

Trends in Metadata Practices: A Longitudinal Study of Collection Federation

Carole Palmer
Graduate School of Library
and Information Science
University of Illinois
at Urbana-Champaign
501 E. Daniel
Champaign, IL, 61820
1-217-244-0653
cpalmer@uiuc.edu

Oksana Zavalina
Graduate School of Library
and Information Science
University of Illinois
at Urbana-Champaign
501 E. Daniel
Champaign, IL, 61820
1-217-265-5406
zavalina@uiuc.edu

Megan Mustafoff
Graduate School of Library
and Information Science
University of Illinois
at Urbana-Champaign
501 E. Daniel
Champaign, IL, 61820
1-217-244-3300
mustafof@uiuc.edu

ABSTRACT

With the increasing focus on interoperability for distributed digital content, resource developers need to take into consideration how they will contribute to large federated collections, potentially at the national and international level. At the same time, their primary objectives are usually to meet the needs of their own institutions and user communities. This tension between local practices and needs and the more global potential of digital collections has been an object of study for the IMLS Digital Collections and Content (IMLS DCC) project. Our practical aim has been to provide integrated access to over 160 IMLS-funded digital collections through a centralized collection registry and metadata repository. During the course of development, the research team has investigated how collections and items can best be represented to meet the needs of local resource developers and aggregators of distributed content, as well as the diverse user communities they may serve. This paper presents results from a longitudinal analysis of IMLS DCC development trends between 2003 and 2006. Changes in metadata applications have not been pronounced. However, multi-scheme use has become less common, and use of Dublin Core remains high, even as recognition of its limitations grows. Locally developed schemes are used as much as MARC, and may be on the increase as new collections are incorporating less traditional library and museum materials, and more interactive and multimedia content. Based on our empirical understanding of metadata use in practice, patterns in new content development, and user community indicators, our research has turned toward identifying metadata relationships between items and collections to preserve context and enhance functionality and usefulness for scholarly user communities.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL '07, June 18–23, 2007, Vancouver, British Columbia, Canada.
Copyright 2007 ACM 978-1-59593-644-8/07/0006...\$5.00.

Categories and Subject Descriptors

H.3.7 [Digital Libraries] – collection, dissemination, standards, systems issues.

General Terms

Design, Management, Standardization.

Keywords

Descriptive metadata, metadata schemes, interoperability, federated digital collections, aggregated services, IMLS Digital Collections and Content Project.

1. INTRODUCTION

Beginning in 2003, the IMLS Digital Collections and Content project has studied the metadata practices of a broad range of digital initiatives. The primary aim of the project has been to develop a resource, specified by IMLS as a collection registry and metadata repository, to provide integrated access to the digital content developed through the IMLS National Leadership Grant (NLG) program and some Library Services and Technology Act (LSTA) grants. During the course of the project, we have developed a collection description metadata schema¹ and built a centralized collection registry and metadata repository (hereafter referred to as the “IMLS DCC”) based on the Open Archives Initiative Metadata Harvesting Protocol (OAI-PMH).² To inform development of the IMLS DCC, we conducted complementary empirical studies to examine how collections and items can best be represented to meet the needs of both service providers and divergent user communities. In this paper, we report longitudinal results from this investigation that update our previous baseline reports [e.g., 14, 22].

¹ DCC collection description metadata schema is available at http://imlsdcc.grainger.uiuc.edu/CDschema_overview.asp.

² For background on the OAI approach see Shreeves, Kaczmarek, & Cole (2003) [24].

The research questions associated with the project were multifaceted³ and required study of metadata applications, interoperability challenges, and the roles of federated collections. The intent was to understand the range and evolution of metadata and interoperability issues encountered by IMLS digital projects over time and how problems can be resolved through assistance to content providers and the development of repository tools. As we developed and implemented a collection level metadata scheme and studied metadata practices more generally, we learned about the various ways that resource developers conceive of collections and what attributes they find most important in describing their collections [14]. Different “cultures of description” were apparent among the many types of participating institutions in the IMLS DCC, which include academic and public libraries, museums, archives, botanical gardens, historical societies, and other organizations [22].

Both item and collection level metadata have been essential for providing different types of discrimination amidst the aggregation. Item description supports retrieval of objects with the same attributes, but quality issues related to richness and consistency of application emerged and also required further study (see, for example, [25]). Collection description provides a broader context for understanding items and their contribution to the intentional aggregations built within the various libraries and museums participating in the IMLS DCC.

Our work developing a collection-level metadata scheme for the IMLS DCC led us to questions about the role of the collection as a defining or organizing unit in the digital environment [23]. We found that many resource developers did not have a firm idea of how many collections they were creating. Collection boundaries were often blurred, and a given digital project was sometimes simultaneously thought of as producing single and multiple collections. As seen in related studies (e.g., [9]) and in our usability testing, collection and subcollection description can help users ascertain features like uniqueness, authority, and representativeness of the objects retrieved and reduces the confusion that users sometimes experience searching large-scale federations.

We have also been interested in better ways of providing subject access to federated collections. Preliminary results on this part of the project have been reported in a paper analyzing user queries and their semantic similarities with controlled vocabulary terms [27]. The study indicated subject search prevalence at the collection-level and the need to supplement the broad subject scheme used for browsing with a more detailed controlled vocabulary to facilitate collection-level retrieval.

Results from the DCC project to date have been shared with practitioners and have informed community efforts to define best practices for metadata sharing and implementation of distributed digital libraries.⁴ Our most recent analysis, presented here, stems from our interest in trends in metadata practices and their implications for ongoing federation efforts more generally. This

paper covers primarily survey results on item level metadata applications, to identify the changes (if any) in metadata selection and application trends among IMLS DCC digital collections between 2003 and 2006.

2. BACKGROUND

Until recently, research and development in interoperability were largely concerned with technical implementation and problems associated with integrating large sets of heterogeneous digital objects,⁵ with few studies of the practices of digital library creation and federation. However, in recent years research and discussion has begun to address additional issues and challenges faced in digital collection building and federation (e.g., [1], [3], [5], [13], [16], [17], [18], [20]). An impressive level of international adoption of the OAI-PMH over the past three years has been documented [7]. But, at the same time, there has been growing awareness of the limitations of both OAI-PMH and the Dublin Core metadata scheme that serve as the basis for a significant part of current repository activity. Resource aggregators, particularly in the scientific domain, have faced challenges with OAI-PMH, Dublin Core, and metadata quality, finding the “seemingly modest architecture based on metadata harvesting” to be difficult to manage [17]. This experience has suggested a need to move from a metadata-centric to resource-centric architecture to better focus on creating and expressing context for resources. However, the metadata aggregation problems encountered in the IMLS DCC project have been moderate and have not caused bottlenecks like those experienced in NSDL development [17].

There is now a recognized need for a model and mechanisms to handle “complex objects held in repositories . . . in a more fully automated and interoperable way” [11]. However, while libraries and museums have produced thousands of successful, independent digital resources, most have not planned for long-term coordination of their digital programs with existing or future digital collection aggregations [5]. Moreover, digital repositories differ from other digital collections in important ways. As outlined by [10], primary repository functions include enhanced access to resources, general and subject access, preservation, new modes of dissemination, institutional asset management, sharing, and re-use. Whatever the approach to aggregation, the act of providing access to a large mass of distributed digital content and achieving asset management requirements do not necessarily produce collections of value to user communities. Other layers of development, for instance, describing resources in multiple ways for potential use in many contexts [2], are needed to create meaningful, functional aggregations that support user communities of interest.

Differences in metadata standards reflect the various aims and practices of resource developers and their constituent user communities. In the library profession where digital metadata has regularly been applied to both digital and nondigital works, MARC and Dublin Core have been widely adopted [4]. Those working in particular subject domains are also building

³ See <http://imlsdcc.granger.uiuc.edu/researchplan.asp> for an outline of the research questions guiding the project.

⁴ A complete list of the project’s publications and presentations can be found at <http://imlsdcc.granger.uiuc.edu/about.asp>.

⁵ See Brogan (2003) [6] for a review of digital library aggregation services and Hunter (2003) [12] for a review of metadata research, with a section devoted to interoperability and coverage of technologies for integration, sharing, and exchange.

sophisticated infrastructures, schemes, and guidelines to support their metadata requirements. For many standards, user communities have informed or been participants in the development of metadata specifications (e.g., GEM represents educators, and TEI was developed by humanities scholars).

A long-term goal of the DCC project has been to determine how to optimize metadata in federated systems to support users' practices and needs. The content and users of most digital collections developed by libraries and museums are not homogeneous. It has always been difficult to build heterogeneous collections that support the interests of diverse user communities, and this remains one of the greatest challenges in the digital environment, where collections are distributed and idiosyncratic. As [15] suggests, traditional library user-based collection criteria need to be extended to distributed digital collection services. Similarly, as we gain in interoperability, we do not want to lose advances that have been made in adaptation and access for communities of users at the local level. Thus, there is much to be learned by examining the practices, problems, and achievements of digital projects at the local level, to inform our ongoing work to develop best practices and principles for federation and building compatibility among local and global requirements.

3. METHODS

Development of the IMLS DCC has been informed by several stages of research applying multiple methods, including surveys, interviews, and case studies. Additional data were collected to investigate item and collection description and subject access issues through content analysis, focus groups, and usability studies. The multimethod approach allowed us to perform analysis across a large sample of projects to address general research questions while studying smaller, representative samples in more depth for fuller analysis. This report covers the survey component of the project, which is a subset of the larger body of data collected throughout the project. This is the first phase of analysis on the survey data, and where appropriate the self-reported survey responses are complemented with selected metadata repository measures and interview data to help bring further perspective to the survey results. A more comprehensive analysis integrating the findings presented here with the other research components is ongoing and will appear in later publications. Earlier reports to date, some of which were identified above, are listed on the IMLS DCC webpage (<http://imlsdcc.grainger.uiuc.edu/about.asp>).

The survey data were collected from NLG awardees in the first and fourth years of the DCC project to monitor progress and change in metadata practices and perceptions over time. In 2003 a two-part survey was administered to project managers of 122 NLG digital collection projects awarded between 1998 and 2003. The response rates were 76% for the first part and 72% for the second part. Part one consisted primarily of closed questions on 1) the type of material in the digital collection, 2) metadata schemes used, 3) the intended audience, and other specifics about the digital collection and its technical implementation. Part two consisted of open-ended questions soliciting information about 1) how the institution would use a central registry of IMLS-funded digital collections, 2) what elements should be included in a collection description scheme and why, 3) issues or problems

encountered applying metadata, and 4) issues or problems encountered in trying to share metadata.

Figure 1 presents a profile of survey respondents by institution type for the longitudinal results reported in section 4.2. In 2003 the breakdown was academic libraries (54.1%), museums and galleries (11.0%), library consortia (7.3%), public libraries (6.4%), historical societies (5.5%), botanical gardens (4.6%), archives (4.6%), state libraries (2.8%), non-profit organizations (2.8%), and zoological societies (0.9%).

The second survey was distributed to the respondents of the original 2003 survey in early 2006 to trace changes in metadata practices over time. This follow-up round covered generally the same questions as the 2003 survey to identify changes in types of materials, metadata applications, intended audience, and other aspects of technical implementation. The response rate was 72%, and the respondents on this second round were distributed as follows: academic libraries (45.9%), museums and galleries (14.9%), library consortia (9.5%), public libraries (5.4%), historical societies (5.4%), botanical gardens (5.4%), state libraries (4.1%), archives (4.1%), non-profit organizations (4.1%), and zoological societies (1.4%).

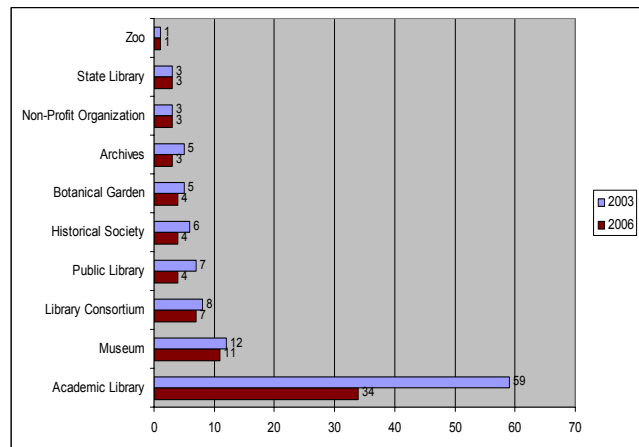


Figure 1. Number of Responding Institutions by Type, 2003 and 2006.

Like most longitudinal panel studies, this study suffered from panel mortality with 32% (n=109) of the projects that responded in 2003 not responding in the 2006 follow-up survey. The following analysis looks at the trends over time by using all projects that responded in the 2003 study and all projects that responded in the 2006 study, not just projects that responded for both years. While the results presented here are largely derived from the comparative analysis of the 2003 and 2006 survey, we also draw from other data, especially the interviews, to provide additional details and further explicate the quantitative findings. For quantitative measures, Ns representing the number of projects that responded to the question are consistently noted to assist in interpretation. Where results are noted as significant, this refers to the p-value of the chi-square test being at or below the .05 level. For odds ratios measures, a significant finding has a confidence interval that does not contain 1.

4. FINDINGS

The survey analysis indicated a number of interesting changes in IMLS DCC development over time, including an increase in intended application of OAI-PMH for NLG-funded digital projects from 15.8 % (n=94) to 19.7% (n=66) between 2003 and 2006. Academic libraries lead (35%, n=17) among the 2006 survey respondents planning to apply OAI-PMH either at the broader institution level or for the particular NLG projects, followed by state libraries, library consortia, academic museums and archives (12% each), botanical gardens, public libraries, and academic departments (6% each). Moreover, the project's harvesting measures show that from October 2003 through March 2007 the percentage of registry projects actually contributing to the repository has risen from 24% to 27%. Below we report results that put this trend in perspective of the metadata practices of IMLS DCC participants, more generally, but first we provide an overview of all the institutions and content within the IMLS DCC, not just those that responded to the survey.

4.1 IMLS DCC Profile

4.1.1 Institutions

While many different types of institutions are represented in the IMLS DCC, academic libraries are the most substantial proportion, followed by museums. However, it is important to note that there is considerable variety within these two categories, especially in terms of size and scope of museum operations. Overall, 343 institutions were listed in the 1998-2003 NLG proposals, either as the primary institution or a partner in the 122 projects. Figure 2, below, shows the breakdown of the number of the institutions by type.

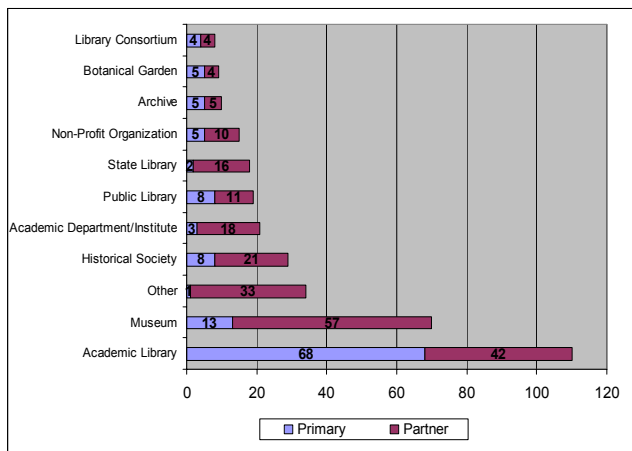


Figure 2. Number of Participating Institutions by Type, 2003.

One hundred and ten academic libraries participated, greatly outnumbering the other types of institutions. In fact, only 42 projects did not involve an academic library, academic department, or a museum based in a university. The next largest category was museums (70), followed by "Other" (34), and historical societies (29). The "Other" category includes nine government institutions, nine school districts, four special libraries, three Native American tribes, two herbariums, a law firm, a Web design company, a theater, a natural history site, and a library system. These frequencies do not take into account all contributing institutions. Some projects are statewide initiatives, like the Maine Memory project, which was awarded to Maine

Historical Society in 2002 and had over 100 contributors in 2003, and over 160 in 2006. Further, each library consortium represents many libraries, one of which is a group of 300 public libraries.

By 2006, 44 additional institutions were contributing to the IMLS DCC, for a total of 387, either as the primary institution or a partner in 136 NLG-funded projects. Figure 3 shows the breakdown of the number of institutions by type. The greatest increases were in "Museum" and the "Other" categories adding nine institutions each. The "Academic Library" category added six institutions. Only a few additions were made in historical societies, academic departments or institutes, and public libraries.

In our overall analyses, we were surprised to find that there were no important differences between university and non-university institutions based on metadata schemes used, mapping, or types of content. For example, there was no statistically significant difference observed between collection types, institution types, team expertise, or audience when it comes to metadata scheme usage in 2006 or between 2003 and 2006. The similarity among the various types of institutions was further reinforced in the case studies, in which we documented common experiences and challenges among resource developers at different types of institutions. Moreover, with the large number of multi-type collaborations among the projects, we have observed that project managers based in academic libraries are highly aware of and easily discuss the perspectives of their museum and archives partners.⁶

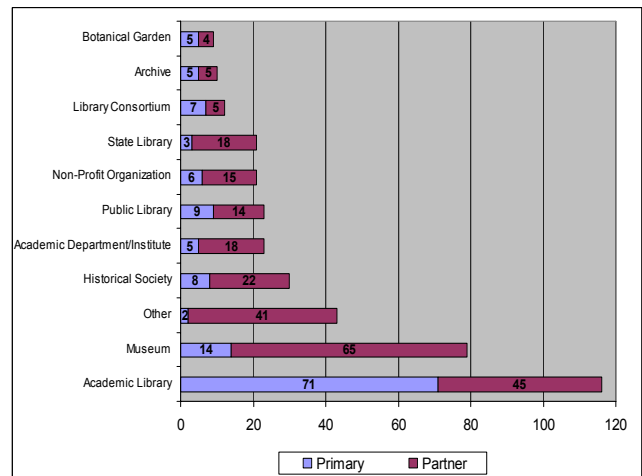


Figure 3. Number of Participating Institutions by Type, 2006.

4.1.2 Content Subjects and Formats

In early 2007, analysis of the 169 collection level records indicated two major subject strengths in the IMLS DCC: 1) Social Studies (80% of collections), including U.S. history, state history, world history, U.S. government, urban studies, anthropology, geography etc.; and 2) Arts (46% of collections), including visual arts, photography, popular culture, architecture, music, history of art, etc. At the item-level, the top five subjects are United States, people, songs with piano, trees, and archeology of the United

⁶ For an introduction to these perspectives as they relate to collection description see Dunn (2000) [8] on museums and Sweet and Thomas (2000) [26] on archives

States. Images are the most common format, identified in 80% (n=169) of the collection level metadata records, followed by text (68%), physical object (29%), sound (20%), interactive resource (10%), moving image (7%), and dataset (4%).

In the survey, projects were asked what types of content they had in 2003, and then in 2006 they were asked what type of content had been added. Of the 39% (n=72) of projects that had added new types of content, most (46%, n=28) added text. Thirty-nine percent (n=28) added images, and 39% (n=28) also added interactive resources. A comparative measure of respondents to both 2003 and 2006 shows change in types of content held by projects, not just that which was added. As shown in Figure 4, from 2003 to 2006 there was an increase in all types of content held by projects, however only interactive resources showed a significant increase from 12% to 25% (n=69) over the three year period [Figure 4].

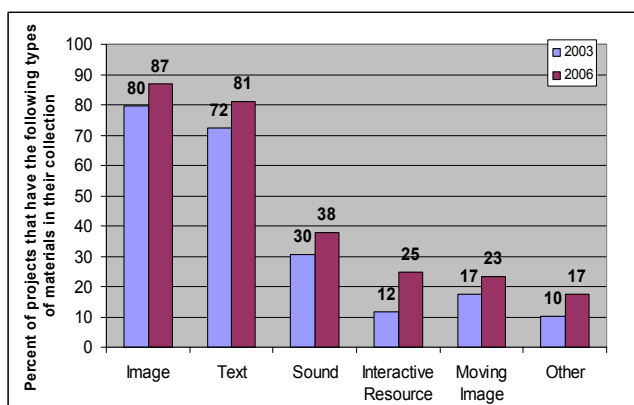


Figure 4. Material Types in Collections 2003-2006, (n=69).

4.2 Resource Developer Responses

4.2.1 Audiences

Overall, most projects described scholars (88%), the general public (83%), or undergrads (82%), as a primary audience (n=72). Slightly fewer, 79% (n=72), reported high school students as a primary audience for their collection. Projects designating high school students as a primary audience increased considerably. Fifty-nine percent (n=94) of projects had high school students as their primary audience in 2003, up to 80% (n=72) in 2006. Designation of the K-12 audience also increased from 65% (n=94) in 2003 to 75% (n=72) in 2006. There was a smaller increase in projects reporting scholars as their primary audience, from 84% (n=94) in 2003 to 88% (n=72) in 2006.

In-depth discussion on users in the recent set of interviews afforded insights into how user groups are understood by resource developers. Many developers only had anecdotal evidence of who used their collections. But, it was common for users to be discussed both as the audience(s) that developers hoped for and the audience(s) that actually seemed to be using the resource. Several respondents noted that they were currently focusing on studying who their audience is and how to better serve them.

4.2.2 Metadata Application

Metadata has been an important component of most NLG projects, however in the 2003 survey about 18.1% (n=94) of respondents indicated that they did not apply any metadata schemes. We expect this is because some resource developers did not realize that metadata was part of their descriptive practices or did not call it metadata. In addition, in the early stages of the projects metadata decisions may not have been made yet. The general breakdown of schemes applied is presented in Figures 5 and 6, which give an aggregate measure of multi-scheme use for 2003 and 2006, offering a percentage view of metadata application that adds up to 100%. Figure 7 provides an itemization of the individual schemes applied by multi-scheme projects.

4.2.2.1 Dublin Core and MARC

In 2003, more than three quarters (n=94) of the projects used MARC or Dublin Core to describe digital objects in their collections; this proportion was up to 87% (n=59) in 2006 [Figure 7]. In our interviews, participants expressed a preference for MARC's field richness, while Dublin Core was valued for its perceived ease of application. But, nonstandard use of fields seemed to be more prevalent with Dublin Core. For example, in one case the source field was appropriated to provide information about the original object that had been digitized, and in other projects the data placed in the description field had been extended to compensate for the lack of appropriate fields in Dublin Core. Dublin Core use for projects using only a single metadata scheme increased from 11% (n=94) to 30% (n=59) of all projects [Figures 5 and 6]. However, when those using multiple schemes is broken out, the increase overall is much less pronounced. In 2003, 50% (n=94) reported use of Dublin Core either alone or in combination with another scheme(s); in 2006, 58% (n=59) reported use of Dublin Core either alone or in combination with another scheme(s) [Figure 7].

Use of MARC as a single scheme also increased from 4% of all responding projects (n=94) in 2003 to 8% (n=59) in 2006 [Figures 5 and 6]. However, much like Dublin Core, use of MARC is notably different when those using multiple schemes are broken out. Using this scenario, MARC sees a slight decrease in use from 29% (n=94) to 27% (n=59) from 2003 to 2006 respectively [Figure 7].

In 2003 and 2006 there was a significant difference between the use of MARC and Dublin Core for multi-scheme projects as compared to single-scheme projects. In 2003, nine percent (n=44) of single-scheme projects used MARC, while 48.0% (n=50) of multiple-scheme projects used MARC. Similarly, only 22.7% (n=44) of single-scheme projects used Dublin Core, while 76.0% (n=50) of multiple scheme projects used Dublin Core [Figure 5*]. In 2006, thirteen percent (n=37) of projects using a single scheme used MARC, while 54.5% (n=22) of projects using multiple schemes used MARC in combination with one or more other schemes [Figure 6*]. Similarly for Dublin Core, less than one-half (47.3%; n=37) of projects using a single scheme used Dublin Core, while more than three in four (77.3%; n=22) of the projects using multiple schemes used it [Figure 6*].

Over the course of IMLS DCC development, only four projects that used multiple schemes did not incorporate MARC or Dublin Core as one of the schemes. In 2003, 4 out of 50 projects using

multiple schemes did not use MARC or DC. In 2006, that number fell to only one out of 22 projects.

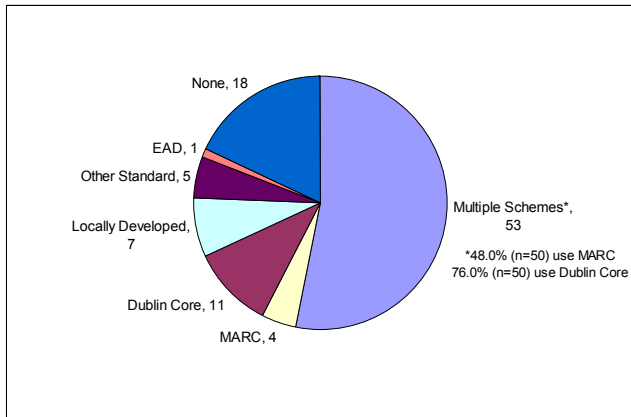


Figure 5. Percentage of Projects Using Single or Multiple Schemes for Item Level Description, 2003 (n=94).

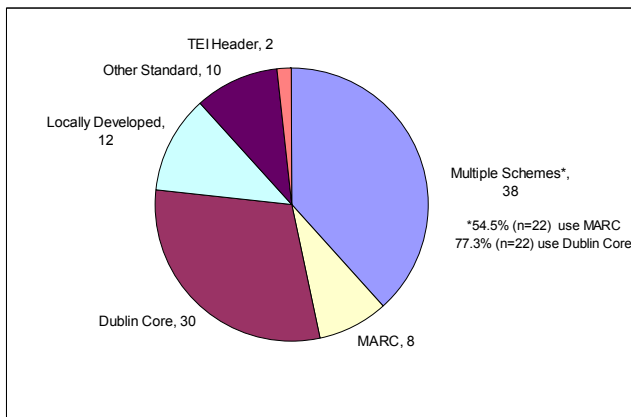


Figure 6. Percentage of Projects Using Single or Multiple Schemes for Item Level Description, 2006, (n=59).

4.2.2.2 Other Multi-Scheme Changes

Multi-scheme use is common, and, contrary to our early preliminary analysis in 2003, fairly equally divided among collaborative and non-collaborative projects. In 2003, 53% (n=94) of the projects proposed to use or were using multiple metadata schemes [Figure 5]; in 2006, projects were significantly less likely to do so with only 38% (n=59) using multiple schemes [Figure 6]. Using an odds ratio measure, projects were half as likely to use multiple metadata schemes three years later. Moving from multi-scheme use to a single-scheme use was not significantly different for various institution types, audiences, collection types, or type of initial and target scheme(s).

Fifteen, or 34%, of projects using multiple schemes chose to use three or more schemes in 2003; eight, or 39% of projects using multiple schemes, reported using three or more schemes in 2006. There was also variation of the schemes applied within multi-scheme projects. Although TEI header and EAD were rarely used alone, they were applied in combination with other schemes both in 2003 and in 2006. Slightly less than one third (32%; n=50) of the projects with multiple schemes used or proposed to use TEI in

some form in 2003, and the same percentage (32%; n=22) of respondents using multiple schemes used TEI in 2006. Just under a quarter (22%, n=50) used or proposed to use EAD in 2003 as one of the multiple schemes, and 23% (n=22) of respondents using multiple schemes used EAD in 2006. Of particular interest in multiple schemes use is of the projects using MARC, 17% used EAD in 2003, and this proportion significantly increased to 42% in 2006 [Figure 8].

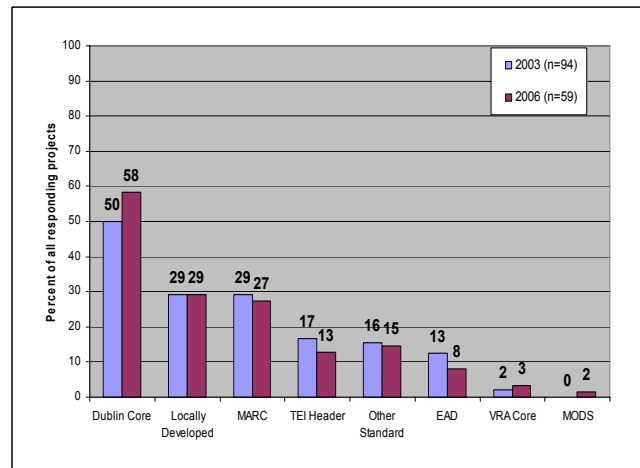


Figure 7. Percentage of Projects Using Schemes for Item Level Description (multiple schemes broken out), 2003-2006.

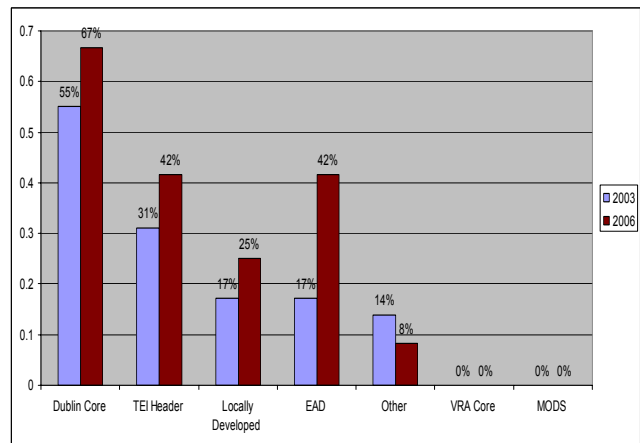


Figure 8. Scheme Use Percentage for Projects Using MARC and at Least One Other Scheme, 2003-2006.

4.2.2.3 Locally Developed Schemes

Whether used as single or with multiple schemes, 29% of projects applied locally developed schemes in 2003 (n=94) and 2006 (n=59) [Figure 7]. There is a significant difference in the distribution of locally developed schemes used as a single scheme as compared to use with other schemes. Forty-six percent (n=22) of multi-scheme users applied a locally developed scheme, while only 18.4% (n=37) of projects using only one scheme applied a local one in 2006. Seven percent (n=94) of the projects applied a locally developed scheme exclusively in 2003, several of which were derived from Dublin Core; similarly, 12% (n=59) of respondents used only a locally developed metadata scheme in 2006 [Figures 5 and 6]. Of all projects that used multiple schemes

(n=50), only one project used a local scheme in addition to Dublin Core or MARC in 2003; this increased to over 15% (n=22) in 2006.

Supplementary data indicated that projects chose to apply a local scheme for a number of reasons: customization was needed to capture information unique to the materials, information already recorded in a database or some other local information source was to be imported, or existing standards did not allow projects to adhere to their goals. All 100% (n=17) of the projects using locally developed schemes, indicated access as the primary purpose of their project in their grant proposals, while only 56.9% of other projects (n=51) listed access as a primary purpose of their grant.

4.2.2.4 Schemes for New Content

There are some significant differences with respect to scheme use between projects that have added new types of data and those that have not. Use of Dublin Core was more frequent than MARC for projects adding new content. Although the number of projects is small, EAD was also applied more by those that had added new types of content (11%, n=27) than projects that had not added new types of content (5%, n=42) [Figure 9]. A question on mapping of metadata was introduced in the 2006 survey. As would be expected, the 36% (n=42) of projects that had not added new types of content were significantly more likely not to have done any mapping, as compared to the 19% (n=27) of projects that had added new types of content [Figure 9].

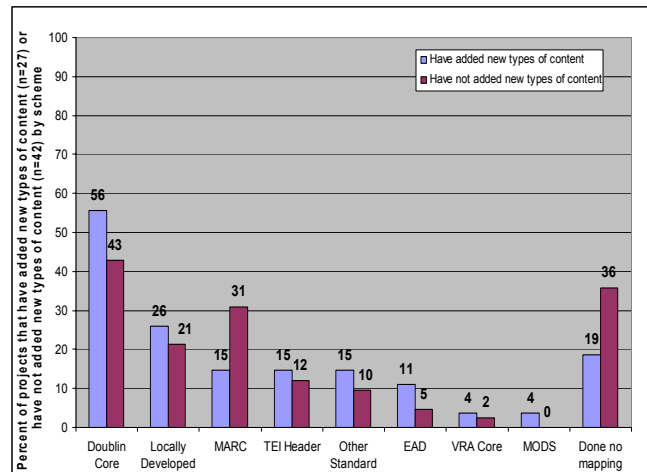


Figure 9. Scheme Use for Projects That Added (n=27) or Did Not Add (n=42) New Types of Content, 2006.

4.2.2.5 Mapping

The mapping question introduced in 2006 asked respondents to specify scheme(s) to which they mapped or planned to map their metadata. Overall, 63.4% (n=56) of projects have mapped their metadata. Dublin Core was mapped to most often with 63% (n=35) of projects mapping to Dublin Core. Twenty six percent (n=35) of projects have mapped to MARC [Figure 10]. “Other Standards” and MODS were the next highest at 14% and 12% (n=35), respectively. Overall, 41% (n=35) of projects that have done mapping have mapped to multiple schemes. Of the 9 projects that mapped to MARC, all 9 projects also mapped to at least one other scheme. Seven out of the nine projects that

mapped to MARC also mapped to Dublin Core. A significant number of projects that used TEI or EAD had already mapped or have plans to map to MARC. Specifically, of the seven projects using TEI, six have mapped or will map to MARC. Two out of the five projects using EAD will map or have mapped to MARC. It is also of note that all six projects adding sound planned to map to Dublin Core, and 2 of those that added sound plan to map to MARC in addition to Dublin Core. The most common purpose given for mapping to MARC was “automation;” other more specific responses included aggregation into other online collections and portals, and providing access to digital collections through individual library OPACs and WorldCat.

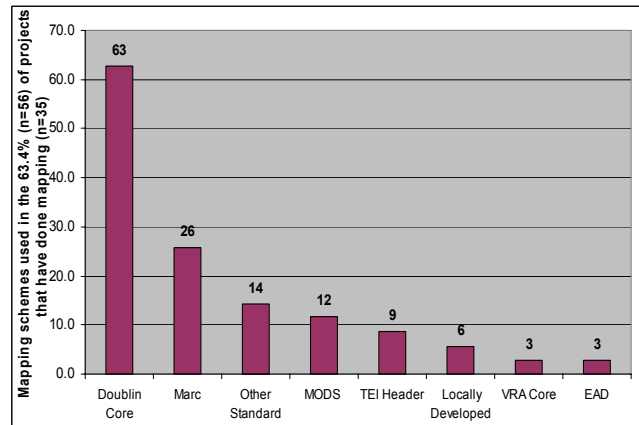


Figure 10. Percentage of Projects That Have Mapped or Will Map Their Metadata to Schemes, n=35, 2006.

4.3 Decision Factors and Problems

Choice of metadata scheme(s) was influenced by factors that might be expected. The overall degree to which a standard had been adopted by peer institutions was an important consideration, as was the compatibility with local systems. For example, one of our recent interview respondents had to drop plans to use MODS as the major scheme due to a new content management system.

In our grant we promised MODS; we started planning for MODS; I had a MODS file specific for our very small collection – it’s a two thousand item collection. And just a few months into the grant, right about when we solidified the profile, and we’d really rolled in on entering metadata, we discovered that the new content management system that we acquired was very difficult to customize for MODS. So we just did not have resources in place, and we made the move to qualified Dublin Core, and the mapping was a little painful. (MF060215, interview, February 15, 2006).

Content management systems also influenced text encoding decisions.

We also planned at the outset of the project, because that software could not handle TEI, to do very minimal TEI encoding of our manuscripts for archives, and then later as the system grew we’ve lost that. We’ve actually just in the last month or so decided to go with a different content management system. So we are upping our TEI on encoding level, from very minimal to a level three

encoding. So, there is another example of a give and take. (MF060215, interview, February 15, 2006).

In addition, several librarians reported that their choice of MARC was due to their OPAC's inability to handle Dublin Core records. Many library-based digital collection developers chose MARC because it allowed for more granularity in description than Dublin Core while also being the easiest to implement since their staff were already proficient using MARC.

The three most commonly reported problems with description were: consistent application of the chosen metadata scheme within a project, identification and application of controlled vocabularies, and integration of sets of data, schemes, and vocabularies either within an institution or among collaborators. In addition, there were clear tensions between local practices and what was perceived as the best for interoperability. One project that began with Dublin Core decided against using it part way into the grant, favoring MARC and TEI for representing the texts in their collection. Later they ended up mapping their metadata back to Dublin Core for OAI interoperability.

Some of the unique content in digital collections in the IMLS DCC cannot be adequately described by existing metadata schemes. For example, interactive resource content has increased significantly in the IMLS DCC since 2003, but resource developers face challenges in description of such content:

I don't feel like we really know how to issue a good metadata standard for a draft of gaming. That's really out there. I've read a few people who've talked about it, and they are just saying 'Do you want to catalog a big game, how people play the game, do you want to actually have a camera and videotape, so playing the game is ... part of the game experience?' (MR060216, interview, February 16, 2006).

In cases such as this, resource developers often look for examples of how similar content has been described by other projects:

There seems to be a lot available for the actual objects. There is not a lot available whether a standard or subject headings for documentation of [interactive] objects. ... It's exciting and novel and nerve-breaking because we don't really have a lot of models to follow, ... we are kind of winging it a lot of the time, but in consultation with ... some others in the field we are hoping that in the end we'll be able to come up with something. (JH060216, interview, February 16, 2006).

Local, home-grown metadata schemes often are developed when no suitable standard can be identified. Folksonomies and social tagging collected from the end-user community were named as one of the term sources for such home-grown schemes.

Smaller digital collections that do not have resources to develop local schemes sometimes end up compromising the richness of description to implement Dublin Core:

A lot of things that we recorded ... don't have a good place [in Dublin Core]. A lot of this is the contextual data that we are getting from the institutional memory of our directors. We are going to things like annotated notes fields that don't really give it any direction or lead to actually making them most accessible. (AC070301, interview, March 1, 2007).

The unstable standards environment has made it difficult to advance without shifts, reconsiderations, and adaptations in original metadata plans to support interoperability and shareability of metadata. One resource developer provided an overview of the evolution of the state of practice at their institution which had been creating digital collections since 1995:

First of all, every single text ... anything that can be considered text on the web is encoded in TEI. And of course, every TEI text has TEI header and that's an excellent metadata. So we use TEI X Lite, or we are starting with that and of course looking forward to diverting to XML and we use XML. Then, every electronic text is cataloged by catalogers: we have two people in the cataloging department dedicated to us and working with us for ten years now. So everything is MARC-cataloged. Then, of course, in a due course of our development when Dublin Core was introduced we've created Dublin Core for our collections. And it's all OAI-compliant. That was how we've developed and evolved. We do not jump on every single metadata standard or what would be or could be or should be standard. We do not jump right away but if it becomes standard we definitely want it to be compliant. (NS060216, Interview, February 16, 2006).

In comparison to the DCC projects overall, this was a well-established and experienced group with a high level of staff support and expertise.

4.4 Units of Description

Over the course of this project we have become increasingly aware of the value of collection description in federated collections. And, previously we have discussed the growing number of collection-like aggregations that need to be represented in a federated environment, such as exhibits, tours, and lessons [23]. Resource developers are also beginning to articulate new distinctions in what constitutes an object or item for the purposes of collections with these alternative identities. The developers of an art gallery digital collection discussed their efforts to figure out how to represent "events" as items.

It's more like an institutional history—showing what the work looked like in the exhibition, because not everything that was in the exhibition ... not every piece of artwork or every performance is documented in and of itself, it's documented as a whole (JH060216, interview, February 16, 2006)

One of the main description difficulties was that the item content could change on a daily or even an hourly basis during the event. In such a case, an item record would need to describe a "show," which might represent a period where an artist "lived in this gallery ... for four weeks, and produced their art in the gallery space." Not surprisingly, the developers have not been able to identify peers projects or existing relevant models for description.

5. CONCLUSIONS

One of the IMLS DCC project's primary aims was to encourage and support content providers in the development of sharable metadata. Thus, the increase in the application of OAI-PMH was an important, positive trend among the contributing institutions. In tracking other metadata patterns, there were no pronounced changes in the use of any given metadata scheme between 2003

and 2006, when multiple scheme use is itemized. Multi-scheme use in general, however, has become less common. Use of Dublin Core remains high, even with wide recognition of its limitations. One of the possible reasons for the stable position of Dublin Core is the desire to make digital content shareable combined with a wide-spread misapprehension that compliance with OAI-PMH requires Dublin Core use. Another obvious reason is the perceived ease of simple Dublin Core application. At the same time, the application of MARC and locally developed schemes has essentially remained the same. MODS application has remained minimal, apparently at least in part due to constraints of content management systems. However a fair amount of mapping to MODS was documented. Locally developed schemes are used as much as MARC, and may be on the increase as new collections are incorporating less traditional library and museum materials, and more interactive and multimedia content, which is not easily described with available standards.

Metadata records harvested through the course of the IMLS DCC project are currently being analyzed to determine actual application trends in Dublin Core usage. Results will determine if metadata has become more homogeneous and if use of Dublin Core fields and mappings have changed as projects mature. In addition, we need to investigate further if turnkey content management systems are influencing metadata application in ways that are not optimal for resource developers. The increase in dynamic content and emerging conceptions of the unit of description raise interesting and important questions about the granularity of representation and standards for describing complex objects. As seen in the case of a gallery event, it is not evident what information needs to be captured in an item record. For example, what is required to adequately describe, not just the art object as it is being created, but also the people, materials, and processes involved in the act of creation?

There is still much to be learned about user communities for federated resources. While more resource developers recognized high school students as one of their user groups, scholarly audiences remained the most widely designated. Based on this trend and the obvious academic emphasis in the institutional profile and the research value of the content, there is tremendous potential to further develop the IMLS DCC as a resource for scholarly research.

Federation of digital collections is a viable strategy for increasing the value and use of the growing body of digital content. And, while federated collections can conceivably be built to offer more than the sum of their parts, these aggregations may also lose important context and meaning inherent in individual collections. Based on our grounded understanding of metadata use in practice, our recent work has turned to identifying metadata relationships between items and collections to preserve context, and to enhance functionality and usefulness for scholarly user communities. Users of digital research library collections are interacting with a context that includes physical, institutional, and intellectual features [19]. At present, this context tends to be scattered and poorly aligned with scholarly practices. Thus, we have adopted a “contextual mass” [21] model of federated collecting. This approach aims to systematically aggregate sources and build functionality that is aligned with the work that scholars do, exploiting meaningful interrelationships among items and collections.

6. ACKNOWLEDGMENTS

This research was supported by a 2002 IMLS NLG Research & Demonstration grant. Project documentation is available at <http://imlsdcc.grainger.uiuc.edu/about.asp#documentation>

7. REFERENCES

- [1] Agogino, A.M. *Enhancing Interoperability of Collections and Services*. Final Report, Dec. 2004. http://best.me.berkeley.edu/%7Eaagogino/papers/Final_Report_SMETE.pdf
- [2] Arms, C. R. Available and useful: OAI at the Library of Congress. *Library Hi Tech*, 21, 2 (2003), 129-139. <http://memory.loc.gov/ammem/techdocs/libht2003.html#exploit>
- [3] Bailey-Hainer, B., and Urban, R. The Colorado Digitization Program: A Collaboration Success Story. *Library Hi Tech* 22, 3 (2004), 254-262.
- [4] Besser, H. The next stage: Moving from isolated digital collections to interoperable digital libraries. *First Monday*, 7, 6 (2002). http://firstmonday.org/issues/issue7_6/besser/index.html.
- [5] Bishoff, L., and Allen, N. *Business Planning for Cultural Heritage Institutions*. Council on Library and Information Resources, Washington, DC, 2004. <http://www.clir.org/pubs/reports/pub124/contents.html>
- [6] Brogan, M. L. *A Survey of Digital Library Aggregation Services*. Digital Library Federation, Washington, DC, 2003. <http://www.diglib.org/pubs/brogan/>
- [7] Brogan, M. *Contexts and Contributions: Building the Distributed Library*. Digital Library Federation/Council on Library and Information Resources, Washington, DC, 2006. <http://www.diglib.org/pubs/dlfl106/>
- [8] Dunn, H. Collection level description—the museum perspective. *D-Lib Magazine*, 6, 9 (Sept. 2000). <http://www.dlib.org/dlib/september00/dunn/09dunn.html>.
- [9] Foulonneau, M. et al. Using collection descriptions to enhance an aggregation of harvested item-level metadata. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries* (Denver, CO, June 7-11, 2005), 32-41.
- [10] Heery, R., and Anderson, S. *Digital Repositories Review*. UKOLN, AHDS. (Feb. 19, 2005). http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf.
- [11] Heery, R., and Powell, A. *Digital Repositories Roadmap: Looking Forward*. UKOLN, Eduserv Foundation. (April 2006). http://www.jisc.ac.uk/uploaded_documents/rep-roadmap-v15.doc
- [12] Hunter, J.L. A survey of metadata research for organizing the web. *Library Trends*, 52, 2 (2003), 318-344.
- [13] Kastens, K. et al. Questions & challenges arising in building the collection of a Digital Library for Education: lessons from five years of DLESE. *D-Lib Magazine*, 11, 11 (Nov. 2005) <http://www.dlib.org/dlib/november05/kastens/11kastens.html>
- [14] Knutson, E., Palmer, C.L., and Twidale, M. Tracking metadata use for digital collections. In *Proceedings of the*

International DCMI Metadata Conference and Workshop (DC'03) (Seattle, WA, Sept. 28–Oct. 2, 2003), 243-244.

- [15] Lagoze, C., and Fielding, D. Defining collections in distributed digital libraries. *D-Lib Magazine* (Nov. 1998). <http://www.dlib.org/dlib/november98/lagoze/11lagoze.html>
- [16] Lagoze, C. et al. What is a digital library anymore, anyway? Beyond search and access in the NSDL. *D-Lib Magazine* 11, 11 (Nov. 2005). <http://www.dlib.org/dlib/november05/lagoze/11lagoze.html#n3>
- [17] Lagoze, C. et al. Metadata aggregation and “automated digital libraries”: A retrospective on the NSDL experience. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (Chapel Hill, NC, June 11-15, 2006), 230 – 239.
- [18] Lagoze, C. et al. *Representing Contextualized Information in the NSDL*. 2006. Preprint available from <http://www.arxiv.org/ftp/cs/papers/0603/0603024.pdf> .
- [19] Lee, H. What is a collection? *Journal of the American Society for Information Science*, 51, 12 (2000), 1106-1113.
- [20] Marmor, M. Six lessons learned: An (early) ARTstor retrospective. *RLG DigiNews*, 10, 2 (April 2006). http://www.rlg.org/en/page.php?Page_ID=20916#article0.
- [21] Palmer, C.L. Thematic research collections. In *A Companion to Digital Humanities*. Blackwell, Malden, MA, 2004, 348-365.
- [22] Palmer, C.L., and Knutson, E. Metadata practices and implications for federated collections. In *Proceedings of the 67th ASIS&T Annual Meeting* (Providence, RI, Nov. 12-17, 2004).
- [23] Palmer, C.L., Knutson, E., Twidale, M., and Zavalina, O. Collection Definition in Federated Digital Resource Development. In *Proceedings of the 69th ASIS&T Annual Meeting* (Austin, TX, Nov. 3-8, 2006).
- [24] Shreeves, S.L., Kaczmarek, J.S., and Cole, T.W. Harvesting cultural heritage metadata using the OAI protocol. *Library Hi Tech*, 21, 2 (2003), 159-169.
- [25] Stvilia, B., Gasser, L., Twidale, M., Shreeves, S.L., and Cole, T.W. Metadata quality for federated collections. In *Proceedings of ICIQ04 - 9th International Conference on Information Quality*. (Cambridge, MA, Nov. 5-7, 2004), 111-125.
- [26] Sweet, M., and Thomas, D. Archives described at collection level. *D-Lib Magazine*, 6, 9 (Sept. 2000). <http://www.dlib.org/dlib/september00/sweet/09sweet.html>.
- [27] Zavalina, O. Collection-level user searches in federated digital resource environment. Forthcoming. In *Proceedings of the 70th ASIS&T Annual Meeting* (Milwaukee WI, Oct. 18-25, 2007).