12-1-2008

Evaluating Descriptive Richness in Collection-Level Metadata

Oksana L. Zavalina

Carole L. Palmer

Amy S. Jackson

Myung-Ja Han

Evaluating Descriptive Richness in Collection-Level Metadata

Oksana L. Zavalina, Carole L. Palmer, Amy S. Jackson, Myung-Ja Han

ABSTRACT

When many collections are brought together in a federation or aggregation, the attributes of the original collections can become difficult to discern. Collection-level metadata has the potential to provide important context about the purpose and features of individual collections, but the qualitative aspects of collections are difficult to describe in a systematic way. This paper reports on a content analysis of collection records in the Digital Collections and Content (DCC) aggregation, conducted to analyze the kinds of substantive and purposeful information represented across 202 cultural heritage collections. We found that the free-text *Description* field often provides more accurate and complete representation of subjects and object types than the specified fields; it consistently represents properties such as uniqueness, importance, comprehensiveness, provenance, and creator of items in digital collection, and other vital contextual information about the intentions of collectors and the value of collections for scholarly users. The results show that free-text collection metadata can be both concise and semantically rich, and can provide a valuable source of data for enhancing and customizing controlled vocabularies.

Keywords: descriptive metadata; metadata aggregation; federated digital collections.

Oksana L. Zavalina (zavalina@illinois.edu), MLS, is a Doctoral Student and a Research Assistant, Carole L. Palmer (clpalmer@illinois.edu), PhD, is an Associate Professor and Principal Investigator, and Amy S. Jackson (amyjacks@illinois.edu), MLS, is a Project Coordinator in the IMLS-funded Digital Collections and Content Project at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. Myung-Ja Han (mhan3@illinois.edu), MLS, is a Metadata Librarian and Assistant Professor at the University of Illinois Library. The authors' mailing address is 501 E. Daniel str., Champaign, IL, 61820.

1. INTRODUCTION

Cultural heritage institutions have conceptualized and developed digital collections in many different ways. They may create a collection to showcase one or more larger physical collections, or they may compile a new, thematic whole from materials previously scattered across multiple institutions. Digital resource developers assemble collections purposefully, carefully selecting and arranging items to create groupings of objects that have significance beyond the aggregated features of individual members, to meet an aim or play a particular role. For example, they may be conceived of by their creators as "displays", "tours", "tools", "lessons", or the record of a cultural event (Palmer et al., 2006)[1]. However, when many collections are brought together in a federation or aggregation, the attributes of the original, deliberately built collections become difficult to discern. The individual items tell us little or nothing about the purpose or distinctive features of the collection from which they originated. Nor can collection features generally be inferred from groups of items retrieved in a search. Collection-level metadata has the potential to provide important context about the purpose and features of a parent collection and why the items may be of value to users, but the qualitative aspects of collections are difficult to describe in a systematic way, as they may embody a good deal of intellectual intent, and, compared to items, they tend to be highly complex and mutable.

This paper presents results from an investigation of how best to retain collection context to support scholarly use of large-scale heterogeneous digital aggregations, as part of the Institute of Museum and Library Services (IMLS) Digital Collections and Content (DCC) project. Over the past five years, the DCC development team has focused on providing integrated access to over 200 digital collections funded by IMLS National Leadership Grant awards, through a centralized collection registry and metadata repository. Concurrently, the DCC research team studied how collections and items can best be represented to meet the needs of both service providers and

diverse user communities. Findings from the project to date have been communicated to practitioners and have informed community efforts to define best practices for sharable item-level metadata.[2] In the new phase of the project beginning in October 2007, the research and development teams are undertaking a series of assessments and investigations to inform expansion and enhancement of the DCC for both academic and independent scholars (e.g., lay historians and genealogists).

The results presented here complement our previous analysis of trends in item-level metadata application (Palmer, Zavalina, & Mustafoff, 2007). Earlier DCC studies have also reported on collection-level concerns, identifying the various ways that resource developers conceive of collections and the attributes they find most important in describing their collections, and the different "cultures of description" evident among libraries, museums, archives, and historical societies (Knutson, Palmer, & Twidale, 2003; Palmer & Knutson, 2004). Preliminary usability studies have also suggested that collection and subcollection descriptions help users ascertain features like uniqueness, authority, and representativeness of the objects retrieved and lessen the confusion sometimes experienced in searching large-scale federations (Foulonneau et al., 2005; Twidale & Urban, 2005). This analysis extends our understanding of the role of collection description through a systematic content analysis of collection records to identify the range of different kinds of substantive and purposeful information about collections available within the DCC Collection Registry and to begin to assess its role and value for users. It is a baseline stage in our longer-term investigations of the relationships between item-level and collection-level metadata (e.g., Renear et al., 2008) and the value of collection description for enhancing the user experience with aggregated digital resources.

2. BACKGROUND

Characterizations of digital collections vary widely in the literature. Our concern with the purposeful nature of collections is reflected in the definition offered in the CIDOC object-oriented conceptual reference model (International Council of Museums/CIDOC, 2007): collections are "aggregations of physical items that are assembled and maintained … by one or more instances of Actor over time for a specific purpose and audience, and according to a particular collection development plan. Items may be added or removed from a Collection in pursuit of this plan." This statement stands out in its explicit attention to the intentions and activities of collectors.

Other definitions specify potentially important aspects of collections, as well. Johnston and Robinson (2002) state that "any aggregation of individual items (objects, resources)" qualifies as a collection, with no limitations as to the form and nature of items in a digital collection—either digital items as surrogates of physical items or "born-digital" content objects. Their view includes catalogs as tantamount to a collection, yet they are neutral on collection size, which can be as small as one item. They also emphasize the transient nature of digital collections and the fact that items are often dispersed across multiple physical locations. The layered nature of collections, acknowledged by Lee (2000), is increasingly evident as digital subcollections are created and as aggregations become more common. And DCC developers have suggested criteria for operationalizing the definition of a digital collection (Cole & Shreeves, 2004), based on dimensions such as thematic cohesiveness (e.g., by topic area, holding institution, type of materials), searchability as a distinct collection, and a unique point of entry (URL). But, traditional user-based collection criteria are still valid and necessary (Lagoze & Fielding, 1998).

It has long been recognized that contextual collection-level metadata is important for facilitating access to documents in archival and museum collections (e.g., Bearman, 1992; Sweet & Thomas, 2000; Dunn, 2000). Digital collections have come to be understood as information

seeking contexts (Allen & Sutton, 1993; Lee, 2000) but they can also be understood as a body of raw materials made available for further interpretation and presentation (Lynch, 2002). Among the developers of the collections contributed to the DCC, there is an interesting ambiguity in how they describe the nature, scope, and organization of what they are creating (Palmer et al., 2006). Many do not have a firm idea of whether they are building one whole or a number of differentiated collections. Not surprisingly, they tend to relate more to "projects" than "collections", and the relations between the two entities are not always clear (e.g., one-to-many or many-to-one). Collection development policy also tends to be conflated with digitization selection criteria. At the same time, some conceptualizations of collections seem to be defusing across professional orientations. For instance, notions of "artificial" and "organic" collections are retaining relevance beyond the archival community, and "exhibit" has been adopted by institutions other than museums and galleries.

The lack of empirical studies on the influence of collection structures, such as components and the organization among the components, has resulted in two significant problems, according to (Lee, 2003):

- considerations for structuring collections are often based on administrative or political factors, rather than on a user-centered approach
- the lack of understanding of requirements for different formats and media impedes effective system and service design.

Information professionals' and users' criteria for conceptualizing and structuring collections differ (Lee, 2000; 2003; 2005). For example, academics have been shown to benefit from the usefulness of collections and subcollections, even when certain subcollections are not explicitly defined by the library as distinct structures. Other important functions provided by collection

structures include: collocation, selectivity, narrowing the search scope to increase precision and ease of use, presenting choices, and assisting in information need clarification.

Collection metadata has a vital role to play in facilitating access, and its importance continues to increase in the digital environment. Macgregor (2003) defined collection metadata as "a structured, open, standardized and machine-readable form of metadata providing a high-level description of an aggregation of individual items." This level of descriptive granularity adds important relational (Macgregor, 2003) and contextual information (Miller, 2000), functional for both users and institutions.

Collection description can be further distinguished as "unitary", which "consists only of information about the collection as a whole and does not provide information about the individual items within it", and "analytic", which "consists of information about the individual items within [a collection] and their content" (Heaney, 2000). More recently, best practice recommendations for OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) data provider implementations and shareable metadata stress the importance of retaining context when aggregating item-level metadata and the necessity of expressing and sharing descriptions of the collections to which items belong (Digital Library Federation/National Science Digital Library, 2005; Shreeves, Riley, & Milewicz, 2006).

As digital content continues to grow and be reconfigured, relational attributes in collection-level metadata specifying associations between a given collection and its various sub- , super- and otherwise related collections will be essential, not only for discovering resources within single repositories, but also across institutions, and across different domains. Foulonneau et al. (2005), Geisler et al. (2002) provide supporting evidence from a study of metadata harvested

from Committee for Institutional Cooperation (CIC)[3] institutions, showing that linking item-level and collection-level metadata can:

- produce higher retrieval rates for item-level descriptions,
- re-contextualize orphaned items by including key access lacking in item-level metadata,
- facilitate browsing behavior familiar to humanities scholars.

Free-text metadata — particularly the *Description* field, defined by the Dublin Core Collection Description Application Profile (DCCAP) as a required "free text summary description of the collection"[4] — has been an integral part of collection-level metadata, providing important human-readable contextual information for users. DCCAP does not prescribe what should be included in collection-level free-text *Description* field, however subjects of a collection are suggested as possible content: "Although a description might contain detailed subject-specific information, at least part of the description should be understandable by an end-user with no specialist knowledge of the subject area." The Dublin Core Metadata Elements Set for item-level metadata[5] provides a slightly more detailed definition and some guidelines as to the contents of the mandatory *Description* field: "An account of the content of the resource", "may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content." The Dublin Core Usage Guide[6] recommends limiting the length of *Description* field to "a few brief sentences."

The usage guides created by different communities for their own needs suggest that collection- and/or item-level *Description* information should "be helpful to users attempting to discern the usefulness of a resource to their research needs" (*NCSU Libraries Core 1.0 Metadata Element Set Best Practices*, 2007), and provide information that is not covered by other metadata elements or "supplement, qualify, or explain" information in other metadata

elements (*Cataloging Cultural Objects*, 2008). Usage guides have recommend providing information about:

- "salient characteristics and historical significance of the subject, function, and significance of the work", work's "relationship to other works, its style, and any aspects of it that might be either disputed or uncertain" (*Cataloging Cultural Objects*, 2008);
- types of materials included in collection, associated dates, "names, dates, and biographical identification of persons and names of corporate bodies significant (by quality and/or quantity of material) to the collection", specific phases of career/activity of the major person/body responsible, geographical areas, events, topics, and historical periods with which the materials in the collection deal, and "particular items of extraordinary interest" (Webform for creating collection records in *National Union Catalog of Manuscript Collections*, 2008).[7]
- "provenance, distinguishing features, inscriptions, the nature of the language of the resource, and/or history of the work" (*OSU Knowledge Bank Metadata Application Profile*, 2006).

The broader cataloging/metadata community has developed detailed guidelines for creating descriptive summary notes in MARC-format item-level records, which might be useful in thinking about encoding of the collection-level *Description* field content as well. The guidelines created by OLAC Cataloging Policy Committee (2002) recommend including such elements as "unique features" or "distinguishing features", "user interaction", "specific effects" (e.g., laser display or animation), and "history of the work", when describing individual items. These guidelines also mention including audience information when creating summary notes in item-level records for motion pictures and video recordings. For describing archival materials –

normally represented as collections – OLAC guidelines recommend inclusion of summary note information about "specific types and forms of materials present", "reason and function of the collection", "significant people, places, events and topics covered," "span of dates covered by collection," "typical and unique characteristics of the collection," and "consequences, products, and results of the events documented."

Overall, among the wide range of free-text metadata components suggested by the existing guidelines, topic coverage, geographic and temporal coverage, and object types are the most consistently recommended.

3. METHODS

The analysis presented here builds on research and development conducted over the previous five years of the DCC project.[8] As stated above, a content analysis of all DCC collection records was conducted to identify the range of different kinds of substantive and purposeful information about collections available within the DCC Collection Registry. We were also interested in determining patterns in representation, the efficacy of the records, and the adequacy of the collection schema (discussed further below) for representing the richness and diversity of collections in the aggregation. This required identifying redundancy within records but also detailing what was being represented in free-text fields. The analysis has also been an important, empirically-grounded step in the DCC research team's ongoing efforts to better understand collections as entities. That is, to specify the ways in which collections are more than a sum of their parts, in terms of both the intentions of collection creators and value for scholarly users.

The results presented here are based on a systematic, manual content analysis of the 202 collection-level records in the DCC Collection Registry. We addressed our research aims by identifying patterns in the data provided in free-text fields, focusing primarily on the *Description* field and other selected free-text and controlled vocabulary fields. It is important to note that the

collection records have been created by the Project Coordinator for the DCC development team, with the content being drawn directly from documentation provided by the local developers of the individual collections. This process is discussed further in the section that follows.

There is considerable variation in the length of the *Description* field, with a range of 5 to 429 words. Figure 1 below shows the frequency distribution of the *Description* field length values, defined as the number of words per *Description* field, for all 202 collections. The average length was 91.93 words; the majority (66%) of collection records had a *Description* field with 100 or less words, 23% had between 101 and 200 words, and only 5% had more than 200 words.

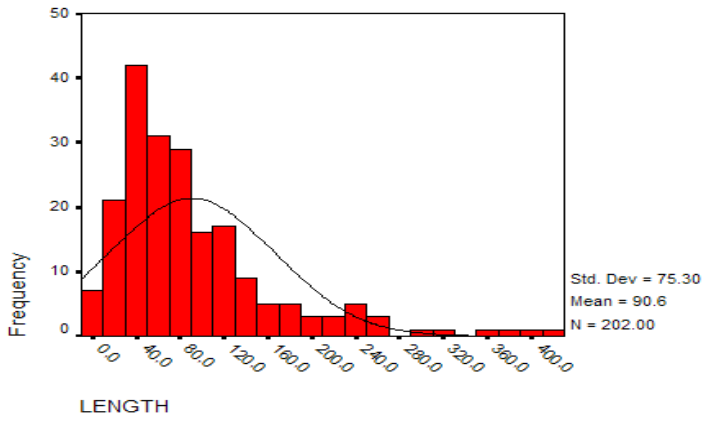

Figure 1 Distribution of *Description* field lengths (number of words)

Our preliminary review of the records suggested that the free-text *Description* field provided essential information, including subjects of digital collections, types of objects represented by collections, collection size, audience, particular collection strengths, etc. Through a full, systematic coding of the content we expected to see free-text *Description* information complementing rather than repeating information found in other fields.

The free-text in the *Description* field was both qualitatively and quantitatively analyzed through direct examination and coding to identify:

- Types of information provided about a digital collection, especially that which was not represented elsewhere in the collection-level record;
- Degree of agreement between information provided in the free-text *Description* field and relevant information found in other fields of the collection-level record;
- Co-occurrence of different types of information;
- Field length and its association (if any) with the richness of information contained in it.

Hereafter, we use the term "collection properties" to refer to the types of information identified in the collection records.

No predefined list of categories was used for analysis. The categories emerged from coding performed by two coders who are authors on this paper. Through iterative review and discussion, the coders developed agreement on the categories represented and the terminology used for the categories. A test of intercoder reliability showed 80.4% agreement in assigning the codes to specific cases.

Additional analysis was conducted on four fields intended for subject indexing in the collection registry (*GEM Subjects, Subjects*—for alternatives or supplements to GEM*, Geographic Coverage,* and *Time Period*), a field describing types of objects in digital collections (*Objects Represented*), and others that matched properties that emerged out of analysis of the free-text description content, such as *Size* and *Collection Development Policy*.

The results of the content analysis are supplemented with longitudinal data documenting modifications made to collection descriptions since February 2005[9], when the DCC Collection Registry was first populated with collection-level metadata. The modification data was brought in to triangulate findings of the content analysis and provide additional context for the discussion, whenever appropriate. Before presenting the findings from the analysis, below we give an

overview of the collection description schema developed by the DCC project and the process used to populate the DCC Collection Registry.

3.1 DCC schema and descriptive practices

The DCC collection description metadata schema was based largely on the Dublin Core Collection Application Profile[10] and the UKOLN RSLP schema[11] (Heaney, 2000). The schema describes four entities: the digital collection itself, the grant project responsible for collection, the institution responsible for the collection, and the person(s) responsible for administration of collection.

For describing the collection per se, the schema provides 17 general attributes (e.g., collection title, size, objects represented, language, etc.), 4 topical attributes (topic, [free-text] description, geographic coverage, and time period), 4 attributes describing relationships with other collections (parent collection, sub-collection, source physical collection, and other associated collection), and 4 attributes describing relationships with projects, institutions, and administrators (grant project, hosting institution, contributing institution, and administrator). The project entity is described in the schema with 5 attributes, the institution entity with 6 attributes, and the administrator entity with 7 attributes.[12]

The information used to create collection records is initially supplied from administrators of individual digitization projects who complete a survey about their collections which is reviewed by the DCC Project Coordinator. The survey collects basic information about the grant project (e.g., title and URL), information about the collection (e.g., time periods covered, types of objects represented, targeted audiences), and technical information (e.g., types of controlled vocabulary, digital library management system used, and availability of OAI-PMH). Additional information is also gathered by a manual review of the collection's website or portal. The free-text

Description field is generally constructed from text provided on the website or in the grant proposal submitted to IMLS. Once the initial record has been created, and before it is made viewable through the public interface, collection administrators review the record and can update, change, or add information or links to related collections through the internal collection registry record edit interface. Before newly added or edited records are uploaded to the publicly accessible copy of the Collection Registry, records are individually vetted by the DCC Project Coordinator. The limitation of this approach is a lack of first hand knowledge by the DCC Project Coordinator of the collection being described, although errors should be corrected by the collection administrator when editing the record.

Thus, the free-text *Description* field retains the original language and characterizations of digital collections as expressed by resource developers, and oversight is provided by current local collection administrators, who are responsible for reviewing and revising the records. Modifications of the Project Coordinator's initial records have been infrequent, however. For example, the *Description* field was changed in only 14 of the 202 records (6.93%), while *Audience*, *GEM Subjects*, and *Size* were modified in at least twice as many records. Overall, the descriptions are relatively complete and every effort has been made to accurately represent collections based on sources provided by the collecting institution, with local review of records as part of the standard procedure.

The subjects of digital collections in the Registry are indexed with the Gateway to Educational Materials (GEM) subject vocabulary, originally created to describe digital objects in the GEM repository and considered suitable for browsing databases in a cultural heritage domain. At the top level, GEM consists of twelve broad subject headings: Arts, Educational Psychology, Foreign Languages, Health, Language Arts, Mathematics, Philosophy, Physical Education, Religion, Science, Social Studies, and Vocational Education. Each of the broad subject headings

has between 12 and 29 narrower "level 2" headings under it. The second level subject headings for Philosophy and Religion replicate ERIC Thesaurus "Narrower Terms" for these two broad subjects. Several of the "level 2" GEM subject headings—Careers, History, Informal education, Instructional issues, Process skills, and Technology—are facets applicable to each of the twelve broad subject categories.

Digital resource developers participating in the Registry are required to provide top-level GEM subjects (at least one) in their collection records. Use of second-level GEM and subject headings from alternative schemes is not required, but is supported by the collection metadata schema. Some other controlled vocabularies used for describing collections in Collection Registry include the Getty Thesaurus of Geographic Terms, Library of Congress Thesaurus of Graphic Materials - Genre and Physical Materials Terms (LC TGM II), etc.[13] In the process of describing their collections through the edit interface, digital resource developers may select from a list of controlled vocabulary values for the following eight elements: *GEM Subjects*, *Geographic Coverage*, *Time Period*, *Objects Represented*, *Supplementary Materials*, *Audience*, *Interaction with Collection,* and *Frequency of Additions*.

4. FINDINGS

The primary focus of this report is on the data provided in the free-text *Description* field, therefore the analysis covers the 198 (out of 202) collection records that have a *Description* field, with reference to fields containing related and complementary data, including *Subjects* and *Size* fields, which are free-text, and controlled vocabulary fields, including *GEM Subjects, Geographic Coverage, Time Period*, and the *Objects Represented* field.

4.1 Collection Properties in Description Field

Tables 1 and 2 outline the collection properties (types of information about a digital collection) that were identified in five or more collection records, through close reading and coding of the data in the *Description* field by two coders. A total of 197 collection records had between 1 and 9 of these collection properties indicated in the *Description* field, with an average of 4.3.

Table 1 lists the properties found only in *Description* field and not reflected anywhere else in the record. These can be subdivided into three groups. Special claims about collections—Importance, Uniqueness, and Comprehensiveness—are found in a limited number of records, but they are of particular interest as the kind of self-assessed, special claims used to distinguish special collections in libraries, museums, and archives. Two other important properties, for which no specific elements in collection metadata exist—Provenance and Item Creator—belong to the second group. The third group includes two properties—Subject and Objects—for which formal elements do exist but *Description* field provides extensive additional coverage.

Table 2 shows nine collection properties which are not unique to the free-text description field.

Collection Property	***Number of collections***	***%***
GROUP 1		
Importance	20	10.1
Uniqueness	17	9.0
Comprehensiveness	6	3.0
GROUP 2		
Item Creator	78	39.4
Provenance	24	12.1
GROUP 3		
Subjects not represented in formal metadata elements	132	66.7
Objects not represented in formal metadata elements	37	18.7

TABLE 1. Collection properties unique to *Description* field

Collection Property	***Number of collections***	***%***

Subjects	181	91.4
Object types	149	75.3
Collection development policy (explicit or implicit)	102	52.0
Collection title	103	52.0
Size	53	26.8
Audience	34	17.0
Navigation and functionality	32	16.2
Participating/contributing institutions	30	15.2
Funding sources	10	5.1

TABLE 2. Other collection properties in *Description* field.

4.1.1 Special claims about collections

As can be seen from Table 1, a number of collection records in the Registry include indications of one or more of the following three collection properties:

- *Importance* (e.g., "collection of the most important and influential 19th and early 20th century American cookbooks", "materials are significant in their place within the fabric of American history and culture", "creating an archive of unparalleled importance", etc.)
- *Uniqueness* (e.g., "unique historical treasures from ... archives, libraries, museums, and other repositories", "rare historic published monographs and serials", "rare and unique library and archival resources on race relations", etc.)
- *Comprehensiveness* (e.g., "a comprehensive and integrated collection of sources and resources on the history and topography", "the most comprehensive library of manuscripts, rare and contemporary books", "one of the most ambitious and comprehensive effort to date to deliver educational content on the Civil Rights Movement", etc.).

Twenty-six free-text *Descriptions* contain one of these special claims, while 7 contain two, and 1 contains three which brings the total proportion of collection records making special claims about their collections to 17%. Although not prominent enough to include in the table, a

related property, “Strength”, also appeared in at least three records, in reference to collections or sub-areas within the collection. These findings on special claims that developers make about their collections will not be surprising to the metadata community. For example, there has been discussion about the inclusion of a *Strength* element into the Dublin Core Collection Application Profile (DCCAP) to accommodate descriptive information related to aspects such as importance, uniqueness, and comprehensiveness (e.g., Johnston, 2003), while the RSLP collection description schema has an “cld:strength” element for “An indication (free text or formalised) of the strength(s) of the collection.”[14] (e. g., Heery & Patel, 2000).

4.1.2 Provenance

Provenance information was included in 12.1% of the free-text *Description* fields. These three sample excerpts represent the kinds of information provided: “in December 2002, the ... Library acquired the Humphrey Winterton Collection of East African photographs”; “acquisition of these hitherto unknown manuscripts was spearheaded by Edgar J. Goodspeed in the first half of the twentieth century”; “a 1988 bequest of more than 850 landscape prints and drawings from the collection of Los Angeles architect Rudolf L. Baumfeld significantly enhanced this wide-ranging and well-studied thematic area”.

The DCC aggregation includes a large number of museum collections and a smaller but substantial group of historical society and archive collections. It seems likely that, if available in our collection metadata scheme, a provenance element might serve even a greater percentage of collections than those exploiting the *Description* field for this purpose. The DC CDAP *Custodial History* element covers provenance information found in our free-text metadata.

4.1.3 Item Creator

Seventy-eight collection records (39.4%) contained names of artists or institutions that created items in the collection. For example, corporate authors may be identified as in, "The Museum Extension Projects of Pennsylvania, New Jersey, Connecticut, Illinois, and Kansas crafted most of the items currently in the collection". Individuals might be specified and further biographical information for them supplied as well (e.g., "images are noted on their mounts as being from Watkins's "New Series".... Watkins was active between 1854 and the late 1890s."). Like the provenance information discussed above, there is no specialized element in the DCC collection metadata schema that could accommodate this type of information,[15] yet it appears of high value as contextual information for users. There are DCC collections related to single or multiple authors that could benefit from more formal representation of item creators. In this case, a new element would need to be specified, since the existing DCCAP *Collector* element is designed to cover creator of the collection, not creator of items in the digital collection.

4.1.4 Subject

Subject-specific information is most prominent in the free-text *Description* field, appearing in 91.4% of the collection records. The content ranges from very specific subject coverage statements (e.g., "cover a broad range of topics, including ranching, mining, land grants, anti-Chinese movements, crime on the border, and governmental issues") to subject keywords scattered throughout the text, as in this example: "During *World War II*, as a member of the *U. S. Army, 252nd Field Artillery Battalion*, he captured over 700 images of *life as a soldier* and unique snapshots of *events of the war*".

In most cases (66.7%), the *Description* field provides more accurate and specific coverage than the fields intended for subject indexing*: Subjects*, *GEM Subjects*, *Geographic Coverage,* and *Time Period*.

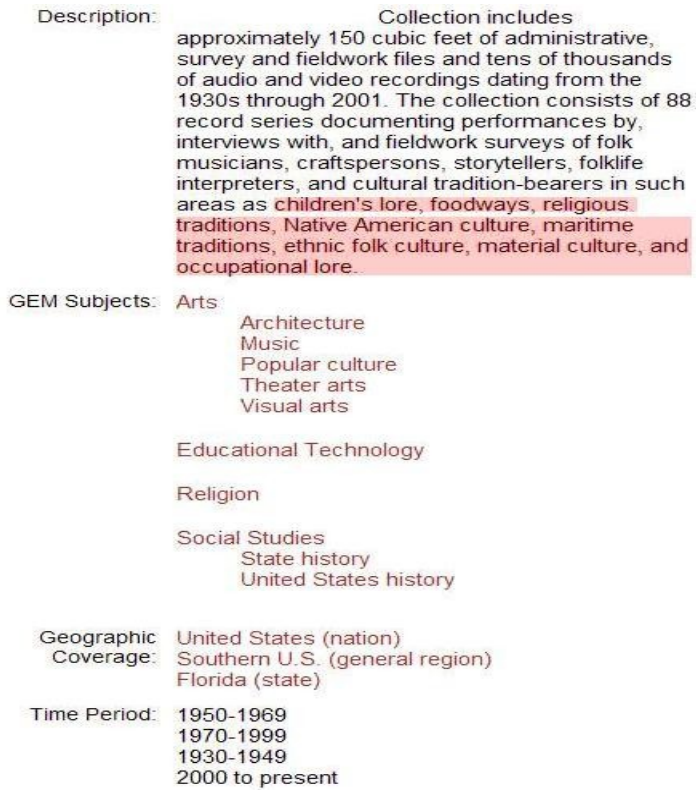

Figure 2 Subject information in *Description* field

As illustrated in Figure 2, free-text often adds essential subject information to a record. In this case, the text includes keywords that provide more accurate and specific coverage than all four fields in the collection records intended for subject indexing taken together (*GEM Subjects,* alternative *Subjects, Geographic Coverage*, and *Time Period* fields). The standard subject vocabulary options are clearly too general and the free-text description is, as one would expect, likely to be more compelling to users.

The *GEM Subjects* field is a required, repeatable field in the DCC collection records. One top-level GEM subject was used by 114 (56%) collections. Seventy-eight (39%) use 2-4 top-level GEM subjects, and only 9 collections (4%) use 5 or more top-level GEM subjects. All but one of the collections that used a top-level GEM subject also used at least one second-level GEM subject. The majority of collections, 128 (64%), used between 1 and 3 second-level GEM

subjects. Eight (4%) collections used between 10 and 20 second-level GEM subjects, and 5 (2.5%) collections used more than 20 second-level GEM subjects, with one of these collections using 67 second-level subjects.

At the same time, our longitudinal analysis of modifications made to collection records by digital resource developers demonstrated that *GEM Subjects* was the second most frequently modified field (after *Audience*), with 27 modifications in 25 collection records. In two collection records, both changes and additions were made. The vast majority of modifications were to add headings, both at the top level (between 1 and 3 headings added for 17 collections) and the 2nd level (between 1 and 54 headings added for 25 collections). It is worth noting that in 6 cases, digital resource developers modifying the *GEM Subjects* field also modified *Subjects* – an element providing optional, alternative topic access to collections through the use of controlled and un-controlled vocabularies other than GEM (e.g., Library of Congress Subject Headings, Art and Architecture Thesaurus, locally-developed vocabularies, keywords, etc.). In 3 cases, the *Subjects* field was modified without modifying *GEM Subjects* field.

A total of 148 (73.3%) collection records in the Registry use the alternative *Subjects* field: one uses it instead of *GEM Subjects* field and 147 collection records use alternative *Subjects* field in addition to *GEM Subjects* field. An overview of some characteristics of the text follows:

- 113 (76.4%) use between 1 and 14 phrase headings (e.g., "Japanese internment", "Louisiana culture", "Atlantic Sea Turtles", etc.) with an average of 2.64 phrase headings per record.
- 60 (40.5%) use between 1 and 28 compound headings (e.g., "industries (lumber, mining, boats, railroads)", "Africa—Rites and Ceremonies", etc.) with an average of 3.47 compound headings per record.

- 35 (23.6%) use between 1 and 11 single-word headings (e.g., “Desegregation”, “Taxonomy”, etc.) with an average of 2.51 single-word headings per record.
- 6 (4%) use between 1 and 2 acronym headings (e.g., “YMCA”, “WPA”, etc.) with an average of 1.16 acronym headings per record.
- 3 (2%) use one free-text sentence enumerating multiple subjects (e.g., “Historical, social, cultural images from the Detroit news photo archives”, etc.).

Longitudinal analysis of record modifications shows that the *Subjects* field has been modified at a lesser ratio than *GEM Subjects*, with revision of seven collections records for a total of 11 modifications. Four digital resource developers made both changes and additions to this field in their collection records. Between one and eighteen subject terms or strings were added by seven resource developers. In three out of four cases, the change was a complete switch to Library of Congress Subject Headings (LCSH).

LCSH is widely used as an alternative *Subjects* vocabulary. Sixty-eight (46% out of 148) collection records explicitly use *Library of Congress Subjects*; nine (6.1%) use subject headings that look like they are LCSH headings (e.g., “Colorado Plateau—History”, “World War, 1939-1945”, etc.). Interestingly, LCSH has been successfully applied across all types of content, some with even highly specialized collections of objects such as “physical specimens (plants / animals / etc.)”, “music (audio files)”, “moving images”, and “prints and drawings”. In fact, among the 68 collections that use LCSH, only 19 included “books and pamphlets”.

Ninety-three (62.8%) of 148 collection records that make use of an alternative *Subjects* field also indicate additional subjects areas in the *Description* field. That is, they articulate subjects beyond those covered in *Subjects* and all other fields that represent subjects: *GEM Subjects*, *Geographic Coverage,* and *Time Period* fields. As can be seen from Table 1, this proportion is

only slightly lower than the percentage of all collection records in which the *Description* field provides additional subject information (66.7%). This finding suggests that although using multiple subject vocabularies for describing collections is beneficial in improving subject access, the free-text *Description* field is still important for enriching subject representation of collections.

In addition to the alternative *Subjects* field, the optional fields for geographic coverage and temporal coverage are widely used. *Geographic Coverage* is used by 174 collection records (86%) with numbers of entries ranging from 1 to 27. Eighty-six (49.4%) collections use 1-2 entries, and 8 (4.6%) collection records use 10 or more entries. Eighty (46%) collection records use between 2 and 9 entries. At the same time, sixty percent of the free-text *Description* fields include indications of geographic coverage of varying granularity (e.g., "Austro-Hungarian Empire"; "Mayan city of Uxmal in Yucatan, Mexico and a Native American Mississippian site, Angel Mounds U.S.A."), often more accurate and specific than in *Geographic Coverage* field.

The *Time Period* field is used by 156 (77.2%) collection records, with numbers of entries ranging from 1 to 10. Sixty-seven (43%) collection records use 1-2 entries, and 41 (26.3%) use 5-10 entries. Forty-eight (30.8%) collection records use 3-4 entries. Fifty percent of the free-text *Description* fields include indications of temporal coverage, ranging from specific dates and date ranges (e.g., "19th century") to known historical periods (e.g., "World War I"; "California Golden Rush").

Longitudinal analysis of modifications made to collection records demonstrates that *Geographic Coverage* and *Time Period* fields were modified in 20 and 17 cases, respectively. A number of resource developers added optional *Geographic Coverage* and *Time Period* fields, which were not originally part of their collection records.

4.1.5 Object types

Object type was the second most widely represented collection property in the free-text *Description* field, with three-quarters of the records describing types of digital objects in a collection. As seen in the case of subjects, above, the *Description* field often (in 18.7% of cases) listed more, or more specific, types than covered by the formal element, *Objects Represented*. General object terms, such as "physical artifacts", were common, as were more specific terms, such as "lanterns, torches, banners".

As seen in Figure 3, physical formats and genres are also frequently specified, as with "pamphlets, leaflets, and brochures", "songbooks", and "political cartoons". Object types and formats are sometimes conflated, even within the same sentence, in the *Description* field, as well as in *Objects Represented*. This lack of disambiguation between object type and format is a known metadata quality problem for digital object description[16] (Jackson et al., 2008; Godby, Smith & Childress, 2003; Park, 2005; Hutt & Riley, 2005).

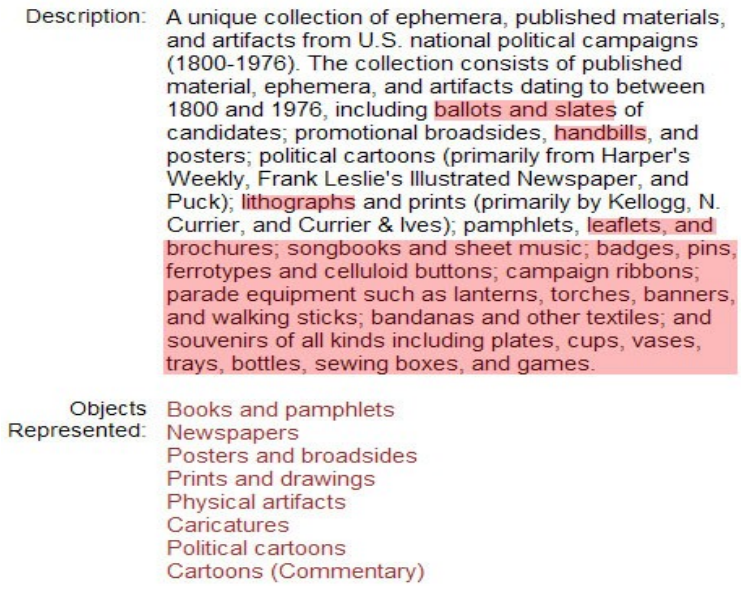

Figure 3 Object types information in Description field

All 202 collections in the Registry use the *Objects Represented* element, with the number of types specified ranging from 1 to 15. Ninety-five (47.0%) collections use 1-2 entries and 11 (5.5%) use 10 or more entries. Ninety-six (47.5%) collections use between 3 and 9 entries. In addition, this field was modified in 15 collection records, with the tendency to include from 1 to 4 new types of objects not previously listed in collection records. Some examples of added object types include sheet music and scores, prints and drawings, maps, posters, and broadsides.

4.1.6 Collection development policy

Collection policies and criteria were rarely encoded for the *Collection Development Policy* metadata element. Only 9 (4.5%) collection records in the Registry made use of this field as of March 2008. However, over half (52%) of the free-text *Description* fields contain either explicit or implicit evidence of certain collection development policies or digitization selection guidelines.

Some of the more specific descriptions offer information such as: "titles published between 1850 and 1950 were selected and ranked by teams of scholars for their great historical importance", to more ambiguous criteria, as in: "a selection of framed items from the collections of the ... Library", or "a sample of the photographic archives". Some descriptions identify plans for future collection development, a potentially significant aspect of collector intentionality, or other locally accessible assets: "in addition to the newspapers, it is planned to provide access to a complimentary collection of Richmond related Civil War period resources"; "additional lesson plans, activities and photo essays designed by teacher advisors and educational consultants will be added in the future". Others explicitly state a purpose: "support global efforts to conserve, study, and appreciate the diversity of palms", or "stimulate the documentation and preservation of ethnic materials and foster a greater interest in the history and cultures of the peoples of the region". These statements are multifaceted, with important data about potential audiences and the intellectual and evidentiary intentions of collectors.

4.1.7 Collection title

One hundred and three records (52%) include collection title information in the free-text *Description* field. While duplicative of the *Title* field, many titles provide concise statements with subject-specific information, as well as information on the types of objects in the collections, which are typical of *Description* field content. An additional 2 records (1%) include collection subtitle only, and 1 record (.5%) uses a collection title's acronym in the *Description* field.

4.1.8 Collection size

Over a quarter of the records (53) had *Description* fields that made statements about the collection size, ranging from quantitative specifications ("13 oversized boxes containing 209 cartoons, 12 Christmas cards, and 3 facsimiles of cartoons") to general orientations (e.g., "hundreds of personal letters, diaries, photos, and maps"). Some free-text *Description* fields also referred to the size of an associated physical collection, such as: "the costume collection at the ... Museum has over 30,000 items of clothing and accessories"; "the physical collection contains over 400 garments"; "physical collection is comprised of several hundred photographs, publications and newspaper clippings", etc.

At the same time, 129 collection records (64%) use the formal *Size* element, including:

- 115 collection records that indicate a specific number of items (e.g., "361 black and white photographs", "7,600 photographs in 75 albums", "approximately 10,000 items", etc.)
- 13 collection records that input the "unknown" value in the *Size* field.

Size specifications may not be straightforward for some collections, as indicated by a collection of "events and primary sources" that encoded the *Size* field with "Timeline of multiple themes".

Eleven records with collection size information in the *Description* field (20.8%) have not utilized the *Size* element and 4 (7.5%) input the "unknown" value in *Size* field. In these cases, the

Description field is the only source of this potentially valuable information for the user. However, out of 53 free-text *Description* fields that indicated collection size, only about half (26 records or 49.1%) match the *Size* field data (e.g., "44,000+ records in nearly 100 collections" and "44000", "plant material for more than 600 of the country's most imperiled native plants" and "600 plant profiles"). In 16 collection records (30.2 %) the size data in the two fields does not match. Eight records report lower collection size in the *Size* field than in the *Description* field. Four records report a conceivably higher number in the *Size* field than in the *Description* field (e.g., "3000" photographs vs. "1000" and "300" photographs; "47,310" and "over 30,000 public documents and 300 publications").

In these cases, there is no evidence that descriptive information about physical collections has been slipped into the record. Instead, these discrepancies seem to reflect, sometimes clearly, the difference between planned/projected and actual current/initial size of the digital collection (e.g., "When finished, the collection guide will consist of well over 100,000 online stereoviews" in the *Description* field and "38254 Stereographic Photoprints" in the *Size* field).

According to our longitudinal analysis of modifications to collection records, 18 additions and 7 changes were made to the *Size* field between February 2005 and September 2007, making it the third most frequently modified field. The majority of those modifications added this optional element to existing collection records, and not surprisingly, changes were to increase the number of items in the collection.

4.1.9 Audience

Audience metadata element was the most frequently modified field in the DCC Collection registry, with twenty-nine records adding anywhere from 1 to 12 new audiences. In the *Description* field, audience information, both broad and specific (and sometimes implicit), were found in 17% of the

collection records. Representative examples include: “Alabama residents and students, researchers, and the general public in other states and countries”; “created especially for middle and high school students”; or the implied general public and educator audience in “provided for personal use or educational presentations”. All but one collection-level records that had audience information in the *Description* field also used the formal element, *Audience*, with 1 to 11 values applied. As illustrated by Figure 4, the *Description* field often complements and clarifies values in the *Audience* field.

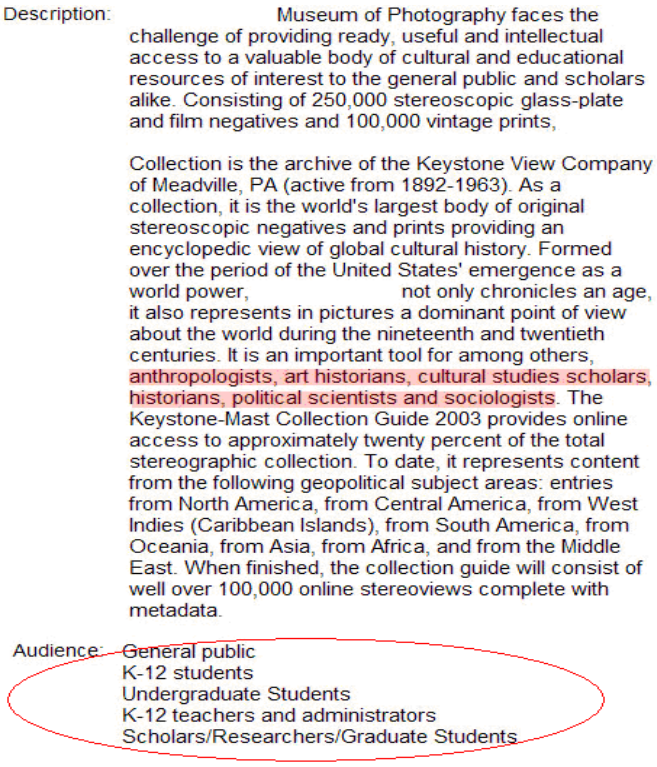

Figure 4 Audience information in Description field

4.1.10 Navigation and functionality

Twenty-three records (11.2%) contained navigation or functionality information in the *Description* fields (e.g., “may be searched or browsed in a variety of ways, including by keyword, subject, creator, title, and date”, “accessed by the scanned county photomosaic or line indexes”,

etc.). Some aspects of the free-text *Description* field information might also be represented in the formal *Interaction with Collection* field (e.g., "accessible by date of issue or by keyword searching" in *Description* and "search, browse" in *Interaction with Collection*). In most cases, information in the two fields was complementary, especially in cases when resource developers used both controlled-vocabulary values and free text in the *Interaction with Collection* field. This excerpt shows the kind of functions associated with a collection of television programs: "video excerpts, searchable transcripts, a select number of complete interviews for purchase, and resource management tools" in *Description* and "search, browse, e-mail select to colleague, create notes with my list favorites, favorite referral (people who liked this also liked....), sort" in *Interaction with Collection*.

Some of the statements in the *Description* field were accompanied by information on how the digital collection is organized for browsing, which was not available anywhere else in the collection-level record. Browsing organization was referred to in 11 (5.6%) of *Description* fields (e.g., "grouped by county", "the overall organization of the database is by tribe", "arranged chronologically by Japanese periods", etc.).

4.1.11 Participating, contributing institutions

Thirty collection-level records (15.2%) provide information about institutions participating in the digitization project and contributing items to digitize (e.g., "project brings Tufts, and the Virginia Center for Digital History together with the University to build a digital repository"; "digital images of archival collections located at three Arizona repositories: the University of Arizona Library Special Collections; the Arizona Historical Society-Tucson; and the Arizona State Library, Archives, and Public Records", etc.).

4.1.12 Funding sources

Ten collection-level records (5.1%) acknowledge funding sources that helped build their digital collections (e.g., "received an IMLS National Leadership grant to create the digital resource", "funds provided by the Institute of Museum and Library Services, under the federal Library Services and Technology Act", "two multi-volume sets digitized as the result of an Illinois State Library FY98 Educate and Automate grant", etc.).

5. DISCUSSION AND CONCLUSIONS

Our findings confirm that the free-text *Description* field provides substance that enriches collection-level records in the IMLS DCC. There is consistent representation of subjects and object types that is more accurate in coverage and offers more detail than that represented in the other fields specified for those purposes. Moreover, notation of what we refer to as "special claims" about a collection—indicated by terms such as Importance, Uniqueness, and Comprehensiveness—add vital qualitative, contextual information about the intentions of collectors and the role the collection plays in the larger universe of related content. These properties are not represented in other parts of the collection record. Provenance and Item Creator properties also emerged as strongly represented within the free-text *Description* field in DCC collection records, with additional complementary contextual information about collection development criteria, collection size, audiences, and navigation and functionality. All of these data represent distinguishing features potentially of interest to scholarly and other research audiences.

The first activity slated for collection record enhancement in the DCC is to realign its collection description schema with the DCCAP, which was released after development of the DCC schema. Certain elements from the DCCAP, including the *Custodial History* field, will accommodate some of the key information currently only found in the *Description* field. A newly

defined field for creators of items in a collection and a specified field for special claims about collections are also under consideration.

The next step in our study of free-text collection-level metadata is a comparative analysis of collection records from sources other than the DCC aggregation, produced by libraries, museums and archives. A broader understanding of the use of the *Description* field in various organizational contexts will be particularly meaningful as we continue to explore the relationship between context and content, and the ways in which collection-level description can complement item-level description.

Due to the varied use of the *Description* field, which includes information on institutions, projects, physical collections, and digital collections, it would be difficult for a registry provider to automate extraction from this field to populate or enhance other elements in a record. However, the *Description* field could more easily lend itself to mining for production of controlled vocabularies customized for use in the DCC and similar aggregations of cultural heritage digital materials. Our intention is to experiment first with improving our existing vocabularies for audience and objects represented, and possibly subject areas for which we have strong concentrations of content. These emergent controlled vocabularies would be more representative of the terminologies being used by current collection developers to clarify the purpose and value of their collections and would provide a more accurate picture of the content included in collections than the overly general controlled vocabularies currently being applied.

Our analysis revealed that a number of the resource properties recommended by existing guidelines for free-text *Description* field (discussed in the introduction to this paper) are being used at the collection-level: uniqueness, significance, provenance, subjects, types of objects represented, navigation and functionality, collection development criteria, and audience. The

findings of this study show that additional collection properties can complement these properties, such as Comprehensiveness, Collection Title, Collection Size, Item Creator, Contributing and Participating Institutions, and Funding Sources in the free-text collection-level *Description* field usage guidelines.

Our analysis has shown that, quite predictably, longer free-text *Description* fields tend to represent more collection properties.[17] However, there are a number of examples in the DCC aggregation that can serve as exemplars of how to construct concise yet rich descriptions. A series of properties can be covered in descriptions of less than 50 words, about half the average length of a *Description* field in DCC aggregation. For example, this description covers at least 4 properties not represented elsewhere in the collection-level record: collection development criteria, objects, uniqueness, and subjects:

> …collects, preserves, and makes available free streaming video of hard-to-find documentary films about American folk or roots culture, giving wider audience to the independent filmmakers and the diverse American artists and groups they have documented… [and] also provides in-depth contextual materials about the films and their subjects.

As "the collection", as a defining or organizing unit, becomes increasingly destabilized in the digital environment (Currall, Moss, & Stuart, 2004; Johnston & Robinson, 2002; Manoff, 2000) we stand to lose essential context for understanding why items have been identified and gathered together in the first place, and, most importantly, how those items, as a group, relate to our cultural heritage. The DCC aggregation contains many unique and highly curated collections, each created with a particular purpose or to make a certain cultural contribution. Rather than presenting only a small window into thousands of items separated from their original contexts, our aim is to learn how to build coherent collections of collections that retain and reflect the original

intentions of their creators, and so better support scholarly inquiry. The DCC and other aggregations can offer an alternative to the decontextualized, uncurated view of the cultural world provided by Google and other resources that do not exploit the valuable properties of "collections" for the advantage of users.

6. ACKNOWLEDGMENTS

This research was supported by a 2007 IMLS National Leadership Research and Demonstration grant LG- 06-07-0020-07.[18] We also wish to thank our colleagues from the Metadata Roundtable[19] for their helpful comments and suggestions on a preliminary draft of this paper.

7. REFERENCES

Allen, B., & Sutton, B. (1993). Exploring the intellectual organization of an interdisciplinary research institute. *College & Research Libraries, 54*, 499–515.

Bearman, D. (1992). Contexts of creation and dissemination as approaches to documents that move and speak. In *Documents that Move and Speak: Audiovisual Archives in the New Information Age: Proceedings of a Symposium held 30 April to 3 May 1990 at the National Archives of Canada*, 140-149.

*Cataloging Cultural Objects: a guide to describing cultural works and their image*s. (2006). Chicago: American Library Association, pp. 245-251.

Cole, T., & Shreeves, S. (2004). Search and discovery across collections: The IMLS Digital Collections and Content project. *Library Hi Tech, 22*(3), 307-322.

Currall, J., Moss, M., & Stuart, S. (2004). What is a collection? *Archivaria*, *58*, 131-146.

Digital Library Federation/National Science Digital Library. (2005). *Best Practices for OAI Data Provider Implementations and Shareable Metadata*. Retrieved October 28, 2008, from http://webservices.itcs.umich.edu/mediawiki/oaibp/?PublicTOC.

Dunn, H. (2000). Collection-level description: the museum perspective. *D-Lib Magazine, 6*(9). Retrieved October 28, 2008, from http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/september00/dunn/09dunn.html.

Foulonneau, M., Cole, T. W., Habing, T. G., & Shreeves, S. L. (2005). 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.

Geisler, G., Giersch, S., McArthur, D., & McClelland, M. (2002). Creating virtual collections in digital libraries: benefits and implementation issues. In *Proceedings of the Joint Conference on Digital Libraries* (Portland, OR, July 14-18, 2002*),* 210-218.

Godby, C. J., Smith, D. & Childress, E. (2003). Two Paths to Interoperable Metadata. *DC-2003: Supporting Communities of Discourse and Practice-Metadata Research & Applications*. Retrieved October 10, 2008, from http://www.oclc.org/research/publications/archive/2003/godby-dc2003.pdf.

Heaney, M. (2000). *An Analytical Model of Collections and Their Catalogues*. Retrieved October 28, 2008, from http://www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf.

Heery, R., & Patel, M. (2000). Application profiles: mixing and matching metadata schemas. *Ariadne*, *25*. Retrieved October 28, 2008, from http://www.ariadne.ac.uk /issue25/app-profiles/intro.html.

Hutt, A., & Riley, J. (2005). Semantics and Syntax of Dublin Core Usage in Open Archives Initiative Data Providers of Cultural Heritage Materials. In *Fifth ACM/IEEE-CS Joint Conference on Digital Libraries 2005* (262-270). New York: ACM Press.

International Council of Museums/CIDOC (2007). *Definition of the CIDOC Conceptual Reference Model: version 4.2.2*, 61. Retrieved October 28, 2008, from http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.2.pdf.

Jackson, A. S., Han, M. J., Groetsch, K., Mustafoff, M., & Cole, T. W. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata, 8 (1).*

Johnston, P. (2003). *Report from Meeting of DC CD WG at DC-2003*. Retrieved October 28, 2008, from http://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0310&L=DC-COLLECTIONS&D=0&I=-3&P=59.

Johnston, P., & Robinson, B. (2002). Collections and collection description. *Collection Description Focus Briefing Paper, 1*. Retrieved October 28, 2008, from http://www.ukoln.ac.uk/cd-focus/briefings/bp1/bp1.pdf.

Knutson, E., Palmer, C.L., and Twidale, M. (2003). Tracking metadata use for digital collections. In *Proceedings of the International DCMI Metadata Conference and Workshop* (Seattle, WA, Sept. 28–Oct. 2, 2003), 243-244.

Lagoze, C., & Fielding, D. (1998). Defining collections in distributed digital libraries. *D-Lib Magazine* (Nov). Retrieved October 28, 2008, from http://webdoc.gwdg.de/edoc/aw/d-lib/dlib/november98/lagoze/11lagoze.html.

Lee, H. (2003). Information spaces and collections: Implications for organization. *Library & Information Science Research, 25*(4), 419-436.

Lee, H. (2005). The concept of collection from the user's perspective. *Library Quarterly, 75*(1), 67-85.

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.

Macgregor, G. (2003). Collection-level descriptions: metadata of the future? *Library Review, 52*(6), 247-250.

Manoff, M. (2000). Hybridity, mutability, multiplicity: Theorizing electronic library collections. *Library Trends, 48*(4), 857-876.

Miller P. (2000). Collected wisdom: some cross-domain issues of collection-level description. *D-Lib Magazine, 6* (Sept.). Retrieved October 28, 2008, from http://www.dlib.org/dlib/september00/miller/09miller.html.

OLAC Cataloging Policy Committee. Summary/Abstracts Task Force (2002). *Summary Notes for Catalog Records.* Retrieved October 10, 2008, from http://ublib.buffalo.edu/libraries/units/cts/olac/capc/summnotes.html.

OSU Knowledge Bank Metadata Application Profile (2006). Retrieved October 20, 2008, from http://library.osu.edu/sites/techservices/KBAppProfile.php#description

Palmer, C. L., Zavalina, O., & Mustafoff, M. (2007). Trends in metadata practices: a longitudinal study of collection federation metadata. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (Vancouver, Canada, June 19-23, 2007).

Palmer, C. L., & Knutson, E. (2004). Metadata practices and implications for federated collections. In *Proceedings of the 67th ASIS&T Annual Meeting* (Providence, RI, Nov. 12-17, 2004).

Palmer, C. L., Knutson, E., Twidale, M, & Zavalina, O. Collection definition in federated digital resource development. In *Proceedings of the 69th ASIS&T Annual Meeting* (Austin, TX, Nov. 3-8, 2006).

Park, J. (2005). Semantic interoperability across digital image collections: A pilot study on metadata mapping. In CAIS/*ACSI 2005 Data, Information, and Knowledge in a Networked World*, edited by Liwen Vaughan. Proceedings of the 2005 annual conference of the Canadian Association for Information Science held with the Congress of the Social Sciences and Humanities of Canada at the University of Western Ontario, London, Ontario, June 2 - 4, 2005. Retrieved October 10, 2008, from http://www.cais-acsi.ca/proceedings/2005/park_J_2005.pdf.

Renear, A. H., Urban, R. J., Wickett, K. M., Palmer, C. L., & Dubin, D. (2008). Sustaining collection value: Managing collection/item metadata relationships. *Proceedings of the Digital Humanities 2008*, 24-29 June 2008, Oulu, Finland. Association for Literary and Linguistics Computing and Association for Computers and Humanities.

Shreeves, S. L., Riley, J., & Milewicz, L. (2006). Moving towards sharable metadata. *First Monday*, *11*(8). Retrieved October 28, 2008, from http://firstmonday.org/issues/issue11_8/shreeves/index.html.

Sweet, M., & Thomas, D. (2000). Archives described at collection level. *D-Lib Magazine, 6*(9). Retrieved October 28, 2008, from http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/september00/sweet/09sweet.html.

Twidale, M., & Urban, R. (2005). *Usability Analysis of the IMLS Digital Collection Registry*. Retrieved October 28, 2008, from http://imlsdcc.grainger.uiuc.edu/3YearReport/docs/UsabilityReport1.pdf.

[1] As our previous research shows, some digital collection developers think about their resources as something other than a project or a collection. The terms applied often related to the background of the developer and the culture of description for their profession. For example, some archivists applied the term archive, and some discussed their archive in relation to "artificial" and "organic" collections, differentiations traditionally used in archival theory and practice. Museum curators frequently spoke of their resources as exhibits, but other developers have also adopted this term, which is evidence of its wider acceptance in the digital library community. The related notions of "display" or "tour" were also applied. To a lesser degree, resources were also discussed as tools and lessons.

[2] The complete list of DCC project team's publications and presentations can be found at http://imlsdcc.grainger.uiuc.edu/about.asp.

[3] http://cicharvest.grainger.uiuc.edu/

[4] http://dublincore.org/groups/collections/collection-application-profile/#coldctermsabstract

[5] http://dublincore.org/documents/dces.

[6] http://dublincore.org/documents/2001/04/12/usageguide/sectb.shtml#description.

[7] Webform for creating collection records in *National Union Catalog of Manuscript Collections* http://www.loc.gov/coll/nucmc/lcforms.html

[8] Described in detail in our five-year report to IMLS available at http://imlsdcc.grainger.uiuc.edu/docs/FinalReport_ResearchMethods.pdf

[9] We have tracked changes made to collection records by resource developers between January 2005 and September 2007. Detailed findings of this longitudinal analysis are presented in five-year project report available at http://imlsdcc.grainger.uiuc.edu/docs/FinalReport_ResearchMethods.pdf

[10] http://dublincore.org/groups/collections/

[11] http://www.ukoln.ac.uk/metadata/rslp/

[12] General overview and detailed description of the IMLS DCC collection description schema are available at: http://imlsdcc.grainger.uiuc.edu/CDschema_overview.asp ; http://imlsdcc.grainger.uiuc.edu/CDschema_elements.asp

[13] Detailed listing is available at http://imlsdcc.grainger.uiuc.edu/CDschema_elements.asp

[14] See http://www.ukoln.ac.uk/metadata/rslp/schema/

[15] DCC collection description metadata schema currently uses dc:creator element in a limited way, to indicate a grant project responsible for creation of digital collection, but overlooks creators of physical items and physical collections.

[16] This problem is mentioned as an example of collection-level metadata patterns revealed by this study; however, the detailed discussion is out of scope of this paper.

[17] A low-to-medium positive association of .420, statistically significant at the 0.01 level, was found between the length of a free-text *Description* field and the number of collection properties indicated in it.

[18] Grant project documentation is available at http://imlsdcc.grainger.uiuc.edu/about.asp.

[19] http://www.isrl.uiuc.edu/~dcc/mdrt.html.