University of Illinois at Urbana-Champaign
Proposal to Extend IMLS Collection Registry and Metadata Repository Project

ABSTRACT

Work to date to implement, study, and develop the IMLS collection registry and item-level metadata repository hosted at the University of Illinois has furthered our understanding of such systems and yielded new insights into their utility for the IMLS community. The collection registry currently includes records for 151 NLG-funded collections; the item-level metadata repository includes nearly 200,000 records from 31 of those collections. We have achieved a better understanding of digital collection-level description attributes and schemas,[1] of the impact of local metadata authoring practice on metadata aggregators,[2] and of current metadata quality (considered in terms of utility for aggregation).[3] Ongoing work will yield findings about use models and interface design approaches for the baseline registry and repository. Major goals of our original proposal will be met by the end of this calendar year. However, work to date also has raised new issues and complexities that can only be addressed through additional experimentation and research. Public accessibility from this summer to the registry and repository will provide opportunities to extend initial findings. To maximize benefits from work so far, and to make best decisions and plans for continuance of the registry and repository after completion of this grant, we recommend that IMLS extend the current grant to the University of Illinois to include added research and experimentation. Specifically:

1. Results to date support importance of uniqueness, authority, and context when aggregating item-level metadata from disparate sources. Collection descriptions can help provide such information.[4] More research on collection identity and ways to better integrate collection-level and item-level metadata is needed to more fully understand implications of descriptive granularity and collection membership.
2. Work to date also suggests a correlation between metadata completeness and an ability to implement services for specific audiences. To adequately assess and lay the groundwork for new, innovative ways that digital resources funded by IMLS can be presented to targeted communities, we propose to experiment with metadata normalization and enrichment methods optimized for an education audience, exploiting community-vetted tools and vocabularies such as those from the GEM Initiative.[5]
3. Ongoing testing of baseline registry and repository portals has highlighted differences in the ways users want to view resource aggregations -- librarians seek biblio-centric views, museum staff think in artifact-centric ways, educators and educational content providers want learning-centric views. To more fully evaluate potential efficacy and utility of registry and repository, we will create and test specialized versions of these applications designed with education users and models of use in mind.
4. Creating initial versions of the registry and repository proved labor intensive. While participation in item-level metadata repository has been encouraging, our survey of NLG projects suggests at least another 30 projects have the wherewithal to participate. We've identified ways to streamline registry and repository maintenance and facilitate repository participation, but more time is required to test efficacy of these approaches. Additional testing with a more diverse group of projects (e.g., with a representative subset of the LSTA projects overseen by the Illinois State Library) also is needed.
5. Results so far have identified ways that IMLS projects can facilitate and improve interoperability, but the process by which new practices are adopted and assimilated within the community of IMLS grantees is complex. Additional research into knowledge diffusion process in this context is needed.

To accomplish outlined additional work, we request to extend current grant until 30 September 2007. Funds remaining from original award as of 30 September 2005 will be carried forward. We also request a supplemental award of $342,982 to cover added expenses. During extension we will work with staff at the Illinois State Library (re inclusion of LSTA grantees) and collaborate with researchers from the GEM Exchange (re metadata enrichment for targeted audience and use of GEM tools and vocabularies).

1. ASSESSMENT OF NEED

Collection Identity & Metadata Granularity: Collection identity has been treated in the library and information science literature as a core problem in the digital environment.[6] Data gathered thus far in our project have confirmed that this is not just a theoretical problem but one that is being contemplated by practitioners and both actively and passively responded to in the daily work of digital library development. The IMLS collection registry and metadata repository (and related federation efforts at state and regional levels) offer an ideal test bed for tracking how collection identity is transferred, transformed, and created in the practice of digital library development. Item-level descriptions support retrieval of individual objects, while collection-level descriptions describe uniqueness, authority, provenance, and context of objects in a collection and support collection discovery. Intermediate levels of descriptive access might enable or enhance other functionalities of federated registries and repositories. Dynamically defined *virtual* collections and arrangements of objects, such as can be created by metadata aggregators, offer the possibility of additional enhancements and functionalities. (Consider reconstituting a set of virtual representations of artifacts once co-located but now distributed across several museums or, alternatively, bringing together a set of virtual representations of artifacts never co-located but appropriate for a particular exhibition, e.g., “All the world’s Vermeers.”) Collection definition and identity will become more important as we move toward less opportunistic development of digital libraries and return to principled collection development and prioritization activities. To leverage collection definition and development, to identify and demonstrate potential advantages of providing description and access at multiple levels of granularity, to understand better the potential of virtual collections, applied research on these issues based on a study of actual practice is needed.

Metadata Normalization, Transformation & Enrichment: While a first priority for the registry and repository was to establish basic keyword search functionality across aggregated metadata, it was always understood that organizing information goes beyond the provision of keyword search. Models articulated by the IFLA FRBR Study Group[7] and others emphasize the importance of functionalities beyond search by query. Limit options, inclusion of context, ranking/arrangement of results are all critical to effective identification and selection of resources. Clustering and co-locating of like items (done virtually and often dynamically in a digital environment) are essential for discovery by simple or associative browsing (e.g., AquaBrowser[8]) and have been described as the "primary act of information organization."[9] Such functionalities require rich metadata well delineated. You can't limit or sort by temporal coverage if coverage information is sometimes in the wrong field or date values are inconsistently formatted. Analyses of item-level metadata aggregated for the IMLS repository demonstrate that metadata harvested from grantees is highly heterogeneous and of variable richness and quality.[10] Some preliminary progress with basic metadata normalization has been accomplished already, but additional experimentation is needed to demonstrate the amenability of IMLS grantee metadata to normalization and enrichment.

For best results, this cannot be done in isolation from intended use or oblivious of target audience. Different users may use different words to describe the same artifacts -- e.g., one can imagine an art historian and an anthropologist using very different terms to talk about the same African mask. Items in a collection have intrinsic meaning, meaning in their collections, meaning in the aggregated collection, meaning in their earlier locations and uses, and many other meanings and interpretations, some contested and some changing. While this has always been the case, it is increasingly so in digital contexts. Yet a natural tendency, and one inherent in many initial approaches to digital library implementation, has been to assume a single "True" or primary meaning for each object that we can aspire to present to the end-user. Initial work on the registry and repository suggests that the weakness in this assumption becomes especially obvious in aggregations of disparate content. The challenge is to design a system that can cope with multiple, shifting and contested meanings and can mediate on behalf of users seeking specific outcomes. Different uses may imply a need for quite different kinds of indexes, facets, and subject classification. This requires metadata optimized for sharing, aggregation, and reuse. While we assume (as does GEM) that the function of metadata creation is readily taken on by collection holders and that the metadata they create is relatively useful for aggregation (albeit optimized for local purposes), the question remains as to how well and easily (and at what cost) metadata created with one audience in mind can be transformed and enriched for use in another context. Thus, experimentation with metadata transformation, normalization, and enrichment for a specific use (e.g., the K-12 community) is needed.

Interface Design for Targeted Community of Users: Following on this same theme, different groupings and classes of users will want to discover and access IMLS funded content in different ways. Different users require different views of the information landscape.[11] As an example, the GEM attempts to describe digital resources and collections of digital content in its Gateway in a manner that can be described as learning-centric -- a classroom and state standards worldview. This can be contrasted with unaugmented uses of simple Dublin Core, intended for more biblio-centric library worldview access. The approach taken thus far in developing search and browse interfaces for the IMLS registry and repository, has been to compromise any targeted views of aggregated content in order to provide general-purpose access. Arguably this makes the aggregation appear less useful to any particular user, since a generic approach will fall short for all users. To complement work done so far we need to investigate specialized, tailored views of registry and repository metadata. User testing of a tailored interface with a targeted group of users (K-12 educators and student teachers) also will allow extended and likely more fruitful experimentation with interface techniques that clarify for the user the distinctions between different collections and sets contained within an aggregation. User studies indicate that people do get confused about such distinctions. We have an opportunity to gain needed practical experience in designs that help users distinguish among collections and between collections and items. Views and access at the collection level is often neglected in digital library applications, but in fact there are many reasons for providing collection-level information and access -- collections have an intrinsically larger meaning than the individual items; there are stories to be told at the collection level; it is often desirable (and appreciated by grantees) to encourage users to access the distinct digital and physical collections in their original context. Additional testing of a tailored interface with a more specialized audience will provide an opportunity to refine preliminary usability testing so far and reinforce conclusions. The questions will remain much the same: Where am I in the interface? What is this resource? Where is it from? What do I need to know to use the resource? But studying the interaction of a targeted audience with an interface tailored for their use will add richness to our observations and complement initial testing on general purpose portals.

Implementation & Testing of Streamlined Processing & Maintenance: Initial implementation of the registry and repository took longer than originally planned. A contributing factor was an overly complex and labor-intensive workflow, especially for OAI metadata harvesting, processing, and indexing. The issues and difficulties we encountered are relatively common across breadth of the OAI community. Our work to date on the IMLS registry and repository (in concert with results from our other OAI projects[12]) has identified software and workflow improvements that should result in more reliable and robust metadata harvesting, processing, and indexing done in a more automated and autonomous fashion (i.e., less labor intensive). Some of these improvements have recently been implemented for this project -- e.g., software features that automatically compensate for invalid record date stamps and bypass invalid records during harvest, but additional testing is required to better assess benefits and savings achieved by these changes. Other identified improvements, especially in end-to-end streamlining of repository and registry maintenance, have yet to be implemented. Concurrently, participation in item-level metadata repository has lagged relative to original expectation, in part because of technical glitches in available metadata provider tools and solutions and underappreciated difficulties with data quality and protocol feature (e.g., character encoding, use of OAI Resumption Tokens). OAI turnkey solutions (e.g., ContentDM) have improved, in part through feedback from OAI projects including this one. Common OAI pitfalls and hazards are better understood and can more easily be avoided. Further experimentation with workflow streamlining and automation will provide a more accurate estimate of cost and level-of-effort to maintain the registry and repository in production and will benefit other OAI service providers. A further period of experimentation with second-generation metadata provider tools and best practices, especially with a broader project participant population (e.g., with inclusion of a sample of LSTA grantees) would allow OAI-based sharing to be more fully tested and benefits better understood. In turn this would give a better and more realistic estimate of likely future levels of participation in an IMLS metadata repository.

Models of Knowledge Diffusion: Even as we better understand implications of current practice in regard to collection identity, metadata quality, interface design, and metadata processing, there is a need to understand how metadata best practices are shared and implemented by practitioners. Our initial results suggest that the informal nature of knowledge diffusion within and among IMLS projects (including our own) can be better fostered and perhaps even formalized to improve library development and professional development. In some ways this is a prototypical knowledge management problem seen in other distributed professional groups and communities. We have an opportunity to use our ongoing analyses and interactions with IMLS grantees and our current OAI and metadata sharing best practice collaboration with the DLF and the NSDL[13] to identify and test ways to disseminate metadata best practices and promulgate community-wide solutions. Local models of knowledge diffusion may exist that can be leveraged more widely. We see real potential, for instance, in the joint Library-Library School Metadata Roundtable begun here as part of this project as a source for developing and advancing the curriculum in metadata for our LIS students. Other local task forces and working groups at other IMLS grantee institutions also might be “repackaged” to facilitate teaching, training, and professional development elsewhere. To assess and help realize such potential there is a need to explore and more systematically describe dissemination and adoption mechanisms for developing and sharing metadata best practices across library, museum, and archive communities and among students and professionals staffing such institutions.

2. NATIONAL IMPACT & INTENDED RESULTS

Increased Competencies & Capacities within IMLS Grantee Community: The report of the IMLS Digital Library Forum (2001) called for increased capacity and better awareness within the IMLS grantee community of opportunities for interaction with nationally-scoped digital library projects. Functioning as a test bed for experimentation with collection-level description and metadata sharing, this project has been effective in helping to accomplish this objective. NLG grantees have developed new competencies and capacities in collection-level description, digital library interoperability, and metadata sharing. Several grantees that implemented OAI-PMH in order to participate in this project's metadata repository are now being harvested by other initiatives such as OAIster and the NSDL. We anticipate continued impact of this kind during the grant extension, amplified since the registry and repository will be publicly accessible. At the same time we also are starting to see impact in how projects approach interoperability collaborations. The roles and responsibilities of the various partners involved in such efforts have not been well defined. What can be expected of a metadata aggregator in regard to metadata normalization, enrichment, and transformation? What are the obligations of a metadata provider in terms of metadata completeness, accuracy, consistency, and fitness for reuse? How do decisions taken by a data provider in regard to granularity and collection definition affect how content can be integrated with other resources? While considerable progress has been made already on this project towards addressing such questions, a project extension, with both the registry and repository fully operational, would extend impact so far achieved and better demonstrate to grantees (and vendors) the impact of their decisions on digital library collaborations. Inclusion of a sample of LSTA grantees will provide a base for comparing and validating preliminary observations and conclusions developed by working with NLG grantees.

Foundation for Future Digital Library Research: Further development of the registry and repository presents a unique opportunity to investigate fundamental problems associated with the federation of heterogeneous digital collections. In the current project we have developed an empirically based understanding of current metadata practices and are now working toward identifying those that impact access and use of collections before and after aggregation. Our findings are generally applicable since they have been drawn from the experiences of a wide range of projects and institution represented by the generation of NLG projects covered in the study. Their cumulated experience coupled with our analysis of the problems and opportunities afforded by metadata federation provides a solid base of knowledge that can inform further research and decision-making and cost-benefit assessment by professionals building digital resources. Our proposed examination of collection identity is a direct outgrowth of our preliminary findings. It is of pragmatic value since it has implications for resource discovery and especially for effective interpretation and use of the resources represented by item level-metadata. However, this area of investigation is also vitally important on a more theoretical level. The digital environment is changing the nature of collections. They are becoming more fluid, dynamic, and amorphous, but we have yet to comprehend the degree to which metadata influences the value of this state of collections. Our research will help determine what aspects of collections need to be captured to satisfy the institutional conceptions of collections that are important to the missions of libraries and museums. For example, in our research thus far we have identified distinct cultures of description that we hypothesize can be exploited in productive ways for increasing the functionality of federated collections. In this area, librarians can learn a lot from archivists who have long relied on collection level descriptions for access to their collections. But, while accommodating institutional needs, it will also be necessary to identify the dimensions of collections that enhance users' research and discovery experiences. It may very well be that there are very different requirements for collection description from the perspective of users or for the development of virtual collections by service providers. As we make further progress on these aims we expect to be able to make recommendations for minimal and optimal metadata requirements supporting the creation of searchable and browsable federated repositories and clarify the tradeoffs involved for metadata providers and aggregators in meeting these requirements. We see this as essential core knowledge for practicing digital resource developers but also for training the next generation of metadata professionals. An additional benefit of this research is our ability to trace the channel through which metadata expertise is developing and being transferred among the community of practice. We intend to continue this line of analysis and assess where and how to best support knowledge transfer and professional development to upgrade metadata expertise in the field and in current LIS higher education.

Based on these theoretical insights, we also believe we can make contributions to a better understanding of both the process of creating and repurposing metadata, and how these data can best be represented in interfaces for a variety of uses, users and contexts. For example, how can we design systems that can cope flexibly and robustly with an institution's changing needs for metadata in light of new research and its gradual dissemination and adoption, new opportunities including federation, and new organizational directions and patron needs? A 'get it right first time and stick with it for years' approach seems naïve in a world of constant change. How can we design interfaces to federated collections including very varied metadata so that patrons can use them effectively and efficiently without being confused by the diversity of origination and initial purposes of the metadata and without needing to be expert librarians?

Foundation for Continuation of IMLS Collection Registry & Metadata Repository: While the registry and repository as currently implemented may not represent an end in themselves, preliminary response from participating grantees and IMLS staff suggest that properly extended and increased in scope these could become useful resources worth maintaining long-term. Both the registry and repository represent unique, useful views of digital content funded by IMLS. As such they are of potential interest to a variety of audiences. Right now the registry and repository remain experimental and impermanent. Any decision to continue them and make them permanent must be based on a reliable understanding of cost and of the potential ways in which they might be used. To reach that point, we need additional time with these resources in broad release (i.e., publicly accessible), we need to investigate additional interface and access options, we need to explore benefits to additional audiences of interest, and we need to better understand the level of effort required to maintain the registry and repository. An important outcome of the proposed extension will be the opportunity to focus on what it would take to continue these resources beyond 2007 and to better describe potential benefits of doing so.

3. PROJECT DESIGN & EVALUATION PLAN

Research on Collection Identity & Metadata Granularity: We will take two approaches to the analysis of collection description and related granularity issues. First, we will examine collection and sub-collection representation from the perspective of the developers of digital collections. This will be done through content analysis of the collection level records submitted to the registry and changes in those records over time documented in systematic snapshots of the database. We will follow up with interviews with staff responsible for making decisions about content of collection records to better understand their intentions and priorities for providing information about and points of access to their collection. We will focus our interviews on those projects where descriptions are particularly rich or sparse, and particularly those that make enhancement to the original base records. Second, we will examine users' interactions with the metadata repository to assess the value and complementarity of collection and item-level metadata. In this segment of the study, we will work especially with practicing educators and student teachers, representing a user community we believe is a key audience for IMLS-funded digital content. (As project resources allow, we may extend this study to include others such as lay historians and museum staff involved in education outreach.) We will conduct observations of user interactions with the repository and document how they interpret records retrieved, the data of most value to them, and what additional content or context could enhance their searching and use of the digital resources. In cases where we need to work with participants that are located at distant sites, we will rely on post-searching interview data on these topics. In addition, we will be analyzing transaction log data on general use of the registry and repository, which will be important for determining search patterns at large and how that may differ from what we observe with our targeted audiences. (Web servers hosting the registry and repository record standard Web transaction logs which include minimal identifying information [e.g., the IP address of client workstations accessing the registry and repository]. Client IP addresses recorded in logs will be used only to establish clustering of log records by session, and then will be discarded. Every effort will be made to protect raw transaction log data and preclude disclosure of any information that might identify an individual user of the registry or repository.)

Metadata Normalization, Transformation, & Enrichment: Work to date has identified metadata quality and consistency issues potentially addressable through post-harvest normalization.[, 10] On other projects[12] we've found qualified DC well suited to retain the kind of element encodings and refinements resulting from more advanced metadata normalization and enrichment. (Currently the IMLS metadata repository indexes simple DC, which is of limited utility for higher-level index and interface functions.) During the extension phase of this project we propose to transform, normalize, and enrich all OAI harvested metadata to a project-customized qualified DC variant. Harvested records in original format will be retained. Search, browse, clustering, export, and similar system functions will rely on an index built from the qualified DC version of the record, while the original record will be used with the qualified DC record for display. Our earlier survey of NLG projects helped us identified controlled vocabularies in use by the NLG projects. In richer formats (qualified DC, MARCXML) providers can (and generally do) label instances of controlled vocabulary terms in metadata records. When used in simple DC, controlled vocabulary terms can sometimes be recognized by automated means (given explicit knowledge of controlled vocabularies being used by that provider). Collection membership, context derived from collection-level descriptions, and special encodings (e.g., use of controlled vocabularies) will be made explicit in indexed qualified DC records, where such information can be automatically determined from original records. Normalization and enrichment will be executed on a collection-by-collection basis (allows for use of collection-specific heuristics) and will be accomplished primarily using an ordered series of general and collection / provider-specific XSL transformations. (XSL transformations are portable and operating system independent.) Scripts will only be used for normalizations requiring extensive string manipulations (e.g., date value string normalization). Type, format, language, identifier, and coverage (both spatial and temporal) attributes (at least) will be normalized and/or enriched and success rates reported. (Work on other projects suggests that the percentage of records containing an indexable temporal coverage date value can double after automated normalization and enrichment.[12]) Post-harvest generic normalization scripts and XSL transformations will be piloted and iteratively developed using item-level records from 3 representative collections. Once proven with that subset of collections, they will be migrated into production use for all item-level harvested collections.

While it will remain beyond the reach of this project even in its extended phase to perform extensive manual in-depth, record-by-record enrichment of all collection descriptions and harvested item-level metadata, we do want to lay the foundation for such work and experiment on a small scale with approaches targeted to make resources more discoverable by educators and education-oriented audiences. To do so we will exploit tools and vocabularies available from the GEM Initiative and we will experiment with ingest of records from the IMLS collection registry and metadata repository into a special test region of the GEM Gateway. In enriching records using GEM utilities we will begin with collection registry records, piloting first manually with a small number of collection records and then automating methodology and extending to touch all collection registry records. We will then experiment with and assess automated enrichment of a sample of item-level metadata records using GEM utilities. (We do not anticipate being able to enrich all item-level records during the course of proposed project extension.) To populate a test region of the GEM Gateway we will implement new export features from both the registry and repository to expose collection and item-level metadata in a qualified DC format suitable for GEM Gateway ingest. The objectives will be a preliminary assessment of the potential (and potential pitfalls) of attempting to enrich IMLS metadata using such methods and an assessment of feasibility of preparing IMLS grantee collection descriptions and/or item-level metadata records for inclusion is utilities such as the GEM Gateway. We will compare records in the IMLS registry and repository to those already in the GEM Gateway, looking at issues related to descriptive granularity, provenance and authority of objects and metadata, packaging of resources, and context. A key objective will be an early assessment of whether descriptions of digital objects and collections designed for use in a local digital library, archive, or museum project might usefully be repurposed for inclusion in a learning-centric resource-discovery application like GEM (and if so, at what expense in time and effort). We will solicit feedback from IMLS grantees at a meeting held in Urbana in spring 2006 and potentially a workshop held in conjunction with Web-Wise 2007. To balance post-harvest work we also will work with a small group of grantee volunteers to encourage and instruct on their use of GEM utilities at time of metadata creation.

Interface Design for Targeted Audience: A key issue to consider is the ways in which an interface to federated content can aid learnability, usability and adoption. A major opportunity but also a challenge is that such a resource offers a functionality that will be novel or at least unexpected to many potential users. The idea of being able to search between different collections located in different institutions and assembled for different purposes, and at the same time search within all the collections at once is powerful, but potentially confusing. Additionally, different people will want to use the information for different purposes, and probably use different words to describe their needs. These different needs and uses must be factored into the interface design, and will require the provision of different emphases and functionalities, and potentially different interface options. For example, for a librarian a federated collection is another resource, essentially another database, that can be used in reference work to help a patron find what she needs. Reference librarians want efficient powerful access, and can be assumed to be skilled searchers. Museum professionals, on the other hand, may want to use the resource to gain inspiration from the work of others, obtain ideas of best practice, find opportunities to cooperate with other institutions, and assess how sharing of digital resources can add value to their own physical and digital collections. Teachers will be looking at the same resources but through the lens of how they can be used to enable learning within a given curriculum. This might involve planning for use before, during, or after a visit to a museum, but it could also be to help with classroom or homework activities as a way to engage students' interest in primary sources. Finally, members of the public bring very many other purposes, interests, and levels of expertise, both with computer systems and with the materials themselves. We need to design interfaces that can cope with this variety of experience and interest.

During proposed extension we will undertake to understand this wider challenge by considering the needs of one specific audience, teachers. In general the collections represented in the registry and repository have not been developed solely with teachers in mind, and so the diversity of the needs of teachers and how those differ from needs of the patrons for whom the content was explicitly or implicitly intended will enable us to explore the need for different interface elements and for different kinds of searching and browsing. Careful user testing, analysis of the causes of user confusions and misconceptions, and subsequent redesign can help us produce an interface that helps people learn about what the system is for as well as how to use it.[14] We will bring to bear experience in usability analysis and interface design in a range of settings including library OPACs, CD-ROMs, online databases, museum websites and digital libraries.[15, 16, 17] We will employ an iterative rapid prototyping approach including small scale in-depth user studies of teachers and student teachers as they search for materials that would help them in their own lesson planning. By careful cognitive diagnosis of confusions and hesitancies users exhibit when using a system, we can obtain very rich data about why the current version of the interface is less than ideal. We can then use that data to develop and further test design improvements that explicitly support the issues needing greater attention. Likely areas of confusion include the nature of federated collections, metadata errors, inconsistencies in metadata, the different context of use of metadata in a federated collection from its home collection, vocabulary, and mismatches between the teachers' needs and what is available. As well as informing interface redesign, these findings will also be fed back to the rest of the project as recommendations for improvements in content, format and representation. The iterative prototyping approach allows the rapid exploration of a range of design innovations and a continual accumulation of a richer understanding of user behaviors in an interaction between their needs, their prior experiences, the data obtained and how this data is mediated via the interface. We expect that both the process and the results to be of interest to wider communities of digital library developers, developers of museum informatics resources (whether federated or not) and the wider HCI community. In particular, our use and refinement of 'extreme evaluation'[18] techniques to be compatible with rapid prototyping, which proved successful with basic museum websites[16] will be of great interest not only to researchers, where it remains controversial, but also to practitioners who often despair of the time consuming and expensive techniques of traditional rigorous scientific usability experiments and consequently fail to do any usability testing at all, or do it too late into a project for the results to be used to improve the design.

Streamlining Workflows & Maintenance: Current metadata harvesting requires frequent manual intervention (providers break, network problems arise, providers disseminate invalid records which can disrupt harvests). Success of all downstream processing and indexing depends on the harvesting step. Early experience suggests a need for more automated ways to track success/failure of harvests and to bypass provider data problems that affect only a small portion of records. Downstream processing needs to be streamlined and managed in such a way that small anomalies upstream are bypassed. Currently, collection registry record creation and vetting and upload of edited registry records is labor intensive. Utilities to streamline and facilitate these processes are needed. The intellectual effort to define tool modifications and added tools necessary to accomplish these objectives has been done. Not all of these upgrades have been implemented, and none have yet been fully tested. In the extension phase of the project we will update harvest and processing software to be more fault-tolerant and will implement a new module to manage harvest, processing (automated normalization and enrichment), and indexing end-to-end. Process logging will be further expanded and new real-time Web views of harvest, process, and indexing logs will be made available to facilitate management of the process and make it less labor-intensive. Templates to facilitate initial collection registry record creation will be implemented, and a more automated notification system will be implemented to facilitate vetting and upload of collection registry record edits. To help effectiveness of these upgrades we will benchmark system performance early in extension phase and then again late in the calendar, after upgrades have been implemented.

Research on Knowledge Diffusion: In the proposed extension, we will develop and extend our work on metadata expertise development and diffusion using survey and interview methodology that has worked well during first 3 years of the project. Through additional interviews with metadata librarians, we will track how they keep up to date and develop their base of knowledge on metadata standards and applications. In addition we will assess how metadata professionals are diffusing their knowledge back into the digital library and museum communities. We will lay foundation for this work by adding a segment on this topic to the large-scale survey being conducted this fall in the final stage of the original grant. Results will provide baseline data on which we can build further in subsequent interviews. The interview and survey data will help identify ways to promote more open and immediate lines of diffusion within practice and LIS education. In addition, our work with two focus groups at Webwise during original grant period proved to be extremely valuable in terms of data collection and in building awareness of central issues in collection federation. In the final year of proposed extension, we expect to conduct a third, capstone focus group with key NLG and invited LSTA PIs that will cover collection and granularity topics and metadata expertise development and best practice diffusion.

Pilot Inclusion of Collection Descriptions & Metadata from LSTA Grantees: While there is overlap between the NLG and LSTA grantee communities, that overlap is not great. In most states LSTA grants tend to be smaller and of shorter duration. LSTA often funds smaller academic libraries and more public libraries than typically funded by NLG. Both programs include a large proportion of digital content creation and development grants. To gain a measure of experience with a sample of LSTA grantees and complement what we've learned working with NLG grantees, we propose during the extension phase of this grant to work with a subset of Illinois State Library LSTA grantees funded to create or develop digital content. We will add LSTA grants in two stages. In academic year 2005-06 we will work intensively one-on-one with 5 or 6 exemplar LSTA grantees (identified with the help of **Alyce Scott** and **Joe Natale** of the ISL). We will add descriptions of their LSTA-supported digital collections to the collection registry and where feasible add item-level metadata to the repository. (Candidate collections we would like to approach include the *Digital Past* [North Suburban Library System], *Windows on Our Past* [Chicago Public], *Upper Mississippi Valley Digital Image Archive* [Augustana College], *Abraham Lincoln Historical Digitization Project* [Northern Illinois University]. At least 2 of these collections make use of ContentDM which will facilitate harvesting of item-level metadata.) While we do not propose funding any of these projects directly for their participation, we will budget to allow site visits by project staff and to bring personnel from these libraries to Urbana for a face-to-face meeting in mid 2006. In academic year 2006-07, we will then open up LSTA participation to as many more Illinois LSTA grantees as can be accommodated. Additional site visits will be arranged as appropriate. Both the collection registry and item-level repository will be accessible in a mode that excludes LSTA content (i.e., an NLG only view), although we anticipate that the default view by the end of 2006 will include both NLG and LSTA content integrated together. This work will allow improved extrapolations of requirements to support inclusion of LSTA digital collections and content nationwide and will identify issues unique to LSTA participation

Evaluation: Many of the goals for the proposed extension phase of the project will have measurable indicators of success and accomplishment. As with the initial phase of work, a primary measure will be the level of participation and inclusion in the registry and repository. We have achieved near 100% participation in the collection registry as measured in number of collections represented. Survey responses and direct viewing/editing of registry records by NLG projects has also been high. An indication of success during the proposed extension will be an ability to maintain these high levels of participation in the collection registry. Participation in the item-level metadata repository is about half of what we think feasible, based on our assessment of NLG grantees to provide item-level metadata via OAI. An indication of success during the extension phase of the project would be an increase in participation in the item-level repository. Other measurable outcomes will include improvements in metadata harvesting, processing, and indexing performance (time per record, end-to-end completion rate without manual intervention) and percentage of records normalized or enriched with regard to specific attributes. By the nature of this project, evaluation will also be provided by the IMLS grantee community. To help provide this evaluation we would like to extend the recharge of the project Steering Committee, with membership changes as necessary given availability, changes in jobs, etc. We also would propose hosting a grantee workshop early in 2007, in conjunction with Web-Wise, to garner feedback on project results and the utility (or lack thereof) of the registry and repository and input on what should be done next with these resources.

4. PROJECT RESOURCES

Personnel Resources: The project PI (**Timothy W. Cole**) and 3 of the original co-PIs (**Carole L. Palmer**, **Michael Twidale**, and **William H. Mischo**) will continue for the extension phase of the project. **Sarah Shreeves** (former project coordinator for this grant) and **Nancy O'Brien** (Head of our Education and Social Sciences Library) will be added as co-PIs. We also will add for the proposed extension the involvement of **Alyce Scott** and **Joe Natale**, Illinois State Library, to advise on work with LSTA grantees, and the participation of **Diny Golder** and **Stuart Sutton** of the GEM Exchange, to oversee experiment to ingest IMLS collection and item-level records into a test region of the GEM Gateway, to advise and consult on use of GEM cataloging tools and controlled vocabularies, to help inform IMLS grantees about the possible uses of GEM utilities, and to perform statistical analyses of IMLS grantee metadata to help determine what it will take to make records more useful in education context.

A listing of permanent University of Illinois staff participating in this project during the proposed extension, along with their titles and roles in the project, is provided with the budget narrative. Temporary staffing during the extension phase of the project will be similar to staffing for the first 3 years of the project, including a full-time visiting project coordinator (current incumbent **Jenny Benevento**, Visiting Assistant Professor of Library Administration) with involvement in all parts of the project. Proposed extension includes funding for two library and information science **Graduate Research Assistants** (to be named), one quarter-time and one half-time (initial grant provided funding for 2 quarter-time RAs, change reflects increased emphasis on research on now functional resources). Library science RAs will assist with research activities described above, especially interface design and testing, research on collection identity, and research on knowledge diffusion. Also included in proposed extension budget is one half-time computer science **Graduate Research Assistant** (to be named) for one year (a reduction from a half-time research programmer for 2 years in the original grant, again reflecting fact that registry and repository have been implemented). The computer science RA will assist with implementation of metadata processing and streamlining of service maintenance. The project will continue to be housed both in the Grainger Library, which will continue to host services and provide server disk space and system administration, and in the Graduate School of Library and Information Science Research Lab, which will continue to provide computer support for GSLIS researchers and library science RAs.

Management Plan: As currently, the PI and Co-PIs will direct and supervise all aspects of project work. Cole, Palmer, and Twidale have lead and management responsibilities for specific project components and staff. Benevento, as project coordinator, insures full and continuous flow of information among project staff and between project and grantees. Regular project meetings are held to facilitate communication and share results. The University Library and GSLIS Business Offices work with the University Grants & Sponsored Contracts Office to oversee finances and insure conformance with regulations. This is an ongoing project with a proven and effective management plan already in place.

5. DISSEMINATION

Results will continue to be reported in the scholarly literature and at appropriate scholarly meetings (e.g., ACRL, ASIST Annual Meeting, JCDL, ECDL, Web-Wise, DLF Forum). Project results also have been and will continue to be disseminated through listservs and collaborative wikis (e.g., DLF/NSDL OAI Best Practices wiki), and through publication on the Website of white papers, schemas, and other project documentation. Two of the participants are under contract to complete a text for Libraries Unlimited on the use of OAI-PMH. Since the backgrounds of the participants span traditional library and museum domains and include integral ties to the scholastic arena provided by the GSLIS, they are well positioned to take advantage of numerous forums in which this research can be exposed, exploited, and built upon.

6. SUSTAINABILITY

This work is being performed in response to a specific Request for Proposals issued by IMLS in early 2002. The need for a collection registry and item-level metadata repository dedicated to digital collections and content created or developed with IMLS funding is of value first and foremost to IMLS, the community of IMLS grantees, and consumers of resources developed under the auspices of IMLS. It is not clear that an IMLS-specific collection registry and metadata repository would ever be self-sustaining absent ongoing support from IMLS; however, the technologies and approaches utilized in creating and maintaining these resources are widely adaptable and an IMLS collection registry and metadata repository are well positioned as possible pre-cursor components of more general resources, e.g., a national digital library of cultural heritage. The University of Illinois Library is committed to exploiting technologies and lessons learned on this project for other work, notably the DLF Aquifer project and the ongoing CIC OAI-based metadata sharing collaboration. More directly sustainable is the enhanced capacity and competencies developed by IMLS grantees participating in this project by sharing collection descriptions and item-level metadata. Once implemented for one project OAI data providers can be ported and maintained for other projects relatively easily and at relatively little expense. Currently most IMLS grantees having implemented OAI metadata provider services to disseminate metadata to this project are also being harvested by other non-IMLS projects (e.g., NSDL, OAIster).

[1] Shreeves, S.L. & Cole, T.W. 2003. Developing a Collection Registry for IMLS NLG Digital Collections [Poster Abstract]. In *DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop* p. 241-242

[2] Palmer, Carole L. and Ellen M. Knutson. 2004. Metadata practices and implications for federated collections. In *Proceedings of the 67th Annual Meeting of the American Society for Information Science and Technology* , Edited by Linda Schamber & Carol L. Barry. Medford, NJ: Information Today, Inc: 456-462.

[3] Stvilia, Besiki, Les Gasser, Michael Twidale, Sarah L. Shreeves, and Timothy W. Cole. 2004. Metadata quality for federated collections. In *Proceedings of ICIQ04 - 9th International Conference on Information Quality*. Cambridge, MA : 111-125.

[4] Foulonneau, Muriel, Timothy W. Cole, Thomas G. Habing, Sarah L. Shreeves. Using Collection Descriptions to Enhance an Aggregation of Harvested Item-Level Metadata. In *JCDL 2005: Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries [Denver, June 7–11]*. New York, Association for Computing Machinery (2005, in press)

[5] Sutton, Stuart A. (2004). "Digital Library Infrastructure: Metadata & the Education Domain." In *Metadata in Practice*. D. Hillmann & E. Westbrooks (Eds.). Chicago, IL: ALA Editions, 1-16.

[6] See, for example, Lee, H. (2000). What is a collection? Journal of the American Society for Information Science. 51(12): 1106-1113; CIDOC. (2002). Definition of the CIDOC object-oriented Conceptual Reference Model version 3.4. Available: http://cidoc.ics.forth.gr/docs_/cidoc_crm_version_3.4.rtf; Johnston, P. & Robinson B. (2002). Collections and collection description. Collection description focus briefing paper. No. 1.

[7] IFLA Study Group on the Functional Requirements of Bibliographic Records. 1998. *Functional Requirements of a Bibliographic Record: Final Report*. UBCIM Publications. New Series. vol. 19. Edited by Marie-France Plassard. München: K.G. Saur Verlag GmbH & Co. KG.

[8] http://www.medialab.nl/index.asp

[9] Svenonius, Elaine. 2001. *The intellectual foundation of information organization*. Cambridge, MA: MIT Press.

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

[11] Michael Heaney (2000). *An Analytical Model of Collections and their Catalogues*. Available: http://www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf [accessed 18 May 2005].

[12] Foulonneau, Muriel and Timothy W. Cole (in press). Strategies for reprocessing aggregated metadata. In *9th European Conference on Digital Libraries, ECDL 2005*, September 18-23, 2005, Vienna, Austria. (Proceedings Series: *Lecture Notes in Computer Science*.) Heidelberg: Springer-Verlag.

[13] Digital Library Federation and NSDL OAI and Shareable Metadata Best Practices Working Group: http://oai-best.comm.nsdl.org/cgi-bin/wiki.pl?OAI_Best_Practices

[14] Crabtree, A., Nichols, D.M., O'Brien, J., Rouncefield, M. and Twidale, M.B. (2000). Ethnomethodologically-informed ethnography and information systems design. JASIS 51 (7) 666-682.

[15] Crabtree, A., Twidale, M.B. O'Brien, J. & Nichols, D.M. (1997). Talking in the Library: Implications for the Design of Digital Libraries. Proceedings of ACM Digital Libraries '97, (Eds.) Allen, R.B. & Rasmussen, E., Philadelphia, PA, 221-228.

[16] Marty, P.F. & Twidale, M.B. (2004.) Lost in gallery space: A conceptual framework for analyzing the usability flaws of museum Web sites. First Monday 9(9).

[17] Twidale, M.B. & Nichols, D.M. (1998). Designing Interfaces to Support Collaboration in Information Retrieval. Interacting with Computers, 10(2), 177-93.

[18] Marty, P.F. & Twidale, M.B. (2005). Extreme Discount Usability Engineering. Technical Report ISRN UIUCLIS--2005/1+CSCW.