IMLS National Leadership Grant LG-02-02-0281
"Proposal for IMLS Collection Registry and Metadata Repository"

This white paper and attached pre-prints are part of the three-year interim project report for the IMLS Digital Collections & Content Project, summarizing major findings October 2002 through September 2005. Project is hosted at the University of Illinois at Urbana-Champaign. Project Director is Timothy W. Cole (t-cole@uiuc.edu). Full report is available at http://imlsdcc.grainger.uiuc.edu. The material in this report is based upon work supported by the Institute of Museum and Library Services under IMLS National Leadership Grant Award No. LG-02-02-0281. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Institute of Museum and Library Services.

Collection Definition in Digital Resource Development

White Paper

Carole L. Palmer, Ellen M. Knutson, and Michael Twidale

Digital Collections and Content Project
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

21 September 2005

Introduction

In the IMLS Digital Collections and Content (DCC) project we have created a registry of digital collections produced by IMLS NLG initiative grantees, while development of an item level repository is ongoing. That work included development of a collection description schema.[1] Concurrently, the DCC research team has been examining metadata practices of the NLG projects and their implications for collection federation. The primary aim of this research has been to investigate how materials can best be represented in the DCC collection level registry and the item level repository to meet the needs of both service providers and divergent user communities. This line of inquiry naturally raises fundamental questions about the role of "the collection" as a defining or organizing unit in the digital environment. This white paper provides preliminary analysis of the data we have collected related to collection definition and provides a sketch of our next steps as we continue research in this area.

[1] For an overview of the collection schema see http://imlsdcc.grainger.uiuc.edu/CDschema_overview.htm for an overview of the collection. For the collection registry, go to http://imlsdcc.grainger.uiuc.edu/collections/.

Collection definition has been only one of several areas of emphasis in our study, and therefore the scope of our inquiry has been necessarily limited. During the course of the project we have assembled a rich body of data using several different techniques, including interviews, surveys, and focus groups with those serving as project managers and metadata librarians on the projects, and content analysis of the collection records created by NLG projects. (See Palmer and Knutson, 2004 [pre-print attached] for further details on our sample and methods.) However, only a subset of these data relate to directly to the problem of collection definition. Moreover, because these data were collected during the development stages of the collection registry, most of it represents the participants' ideas and impressions before they were exposed to the schema or the prototype repository. This report does include some early analysis of initial revisions projects chose to make in the collection records provided by the registry, but all other data were collected before implementation of the registry or review of collection records. As such, this phase of research provides us with an a priori view of how participants see the role of the collection in the development of their own, individual digital resources.

While our respondents did not yet have a working knowledge of the registry, their views and actions prior to implementation will serve as an important baseline as we move into the second phase of our research. In this next phase, collection definition in the federated environment will be a major focus, and we will have a functioning prototype of the registry and metadata repository for participants to test, use, and make available to their user communities. Our participants will also have further experience reviewing and revising collection descriptions and be able to assess how their collections and items "behave" after they are federated with over 100 other collections.

Conceptions of collections

There is no doubt that the notion of what constitutes a collection is evolving in the digital era, and that our collection development practices and principles may be changing as well. Characterizations of digital collections vary a great deal in the literature. Permanence and the actors in the collection process have been emphasized by some (International Council of Museums/CIDOC, 2002); others have taken a neutral stance toward transience and even size (Johnston & Robinson, 2002). Lee (2000) proposed that collections are best understood as information seeking contexts (Lee, 2000), and Lynch (2002) suggested that the aim of digital collection building should be to develop bodies of raw materials available for interpretation and presentation (Lynch, 2002). More traditional user-based collection criteria have also been considered the key to successful digital collection services (Lagoze and Fielding, 1998).

In the DCC project, our interviews with those producing digital collections in libraries, museums, and archives also show little consensus in how collections are being conceptualized. Practitioners working in the day-to-day world of digital collections are grappling with the "collection" construct, its role and its relevance in the digital environment. For example, while some participants questioned the need for such structures or differentiation, many continue to place a high degree of importance on the uniqueness, history, and name of their collection. In discussing and describing their collection, project developers stress the physical collections from which their digital resources are derived and the need to recognize the individual institutions that own the materials. At the same time, while many of the NLG projects were completed some time ago, few project developers had given consideration to collection description issues, beyond perhaps a response in their grant proposal to a standard request from IMLS for an indication of their plans for sharing collection-level descriptive records.

Some collection ambiguity seems to have been spawned by the processes that are essential for getting external support for digital collection building. Crafting interesting new projects in the form of a competitive grant proposal, and the actual work of managing those projects once they are funded, blur the notions of project and collection. For instance, often there is not a one-to-one relationship between a project and a collection. A project may produce or coordinate production of multiple collections or a collection may consist of distinct subcollections. And, of course, in the digital environment, materials digitized for a particular project can be added to more than one collection. Moreover, the content of some projects may not be considered collections, per se, by their developers, but rather exhibits, learning modules, or multimedia compilations.

We have also documented a number of cases of uncontrolled or developmental changes in how collections are developed and represented. Plans for organizing materials for browsing may change due to technical problems or changes in available technology, and inclusion of items can also be influenced by technical constraints or opportunities for expansion that arise as projects progress. Copyright constraints were, of course, a common factor in what materials were included in many digital collections.

Collection differentiation

Interview respondents frequently did not have a firm idea of how many collections they were creating. The form and parameters of their collections seemed to be undefined, suggesting that they may not have formal or firm collection policies or guidelines beyond what was outlined in their grant proposals. Some found it difficult to answer questions about how they define their collection, as represented in the following statement excerpted from an interview: “We have a problem with that word collection. We fought about that word, so when you use it what do you mean?” (Interview RD031217).

A slight majority of participants considered their digital resource to be multiple collections. For collaborative projects, the delineation of collections was generally made by institution, while for non-collaborative projects format was the most common distinction. Others described their resources as a new integrated whole that should not be parsed into smaller structural units. In the course of discussing the matter some tended to fall back into thinking in terms of identifiable collections or just present both sides, as seen in the excerpt below.

> I guess coming from an archivist point of view I would consider those multiple collections, but … I’m familiar with OAI and those types of activities. I guess that it could be considered a single collection as well in the larger scheme of what type of materials are being collected and how these digital collections are being created, because what we’re creating is actually something very new and I guess I would take a wider viewpoint than strict archival provenance. (Interview MS030905)

However, when surmising the user’s perspective, respondents tended to think that the collection construct was not useful for describing or organizing digital materials. In one focus group there was agreement that grouping material into subcollections limits the user by presenting “too chosen” a view of materials. The discussion among the focus group attendees emphasized education themes, exhibits, institutions, material type, and topics of scholarly or historical interest rather than collections per se.

The survey data also revealed some uncertainty about subdivisions within collections. More than 75% of respondents reported that their collections are divided into subcollections, with most listing between 3-10 units. However, several specified much larger numbers between 20 and 50, and above. We suspect that some of these responses are actually reporting the number subject areas or contributing institutions represented in their collection, which may or may not be intended as differentiated subcollections in the completed resource. The most common divisions were by topic and type of material, with geographic categorization emphasized by many, as well as time period, and to a lesser degree, audience. Many also differentiated administrative units, with categorizations for both owning institution and subunits within institutions.

To get another data point on collection differentiation, we compared the survey responses against the registry collection description records, which were accepted and often expanded or refined by the projects. The IMLS DCC collection registry was initially populated with data gathered from the surveys and collection websites. Once the registry was made live, collection administrators were invited to edit the entry for their collection and to add child or subcollection records for the main collection record. Two collection administrators were asked to test the interface of the registry before it went live; therefore the changes they made were not tracked. One of these administrators had a parent collection and two subcollection records.

In comparing the survey and registry records we found a number of discrepancies, especially in how subcollections are distinguished. Of the collection records that had been reviewed by collection administrators, only 4 collections out of 24 that said they had subcollections in the survey listed subcollections in the registry. One of these had a different number of subcollections in the two places--40 on the survey and 26 in the registry. None of the projects that indicated that they did not have subcollections on the survey created subcollections in the registry, but one collection that had left the subcollections section blank in the survey had one subcollection in the registry. That said, we speculate that two description fields provided in the schema may have been used as alternative as alternative ways to subcollection differentiation. The first is the contributing institutions field. In the survey many projects indicated that their subcollections were divided by administrating institution, and the institution field my have been considered sufficient for capturing this information. The second is the associated physical collections field. This may have been considered adequate for mapping digital materials back to the physical collections that were considered viable subcollection differentiations at the time of the survey. In the next phase of our research, we will be closely tracking changes made to the collection registry records and conducting interviews with participants to further explore how and why specific description decisions were made.

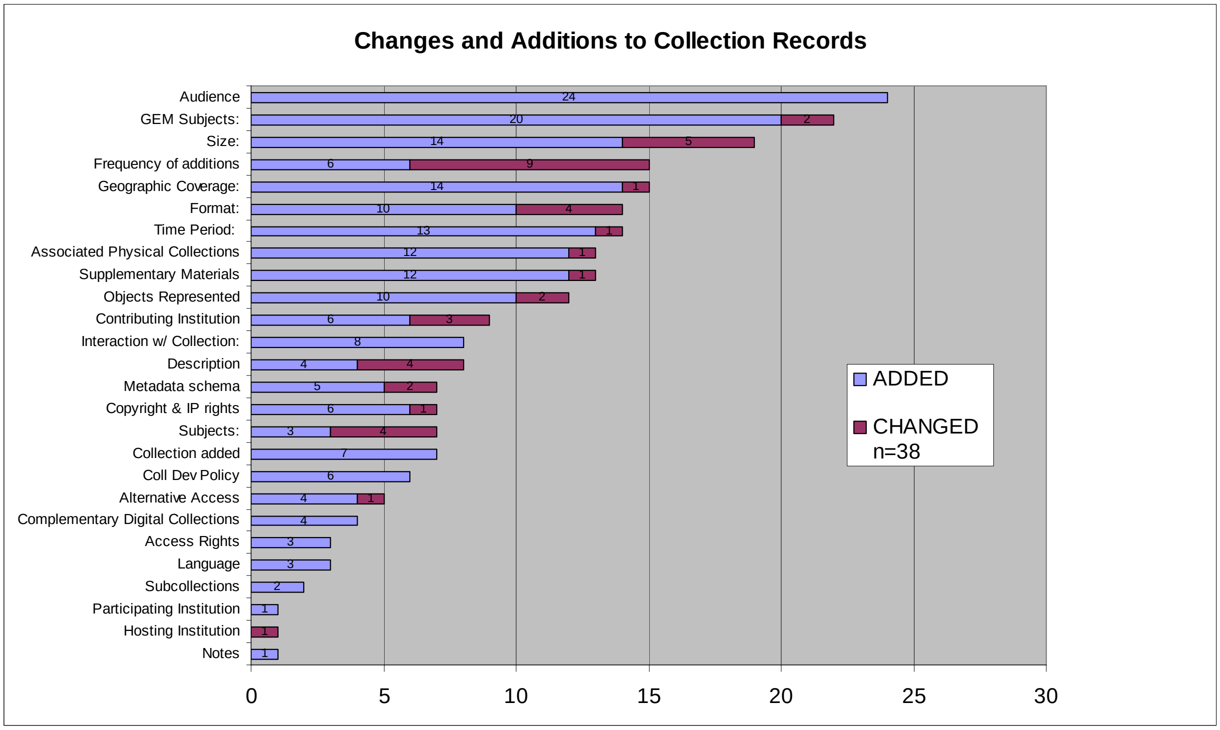
Definition through description

Conceptualization, definition, and internal differentiation of collections are being made more explicit in the individual collection records as they are reviewed and revised by the projects.[2]

[2] It is also important to note that in our analysis we do not differentiate between parent and subcollection records because they are equally detailed. In other words, when we indicate that a collection record was changed it could be a record for a parent or a child collection.

Between January and May of 2005, 38 collection administrators made a change or addition to their collection record—two came back a second time to make additional modifications. Five people reviewed their entry but made no additions or changes. Seven complete collections were added. Most of these modifications were made within one month of the initial call for editing. The vast majority of the changes were actually additions to the collection descriptions not a change of what was filled in based on survey and website information. If changes or additions were made, between 2 and 14 fields were modified. More than half modified more than 5 fields.

Figure 1 shows the number of records in which a field was modified. In general, those who reviewed their records enriched rather than refined the information provided. The field that participants modified most frequently was audience with 24 administrators adding anywhere from 1 to 12 new audiences. Most added 1 or 2 new audiences; it was only when there were no audiences pre-filled in that larger numbers were added. The next field most frequently modified was GEM subject heading. Twenty-two collection administrators made modifications to this field. Again the vast majority of the modifications were adding additional headings both at the top level (1-3 headings added) and the 2nd level (1-54 headings added) headings. In the up coming months we will be doing further research on how GEM subject headings are being applied in the collection descriptions, the adequacy of the GEM vocabularies for covering the materials in NLG projects, and the actual use of subject terms for searching based on user log data.



In the revisions to date, more than 10 collection administrators modified the following fields: size, frequency of additions, geographic coverage, time period, format, associated physical collections, supplemental materials, and objects represented. With size and frequency of additions we also saw people making changes to what was already in the registry. Not surprisingly the trend in size was to increase the number of items. With frequency of additions the trend was to change to a more frequent schedule; weekly to daily, for example. Three collection records were changed to remove "irregular" from the frequency of additions field, without replacing it with anything. Perhaps these collection administrators felt that it looked better to have this field blank than to show uneven growth or suggest that the collection was perhaps not growing at all. Another possible reason might be time sensitivity. Collection administrators may think that this information could go out of date and that they would be unlikely to update their record.

The changes in time period, geographic coverage, format, and objects represented were all additions in the following ranges: time periods (1-10), geographic areas (1-24), formats (1-2, though one collection added 4) and objects (1-4). The addition of associated physical collections may indicate the evolving nature of the relationship between physical and digital collections. As mentioned above we suspect that this is one way that collection administrators indicated what they considered subcollections in their survey responses. One collection record had 24 associated physical collections added to its record. For this particular collection the survey noted 16-18 subcollections with more to come. In the supplementary material field collection administrators added from 1 to 4 types of supplementary materials. Related websites, bibliographies, biographical information, essays, teacher and student resources, and contextual information were the types of material most frequently included.

Given the number of collections that listed subcollections in the survey, it was surprising that only four collections included subcollections in their registry record. We also expected to see greater use of the notes field. It is possible that collection administrators felt the schema was sufficient in describing their collection with out using an open-ended field such as notes, but this and our other speculations about description decisions will need to be explored in follow-up work.

Collection federation

Clearly, resource developers' ideas about collections are still being formulated as projects progress. Therefore, it is not surprising that we found their views on the IMLS federated collection to be amorphous. In both the surveys and the interviews, many respondents were unsure of the role of the DCC repository. As might be expected from this group of respondents, there was considerable interest in the repository as a source for information on up-to-date practices for digital projects and IMLS funding trends. But, clearly this type of current awareness could be achieved with a simple project directory and would not require building a formal repository. To some extent, this narrow view may be an artifact of the strong association of the DCC project with the IMLS NLG grant program, and may dissipate over time. Different approaches to branding the resource and new marketing approaches may also be needed. Our current results show, however, that only 40% of respondents recognized how a federated resource could benefit reference and research services at their institutions, and few perceived it as a helpful tool for end users. Moreover, there were scarcely any comments about the repository's potential for supporting programmatic resource sharing or the creation of new configurations of collections.

We expect that once respondents have a chance to use the DCC repository, they will be more likely to identify how it can be used and recognize its contribution to digital library development. Nonetheless, it is important to note that this population of digital resource developers represents professionals that are very likely to have a higher degree of awareness of trends in digital resources than the typical library, museum, or archive professional. Their inability to imagine applications for their institutions or their users suggests a need for wider professional development on the potential and future role of such resources in digital library development. Our project is increasing awareness of trends and technical issues associated with federation, at least among NLG grantees, and more widely as we disseminate our results in various professional venues. But, our perspective is that the implications of federation for providing new information resources and services should be part of the current professional discourse and part of the knowledge new LIS graduates are bringing to the field.

The importance of the collection concept in digital resource development

As this white paper has illustrated, the concept of collection is complex, diverse, and for many ambiguous or even problematic. The complexities and hesitancies outlined above are particularly noticeable given the expertise of our respondents and their very clear-cut opinions about other issues that we raised with them. This emphasizes that there are far greater ambiguities about people's perceptions about the concept of collection in the digital age than might at first be imagined, and that this applies to professionals directly involved in the process of digitizing collections, let alone to intended end users.

We believe it is important to investigate these concerns, ambiguities, and rapidly changing conceptualizations, because they are likely to have a direct impact on the adoption, tailoring, and development of systems in complex and perhaps undesirable ways. It is a truism of software engineering that requirements capture is difficult even when all stakeholders are involved and have clear ideas of what the intended system is for. Where there is ambiguity, and rapidly changing perceptions, it is all too easy to build structures into the underlying system architecture that militate against graceful adaptation of use and meaning. Similarly, where assumptions about use and purpose remain vague and implicit, they can get built into a design in ways that make it very expensive to recover from when they are finally identified.

As we move forward, we will work toward a better understanding of what "the collection" can contribute to functionality, interpretation, and use of digital resources. We will take a number of steps, including conducting interviews to document why particular changes have been made in metadata records and why certain fields have not been widely used or used in alternative ways. It is clear from the initial modifications to the collection records that many projects now consider their intended audience to be more inclusive than originally indicated. We will be examining what is driving this trend, as well others patterns in the data related directly to how collections are being conceptualized for presentation and use. For example, quite a few projects added information about associated physical collections, which suggests that traditional notions of collection are still important in articulating the context for items in a collection. But there is still much to learn about how collections function as “contexts” (Lee 2000) or “raw materials” (Lynch, 2002) from the user perspective. In previous research (Brockman, Neumann, Palmer, and Tidline, 2001), we found that humanities scholars identify strongly with library collections at their institutions. They perceived them as critical capital that makes it possible to do quality research and adds value to being a member of an institution. The second phase of our research will allow us to look at these and other aspects of the value of collections in the digital realm.

The work reported here is inevitably preliminary, but it enables us to establish a baseline for our future more detailed analysis which will lead to recommendations for systems design, education, and outreach. Only by facilitating a richer understanding of how the meaning of collections can evolve under processes of digitization and federation can the full value of these innovative NLG project pioneers be realized.

References

Brockman, W.S., Neumann, L., Palmer, C.L., & Tidline, T.J. (2001). Scholarly work in the humanities and the evolving information environment. Washington, DC: Digital Library Federation/Council on Library and Information Resources.

International Council of Museums/CIDOC. (2002). Definition of the CIDOC object-oriented Conceptual Reference Model version 3.4. Retrieved May 12, 2003 from http://cidoc.ics.forth.gr/docs/cidoc_crm_version_3.4.rtf.

Johnston, P., & Robinson, B. (2002). Collections and collection description. Collection description focus briefing paper. No. 1. Retrieved May 12, 2003 from http://www.ukoln.ac.uk/ cd-focus/briefings/bp1/bp1.pdf.

Lagoze, C., & Fielding, D. (1998). Defining collections in distributed digital libraries. D-Lib Magazine, (November). Retrieved January 23, 2003 from http://www.dlib.org/dlib/ november98/lagoze/11lagoze.html.

Lee, H. (2000). What is a collection? Journal of the American Society for Information Science, 51(12), 1106-1113.

Lynch, C. (2002). Digital collections, digital libraries, and the digitization of cultural heritage information. First Monday 7(5).

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