Proposal for an IMLS 2-year (10/04-9/06) research track project

The Distributed Library: OAI for Digital Library Aggregation

National Impact

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has proven itself as a straightforward and functional mechanism for sharing metadata across systems for the purpose of building focused services. We have collectively proven the protocol and its utility; however, as Martha Brogan notes, "there are numerous practical, technical and philosophical impediments to the full realization of OAI-based services…." The multi-stranded research proposed in this grant will address these various impediments, and move us closer to the goal of federated collections across institutions, and the ability to create richer services for our library users.

This issue of a "distributed library" made up of multiple collections in a much more malleable, easily accessed "service landscape" is of sufficient importance to the major research libraries who make up the DLF that – after a strategic planning session and eight months of discussion and planning (Feb-Oct 2003) – we voted unanimously to make the creation and sustenance of a distributed, open, digital library the organization's overarching strategic goal, taking us full circle to the first item on our founding 1995 charter. This endeavor, we believe, have far-reaching impacts inside and outside the DLF institutions, and the work done during this grant will be of fundamental importance to our ambitions for a comprehensive finding system.

Our team of partner organizations draws on very active OAI practitioners: the University of Michigan will host an experimental harvesting service and metadata search portal for the purposes of the grant, with the harvesting service functioning as a test-bed for our research and as an encouragement for broad participation within DLF; UIUC will develop a registry to allow closer relationships between harvesters and providers; Emory will oversee the training and consultancy portion of our work, and we expect to see their thriving and established *American South* collaboratory as a major subject-focused "proving ground" during the IMLS grant period.

Workflow and training

One impact will be to learn how to ingrain the creation of item-level metadata into the digitizing workflows of the DLF – we are keen to know how to do this and will look to the workflow evaluation and subsequent training provided here by Emory to do this. The barriers to the routine provision of sharable metadata are – we believe – not substantial, but they are real and not clearly understood. By canvassing the DLF member institutions we expect to learn a lot more about what changes in local day-to-day practices and our collective sense of mission need to occur for us to share digital object metadata readily, and start to rely on its existence in our local service provision. Our ambitions for richer "deep

sharing" and more tailored local library services have at their foundation the need for a trustworthy and well-populated finding system for our distributed content.

Enriching the OAI record

The ability to examine the form and use of collection-level OAI records, and to build on work currently being funded at Illinois, is another important area, as is the question of when we need a metadata "base" that is richer than the "unqualified Dublin Core" mandated by OAI. We feel confident that digital library providers of records for digital library aggregations and services will be willing to follow rather more nuanced and prescriptive Best Practices for the creation of the "digital library" OAI record; indeed, we are typically avaricious for such guidance and advice, as the success of past Best Practice guides from DLF, RLG, IMLS, and elsewhere attests. This IMLS grant will cull the lessons learned from the first wave of harvesting services and reflect back our collective observations with regard to the formation of the catalog records. Early harvesters have spent inordinate amounts of time normalizing and completing the records in order to build the services they wish to provide, and we must improve that situation if a distributed finding system is going to scale and thrive.

Coordinating Services and Providers

In her conclusions to her *Survey of Digital Library Aggregation Services*, Martha Brogan highlights the lack of any comprehensive registry of OAI service and metadata providers, "it is difficult (at best) for users to know the extent of services available." Recent work at the University of Illinois and within the European Union, along with tools like Virginia Tech's "Repository Explorer," are helpful in this regard, but clearly more research is needed to enable OAI service providers to better discover and learn about OAI metadata providers (and vice-versa).

As part of the proposed research and using as a starting point the experimental registry of OAI metadata providers already developed by the University of Illinois at Urbana-Champaign (UIUC) (http://oai.grainger.uiuc.edu/registry), we propose to investigate and experiment with potential methods to enable better discovery of OAI service and metadata providers. The outcome will have positive national and international impact on how OAI service and metadata providers discover each other and learn about available metadata and services.

Adaptability

OAI PMH is based fundamentally on omnipresent standards and protocols such as HTTP, XML, and the Dublin Core metadata schema. It has proven itself as a functional and stable protocol, and is increasingly widely used. What we provide by way of a prototype registry, or services crafted with deep input from users, or guidance on the creation of records more amenable to use in digital library services, will be usable far beyond the confines of the DLF institutions who

comprise the test-bed collections and user communities. The ultimate impact, we believe, will be that many other libraries will follow the example of the DLF and its recommendations for OAI metadata tailored for digital library services; there is a strong history of DLF recommendations and best practices having a wider influence than the DLF membership. As with other project components, software and schemas developed as part of the upgrading of the UIUC registry of OAI metadata providers will be made available under OpenSource license. This work will exploit collection level description being done by the UIUC IMLS Digital Collections and Content project, the UK's RSLP (and related) projects, and the Dublin Core Collection Description Working Group. Schemas developed as part of the OAI registry component enhancement work will be especially adaptable for use in sharing information about collections generally and OAI metadata providers in particular.

Design

Context. DLF has been actively engaged with the OAI community since its beginning, providing some of the funding for the early work (with CNI) and through its members having an early commitment to providing OAI records and harvesting services. California Digital Library, Carnegie Mellon University, Cornell University, University of Pennsylvania, UIUC, University of Tennessee, Knoxville, UC Berkeley, MIT, the University of Virginia, Michigan, Emory, Indiana, Chicago, the Library of Congress, the University of Washington, and NCSU all show up as providers of OAI records currently, and as the letters of support attached to this proposal demonstrate, other DLF institutions are committed to much deeper engagement with sharable metadata and the “next-generation” services they allow.

Audience. The material to hand most easily for this research and demonstration project is that which has already been digitized in DLF libraries. Typically, to date, this biases heavily towards text and images, and predominantly in the humanities and social sciences (so much so, in fact, that in our discussions of the DLF distributed library we have raised again the notion of using it as the seed for a national library for the humanities). Brockman *et al,* in the abstract to *Scholarly Work in the Humanities and the Evolving Information Environment*, have reminded us “in particular, the findings emphasize how important it is for libraries to chart their evolutionary course in close consultation with scholarly user communities.” Working with our scholarly advisory group (see below) we plan to extract out of the larger collection of records some area of areas of subject focus – The American South and/or American literature would be likely candidates. This builds on existing communities of interest and our own sense of what is in our collections and – if at all possible – would allow us to extend our examination of how scholars and teachers use a metadata-based finding system into some ongoing research and classroom projects. The DLF collections registry, currently being updated with material up to December 2003, will be of considerable aid in this work (http://www.diglib.org/pubs/techreps.htm).

Assessment of needs. From their own work and that of close colleagues, the participants from Illinois, Emory, and Michigan have a firm grasp of what relevant work has been done to evaluate OAI-based services, and what is going on to improve them (sample publications of theirs are included in the *References* list).

In addition, the DLF commissioned Martha Brogan to undertake a *Survey of Digital Library Aggregation Services*, which provided an overview of a diverse set of more than thirty digital library aggregation services, organizing them into functional clusters, and then evaluating them from the perspective of an informed user. Most of the services under review rely wholly or partially on the Protocol for Metadata Harvesting of the Open Archives Initiative (OAI-PMH), and the conclusions of this study are ones that will guide part of our research. Her findings urge cautious optimism but show plenty of room for major improvement in precisely the areas we concentrate on here.

Most important is the current IMLS-funded work at Illinois, *Digital Collections and Content (DCC)*, a three-year effort at the University of Illinois to build a national infrastructure for adaptable, interoperable, and sustainable digital collections, which uses OAI-PMH to harvest metadata from current and past National Leadership Grant (NLG) awardees with digital collections. The current DLF submittal to the IMLS NLG program proposes explicitly to build on and extend important facets of the current University of Illinois project. Specifically, this new and deliberately complementary proposal focuses on a more homogeneous universe of digital collections and institutions, which will allow the DLF researchers to go further in developing more advanced standards, recommended best practices, and training approaches expressly tailored for academic research libraries. Development of these standards will build on research undertaken on our broader based project at Illinois into metadata schema implementations and authoring practices within the community of IMLS grantees, and will inform and be informed by the ongoing involvement of many DLF members in the evolution of metadata standards like Dublin Core, MODS, EAD, and METS. This work promises to further advance and prove in yet another domain the usefulness of OAI as a technology for sharing descriptive metadata.

The work proposed by DLF also will extend and make more valuable the work being done at Illinois on collection-level description. The DLF proposal promises to take the collection-level description best practices developed at Illinois in support of the collection registry component of its project and identify a range of standard ways in which such rich collection-level description can be included in OAI metadata provider services (e.g., through use of more complex, multi-faceted metadata schemes such as METS). This is the logical next step towards more closely coupling technologies used for the dissemination and management of item-level metadata and collection-level descriptive metadata, and is a necessary prerequisite towards merging human-generated metadata about collections with descriptive metadata derived by automated text mining. More seamless joining of collection-level description, item-level description, and automated full-text retrieval technologies is essential in the long run to enable capable end-user portals to our digital libraries.

Implementation

Based on this range of experience, knowledge, and expertise, we will undertake this research and demonstration by performing the following tasks, expressed here in four 6-month blocks. More details on the work of the individual partners can be found in *Appendices I, II,* and *III.*

Phase One -- months 1-6:

- set up technical components for OAI-PMH harvesting [Michigan team]
- harvest all current available data from DLF institutions [Michigan team]
- convene *DLF OAI Best Practices team* to draft documents of best practices for OAI records and collection descriptions, drawing heavily on the collective experience of DLF members who have already created OAI harvesting services and are actively researching the next steps, including Illinois, Michigan, Emory, and OCLC
- establish the research project's *Scholarly Advisory Group*, and provide them with the means and encouragement to inform and challenge our decisions and assumptions throughout the project. Note: prior to the beginning of the grant, the DLF will have convened a meeting of scholars who are deeply involved in digital projects that draw heavily on digital library resources; we expect that the Scholarly Advisory Group will grow out of that focus group we will establish in the spring of 2004.
- work on export of registry records [Illinois Team]
- experiment with addition of classification attributes to registry records [Illinois Team]

Phase Two -- months 7-12:

- test the harvesting mechanisms and resolve any remaining technical problems [Michigan team]
- plan the training needed to ease the transition of DLF institutions to the provision of OAI metadata records for harvesting [Emory team]
- Provide draft versions of best practices documents for the creation of new OAI records that are explicitly tailored for digital library services [DLF OAI Best Practices team]
- Convene a DLF Finding System Research team, drawing from across the DLF institutions; their first task will be to identify tools used for creation of OAI records at local institutions, in order to inform the training and guidance provided to new providers
- sample metadata records to look for ways to enrich registry records [Illinois Team]
- implement experimental search features that use collection-level description metadata [Illinois Team]
- Expect to be working with some new providers by this point

- Commission Martha Brogan to review and revisit the work she did in 2003 for the DLF in surveying digital library aggregation services, in part to inform the service design in the second half of the grant

Phase Three -- months 13-18:

- Expect to see the bulk of new OAI providers by this phase of the grant
- Through significant use of the Scholarly Advisory Group, design a functional OAI-based finding system that will form the initial prototype for the DLF's distributed library
- Work with the Scholarly Advisory Group to uncover focused collections within the mass of the aggregated records, and identify individual scholars and/or teachers who will use that material in their work.
- explore what level of normalization of the harvested data is needed [DLF Finding System Research Group]
- experiment with exposure of the records to Google, and report on the results [Michigan team]
- experiment with the use of OCLC web services tools for name authority provision, and report on the results [Michigan team]
- research techniques to automatically characterize metadata records available from given OAI providers [Illinois Team]
- expand current registry to include OAI service providers [Illinois Team]

Phase Four -- months 19-24:

- experimentation with existing portals such as Oaister, Internet Scout portal, iVia [Michigan]
- Explore the collection development potentials of the prototype [All]
- Gather feedback from across the IMLS-funded and DLF-funded teams that have contributed to this research and prototyping [All]
- Revision, documentation, and promotion of results [All]
- Liaise with the DLF teams (and others) that will be moving forwards after this research to implement a finding system informed by our research [All]

Management

The Principal Investigator, together with Co-Principal Investigators, shall provide the overall direction of the work undertaken here.

Principal Investigator David Seaman will provide coordination and oversight for the whole project.

The three Co-Principal Investigators -- one each at partner institutions Emory, Michigan, and UIUC -- have all been involved in successful digital library projects for funding agencies including NEH, IMLS and the National Science Foundation. The University of Michigan Library will host an experimental OAI metadata harvesting service and metadata discovery portal(s) for the aggregation of DLF-

member contributed metadata. This harvesting service and associated portal(s) will serve as a test bed for many of the research investigations outlined in the proposal narrative. UIUC's Grainger Library will host the collection registry and metadata repository services created as part of this project. Grainger Library currently hosts both the Illinois OAI Metadata Harvesting project and the TDC project.

Budget Narrative

The largest part of the budget is for personnel. The staff time to be devoted to the activities described above is based on the past experiences of the investigators. Details regarding allocation of personnel time and other expenses are provided in the budget notes.

Contributions

The DLF requests no funds for its central administration, but only to its member institutions who are partners in this grant: Michigan, UIUC, and Emory. This significant cost share is a tangible expression of how important this OAI research is to the DLF as a whole, and we believe to the much wider library profession.

As a cost-share contribution, DLF will provide significant oversight and coordination, including the close management of long-distance collaborative work (with which we have much experience): we will pay for the necessary consultants, cover travel costs to allow the project teams to meet in person twice a year, convene and fund the scholarly and technical advising groups, and pay all costs of the publications to come from the work.

The University of Michigan is contributing 10% of Perry Willett's time to manage and insure timely completion of the Michigan part of the proposal. The UIUC Library is contributing 5% of Thomas G. Habing's time and at least $3,000 towards purchase of a server. Emory is contributing $3,000 in travel funds and $1,288 in indirect costs.

Personnel

Digital Library Federation, Council on Library and Information Resources
David Seaman (Principal Investigator): has previously run grants from the Andrew Mellon Foundation, NSF (international collaboration grant with the Deutsche Forschungsgemeinschaft), NEH (a national challenge grant), and has been a contributing partner on an IMLS grant (*The Philip S. Hench/Walter Reed Yellow Fever Collection*. He has taught, lectured, and written frequently for the past twelve years on various aspects of digital library and humanities computing, in particular the use of SGML and XML texts in large-scale digital library aggregations.

University of Michigan
Perry Willett (Co-Principal Investigator): the new Head of the Digital Library Production Service, University of Michigan, Perry Willett was before that Assistant Director for Projects and Services, Digital Library Program (DLP), Indiana University and Head of the Library Electronic Text Resource Service (LETRS)
Kat Hagedorn: Work on metadata standards, interface design and reporting on current practices will be managed at the University of Michigan by Kat Hagedorn, the Metadata Harvesting Librarian at the University of Michigan. She managed the OAI project funded by the Andrew W. Mellon Foundation at the University of Michigan, OAIster. Her work on OAIster and DLXS will provide the basis for the development of the new service envisioned in this proposal.
José Blanco: Experimentation with different search engines, and work on technical issues involved with implementing new metadata standards for OAI harvesters will be managed by José Blanco, Senior Programmer/Analyst in the Digital Library Production Service at the University of Michigan. He has worked extensively with OAI, and provides extensive technical expertise that will be required in this project.

University of Illinois
Thomas G. Habing (Co-Principal Investigator): Work on the experimental registry of OAI service and metadata providers will be managed at the University of Illinois by Thomas G. Habing and carried out by a library school graduate research assistant under his direction. Mr. Habing has been active as a developer of OAI-related software and services since the alpha testing phase of the protocol. He is a co-author of an XML schema for Qualified Dublin Core and the creator of the UIUC experimental registry of OAI metadata providers (http://oai.grainger.uiuc.edu/registry).

Emory University
Martin Halbert (Co-Principal Investigator): Martin Halbert has been Director for Library Systems at the Emory University General Libraries since 1996, and has extensive experience in planning and coordinating metadata harvesting projects. He is currently principal investigator and executive director for the projects of the *MetaScholar Initiative* (http://www.metascholar.org). A recognized authority on metadata harvesting services, he has spoken on the topic at a number of national and international conferences.

Project Evaluation

Two advisory committees will be formed early in this research project to assist with ongoing evaluation and assessment and to help ensure usefulness and sustainability potential of this work.

The first, to be chaired by project PI David Seaman, will be comprised of selected staff from DLF member libraries and digital projects. As representative of librarians and content providers, this advisory committee will help set project

priorities and research agendas and will provide guidance on the resolution of high-level project issues. This first committee also will assess, evaluate, and provide ongoing qualitative feedback regarding what the research results tell us about the potential uses of metadata repositories to support the provision of advanced library services and their effectiveness as utilities for resource sharing and digital collection interoperability. Timothy W. Cole (University of Illinois at UC), Martin Halbert (Emory University), and Perry Willet (University of Michigan) have all committed to serve on this first advisory committee should this proposal be funded. Additional members will be selected at project start to fill out this group. Members will be selected for their familiarity with digital library systems and interoperability issues, and to represent a diverse set of subject domains and digital information resource knowledge. We anticipate that this committee will meet face-to-face twice annually during the project, and will confer by email and conference calls as necessary at other times during the project.

A second advisory committee will be comprised of approximately 8 DLF member institution teaching faculty who are well positioned to speak to scholarly user needs and interests, across all levels of university constituencies (i.e., undergraduate, graduate, faculty, and staff). This committee as well will advise on project priorities and agendas on an ongoing basis, but this second group also will be especially well-positioned to comment from an end-user perspective on outcomes of prototyping activities and usability testing of experimental systems and services developed. This scholarly advisory committee will serve not only as a key resource for evaluation but also as an aid to dissemination of results and as a group that can help facilitate sustainability and transformation of experimental services developed here into the longer term DLF distributed library. In this role they will convene an end-of-project conference to disseminate and research results and solicit user community input into next steps. The scholarly advisory committee will also contribute substantial subject expertise during the content selection/collection development phase of the project.

Dissemination

The Digital Library Federation has both its semi-annual Forums and its listserv, Dlf-announce, for dissemination, and we will be publishing several reports out of this project. In addition, we will actively seek out appropriate electronic forums, such as listservs and online discussion lists, in which to alert the wider library and museum communities of the progress of this project. The participants fully intend to generate journal articles, conference papers, and avail themselves to professional presentations designed to disseminate the findings of the project.

Sustainability

The ultimate goal of creating a finding system for DLF holdings is to research the usefulness and viability of sharing collection-level and item-level metadata in the context of digitization projects within DLF. The work we propose will accomplish

those goals and will lay the foundation for further exploitation of these technologies and approaches.

All software, documentation, training modules, and best practice recommendations developed for the distributed library service will be publicly available.

With regard to item-level metadata sharing, a primary objective of the proposed work is to engender a commitment on the part of DLF institutions and others to implement and maintain metadata provider services so that metadata may be harvested not only by other DLF members but by all such interested parties.

The DLF is committed to long-term research, creation, and support of a range of elements that will go to make up the open, distributed library that we are ambitious for, with the richer services and better scholarship that will engender.

Conclusion

Recently, the DLF's Steering Committee has unanimously renewed the organization's original commitment to an open, distributed, digital library, which means that we start this research at a moment when the DLF membership is prepared to commit to build more OAI records, fund related initiatives, and to implement and maintain a permanent finding system (a prototype of which this grant will provide); it also means that we are avaricious for research and demonstration that -- for example -- encourages new behaviors such as the routine provision of publicly-available item-level metadata records as part of our daily production processes, and that provides a richer knowledge of how users expect finding systems to behave.

DLF members are committed to developing the prototype into a large-scale, long-term tool for discovery and re-use of their rich but scattered digital holdings. This in turn, we hope, quickly expands beyond the DLF to much wider participation, and we will do everything we can to promote, publicize, and empower larger distributed libraries.

During this research period we will coordinate closely with our colleagues in other large-scale library aggregations, especially the National Science Digital Library (NSDL), and with ongoing research at UIUC. Already we see ways in which the practical experimentation undertaken here will inform the emerging discussions about national cyber-infrastructures for both the sciences and the humanities, and will contribute to the discussions about a National Digital Library for the Humanities.