

Next Generation Digital Federations: Adding Value through Collection Evaluation, Metadata Relations, and Strategic Scaling

1. Assessment of Need

As efforts to integrate and federate digital resources proceed apace, we are learning more about the problems that emerge at different levels of scale and granularity. Building on prior work of the Digital Collections and Content project (DCC), we propose to investigate and implement a systematic approach that confronts these problems and offers robust means for adding value and improving access to existing digital aggregations. Based on over four years of development and research experience, we have identified new core initiatives that can substantially upgrade the quality of the IMLS digital collection registry and metadata repository for users and advance the current base of knowledge and practice in the federation of digital collections. Our work to date has positioned us to move into a more highly informed level of development and theoretical work as we continue to test and ground advances in the engineering pragmatics of building, learning, and documenting.

This work is timely as it aims to leverage the progress made by the many dozens of pioneering digital projects sponsored by IMLS in the recent past to address new challenges of maturity. That is, as a result of the earlier projects' successes in proof of concept and widening access, new and difficult problems have emerged related to scaling and granularity of content, and evolving usage and user expectations. Individual digitization projects vary in scale, scope, and technical complexity. The choice of the process of digitization, the metadata scheme or schemes used, and the amount of detail and access provided for items is a product of complex interactions among the evolving state of the art, best practices, available skill-sets, funding, resources, and project priorities. Federation raises many issues of interoperability of content, as our prior work has shown. But a larger, overarching concern still prevails: how to systematically build networks of digital content in ways that ensure they will be useful, meaningful, usable, and indeed used. To get the cumulative value out of many years of collective investment in generating digital content, we must now improve how people can work with expansive and potentially complementary stores of digital resources.

Federation of digital collections is a viable strategy for increasing the value and use of the growing body of digital content. And, while federated collections can conceivably be built to offer more than the sum of their parts, these aggregations may also lose important context and meaning inherent in individual collections. Even in the seemingly flattened federations of research papers already provided by digital libraries, repositories, Google Scholar etc., collection and collection-like information remains highly useful. Individual journal papers may be distinguished by their inclusion in a special issue, or a technical report's contribution may only resonate in relation to a series of related working papers from a particular research lab. Such information is about the 'collection' that the paper comes from, and this kind of context is arguably even more essential for interpretation of digitized artifacts from highly purposeful library, archive, and museum collections.

The key to preserving this context, we believe, lies in collection-level representation and in building relationships between collections and items that not only retain context but enhance functionality and usefulness. Of course, many kinds of people use digital content for many different reasons. Designing for this diversity is extremely challenging, particularly as uses and kinds of users keep changing. In this project, we will concentrate on the needs of scholars for two reasons. Our previous studies of the DCC collection registry and metadata repository (hereafter referred to as the IMLS DCC) established that scholarly users are one of the largest and fastest growing target audiences for collections in the registry, and our project team has expertise and a track record in this increasingly important area of user studies. Our definition of scholar, however, is quite broad, going beyond academic faculty and graduate students to include professionals such as architects and journalists, and other researchers in the general public such as lay historians and genealogists.

The overarching questions to be addressed in the proposed project are:

1) How can the existing IMLS DCC be expanded and enhanced to be more useful to scholarly users?

2) What approaches can be specified to assist the community of digital resource developers to consistently work toward more useful large-scale federated digital collections?

The work will be aimed at five objectives that we believe are essential, interrelated steps in answering these questions. The first objective provides the foundation for the four new core initiatives that follow.

- A. Maintain the IMLS DCC as a resource and testbed to benefit users and researchers of digital content.
- B. Articulate content evaluation and development guidelines specifically designed for strategic, use-focused federation. Conduct a formal evaluation of the IMLS DCC content and based on the outcomes, expand the collection for targeted scholarly communities.
- C. Analyze relationships between collection-level metadata and item-level metadata to better preserve context and enhance functionality for communities identified in Objective B.
- D. Experiment with and test metasearch capabilities with primary and secondary sources, exploiting findings from Objective C.
- E. Improve interface representation of content and context for scholarly use.

Studies of scholarly information use have demonstrated the value of the collection, and its inherent context, as an essential aspect of inquiry, especially for historians and humanities scholars (Brockman et al., 2001; DLF, 2005; Palmer, 2005). As Duff and Johnson's (2002) research has shown, "The totality of the records provides information that no individual record can. Historians must comprehend the records in their context rather than as separate disembodied items. Without this context information, the historian could easily misinterpret the meaning or significance of the information in an individual record." Library collections, be they digital or physical, are fundamentally resources for inquiry (Buckland, 1999) that have been intentionally created for that purpose (Currall et al., 2004). This project will address the challenge of creating federations with "contextual mass" for scholarly inquiry (Palmer, 2004). This approach sets priorities beyond size to include principled selection and integration of sources, metadata, and tools that work together to provide a supportive context for researchers. Implementing this approach requires analysis of the materials and activities involved in the practices of target research communities and a commitment to design collections around what scholars actually do and value. When a person uses a digital research library collection they are interacting with a context that includes physical, institutional, and intellectual features (Lee, 2000). At present, this tends to be a grand and scattered context that is not well aligned with scholarly practice, even in physical research libraries and most certainly in the digital realm. Collections built on a contextual mass model create a system of sources, with meaningful interrelationships between different types of materials and subjects, that work together to support the inquiry for a research area. In this project we will extend this approach to the federation of collections, developing the IMLS DCC as a collection of collections applying a contextual mass model.

As indicated above, we believe that collection metadata is the key to retaining context and building richer collections to support scholarship and that metasearch is a viable approach for extending the base of scholarly resources in federated collections (see, for example, Bieber et al., 2005). We recognize the importance of current work such as the Object Reuse Exchange (ORE) and the Pathways project (Warner et al., 2006) that emphasize objects over metadata for networking distributed, heterogeneous content. However, our experiences developing the IMLS DCC and working with cultural content and research artifacts has lead to a different approach. For instance, we have not encountered the same metadata bottlenecks as those reported in NSDL development (Lagoze et al., 2006), and, more importantly, we believe federation offers a better platform for adding value for scholarly users.

While we work to maintain, evaluate, and purposefully add content to the IMLS DCC, our research aim is to make transformative advances in how to improve access and functionality for users. To fully exploit the two major components of the resource, the registry and the repository, and meet our stated objectives, it is

necessary to address theoretical questions about the metadata and content relationships that control users' ability to search, discover, interpret, and apply digital materials in their work.

1.1 Core Initiatives to Address Need

Resource and Testbed Maintenance

The IMLS Digital Collections and Content (DCC) project has created a publicly available registry of IMLS NLG digital collections and a repository of item-level metadata available from these collections. The initial round of the project (2002-2005) examined shareable metadata, collection description, and the scalability of OAI-PMH. An extension of the project (2005-2007) included an examination of the necessary requirements for extending these resources beyond 2007 and the possible benefits for long-term maintenance of the portal. The proposed continuation of the IMLS DCC will maintain and enhance the resources developed in previous grant cycles, including continued harvesting of metadata and refreshing of the database, while answering research questions encountered during previous IMLS DCC development. Registries of government-funded digital library collections have become commonplace in the international library community (e.g., Project Michael). The IMLS DCC registry is the only integrated resource where one can find information regarding past and ongoing NLG-funded digital collections. In addition to providing integrated information about IMLS NLG projects, the existing registry also serves as a resource for the entire IMLS NLG community, and the metadata repository encourages projects to develop OAI-PMH compliant resources. IMLS DCC staff also provide consultation on shareable metadata, best practices, and implementation of OAI-PMH with other NLG projects. Loss of these resources would prove detrimental to the community.

Evaluation and Content Development

Results from the DCC project show that many service providers do not see the IMLS DCC as a valuable resource for their end users or as an effective reference tool in their institutions. This situation is not due to a lack of rich content or poor usability, the usual areas of concentration in digital resource development. In fact, the IMLS DCC is strong in both these areas. Instead, the problem stems from the nebulous nature of large-scale aggregations. The strengths and potential of the content are not evident to users. For example, our preliminary analysis of subject coverage in the IMLS DCC shows a clear strength in U.S. and state history and a secondary emphasis in the visual arts (see Appendix A). However, additional analysis is needed to determine how these and other more minor concentrations might complement each other and how these emphases can be further enhanced in alignment with the needs of targeted user communities.

In terms of potential, there is a need to assess how the independent subject collections produced by NLG projects can be rebuilt to provide a valuable resource for scholarly use. Additionally, we aim to determine how to coordinate development with other content providers and federated resources such as Getty's digitized collections or Library of Congress American Memory collections to better integrate complementary content. In particular, the IMLS DCC is well positioned to become a repository of information for and about LSTA funded digitization projects. For example, the Illinois State Library awarded 100 grants for digitization projects between 1998 and 2007. All the projects focus on the state of Illinois and its history, and more than forty were given to public libraries. LSTA projects are the source of unique and valuable local history collections, material of particular value to historians, both academic and non-professional. One important outcome of this project would be to integrate and present LSTA content in ways that maintain the local context that is important to institutions and scholars, but to also create larger aggregations that are greater than the sum of their parts.

Collection and Item Metadata Relationships

There is good reason to believe that even within the digital environment collections remain important to the research process, important, that is, *as collections*, not just as aggregations of items. Realizing full value from

federation depends in part on understanding how collections themselves provide value beyond that represented by item-level metadata. (Palmer, 2004), and ensuring that this value is supported and enhanced within the federation environment. This will require an improved understanding of collection-level metadata, and, in particular, of the systematic semantic differences among categories of collection-level attributes with respect to their complex relationships to items and item-level attributes. In some cases the relationship is straightforward. If a collection is described as a collection of paintings we can conclude that each item is a painting. However here, while it is important for users to find or identify an item as a painting, the alternative possibilities of representing items individually as paintings vs. representing a collection as a painting collection are of little interest to the user, and the propagation of the appropriate property from collection to item a simple matter. Other cases are more complex. As discussed in Wendler (2004) and Foulonneau et al. (2005), collection level may usefully combine with item level metadata in subtle but systematic ways: in Wendler's now classic example, if a collection is about Theodore Roosevelt and an item is described as "on a horse," we can perhaps conclude that it is Roosevelt who is on a horse. Here the logic of propagation (and our confidence in it) is more complicated than of the case of paintings, and more difficult to support, but it is undeniably of importance in typical search scenarios.

In still other cases, collection-level metadata has no direct implication at all for the individual item, but nevertheless has critical implications for the significance of the item within the research process. For instance, collections can be described as complete or incomplete, large or small, representative of a period or style, developed according to some systematic method (or not), heterogeneous with respect to genre or type of object, etc. In such cases, unlike the preceding example (painting), one cannot draw any conclusion at all about an individual item from the fact that it comes from a collection that is (in some way) representative. Nevertheless, it is precisely this sort of collection-level metadata that often establishes the scholarly significance of the search results. We might summarize these observations by saying that collection level representations such as "collection of paintings" are "item-convertible" and that collection-level descriptions like "representative with respect to..." are not. Interestingly, it is these the non-convertible attributes that would seem to make collections as collections (and not just groups of items) important to users, especially scholarly users. We must accommodate such attributes and support the appropriate relevant inferences if federations of collections are to provide full value. Relationships such as "item-convertible" are fine as far as they go, but they do not go far enough - the variety of possible collection-level / item-level relationships has never been catalogued, let alone explored in user studies. Although even bibliographic cross-level attribute relationships have proven difficult to articulate correctly (Renear & Choi, 2006; Renear & Dubin, under review), new more sophisticated frameworks are promising (ICOM/CIDOC 2005).

We will develop and empirically confirm a taxonomy of these relationships and an exploratory implementation, and then work with projects to enrich IMLS DCC collection registry records to reflect these distinctions in order to better support user tasks such as find, identify, select, interpret, and navigate. Enrichment will initially be done manually for testing purposes, while working toward the ultimate aim of automated propagation. We will empirically confirm the taxonomy and assess impact with end-user testing.

Metasearch Capabilities

Existing digital libraries tend to be homogenous collections of either primary source or secondary source materials. Thus ARTstor is composed of digital surrogates of works of art (primary sources), while JSTOR is composed of digitized scholarly journal articles (secondary sources), including some articles related to art. Both OAI-PMH and metasearch can be used to search, discover, and use resources from multiple repositories. However, the association of OAI-PMH with open access means that OAI-PMH is most often used to share and aggregate metadata describing digital surrogates of primary resources (Library of Congress American Memory, NSDL, IMLS DCC) while metasearch facilitates cross-database searching of licensed databases. Metadata

describing a smaller number of open access scholarly journals and institutional repository-held resources is available through OAI-PMH. While considerable progress has been made in the development of digital library systems focusing on each class of scholarly resource independently, little work has been done on implementations that provide simultaneous, integrated access to both primary and secondary content in concert. The IMLS DCC metadata repository currently contains almost exclusively primary source, open access resources. Related secondary source materials, especially those under restricted access control, must be discovered through separate portals. However, previous studies analyzing information seeking behaviors of interdisciplinary humanities scholars indicate that these researchers and faculty use heterogeneous information gathered from disparate locations (Palmer & Neumann, 2002). A faculty member teaching art history may want students to view digital representations of an artwork (primary source) alongside commentary about that artwork published in scholarly journals (secondary source). A social historian researching an aspect of Abraham Lincoln's career may have simultaneous need for both copies of Lincoln's correspondence and previously published research about those letters. Existing digital library infrastructures do not meet the needs of these researchers (Dempsey, 2003), and a digital library service that provides coordinated simultaneous access to both primary and secondary digital resources would better serve users (Borgman, 2003).

As described elsewhere in this proposal, there remain significant challenges to using harvested aggregations to facilitate end-user discovery and end-use of resources. Research suggests that metasearch is subject to many of the same difficulties encountered when searching over an aggregation of heterogeneous metadata. Current discussions involving metasearch technologies emphasize the need to provide options to the users outside the constraints of the individual databases (Dempsey 2005), simplify the search process (Hickey, 2005), and to identify the best approaches for efficient development of these tools (Reese, 2006; Christenson & Tennant, 2005). The proposed project will explore the usefulness of combined metasearch and harvested resources, issues inherent in combining metasearch and harvesting, and the applicability of lessons learned in our collection analysis and identity work on harvested metadata to metasearch implementations.

Building on the collection evaluation and metadata relations segments above, our research questions include: Would users find a DL that provided integrated access to primary and secondary scholarly resources useful? How so? Would this type of access help achieve essential contextual mass and better match scholarly end-users approaches to research and instruction? Would the perceived value and usefulness of both primary and secondary digital resources be increased if made available together through a single DL service? Would the perceived value and usefulness of resources indexed in the IMLS DCC be enhanced if viewed through a portal providing simultaneous access to related secondary content, both free and licensed? Is it useful and/or necessary, and to what degree, to characterize metasearch targets as collections (i.e. equivalent to collections) defined in the course of aggregating item-level metadata? Can collection description schema used to characterize collections within item-level metadata aggregation be extended or adapted for use with metasearch targets? Can "collection" information about metasearch targets inform metasearch strategies? Can this process be as dynamic in metasearch as when searching across a pre-coordinated/pre-analyzed metadata aggregation?

Interface Representation

How do we develop a more useful and usable tool for exploring collections and items in the IMLS DCC? According to Assessment of End-User Needs in IMLS-Funded Digitization Projects (IMLS, 2003), "project leaders and members of digitization teams should be flexible and adhere to the users' needs in designing interfaces." This advice also holds true for aggregation of digital materials. The current IMLS DCC interface is heavy on text and does not have many browsing options for the end-user. Studies have shown that better visual representation of content provides end users with more useful results (Myaeng & Song, 1999). Digital libraries have developed excellent searching abilities through the proper creation of metadata and querying of database fields. However, browsing abilities for items in these collections are still rudimentary and in need of additional

research (Twidale et al., 2007). Current explorations include portal views based on subject area (Tanase et al., 2006), and visual browsing of collection registries (Project Michael, www.michael-culture.org.uk). As noted above, scholarly activity particularly in the humanities is not simply a matter of extracting particular items from a large dataset. It is also a matter of finding whole sets of related items, and uncovering new sets of relationships between items. It involves the integration of both primary and secondary materials (such as particular artifacts and what people have said about those artifacts). It involves assembling – indeed “collecting” (Palmer, 2005), grouping, and annotating results to help make sense of a situation and to create new ways of looking at an issue. It involves annotation, writing, and debate. These activities are performed by scholars making the best use of the tools they have at hand, but frequently those tools fail to explicitly support these activities, imposing additional tedious effort to transfer resources between different media or applications that support different parts of the process. How might a system that more explicitly supported the larger context of scholarly activity enable more effective use of the available resources? What needs to be represented? How can we best represent it? How does interface design influence retrieval results? How can the different activities of searching for items, collecting and grouping them, considering related items and related secondary resources, and annotating these be best supported?

2. National Impact and Intended Results

Previous work on the IMLS DCC project has created positive national impact through the public availability of the IMLS DCC and consultations regarding shareable metadata best practices and implementation of OAI-PMH. An extension of the project will ensure the continued availability of these resources, while usefulness and usability of the site is enhanced. Moreover, answers to the above research questions will provide important results on applying collection development and description approaches, metasearching and harvesting techniques, and interface design features that will benefit the larger community of researchers and practitioners working toward federating and integrating digital content. To extend our growing base of knowledge, we have arranged for an ongoing "research exchange" with OCLC researchers working in related areas during the course of the project. We will coordinate our work with the three OCLC metadata projects described in Appendix B to exploit complementary efforts among our respective research teams. In addition, as we make progress on collection, metadata, metasearch, and interface research and development, our first-hand knowledge of problems and solutions will be integrated into professional LIS education through the courses and study groups directed by the project PIs and their associates.

Resource and testbed maintenance of the IMLS DCC provides a sophisticated platform for ongoing research on digital collections and for efforts to integrate IMLS-funded content with other nationally scoped initiatives. Our previous work has fostered participation in NSDL and opened collaborative exchanges with GEM. The current resource has generated interest among librarians and other scholars as indicated by general publications (Chronicle of Higher Education article) and blog entries. Other national/international impact that the DCC includes presentations at CERN, the NSDL Shareable metadata best practices, the DLF collection registry, and the DLF Aquifer project. **Intended results:** Continued availability of the IMLS DCC for end users and researchers and continued engagement with the community of developers around its development.

Evaluation and content development concerns identified above are not unique to the IMLS DCC initiative. On a national scale, digital projects have been aimed at worthy objectives, of interest to both local institutions and national funding agencies. But explicit, long-term coordination of content has not been a guiding principle. Libraries and museums have produced thousands of successful independent resources, but most have not planned their digital programs in ways that anticipate complementary aggregation with existing digital collections or future federation (Bishoff & Allen, 2004). The act of bringing together and providing access to a

large mass of distributed digital content is only a first step in producing a valuable, working federated collection. Other layers of development are needed to create meaningful, functional aggregations that support user communities of interest. Our proposed plan of integrating with other large federations and state libraries with LSTA digital projects will provide additional value for and awareness of the IMLS DCC. **Intended results:** Expanded, complementary DCC content; articulated principles for evaluation and development of federated collections.

Collection and item metadata advances in the digital community have been impressive. For federation purposes the analysis behind the Dublin Core Collection Description Application Profile (DCMI, 2006) stands as a significant achievement. Nonetheless, there is still no general framework that systematizes the possibilities, no specific framework for categorizing collection-level attributes with respect to features such as item-convertibility or exploring attribute propagation techniques, and little empirical data on the value of nonconvertible collection-level metadata in the research process. Other important advances have not been adequately analyzed for their application to collection federation, especially the CIDOC/CRM and other object-oriented approaches to metadata, as well as practices in contemporary information modeling (particularly the concepts of aggregation and composition), and ontologies (e.g., mereology). We note that any positive outcomes will further demonstrate the need for the collection registry and collection descriptions. **Intended results:** A taxonomy of collection-level attributes; revised collection-level schema differentiating between kinds of collection level attributes; pilot XML bindings; and recommendations for further work in standards development, systems design; and recommendations for further empirical research.

Metasearch techniques for integrating end-user access to distributed resources are a topic of current interest to many libraries and an active area of ongoing investigation, as evidenced by the continued proliferation of digital library portals. The IMLS DCC, extended and expanded as described above, represents a unique view of high-quality digital resources. By investigating the means and impact of enriching the core, metadata harvesting-based IMLS DCC, with complementary collections and commercially available content, we will provide research results that have broad adaptability and impact. Techniques developed and proven in the context of this project will inform future evolution of DL portal design and architecture. Additionally, by matching related resources to content digitized by IMLS funded projects, the portal will increase the impact of previous and ongoing NLG and LSTA digitization projects, and encourage further development of similar resources. **Intended results:** Assessment of viability of metasearch techniques and implementation challenges.

Interface design and search option enhancement of the existing resource will improve value for scholars and librarians at large. We intend to make the IMLS DCC a nationally recognized portal for humanities scholars. Additional lasting results of this project will be to better understand the browsing needs of the humanities scholar. **Intended results:** Report on value of alternative representations of collections for academic, professional, and lay humanities scholars and historians and improved browsing interface.

3. Project Design and Evaluation Plan

Stage 1:

A) **Collection evaluation and development:** Perform a formal conspectus-style evaluation of the IMLS DCC as a whole, analyzing subject and format strengths and weaknesses in terms of size, scope, depth, and significance of the contributing projects. Identify bridge areas and possible synergistic strengths. Work from standard collection evaluation guidelines, adapting them to address the special case of federated collection development. Begin collaboration with state libraries with LSTA funded digital projects. We will start with a sample of 5 nearby Midwestern states to determine the scalability of the model to a national level. Upon initial

examination of state libraries and LSTA funded digitization projects two models emerge: centralized projects (Missouri and Nebraska) and decentralized projects (Indiana and Wisconsin). (Illinois has only recently centralized their projects.) Ensure that the initial sample includes at least one of each type so that we can explore the needs of both models. (See Appendix C for list of potential targets.)

B) Collection metadata: Following a review of prior related work, assemble candidate metadata features from DCMI, IMLS DCC, and other relevant standards and projects for further analysis. Analyze a subset of IMLS DCC collection records from target subject areas determined in part A to assess "context" captured in the free-text description element. Develop a taxonomy of collection-level and item-level features with respect to relationships such as convertibility, inheritance, and propagation, including rules for defeasible inferences. Particular attention will be paid to i) context preserving inferences and ii) collection-level metadata that can neither be inferred from item-level meta-data, nor converted to item-level metadata, and which are therefore most likely to represent the distinctive value of the collection as a collection. Publish preliminary taxonomy and propagation rules to elicit early comments.

C) Metasearch: Based on the conspectus developed in part A, define focused, cohesive views (subsets) of IMLS DCC metadata to inform the selection of complementary metasearch and non-IMLS OAI-PMH targets. Ubiquitous commercial metasearch targets (likely, only those licensed by the University of Illinois at UC Library) will be considered for restricted access views which will be tested with UIUC scholars and students. While this will constrain breadth of testing, the applicability and adaptability of results will be enhanced by using appropriate scholarly targets (including commercial). Metadata and interface mappings (e.g., "Rosetta stone code") for each metasearch target will be developed as targets are identified.

D) Interface: Identify the elements of searching and use of search results in scholarly activity. Survey best practice in techniques to support these in digital library functionalities and interfaces. Prioritize the development of particular features to add and interfaces to revise.

Stage 2:

A) Collection development: Exploit collection strengths by prioritizing and targeting these thematic areas for acquiring new content for the IMLS DCC. Collaborate with relevant local, regional, and national initiatives, including LSTA funded digital projects and nationally scoped initiatives such as Library of Congress American Memory, DLF Aquifer, and other repositories to work toward thematic integration across resources. Explore inclusion of related "published" and commercial content. Conduct a site visit to the Illinois State Library for an overview of LSTA-funded digitization activities. Provide information about the DCC project and identify areas of expertise (metadata, OAI-PMH, digitization, etc.). Initiate planning for continued collaboration. Collect contact information for LSTA projects and add existing collections to the IMLS DCC.

B) Collection metadata: Test the intuitive plausibility of the relationships identified by surveying scholars, cataloging and metadata librarians, and catalog system developers. Revise taxonomy and analysis accordingly. Conjecture mappings of features to user needs, use FRBR Chapter 6 and DCMI DC AP as models.

C) Metasearch: Create one or more integrated portals to primary and secondary resources by mixing harvesting of NLG and complementary projects and metasearching of commercially available resources. Portal scope (and therefore the range of resources it provides access to) will be constrained to make the portal attractive to scholarly users working in particular domains. Thus, a portal might be created that features American Social history from founding of the U.S. through 1950. This will allow testing with targeted user groups. Various

techniques will be developed to deal with specific facets of mixing aggregated metadata and metasearch approaches. For instance, a facility will be included to allow users to initiate batch processing of specific queries to provide de-duped, ranked results.

D) **Interface:** Hire research programmer to create a browsing, collecting, and annotating interface and a graphic design student to collaborate on development. Options to be explored include subject clustering, word clouds, and timelines. Additional work will be conducted on the search abilities, and concatenation of collection and item-level descriptions. Build and test specialized views of the interface based on subjects and audience, with comparative analysis of displays for high-end scholars and "scholarly" user groups represented in the more general public. Collaborate with others for better tools.

Stage 3:

A) **Collection development:** Formalize prospective collection criteria and policy building guidelines for diverse, distributed federated collections. Test the guidelines through a Delphi-like method with a panel of experts. Refine and disseminate to the larger digital cultural heritage community. Attend state-wide LSTA digitization meeting to demonstrate the IMLS DCC. At the end of year 3, reanalyze collection to document advancements in strength building and to identify new target areas for development.

B) **Collection metadata:** Apply object-oriented approaches borrowed from CRM CIDOC and related standards, and other branches of computer/information science, to develop implementations of collection-level descriptions based on the preceding analyses and taxonomy, focusing on features likely to be of substantial value. Collaborate with projects to test strategies for enriching IMLS DCC collection registry records. Conduct empirical studies to confirm value of upgraded collection registry records to better support user tasks such as find, identify, select, interpret, and navigate. Publish results on user studies for early comments. Present working results to DCMI and the CRM/IFLA Object-Oriented FRBR Working Group. Develop pilot XML bindings and make recommendations to support further experimentation and adoption.

C) **Metasearch:** Explore the viability of this type of resource, and this resource available to a larger audience. Conduct an in-depth study of a small selected group of high-end scholars (e.g., those publishing in a domain) on their work with online resources to assess if an integrated collection is useful. Study transaction logs. Disseminate results encompassing technical discussions involving the techniques and process required for integrating primary and secondary resources as well as theoretical discussions involving the magnitude of achieved benefits and potential for further developments.

D) **Interface:** Conduct usability tests with scholars; determine usefulness for end-users; refine interface.

Throughout the three stages, the project team will use the ongoing experiences and knowledge base of the IMLS DCC project, as appropriate, as teaching case material in the courses they teach in the areas of digital collections, ontologies, interface design, and metadata. The project will also continue to host and direct the weekly Metadata Roundtable colloquium for students, practitioners, and researchers across campus. This forum promotes professional exchange and the integration of research and practice on topics related to metadata, digital collections, and data curation of immediate importance to the library and museum digital community.

4. Project Resources

Personnel and Computing Resources: The project PI (Carole L. Palmer) and 5 co-PIs (Timothy W. Cole,

William H. Mischo, and Sarah Shreeves from the University Library, and Allen H. Renear and Michael Twidale from GSLIS) will be collaborating on the project. Other key personnel will be the Project Coordinator (Amy Jackson, current incumbent on the DCC project, will continue and provide important continuity), a faculty researcher, Dave Dubin, who has valuable expertise in metadata and quantitative analysis, a programmer (to be named), and a data analyst from the Library Research Center to assist with managing and analyzing data from the empirical components of the research. Two one-half time GSLIS Research Assistants (to be named) will be assisting with research activities described above, especially the collection evaluation, metadata analysis, and all data collection from scholarly user groups. An hourly graduate student from the Art & Design department will be brought in for interface design expertise. The project will be housed both in the Grainger Library, which will continue to host services and provide server disk space and system administration for maintenance of the IMLS DCC, and in the Graduate School of Library and Information Science, Library Research Center, which will provide computer support for GSLIS researchers and RAs. Jean Godby will serve as the primary OCLC contact and work with the PI to coordinate complementary IMLS DCC and OCLC metadata work.

Management Plan: The research work will be based and managed out of the Library Research Center, which is directed by the PI. The PI and Co-PIs will direct their component parts of the project and supervise the associated research assistants: Palmer on collection evaluation and development, Renear on metadata relations, Cole and Mischo on metasearch, and Twidale and Shreeves on interface. Shreeves will be involved with Palmer on monitoring progress and integrating results across the components. The project coordinator will be responsible for the IMLS DCC maintenance activities and for coordination with all other projects and institutions involved in the ongoing work. She will also provide the major coordination among the development components to insure full and continuous flow of information among project staff. She will work with the PI to supervise the data analyst and Art & Design graduate student contributions to the project and coordinate their work with the ongoing needs of the project. Regular project meetings will be held to facilitate communication and share results. GSLIS and the University Library Business Offices work with the University Grants and Sponsored Contracts Office to oversee finances and insure conformance with regulations.

5. Dissemination

Results will be reported in the scholarly literature and at appropriate scholarly meetings (e.g., Digital Humanities, Museums and the Web, ACRL, ASIST Annual Meeting, JCDL, ECDL, DCMI, SAA, CDOC, Web-Wise, DLF Forum). Project results will be disseminated through publication on the project Website and by deposit into the UIUC Institutional Repository (IDEALS), including white papers, taxonomies, guidelines, and other project documentation. Findings will also be published in journals such as D-Lib, Literary and Linguistic Computing, JASIST, etc. Since the backgrounds of the PIs span traditional library and museum domains and include integral ties to the scholastic arena provided by GSLIS, they are well positioned to take advantage of numerous forums in which this research can be exposed, exploited, and built upon.

6. Sustainability

Through our extensive partnerships with state projects, large-scale collection efforts, OCLC and others, we will be able to use these results to improve the usefulness of the IMLS DCC and to help IMLS make judgments about its long-term value and viability. In addition the approaches and technologies developed in this project are easily adapted to support other digital initiatives, and the research outcomes will be highly relevant to other continuing and future federation and integration efforts. In particular, the results on collection metadata will inform future development of the DCMI collection application profile and metasearch can inform related NISO activities.

References

- Bieber, M., Im, I., Oria, V., Sweeney, R., & Wu, Y-F. (2005). Lightweight integration and recommendation of documents and services. New Jersey Institute of Technology, April 2005.
<http://web.njit.edu/~bieber/presentations/dlii-gre-0405.ppt>
- Bishoff, L., & Allen, N. (2004). *Business Planning for Cultural Heritage Institutions*. Council on Library and Information Resources, Washington, DC.
- Borgman, C. (2003). The invisible library: paradox of the global information infrastructure. *Library Trends*, 51(4), 652-674.
- Brockman, W. et al. (2001). *Scholarly Work in the Humanities and the Evolving Information Environment*. Washington, DC: Digital Library Federation/Council on Library and Information Resources.
- Buckland, M. (1999). *Library Services in Theory and Context*. 2nd ed.
<http://sunsite.berkeley.edu/Literature/Library/Services/index.html>
- Christenson, H., & Tennant, R. (2005). *Integrating Information Resources: Principles, Technologies, and Approaches*. http://www.cdlib.org/inside/projects/metasearch/nsdl/nsdl_report2.pdf
- Currall, J., Moss, M., & Stuart, S. (2004). What is a collection? *Archivaria*, 58, 131-146.
- Dempsey, L. (2003). The recombinant library: portals and people. *Journal of Library Administration*, 39(4), 103-136.
- Dempsey, L. (2005). From metasearch to distributed information environments. *Lorcan Dempsey's Weblog* (October 9, 2005). <http://orweblog.oclc.org/archives/000827.html>.
- DLF. (2005). *The Distributed Library: OAI for Digital Library Aggregation: OAI Scholars Advisory Panel Meeting*, June 20-21, Washington, DC.
<http://www.diglib.org/architectures/oai/imls2004/OAISAP05.htm>
- Duff, W., & Johnson, C. (2002). Accidentally found on purpose: Information-seeking behaviors of historians in archives. *Library Quarterly*, 72(4), 472-496.
- Foulonneau, M., Cole, T. W., Habing, T. G., & Shreeves, S. L. (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*. ACM Press, New York, NY, 32-41.
- Hickey, T. (2005). Metasearch and metadata. *Outgoing: Library metadata techniques and trends* (August, 2005), <http://outgoing.typepad.com/outgoing/2005/08/metasearch.html>
- ICOM/CIDOC (2005). The CIDOC Content Reference Model, version 4.2.
- Institute of Museum and Library Services (2003). Assessment of *End-User Needs in IMLS-Funded Digitization Projects*. <http://www.imls.gov/pdf/userneedsassessment.pdf>
- Lagoze, C. et al. (2006). Metadata aggregation and "automated digital libraries": A retrospective on the NSDL experience. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM Press, New York.
- Lee, H. (2000). What is a collection? *JASIS*, 51 (12), 1106-1113.
- Myaeng, S., & Song, S.-K. (1999). Visualization of retrieved documents using a presentation server. In *Proceedings of the 2nd Asian Digital Libraries Conference*. Taipei.
- Palmer, C. (2004). Thematic research collections. In S. Schreibman, R.Siemens, & J. Unsworth (Eds.). *Companion to Digital Humanities*. Oxford: Blackwell, pp. 348-365.
- Palmer, C. L. (2005). Scholarly work and the shaping of digital access. *JASIS*, 56(11), 1140-1153.
- Palmer, C. L., & Neumann, L. (2002). The information work of interdisciplinary humanities scholars: exploration and translation. *Library Quarterly*, 72 (Jan.), 85-117.
- Reese, T. (2006). Metasearch: building a shared, metadata-driven knowledge base system. *Ariadne*, 47(April).
<http://www.ariadne.ac.uk/issue47/reese/>

- Read, B. (2006). Federal agency unveils database of digital collections from museums and libraries. *Chronicle of Higher Education*, 52(33), A41.
- Renear, Allen H., Dave Dubin. (under review). "The FRBR Group 1 Entity Types are Roles not Types".
- Renear, Allen H., & Yunseon Choi. (2006). "Modeling Our Understanding, Understanding Our Models -- The Case of Inheritance in FRBR." Proceedings of the 69th ASIS&T Annual Meeting, 3-8 November, 2006, Austin, Texas. [http://eprints.rclis.org/archive/00008158/01/Renear_Modeling.pdf]
- Tanase, D., Joiner, D., & Stuart-Moore, J. (2006). Computational Science Educational Reference Desk: A digital library for students, educators, and scientists. *D-Lib Magazine*, 12(9) <http://www.dlib.org/dlib/september06/tanase/09tanase.html>
- Twidale, M.B., Gruzd, A., & Nichols, D.M. (2007). Writing in the library: Exploring tighter integration of digital library use with the writing process. To appear in *Information Processing and Management, Special Issue on Digital Libraries in the Context of Users' Broader Activities*.
- Warner, S., Bekaert, J., Lagoze, C., Lin, X., Payette, S., & Van de Sompel, H. (2006). Pathways: Augmenting interoperability across scholarly repositories. Accepted for *International Journal on Digital Libraries special issue on Digital Libraries and eScience*.
- Wendler, R. (2004). The eye of the beholder: Challenges of image description and access at Harvard. In Hillmann, D. I. and Westbrook, E. L., eds., *Metadata in Practice*. American Library Association, Chicago, IL, pp. 51-69.