IMLS Digital Collections and Content

Grant LG-02-02-0281

Interim Performance Report 4
1 April – 30 September 2004

Submitted by Timothy W. Cole, Principal Investigator, and
Sarah L. Shreeves, Project Coordinator
University of Illinois at Urbana Champaign
October 2004

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 4
1 April – 30 September 2004

Submitted by Timothy W. Cole, Principal Investigator,
and Sarah Shreeves, Project Coordinator
November 2004

Summary

The IMLS Digital Collections and Content (DCC) Project has made progress on a number of fronts over the past six months. After a second distribution of the survey to National Leadership Grant (NLG) projects funded in 2003 as well as those non-respondent projects funded in 1998-2002, the response rate to Survey One was 86% and 77% to Survey 2. A third distribution of the survey is planned in early summer to NLG projects funded in 2004 and to the non-respondents to the first two survey distributions. The collection registry edit and entry forms were finalized and sent to the Office of Management and Budget in mid-September for approval. The first phase of the collection registry browse and search interface is complete, and we are preparing for usability testing. We have continued to work with IMLS funded institutions to set up Open Archives Initiative (OAI) metadata provider services. As of September 2004, the IMLS item-level metadata repository contained 193,677 DC records from 26 OAI-compliant NLG projects. The research team has continued to work through the interviews and surveys it has collected, and completed a preliminary analysis of metadata quality issues.

General Project Activities

Timeline

We are moving forward on our project timeline, though several milestones have shifted from our original timeline. We hope to have approval of the collection registry by the end of November 2004, and will make the collection registry open for vetting by December 2004. The collection registry interface will be made available to participating institutions at the same time. We are now in the process of analyzing harvested metadata and hope to have a new preliminary interface to the harvested metadata to present at the 2005 WebWise Conference in February / March 2005. We have also been collecting information for a report on Barriers to Interoperability and hope to make a preliminary report at the 2005 WebWise Conference.

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.

Publications

Tim Cole and Sarah Shreeves co-edited a special issue of Library Hi Tech on IMLS National Leadership Grant projects with special attention on collaborative efforts within these. Included in this issue is an article on the IMLS DCC project. See: Cole, T.W. & Shreeves, S.L. (2004). Search and discovery across collections: The IMLS Digital Collections and Content Project. *Library Hi Tech* 22(3): 307-322.

Sarah Shreeves wrote a synopsis of the IMLS DCC Collection Description work for *CD Focus News Bulletin*, a UKOLN-based newsletter for collection description work. See: Shreeves, S.L. (2004). IMLS Digital Collections and Content. *Collection Description Focus News Bulletin*. March/April. http://www.ukoln.ac.uk/cd-focus/newsletters/news200404.html.

Presentations

Tim Cole presented the "Rationale for Interoperable Metadata" at the Innodata Isogen Symposium on the Economics of Digitization at the Newberry Library in Chicago, IL on May 17 2004. See http://dli.grainger.uiuc.edu/publications/twcole/newberry2004/.

Sarah Shreeves presented on "Current Work in Collection Description and Terminology Services: A Report on Two Forums" at the E-Text Working Group of UIUC Library on May 20, 2004. See http://imlsdcc.grainger.uiuc.edu/ETextOutline.htm.

Sarah Shreeves gave an invited seminar on "The Basics of OAI: An Introduction to the Protocol for Metadata Harvesting" at the Technology for the Rest of Us Seminar Series at Ohio State University on May 27, 2004. See http://imlsdcc.grainger.uiuc.edu/OAI_Tutorial_Ohio.ppt.

Sarah Shreeves presented the "Basics of OAI" at the Illinois Digitization Institute's Basics and Beyond Track 3 Online and Hands-On Course at the University of Illinois, Urbana, IL on July 27, 2004. See http://imlsdcc.grainger.uiuc.edu/OAI_BasicsBeyond.ppt.

Sarah Shreeves presented "Digitization Training and Metadata: The View from Two UIUC Projects" at the "Truth and Consequences of Digitization" panel at the American Association of State and Local History Annual Conference in Saint Louis, MO, Sept. 30, 2004. The second half of the presentation discusses ways to think about metadata authoring culled from experiences in the IMLS DCC project. See http://imlsdcc.grainger.uiuc.edu/AASLH.ppt.

Other Forums

Sarah Shreeves and Tim Cole are participating in an OAI and Shareable Metadata Best Practices Working Group sponsored by the Digital Library Federation (DLF). This working group, formed as part of an IMLS grant to the DLF, is developing best practices for DLF members and similar institutions. Our experiences from the IMLS DCC work heavily inform the contributions we are making to this group. Sarah is serving as chair and co-editor of this work, and Tim Cole is leading a sub-group on best practices for OAI data provider implementations.

Upcoming Publications and Presentations

Besiki Stvilia, a doctoral student on the project, will present a paper on metadata quality based on data from the IMLS DCC project at the 9th International Conference on Information Quality at MIT, Boston on November 6, 2004.

The IMLS DCC project will be well represented at the American Society for Information Science and Technology (ASIS&T) annual conference in November 2004. ASIS&T has accepted a paper, "Metadata Practices and Implications for Federated Collections" co-authored by Carole Palmer, Ellen Knutson, and Michael Twidale. In addition, Carole Palmer will also be presenting a paper on metadata standards adoption and knowledge-sharing on a panel at the ASIST conference. Besiki Stvilia will present a poster on metadata quality based on data from the IMLS DCC project.

Tim Cole will be presenting the experiences of the IMLS DCC project in a panel with Richard Rinehart of the Berkeley Art Museum and Martin Halbert of Emory University at the Museum Computer Network (MCN) annual conference in November 2004.

We will be discussing the OAI protocol and its potential for use within the VRA Core community at the Visual Resources Association annual conference in March 2005.

Several members of the IMLS DCC team are collaborating on a paper titled "Is 'quality' metadata 'shareable' metadata? The implications of local metadata practices for federated collections" for the Association of College and Research Libraries conference in April 2005.

Sarah Shreeves has written a chapter presenting the basics of the OAI Protocol in an upcoming publication from Libraries Unlimited. See: Shreeves, S.L. (In press). The basics of the Open Archives Initiative Protocol for Metadata Harvesting. Chapter in N. Courtney (Ed.) *Technology for the Rest of Us: A Primer on Computer Technologies for the Low-Tech Librarian*. Westport, CT: Libraries Unlimited.

Sarah Shreeves has co-authored an article on current developments for the OAI protocol for *Library Trends*. See: Shreeves, S.L., Habing, T., Hagedorn, K., & Young, J. (In press). Current developments and future trends for the OAI Protocol for Metadata Harvesting. *Library Trends*.

Tim Cole and Sarah Shreeves are collaborating with colleagues at UIUC to write a book on the Open Archives Initiative which will be published by Libraries Unlimited.

Steering Committee Activity

The steering committee did not meet during this period, although several members were consulted about their own National Leadership Grants and on questions in their area of expertise. The next steering committee meeting is planned for the 2005 WebWise Conference.

Collection Registry Metadata Schema and Service

Survey of IMLS NLG Projects

The web-based Surveys 1 and 2 were announced in May 2004 to both new 2003 grantees and to the 1998-2002 grantee non-respondents. As of September 2004, we had a total 86% response rate on Survey 1 and a 77% response rate on Survey 2. The research team is following up on the survey results with emails and phone calls. We are planning a third round of the survey in

summer 2005 for the 2004 National Leadership Grant recipients as well as non-respondents to the first two rounds of the survey.

Survey 1 continues to yield interesting results. 75% (71) of the respondents have divided their IMLS funded collection into sub-collections based on factors such as topic, administrative unit, type of material or a combination of these. 81% (77) of respondents had item level metadata for the content in their digital collections. Of these 69% (53) were using multiple schemas. Most IMLS funded collections contain a combination of material types. Only 15% (13) of the 89 respondents to the material type question had a single material type in their digital collection. 39% (35) had a combination of image and text. Appendix Two includes a synopsis of these and other Survey 1 results.

Collection Description Metadata Schema

Sarah Shreeves has continued to be an active participant in the Dublin Core Collection Description (CD) Working Group and the NISO Metasearch Initiative Collection Description Task Group. She is currently working on a Usage Guide for the Dublin Core CD Application Profile, based in part on her experiences in the IMLS DCC project.

Development of Collection Registry

The results of the second round of the survey were used to populate the collection registry. 134 collection records were created from the survey results and then edited and expanded through information gleaned from collection websites and other communications. Entry/edit forms were submitted to the Office of Management and Budget (OMB) for approval in September 2004. A browse interface for the preliminary records was developed and is available at http://imlsdcc.grainger.uiuc.edu/collections/ (password protected). When the registry entry/edit forms are approved by OMB we will ask NLG projects to vet the collection registry entries and make the collection registry interface public to these projects. A round of usability testing for the collection registry browse and search interface is planned for late 2004.

Item-Level Metadata Repository

Assisting projects in implementing OAI-data provider services

The survey results and continued discussions with NLG recipients about implementing OAI-data provider services have given us a clearer picture of the landscape. The breakdown of 1998-2003 NLG projects (121 total) in relation to OAI-data provider services is:

Category of 1998-2002 NLG Projects:	**Number / % of NLG Projects:**
Group 1 – Projects with OAI data provider sites for NLG content	27 (22%)
Group 2 – Projects whose institutions have an OAI implementation (not yet being used for NLG content) and projects that have explicitly expressed plans to add OAI functionality	35 (29%)
Group 3 – Projects who meet certain technical criteria – e.g. have item-level metadata and a maintained web site	25 (21%)
Group 4 – Projects with no item-level metadata, no interest in providing metadata via OAI, or whose grants were given up	20 (16%)
Unknown	14 (12%)
Total	121 (100%)

We are in communication with several NLG projects about planned implementation of OAI data providers, including Richard Rinehart at the Berkeley Art Museum/Pacific Film Archive (BAMPFA) for the MOAC project (LL-90130); Patricia Michaelis from the Kansas Historical Society for the Territorial Kansas project (LL-90069); Melissa Watterworth for the Connecticut History Online (LG-30-02-0256); Chris Freeland for the Missouri Botanical Garden (LL-80066); and several others.

In addition, the project worked with Susan Pyzynski from Brandeis University (Honore Daumier Lithographs – "ND-10005') to test their OAI data provider implementation as provided through their content management system, Ex Libris' DigiTool. (We later discovered that this was the first time the Ex Libris data provider had 'gone public' and been thoroughly tested.) We discovered several problems with the Ex Libris data provider (including metadata mapping, resumption tokens, and responses to the "ListRecords" verb) which were communicated back to Brandeis who then passed them on to Ex Libris. Ex Libris has since committed to resolving these issues. This work with Brandeis illustrates the importance of constructive and interactive feedback for data providers. It would have been difficult for the automated OAI validators to provide useful feedback on some of these issues seen in Brandeis' case.

We have continued to contact NLG recipients about implementing OAI-data provider services and to document the barriers to implementation. Our goal by the end of the project remains to have approximately 50% of all NLG projects providing or committed to providing metadata via OAI.

Metadata harvesting and design of item-level repository

We have continued to harvest and index item-level metadata from NLG projects. As of September 30, 2004 we had harvested 193,677 DC records from 26 OAI-compliant NLG projects. The repository is currently available at: http://imlsdcc.grainger.uiuc.edu/search/. Sites and number of records harvested as of September 30th, 2004 is available in Appendix Three.

We have begun to harvest metadata schemas other than simple Dublin Core. For example, Brandeis provides its richest metadata in MARC21 via the OAI protocol. We have harvested the MARC21 records and have mapped them to simple Dublin Core as well as qualified Dublin Core. Our mappings have been approved by Brandeis. We are now looking at mapping the metadata of selected data providers who expose metadata in only simple Dublin Core to qualified Dublin Core and using the qualified Dublin Core as the basis for our repository.

The current interface for the item level metadata repository is an internally developed interface. The project team has begun exploring different interfaces for the item level metadata repository including the Scout Portal Toolkit (http://scout.wisc.edu/Projects/SPT/), as well as University of Michigan's DLXS product, XPat (http://www.dlxs.org/products/xpat.html). An initial exploration of the Scout Portal Toolkit indicates that it most likely would not scale to the size needed (the documentation for the Toolkit indicates 10,000 is its upper limit). It also does not provide much flexibility in terms of metadata formats available.

Research

Data Collection

We are continuing with our analysis of the interviews, focus group and survey data. The transcription of the focus group is complete and transcription of the interviews is nearing

completion. We are waiting for the collection registry to be released to National Leadership Grantees, before we start our second round of interviews.

We sent an email follow up to the survey responses seeking more information on metadata scheme selection, adequacy of scheme for search and discovery, and sub-collections. We received a 70% response rate to the email follow-up.

Our main research focus for this year is on users. Since the development team is still in the process of building the collection registry and preparing it for participant contributions we have temporarily postponed stages of data collection contingent on that work. We have begun a line of investigation to collect and analyze studies of end users already conducted by National Leadership Grantees. We contacted 23 institutions, all of which had either participated in an interview or in the focus group to ask them if they had conducted any kind of user study or are collecting any information on the use of their collection(s) through web logs. We have received information from 11 institutions (web stats, evaluation reports, survey results or some combination) and have begun to consider ways to combine the various types of user data. We are interested in pursuing a closer relationship with several projects that have collected user information through a variety of different means to better situate the user study results already being generated by participating institutions.

Metadata Quality

We have conducted additional interviews with two institutions that are using OAI to expose their metadata. The focus of these interviews was metadata quality. Besiki Stvilia completed his work on analyzing metadata quality of the IMLS DCC aggregated collection. The findings of this research were included / reflected in a paper he will present at 9th International Conference on Information Quality at MIT, Boston. A copy of the paper can be found at: https://www-s.isrl.uiuc.edu/~stvilia/iq/iciq_144.pdf.

Related Activities

Our bi-monthly metadata roundtable continues to bring members of the university library community together with students, faculty and visiting scholars of the Graduate School of Library and Information Science to discuss issues around metadata. Some topics of the last six months include: metadata quality, abstraction and abstract models, and best practices for OAI data providers and service providers. The resources associated with these activities have been compiled into a website for the roundtable members and larger community. This 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.

Appendix Two – Selected Results from Survey One

Number of surveys sent: 118 (representing 121 total projects)
Number of non-active/not-ready projects identified through survey or other communication: 7[1]
Number of surveys sent to "active" projects: 111
Number of respondents to Survey One: 95 (86%)

Sub-collections:

Number of respondents to question: "Is this collection divided into sub-collections (for example, type of material or subject area)?": 95 (100%)

Number of respondents with sub-collections: 71 (75%)

Basis of sub-collection organization:	**Number (%) of respondents with sub-collections:**
Administrative unit only	9 (13%)
Topic only	15 (21%)
Type of material only	12 (17%)
Other basis only	11 (16%)
Based on two factors:	
Administrative unit and Topic	3 (4%)
Administrative unit and Type of material	1 (1%)
Administrative unit and Other	4 (6%)
Topic and Type of material	6 (8%)
Topic and Other	2 (3%)
Type of material and Other	1 (1%)
Based on three factors:	
Topic, Type of material, and Administrative unit	5 (7%)
Topic, Type of material, and Other	1 (15)
Administrative unit, Type of material, and Other	1 (1%)

Selected 'other basis' responses:

- Could also be sub-divided according to certain aspects of the collection, e.g. Ill. State Board of Education Learning Standards or teachers' lesson plans
- Keywords
- Grade level (age) appropriateness
- Time period
- Audience sub-collections; examples: educators, journalists, historical researchers, commercial, staff, museum partners
- Donating Individual or organization
- Taxonomic (biology) description at species level; character sets; image collection
- By county

Number of respondents to question: "How many sub-collections are within your overall collection?": 50 (70% of respondents with sub-collections)

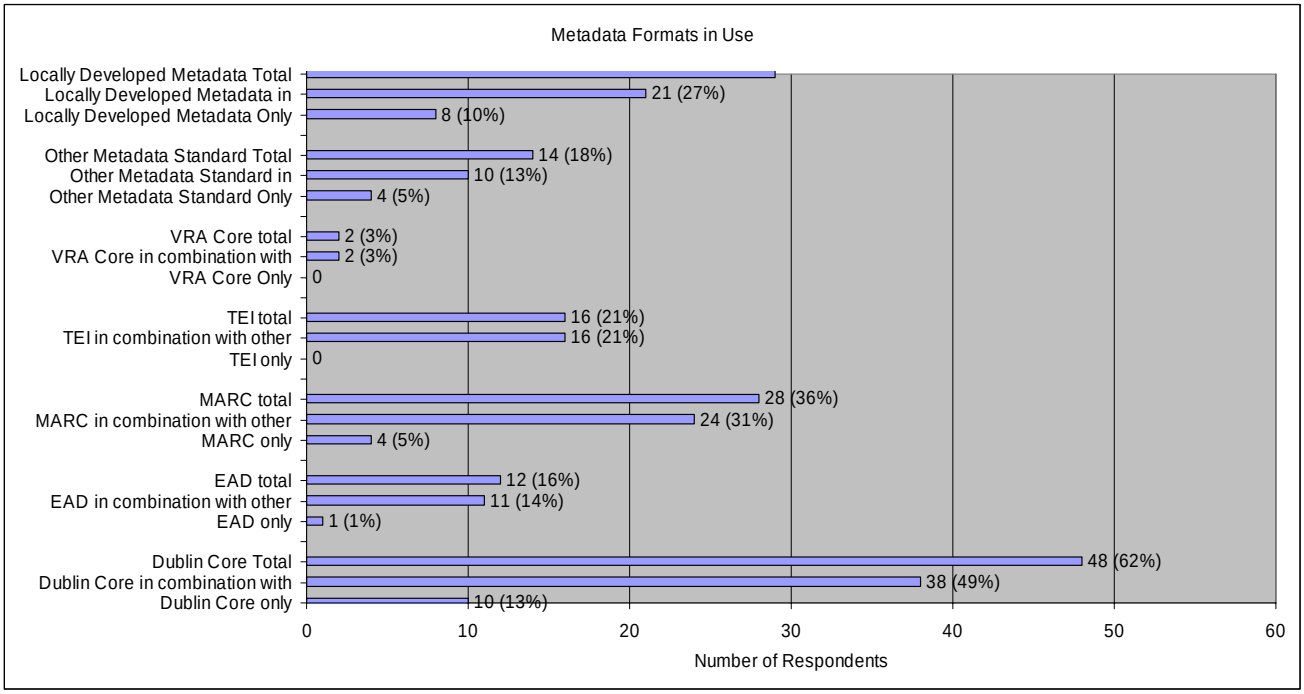
[1] Two of these are active but not yet ready to complete the survey.

Number range of sub-collections	Number (%) of respondents to question
2-5 sub-collections	20 (40%)
6-10 sub-collections	13 (26%)
11-15 sub-collections	5 (10%)
16-20 sub-collections	4 (8%)
21-30 sub-collections	1 (2%)
31-40 sub-collections	2 (4%)
41 or more sub-collections	5 (10%)

Item level metadata:

Number of respondents with item level metadata: 77 (81 %)
Number of respondents without item level metadata: 18 (19 %)

Number of respondents using just one metadata schema: 24 (31% of respondents with item level metadata)
Number of respondents using multiple metadata schemas: 53 (69% of respondents with item level metadata)



Other Standards in Use:

- Mets
- MOA2
- Museum MARC
- Darwin Core
- TDWG-SDD (Taxonomic Data Working Group - Structure for Descriptive Data)
- Western States Dublin Core Metadata Standards

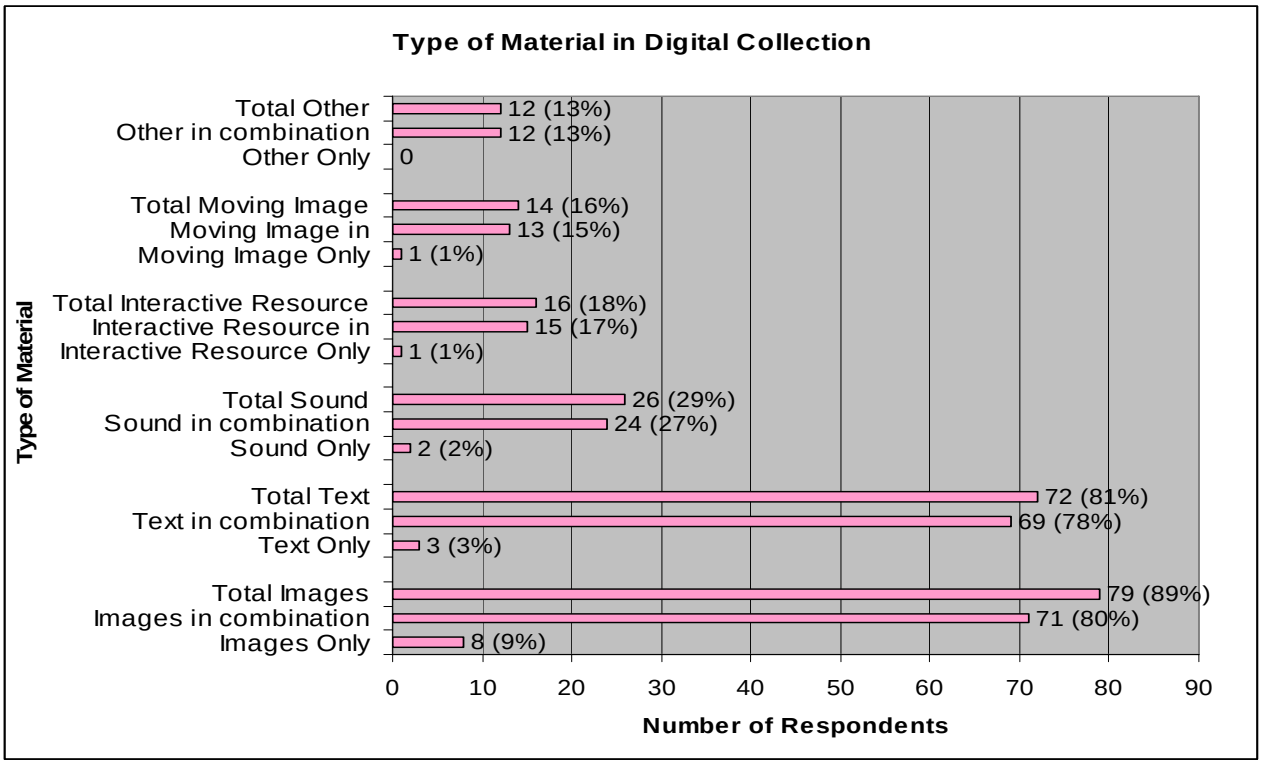
Controlled Vocabulary in Use:

Number of respondents who identified controlled vocabulary in use: 72 (94% of respondents with item level metadata)

Element	Top three used Controlled Vocabulary (% of respondents who identified C.V.)
Subject	LCSH (50%); LC TGM I (19%); AAT (13%)
Format	LC TGM II (7%); AAT (7%); MIME types (4%)
Type	DCMI Type (8%); LC TGM II (7%); AACR2 (7%)
Personal names	LC Name Authority File (47%)
Geographic names	LC Name Authority File (18%); LCSH (15%); Getty Thesaurus of Geographic Names (10%)

Type of Material in Digital Collection:

Number of respondents to question: "What type(s) of material have been digitized or created digitally?": 89 (94%)



Other Material Types in Digital Collections:

- Flash 'movies' (.swf)
- 3-D materials
- Artifact images, historic site views
- 3-D Objects eg: plates, buttons, hats, mugs, jewellry, ribbons, lanterns, textiles, pens, trinkets, ceramics, etc.
- botanic (herbarium) speciments
- maps, music scores, 3-D artifacts (photos)
- Atlas search

- Currency
- maps
- illustrations within texts, fold out maps, charts, book jackets

Three most common combinations of material types:

Combination of Material Type	Number (%) of respondents
Image and Text	35 (39%)
Image, Text, and Sound	6 (7%)
Image, Text, and Interactive Resource	5 (6%)

Other Observations from Survey:

Access restrictions:
95% (90) of respondents answered the question: "Is access to your collection limited to a specific group(s) of users?". Only two respondents placed access restrictions on their collection – and then only on a portion of their collection – specifically, copyrighted materials that could only be used for educational purposes.

Tracking use of collections:
93% (88) of respondents answered the question: "Is your project tracking usage of your digital collection through transaction log data?". Of these, 88% (77) were tracking use of their collections OR were planning to.

Collections developed prior to IMLS grant period:
95% (90) of respondents answered the question: "Was any digital content in collection developed prior to the NLG award?". Of these, 57% (51) had developed content prior to receiving the NLG award.

Continued development of collections after the IMLS grant period:
95% (90) of respondents answered the question: "Has/will digital content be added to collection after the completion of the grant period?". Of these, 82% (74) indicated that they would continue to add content to the collection, although 8 of these noted that additions would depend on additional resources (other grants, etc).

Appendix Three – National Leadership Grant Collections and Number of Records Harvested

193,677 DC records from 26 OAI-compliant NLG projects as of Oct 1, 2004

Academy of Natural Sciences
"American Natural Science in the First Half of the Nineteenth Century" - LL-90013
349 records

Alliance Library System
"Illinois Alive!" - LL-80052
111 records

Brandeis University
"Honore Daumier Lithographs" - ND-10005
3,897 records

California Digital Library
"MOAC: Collections in California Museums" - LL-90130
75,191 records

Colorado Digitization Program
"Heritage Colorado" - LL-90094
25,629 records

Colorado Digitization Program
"Western Trails" - NL-10024
6,253 records

Florida Center for Library Automation
"Florida Environmental Information Online" (Part of "Linking Florida's Natural Heritage" - LL-80016)
1,155 records

Indiana University
"Charles W. Cushman Photograph Collection" - ND-00022
14,425 records

Louisiana State University
"Louisiana Purchase Bicentennial: A Heritage Explored" - ND-00010
714 records

Tufts University
"Bolles Archive of London" - ND-00015
35 records

Tulane University - Amistad Research Center
"American Missionary Association and the Promise of a Multi-cultural America:1839-1954" - LL-90044
3,342 records

University of California, Riverside
"INFOMINE Scholarly Internet Resource Collection" - LG-02-03-0083
81 records

University of Georgia / University of Tennessee
"Southeastern Native American Documents" - LL-90019 and ND-00017
266 records

University of Illinois
"Teaching with Digital Content" - NL-00003
1,559 records

University of Maine
"Maine Music Box" - LG-03-02-0116
11,779 records

University of Michigan
"Flora and Fauna of the Great Lakes" - NL-00034
23,931 records

University of Minnesota
"Summons to Comradeship: World War I and II Posters" - ND-10007
2,306 records

University of North Carolina
"Southern Homefront" - LL-80202
403 records

University of North Carolina
"North Carolina in Black and White" - ND-00031
437 records

University of Tennessee
"Tennessee Documentary History" - ND-10020
1,207 records

University of Tennessee
“Frank H. McClung Museum WPA/TVA Photograph Archive” - LG-03-02-0080
4,011 records

University of Washington / Museum of History and Industry
"King County Snapshots: A Photographic Heritage of Seattle and Surrounding Communities" - NL-10016
11,993 records

University of Wisconsin-Madison
"Africa Focus" - LL-80131
3,650 records

Washington State University
"Columbia River Basin Ethnic History" - NL-10032
774 records

Wisconsin Historical Society
"American Journeys: Eyewitness Accounts of Early American Exploration and Settlement" - LG-03-02-0112
179 records