Library. (2005) *Draft Best Practices for Shareable Metadata*. <http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?PublicTOC>.

DLF Aquifer: Bringing collections to life through the Digital Library Federation. <http://www.diglib.org/aquifer/>.

DLF Aquifer Collections Working Group. (2005). *Aquifer collections development policy*. <http://www.diglib.org/aquifer/Aquifer_CollDevPol_03rev.pdf>.

DLF Aquifer Metadata Working Group. (2005). *MODS implementation guidelines for cultural heritage materials: Draft for public comment*. <http://www.diglib.org/aquifer/DLF_MODS_ImpGuidelines_ver4.pdf>.

DLF Aquifer Services Working Group. (2006). *Aquifer Services Working Group report on the institutional user-services survey report*. <http://www.diglib.org/aquifer/SWGisrfinal.pdf>.

Hagedorn, Kat. (2003). OAIster: a "no dead ends" OAI service provider. *Library Hi Tech* 21(2), 170-181. <doi:10.1108/07378830310479811>.

Halbert, M. (2003). The Metascholar Initiative: AmericanSouth.Org and MetaArchive.Org. *Library Hi Tech* 21(2), 182-198. <doi:10.1108/07378830310479820>.

Library of Congress. Network Development and MARC Standards Office. (2005) *MODS User Guidelines Version 3*. <http://www.loc.gov/standards/mods/v3/mods-userguide.html>.

Shreeves, S.L., Kaczmarek, J.S., & Cole, T.W. (2003). Harvesting cultural heritage metadata using the OAI protocol. *Library Hi Tech* 21(2), 159-169. <doi:10.1108/07378830310479802>.