Institute of Museum and Library Services
Digital Collections and Content
Working toward interoperable digital content.

University of Illinois
1301 W. Springfield
Urbana, IL 61801
Tel 217.244.7809
Fax 217.244.7764

Grant LG-02-02-0281
Interim Performance Report 7
April 2006 – September 2006
Submitted by
Timothy W. Cole, Principal Investigator
Amy Jackson, Project Coordinator
October 2006

Summary

During the past six months, the IMLS Digital Collections and Content (DCC) Project has continued to make progress toward stated goals and objectives. A second survey was sent to the initial 100 projects to track the evolution of digital projects, and we continue interviewing participating projects and adding new collections to the collection registry. A preliminary group of LSTA projects have also been added to the collection registry. A new search interface focusing on individual collections was tested during this period, and we continue to integrate item-level and collection-level metadata searching. Item-level metadata reprocessing and augmentation techniques were examined, with plans to implement these findings during the next performance period. As of September 2006, the IMLS metadata repository contained 245,012 records from 33 OAI-compliant NLG projects, and the collection registry contained records for 167 NLG digital collections. Our team also continues to publish and present findings from research performed as part of this project, and to provide advice on metadata design and implementation.

General Project Activities

Project Staffing

A new project coordinator, Amy Jackson, was hired and began working on the project in September of 2006, replacing Jenny Benevento.

Website and Search Developments

A new interface featuring a subset of DCC collections was tested during the summer of 2006 (available at <http://cicharvest.grainger.uiuc.edu/heritage/collections.asp>). This interface shows both items and collections and allows grouping of item results by collection. We are currently discussing the usability of this approach and scalability to the larger collection registry.

We also investigated item-level metadata reprocessing and augmentation and plan to integrate these findings into the regular workflow in the next performance period.

Timeline

Our project's goals and targets continue to be met as scheduled on the timelines adjusted on previous interim reports. Work on metadata normalization, enrichment, and transformation continues on target, as does streamlining of processing and maintenance. Research into collection identity and metadata granularity continues through analysis of surveys and harvested metadata, and usability of new interfaces is currently being studied. Survey Two is being analyzed regarding development of expertise for collection managers, and methods for inclusion of LSTA data are being discussed. We have also continued work with GEM on ingesting DCC records into their test region (see Appendix Two for GEM interim report). We continue to provide information and assistance to NLG and LSTA projects regarding metadata creation and interoperability and OAI implementation.

Financial Status Report

The Annual Financial Status Report (Appendix One) has been forwarded to the IMLS Grants Administration office from UIUC's Grants and Contracts Office.

Dissemination

The IMLS DCC project staff and investigators have published and presented on the various standards, protocols and research findings from the project in several forums.

In March 2006, Besiki Stvilia defended his doctoral dissertation "Measuring Information Quality" which incorporated analysis of the metadata quality in the IMLS DCC item repository. The dissertation will be disseminated through the university's new institutional repository. Other papers that led to the thesis work have been recorded in earlier reports. A recent paper that informed this work is:

Shreeves, S., Knutson, E., Stvilia, B., Palmer, C., Twidale, M., Cole, T. (2005). Is Quality Metadata 'Shareable' Metadata? The Implications of Local Metadata Practices for Federated Collections. In H.A. Thompson (Ed.) Proceedings of the Twelfth National Conference of the Association of College and Research Libraries. (pp. 223-237). Minneapolis, MN. Chicago, IL: Association of College and Research Libraries.

The project team's paper focusing on collection identity -- *Palmer, C., Knutson, E., Twidale, M., & Zavalina, O. (2006). Collection Definition in Federated Digital Resource*

Development. In Proceedings of the 69th Annual Meeting of the American Society for Information Science and Technology -- was accepted in April 2006 by the review committee for the American Society for Information Science and Technology annual meeting to be held in Austin, Texas in November 2006. The paper was updated according to the reviewer's recommendations and the final version submitted in June 2006.

In June 2006, the analysis of the collection registry transaction logs was submitted as a GSLIS technical report -- *Zavalina, O. (2006). User Searches in IMLS DCC Collection Registry: Transaction Log Analysis. Technical Report UIUCLIS--2006/3+IMLS, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, Champaign, IL* https://netfiles.uiuc.edu/zavalina/DCC_Project/Tech_report.doc -- and reported at the metadata roundtable.

Jenny Benevento presented *IMLS Digital Collections & Content and LSTA Grantees* to the Statewide Digitization Planners Meeting at ALA Annual 2006 in New Orleans, LA on June 24, 2006.

The Chronicle of Higher Education published an article describing the registry – Brock, Read (2006). "Federal Agency Unveils Database of Digital Collections from Museums and Libraries" in *The Chronicle of Higher Education* v. 52 (33), p.41.

Research

Data collection and analysis

The second round online survey of the 100 initial projects that responded to the first survey in 2003 was conducted in February-April 2006. The focus of this survey is on tracing changes over time in the type of material in the digital collection, metadata schemes used, the intended audience, and other specifics about the digital collection and its technical implementation. The data collection is completed; the data analysis is ongoing to further extend our understanding of changes in collection identity perceptions, audiences, metadata schemes and controlled vocabularies used, etc. The data from this round of the survey, as well as from February 2006 interviews with digital collection developers conducted at the Web Wise conference and content analysis of the registry records, has been incorporated into the white paper on IMLS/NISO Framework of Guidance for Building Good Digital Collections.

In April-June 2006 additional content analysis of the IMLS DCC Collection Registry records regarding collection development policies, sub-collection delineation, and changes in Registry records made between January and June 2006 was conducted, with results incorporated into the final submitted version of the research team's collection identity paper that will be presented at 2006 annual meeting of ASIST.

We have continued to analyze the transaction logs of registry use to assess types of searches conducted and identify correlations among subject keywords used by registry

searchers with the GEM subject scheme used in the collection level description as well as with others widely used in the cultural heritage domain controlled vocabularies (e.g., Library of Congress Subject Headings and Art and Architecture Thesaurus). Analysis of about 500 user keyword searches in the collection registry made in February-September 2005 demonstrates a high level of subject searching made at the collection level (70% of the keyword searches represented subject-type search, including concept, class of persons, object, national/ethnic group, place, and event). We have also discovered the lack of semantic match between user queries and subject terms in controlled vocabularies. Only 2.6% of user search terms were matched in GEM subject scheme, while 22.63% were matched in Art and Architecture Thesaurus, and 71.3% in LCSH. GEM as the subject scheme representing only concepts seems to be incapable of meeting wide range of user queries in IMLS DCC collection registry.

In the usability track, we have been focusing on evaluating and testing the next iteration of the item-level repository interface. Our approach is a series of formative evaluations continually informing the ongoing iterative design of the interface. This effort has consisted of two components: a detailed evaluation of several of the most sophisticated, widely known and popular digital libraries/federated search applications -- with an eye towards discerning best practices in the presentation of item-level search results and linking -- and usability testing of the most recent iteration of the search interface, currently under development. Work on the competitive analysis portion is completed and indicates that the latest iteration of the IMLS DCC site now is consistent with best practices in the field, addressing the challenges of not only providing multiple kinds of access to the resources available, but also helping new users to understand the nature of a federated collection consisting of both item-level and collection-level information. User testing with a focus on two important but distinct constituencies of potential users — librarians and teachers — was conducted between June-August 2006. A preliminary analysis of the data collected indicates that the latest iteration of the IMLS DCC interface supports and encourages collection-centric navigation and searching. A substantial portion of users, when presented with tasks, sought to identify the most relevant collections — based first on collection title and second on collection descriptions — in order to search within the those collections. This indicates the desirability of highly descriptive collection titles as well as concise collection descriptions that indicate collection coverage, as users frequently eliminated as irrelevant collections that contained relevant materials. Users generally found the presentation and content of brief search results sufficient, particularly when seeking visual materials. Despite the availability of full metadata records on the IMLS DCC site, users preferred to examine the item within its collection context. One area of ongoing concern is the usability of the "Relevant Collections" box on the search results page. Although the feature is potentially very valuable, its function is unclear to many users and, due to its placement and visual similarity to Google Ads, is frequently overlooked. Further data analysis is ongoing.

Research Plans October 2006 through March 2007

In October 2006 the results of usability testing conducted in summer 2006 will be reported at the metadata roundtable and submitted as GSLIS Technical Report.

We will continue conducting transaction log analysis of the queries submitted to the IMLS DCC Collection registry in 2006 in conjunction with the [same-procedure] analysis of the user queries in the Item Repository.

Recent online survey and Web Wise interview analysis will be incorporated, along with other data, into a more developed paper on some themes of interest that emerged in the course of developing the ASIST paper.

We will also survey some of the LSTA grantees regarding this community's readiness to participate in the collection registry and item-level repository. In particular, we are interested if individuals involved in overseeing creation and management of LSTA projects/collections are emphasizing or including digital component in LSTA programs, and what is the best way to add collection information and item-level records (e.g., at state level or individual project level).

Related Activities

Metadata Roundtable

We continued to hold the metadata roundtable study group with a weekly frequency. We are very pleased that the number of the regular participants in the metadata roundtable has increased. Recent roundtable topics have included Is Metadata Dead?; *The Framework of Guidance for Building Good Digital Collections*; Collection Definition; Calhoun Report; Subject Access to Federated Collections: a Case of IMLS DCC Collection Registry; and Whole-part Relationships and Boundaries in the Context of VRACore Metadata. The website, which includes a full listing of the metadata roundtable topics and background readings, can be found at: http://www.isrl.uiuc.edu/~dcc/mdrt.html.

Report on the Framework of Guidance for Building Good Digital Collections

The project team produced a report on how *The Framework of Guidance for Building Good Digital Collections* is being used by the digital library community. Results of surveys one and two inform this report, and a discussion of the *Framework* was brought to the metadata roundtable. The final version of the report includes sixteen recommendations to NISO suggesting potential ways to improve the impact and/or utility of the *Framework*, and eight related research opportunities that IMLS may wish to consider.

GEM Exchange Interim Report
Grant Code A8808; Sub-award No: 2003-03633-01
Report Period: April 2006 – September 2006
Submitted by
Stuart A. Sutton, GEM Exchange (University of Washington)
Diny Golder, GEM Exchange (JES & Co.)
September 30, 2006

GENERAL PROJECT ACTIVITIES:

During the April 2006 through September 2006 report period, work has been completed on Task 2: ingesting IMLS DCC records into Test Region (Dec. 2005 – July 2006), and progress made on Task 3: analysis of metadata effectiveness and design enrichment. To date, we have loaded 163 collection-level records and 43,000 item-level records from ten collections into the GEM staging area (test region) set up under Task 1.

Project Staffing:

In addition to Stuart Sutton (investigator) and Ryan Laundry (GEM technical lead) at the University of Washington and Diny Golder with JES & Co., Ok Nam Park (PhD candidate) at the University of Washington joined the project to assist with data analysis and other technical tasks. Building on the preliminary work of Hillmann and Phipps during the last report period, Park was responsible for data manipulations (described below) needed to bring the collection- and item-level records into the staging area.

TASK 2: Ingesting IMLS DCC Records into Test Region (Dec. 2005 – July 2006):

Collection-Level Metadata Records:

Based on XSLT transformations developed during the last report period, we have completed the loading of 163 collection-level records into the staging area. As we noted in our last report, an RDF modeling of all of the information in an METS "description sets" (as defined in the DCMI Abstract Model) would include the modeling of a number of separate resource descriptions—a description for the collection resource itself and related descriptions for persons, projects and institutions.

We noted in our last interim report that much of the information in the related resources associated with the collection-level record are particular to the purpose of the UIUC DCC project. At the time of our last report, we were still determining which statements in these related resources would be carried over into the GEM representations of the collection-level record. Near the beginning of this report period, decisions were finalized regarding these open issues. In essence, we determined that we could not justify the extension of the level of GEM description beyond that currently framed in the GEM schemas which focused on resource discovery as opposed to deep description. In other words, the UIUC metadata needed to be

pared down to conform to the GEM schemas while trying to maintain as much of the richness as possible from the UIUC metadata.

GEM RDF does not currently model agents as separate resources and instead represents the values of these agent properties as string values in the resource description. What this means is that UIUC elements describing the characteristics of agents were eliminated in our transformations while those agent elements that comport with the one-to-one rule with regard to the collection description were retained. However, this meant ferreting out the values for the remaining properties from the METS records. Thus, for example, the separate UIUC publisher agent resource with URI "http://imlsdcc.grainger.uiuc.edu/Registry/Person/?723" is transformed from its XML/METS representation of <vcard:ORG><vcard:Orgname>Southern Utah University. Gerald R. Sherratt Library.</vcard:Orgname></vcard:ORG></vcard:VCARD> to "<dc:publisher>Southern Utah University. Gerald R. Sherratt Library.</dc:publisher>" in the GEM record. Where ever we were able to make these transforms within the contexts of both a single METS file and the GEM schema, such transforms were done prior to loading the collection-level resources into the staging area.

These changes are what we are calling Level 1 transformations. In Level 1: (1) we eliminate what is not to be carried forward from the METS documents—(a) elements that are not equivalent to any current GEM or DCMI property; and (b) elements that do not violate the one-to-one rule in terms of the collection description; (2) we "transcribe" values from agent elements to collection description properties (e.g., the dc:publisher example above); and (3) we transform the records from XML to RDF/XML Level I transformation provide us with the lowest level of interoperability with existing GEM metadata records produced by GEM Consortium members including basic faceting capability on critical facets in the Seamark RDF search and navigation system.

Full interoperability with GEM Consortium-produced metadata is dependent on Level 2 refinement. Using the data set from the Level 1 transformations, Level 2 will transform controlled vocabulary value strings (as defined in the DCMI Abstract Model) to value URIs—particularly GEM subjects and, where possible, audience terms that can be mapped to GEM audience URIs. We are currently working on the Level 2 collection-level transformations.

Item-Level Metadata Records:

We noted in our last interim report that the greatest challenge we faced at that time was the lack of an explicit mapping between IMLS collections and their component items. GEM's RDF retrieval engine requires that all relationships between resources—e.g., an item-level record and the collection-level record for its parent collection—be an explicit metadata statement expressing that relationship by means of an RDF property (here, a dcterms:isPartOf property in the item-level record). While these relationships between an item-level record and its parent might be managed programmatically in a conventional data management system, such is not the case with an RDF engine.

Since mid-April, we have had mapping information from UIUC linking unharvested collections to their parent collection records in the UIUC DCC system. Some mappings between the

unharvested, unqualified Dublin Core item-level records and their parent UIUC collection-level record were straightforward while many others would have required substantial post-harvest filtering to derive meaningful collection sets matching the UIUC collection-level records. In order to concentrate our limited resources on the analysis phases of the project, we have limited our item-level harvesting to ten collections where the relationship between items and the parent collection were straightforward and unambiguous. This has given us a current collection of 43,000 item-level records which we think is sufficient for our current analytic purposes. Time and resources permitting, we will harvest additional item-level records demanding more post-harvest processing to isolate collections.

The item-level records also have Level 1 and Level 2 transformations. The Level 1 transformation has already been completed and is made up primarily of: (1) declaring the explicit relationship between the 43,000 item-level records and their parent collections; and (2) transforming the records from conventional XML to RDF/XML. The Level 2 transformations to item-level descriptions have yet to be done and will include attempts to enrich the descriptions for educational purposes through the addition of audience, educationLevel, and instructionalMethods where such statements can be inferred from collection-level information. Additional transformations on keywords may be attempted for keyword transformation to gem:subject. Whether such enrichments can be reliably made will be dependent on a number of factors including the reliability of metadata statements in the associated collection-level records.

TASK 3: Analyze Metadata Effectiveness and Design Enrichment (Apr. 2006 – Jan. 2007):

Analysis of the metadata effectiveness for the Level 1 transformations of both the collection- and item-level records is ongoing and, in general, will continue throughout the second year of the sub-contract. Our current plan is to perform the analysis using two primary technical mechanisms—(1) facet analysis using the built-in functionality of the Seamark faceted retrieval engine; and (2) more detailed content analyses of GEM-critical properties using the Spotfire DecisionSite software. The first of these mechanisms provides us with the means of viewing the effectiveness of the metadata within the context of an advanced search engine using integrated search and browse that supports end-user exploration of the metadata by means of faceted navigation. The GEM staging area for the IUIC/IMLS metadata was set up with all facets enabled with the exception of dc:title and dc:description where we assumed at this point in our research that: (1) faceting would serve no meaningful purpose; and (2) the search engine's keyword functions would be the appropriate mechanisms for discovery and retrieval from these full-text fields. Thus, our assertion is that the metadata effectiveness of the UIUC/IMLS metadata would hinge largely on facet performance—i.e., the adequacy and quality of metadata values in key facets used in the production search engine. Final reporting of our work will include two core judgments—metadata effectiveness within the current GEM environment and metadata effectiveness with use an array of facets available in the metadata that goes beyond those currently enabled in GEM.

Our base assumption at this time is that all enrichments to be accomplished through this subcontract will be within the context of the metadata on hand and that no attempts will be

made to perform enrichments by returning to the original resources beyond a systematic spot-checking against resources to make assessments of metadata quality in terms of intended use. Thus, mechanisms such as automated metadata generation of item-level metadata for purposes of enrichment will not be attempted.

For the facet analysis, we have created three metadata partitions in the staging area: (1) a collection-level partition, (2) an item-level partition, and (3) a partition with a combination of all collection- and item-level metadata records. When work under the subcontract is complete, it is our current intention is to provide analysis conforming to the following matrix:

	Facet Analysis		**SpotFire Analysis**	
	Existing Facets	Expanded Facets	Existing Facets	Expanded Facets
Collection-Level Metadata Alone				
Item-Level Metadata Alone				
Combined Item- and Collection-Levels				

IMLS Grant – University of Illinois at Urbana-Champaign

Project Title: Development of a Registry and Metadata Repository for Digital Collections

Grant Code A8808 must be referenced on every invoice (also use on reports)
Subaward Number: 2003-03633-01

Project Director: Dr. Timothy Cole t-cole3@uiuc.edu
216 Altgeld Hall (MC-382)
1409 W. Green Street
Urbana, IL 61801
Tel 217-224-7837
Fax 217-244-4362

Reports: Progress reports submitted periodically to Project Director
Interim Narrative Performance Reports April 15 and October 15 (assuming that year 2 is optioned and an amendment is executed)
Final Performance Report is Due Sept. 30
Annual financial status report must be attached to the fall report

Invoicing
Not more frequently than monthly
Itemize current and cumulative costs by budget category, per the approved budget
Final invoice due within 45 days of contract expiration
Mail to: University of Illinois
c/o Grants and Contracts – Post Award
Attn: Denise Connour
1901 S. First St., Suite A
Champaign, IL 61820

Certification language on every invoice:

I certify that all expenditures reported (or payments requested) are for appropriate purposes and in accordance with the Agreements set forth in the application and award documents.

Certification must be signed!

Budget (both years identical)
Direct Costs:

Salaries & Wages	18,000
Fringe Benefits (9.45%)	1,700
Travel	2,000
Indirect Costs (18.085%)	3,255

Project Calendar (based upon contract date of Oct. 1)

October 2005 to Mid-February 2006: Test Region in GEM

December 2005 through July 2006: Ingest IMLS DCC Records into Test Region

April 15, 2006: Interim Narrative Performance Report due

April 2006 – January 2007: Analyze Metadata Effectiveness & Design enrichment

October 15, 2005: Interim Narrative Performance Report and financial status report due

October 2006 – September 2007: Statistical Analysis & Long-term strategies to support education use.

September 30, 2007: Final Performance Report and financial status report due