Free-Text Collection-Level Subject Metadata in Large-Scale Digital Libraries: A Comparative Content Analysis

Oksana L. Zavalina
University of North Texas, USA
Oksana.Zavalina@unt.edu

Abstract

Metadata is central for information organization in digital libraries. A growing number of digital libraries worldwide are now generating metadata to describe not only individual objects but entire digital collections as integral wholes. However, collection-level metadata has not yet been empirically evaluated. This paper reports results of the study that used an in-depth comparative content analysis to assess free-text collection-level subject metadata in three large-scale digital cultural heritage aggregations in the United States and Europe. As observed by this study, the emerging best practices include encoding a variety of information about a digital collection in free-text collection-level *Description* metadata element. This includes both subject-specific (topical, geographic and temporal coverage, and types/genres of objects in a digital collection) and non-subject-specific information: title, size, provenance, collection development, copyright, audience, navigation and functionality, language of items in a digital collection, frequency of additions, institutions that host a digital collection or contribute to it, funding sources, item creators, importance, uniqueness, and comprehensiveness of a digital collection.

Keywords: digital libraries; digital aggregations; digital collections; collection-level metadata; subject metadata; free-text metadata; metadata evaluation; metadata quality; content analysis.

1. Introduction and Background

Cultural heritage institutions and funding agencies have invested intensively in digitization projects. Thousands of digital collections produced by numerous digitization projects have made substantial contributions by providing a community broader access to primary materials. Digital aggregations now bring together hundreds of individual digital collections. These aggregations, commonly referred to as digital libraries, operate at the international level (e.g., The European Library, Europeana), national level (e.g., Memory of the Netherlands, New Zealand Digital Library, American Memory, National Science Digital Library, OAIster, Opening History), regional level (e.g., Mountain West Digital Library), or state level (e.g., Texas Heritage Online, Arizona Memory).

Multiple and easily understood access points are essential to the users of digital libraries (Xie, 2006; 2008). Metadata – "structured data about an object that supports functions associated with the designated object" (Greenberg, 2003, p. 1876) – is used in digital libraries to organize information for effective retrieval via search and browse functions. Subject metadata provides important access points to both items and collections as a whole. Metadata is subdivided into two distinct kinds based on how the metadata elements are populated with values: controlled-vocabulary metadata which draws values from formally-maintained lists of terms, and free-text metadata which relies on natural language. In the Dublin Core Collections Application Profile (Dublin Core Metadata Initiative, 2007), which is widely used in digital libraries as a metadata scheme for describing digital collections, the subject metadata is represented by four elements: free-text *Description*, and controlled-vocabulary *Subject*, *Type*, and *Coverage*.

Metadata that describes collections as an integral whole (as opposed to individual items) has a long history. It has been recognized in the archival community as central to facilitating access to documents contained in archival collections (e.g., Bearman, 1992). Collection-level metadata is "a structured, open, standardized and machine-readable form of metadata providing a high-level description of an aggregation of individual items" (Macgregor, 2003, p. 248). It provides an

added level of descriptive granularity: important contextual (Miller, 2000) and relational information (Macgregor, 2003). Such functionality becomes especially important in digital aggregations. Therefore, many digital libraries supply collection-level metadata as means of providing context for the digital items harvested from distributed collections. Moreover, analysis of collection-level metadata is starting to inform collection development policies and efforts in digital libraries (e.g., Palmer, Zavalina, & Fenlon, 2010).

Digital libraries that aggregate metadata from different sources inevitably face problems with metadata consistency, and thus evaluation of metadata, which has not yet become a common practice, gains more and more importance (Hillmann, 2008). At the same time, almost no published research to date has attempted to evaluate collection-level metadata in digital libraries. The only available study (Zavalina, Palmer, Jackson, & Han, 2008) assessed collection-level metadata in a single digital library; that approach limited generalizabilty of its results. Due to lack of generalizable evaluation research results no best practice recommendations exist for how detailed collection-level metadata should be to facilitate collection-level subject access in aggregations of digital collections. To start addressing this research gap, the study presented in this paper sought to examine and compare the free-text collection-level subject metadata (i.e., *Description* metadata element) across multiple digital libraries.

2. Methods

In this study, a combination of qualitative and quantitative content analysis was used for evaluation of free-text collection-level subject metadata in digital libraries. Units of analysis ranged from a phrase or sentence to the entire contents of a free-text *Description* metadata element in collection-level metadata records.

Three large-scale digital cultural heritage aggregations were selected for content analysis: The European Library (http://www.theeuropeanlibrary.org) that aggregates digital collections created by the national libraries in the European Union and neighboring European countries, American Memory (http://memory.loc.gov) developed by the United States Library of Congress, and Opening History (http://imlsdcc.grainger.uiuc.edu/history) developed by the University of Illinois at Urbana-Champaign. At the time of this report (April 2011), these three digital libraries aggregated nearly 1,500 digital collections: 1,089 in Opening History, 199[1] in The European Library, and 140 in American Memory.

A random sample of collection-level metadata records from the three digital libraries was analyzed. The sample included 103 records from American Memory (73.5% of the population of 140), 131 records from The European Library (65.8% of the population of 199), and 488 records from Opening History (44.8% of the population of 1,089). This sample size allows for generalizations with 95% confidence level and 5% margin of error.

The resulting 722 collection-level metadata records were closely examined to determine what kinds of information about the digital collection (hereafter, referred to as collection properties) are included in the free-text *Description* subject metadata element values. The descriptive statistics indicators were measured for the sample of collection metadata records as a whole and for each of the three digital libraries: the average and median number of collection properties encoded in *Description* element and the measures of variability in the number of collection properties (range, variance, and standard deviation). The free-text *Description* element value length (absolute, average, median; range, variance, and standard deviation) was measured. The correlation coefficient (Pearson's r) between the collection-level *Description* element value length and the number of collection properties encoded in it was calculated for each of the three digital libraries.

[1] In addition to digital collections, operationally defined for this study as aggregations of two or more digital objects, The European Library also includes over 40 catalogs and bibliographic databases, which do not contain digital objects per se and therefore were excluded from this analysis.

[2] In particular, the Scope Content element of EAD metadata scheme.

[3] Dublin Core Usage Guide (http://dublincore.org/documents/usageguide/elements.shtml) provides guidelines on

The preliminary list of coding categories used in this content analysis had been developed in an exploratory study of 202 Digital Collections and Content Collection Registry (http://imlsdcc.grainger.uiuc.edu) collection-level metadata records (Zavalina, Palmer, Jackson, & Han, 2008) and had included fourteen collection properties: subjects, object types/genres, creators of items in collection, collection title, size, collection development information, provenance, collection's importance, uniqueness, comprehensiveness, intended audience, navigation and functionality, participating, hosting or contributing institutions, and funding sources. This list was refined in the process of detailed manual content analysis and coding of collection metadata records from the three digital libraries. As a result, the initial "subjects" category was subdivided into three collection properties: topical coverage, geographic coverage, and temporal coverage; three more collection properties were added: copyright information, frequency of additions to collection, and language of items in collection.

A coding manual was developed to aid coders in interpretation of the categories. Intercoder reliability tests were performed on a subset of collection-level metadata records totaling 20% of the main sample. In the pilot study, a subset of 141 Opening History collection-level metadata records was coded by two coders with intercoder reliability of 80.4%. Another sample of 6 metadata records – 2 from each of the three digital libraries under investigation – was coded by eight coders; intercoder reliability constituted 90%.

This study's findings in regards to collection properties encoded in free-text collection-level *Description* metadata element were compared with:

1. available best practice recommendations for *Description* element values in metadata records describing physical collections of manuscripts (National Union Catalog of Manuscript Collections, 2010) and archival materials (OLAC Cataloging Policy Committee, 2002);
2. applicable item-level best practice metadata guidelines for *Description* element derived from sources including Cataloging Cultural Objects (CCO) (Baca et al., 2006), Categories for the Description of Works of Art (CDWA) (Baca et al., 2009), Encoded Archival Description (2002)[2], and OSU Knowledge Bank Metadata Application Profile for Digital Video (Ohio State University Libraries, 2006).[3]

3. Findings and Discussion

Each of the following nineteen collection properties was found in at least one metadata record in the sample: object types/genres, topical, geographic and temporal coverage, creators of items in collection, collection title, size, collection development information, provenance, collection's importance, uniqueness, comprehensiveness, intended audience, navigation and functionality, frequency of additions to collection, hosting, contributing or participating institutions, funding sources, copyright information, and language of items in collection. All nineteen collection properties were found in collection metadata records in the Opening History. American Memory collection-level *Description* metadata elements lacked frequency of additions information, and The European Library collection-level *Description* metadata elements lacked audience information. Across the three aggregations, the average collection-level *Description* metadata element provided information about 6 collection properties. American Memory exhibited the highest average number of collection properties encoded in *Description* element, with between 1 and 12 collection properties (Table 1).

It should be noted that in The European Library, the values in the collection-level *Description* metadata element are presented in 28 European languages. This added level of complexity and resulting practice of shortening values in collection-level metadata elements to simplify

[2] In particular, the Scope Content element of EAD metadata scheme.

[3] Dublin Core Usage Guide (http://dublincore.org/documents/usageguide/elements.shtml) provides guidelines on how to use item-level metadata elements. However, it does not detail what information should be included in *Description*, besides a broad recommendation, "*Description* 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" (Hillmann, 2005).

translation efforts arguably somewhat reduces the richness of values in collection-level *Description* subject metadata elements in The European Library, as demonstrated by lower mean and median numbers of collection properties encoded in free-text *Description* (Table 1). While the average and median *Description* element value length are the lowest in The European Library, the standard deviation is also the lowest, which means the *Description* value length is more consistent in this digital library.

In each of the three digital libraries under investigation, the value length of free-text *Description* was found to have a medium positive correlation with the number of collection properties encoded in this metadata element (Table 1). The highest Pearson r value (.60913) was recorded in the American Memory which had the highest median value length of the *Description* metadata element. This finding suggests that the longer *Description* metadata element values tend to provide richer descriptions of digital collections. American Memory also exhibited the highest average number of collection properties encoded in the *Description* element, with some *Description* elements containing as many as 12 collection properties, which indicates somewhat higher overall richness of free-text *Description* metadata elements in American Memory.

TABLE 1. *Description* metadata element value length and number of collection properties encoded in *Description*

	***Description* element value length**					**Number of collection properties encoded in *Description***					**Length to no. of properties correlation (Pearson r)**
Digital Library	**Range**	**Ave.**	**Med.**	**Var.**	**St. Dev**	**Range**	**Ave,**	**Med.**	**Var.**	**St. Dev**	
American Memory	23-260	97	85	2390	49	1-12	6.58	6	3.30	1.82	.60913
Opening History	5-429	98	83	4861	70	1-11	5.62	6	3.09	1.76	.47125
The European	7-181	39	27	1014	32	1-8	4.63	4	2.39	1.54	.57562

Subject-specific collection properties (types and genres of objects in a digital collection, topical, geographic and temporal coverage) were the most consistently represented in free-text *Description* elements across the three digital libraries.

As seen in Figure 1, object type or genre information was included in *Description* metadata elements the most often: 99% of collection metadata records in American Memory, 89% in Opening History, and 90% in The European Library. Object type terms, such as "physical artifacts," "lanterns, torches, banners," and "cups, vases, trays, bottles, sewing boxes" were common. Genre information was frequently specified, as with "pamphlets, leaflets, and brochures," "songbooks," "political cartoons," and "chronics, letters, annals, official documents."

Topical information was the second most widely represented collection property in the free-text *Description* field. Ninety-seven percent of collection metadata records in American Memory, 96% in Opening History, and 57% in The European Library contained topical information. The content ranged from specific topical coverage statements (e.g., "major topics and issues illustrated include the establishment of the Everglades National Park; the growth of the modern conservation movement and its institutions, including the National Audubon Society; the evolving role of women on the political stage; the treatment of Native Americans; rights of individual citizens or private corporations vs. the public interest; and accountability of government as trustees of public resources, whether for the purposes of development, reclamation, or environmental protection") to broader statements (e.g., "in the fields of culture, education, and academic research") to keywords and noun phrases scattered throughout the text (e.g., "decolonization," "life as a soldier," "American discovery," "drafting and ratification of Constitution," etc.)

Temporal and geographic coverage of a digital collection were the third and fourth most widely represented collection properties in *Description* metadata elements. Indications of temporal coverage in the *Description* were found in 77% of collection metadata records in

American Memory, 65% in Opening History, and 63% in The European Library. These indications ranged from specific dates and date ranges (e.g., "19th century," "covering the period of 1894-1932, with the exception of 1896"), to known historical periods (e.g., "World War I," "California Golden Rush"), to combinations of temporal range and period (e.g., "Lithuanian press ban period, 1864-1904"). Geographic coverage information was found in 81% of collection metadata records in Opening History, 69% in American Memory, and 55% in The European Library. Indications of geographic coverage of varying granularity (e.g., "Austro-Hungarian Empire," "Dutch Indies," "Mayan city of Uxmal in Yucatan, Mexico and a Native American Mississippian site, Angel Mounds U.S.A.") were found in free-text *Description*.

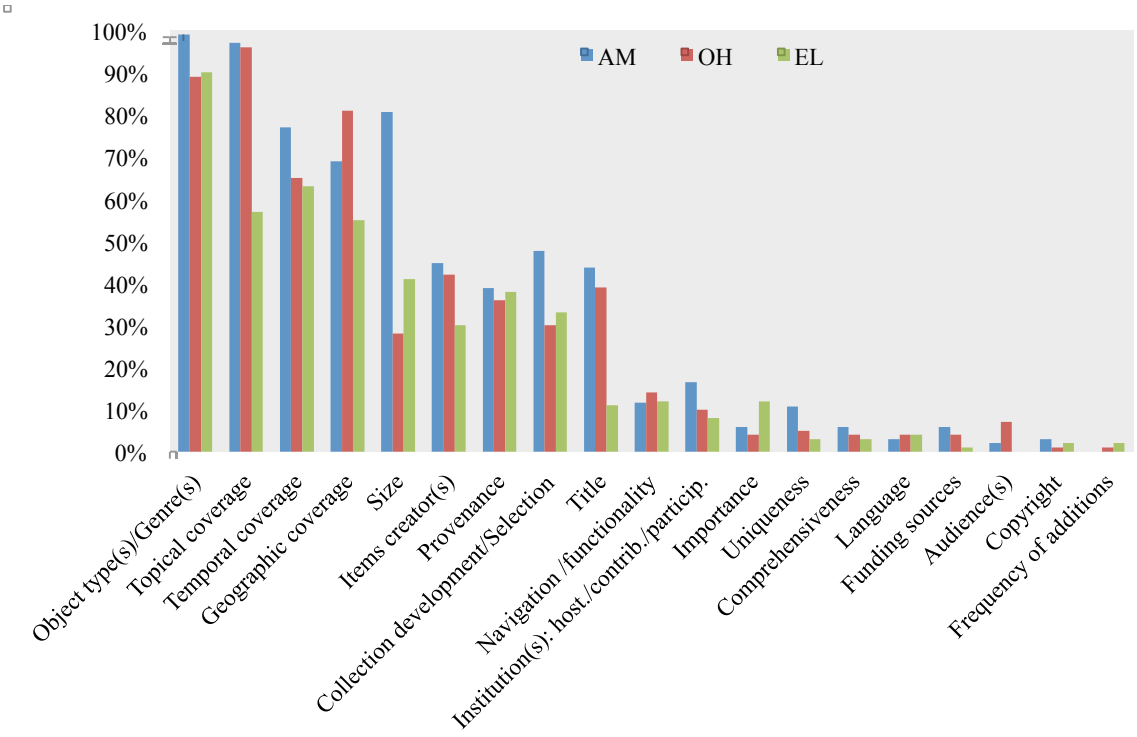

FIG. 1. Distribution of collection properties in *Description*

In addition to the subject-specific information of object type/genre, topical, temporal and geographic coverage, free-text *Description* metadata elements were found to include a variety of other collection properties. Table 2 includes representative examples of these collection properties. Eighty-one percent of collection metadata records in American Memory, 41% in The European Library, and 28% in Opening History contained *Description* metadata element values that made statements about the collection size. Names of artists or institutions that created items in the collection were found in 45% of *Description* metadata elements in American Memory, 42% in Opening History, and 30% in The European Library. Provenance information was included in 39% of *Description* metadata elements in American Memory, 38% in The European Library, and 36% in Opening History. Forty-eight percent of *Description* metadata elements in American Memory, 33% in The European Library, and 30% in Opening History contained more or less specific collection development information, sometimes including information on the purpose or mission of the digital collection. Collection title information was the ninth most often found collection property in free-text *Description* across the three digital libraries (44% in American Memory, 39% in Opening History, and 11% in The European Library). Fourteen percent of *Description* metadata elements in Opening History, 12% in American Memory, and 11% in The European Library contained navigation or functionality information. Seventeen percent of *Description* fields in American Memory, 10% in Opening History, and 8% in The

European Library provided information about one or more institutions hosting the digital collection, participating in the digitization project, or contributing items to digitize. Indications of digital collection's importance were found in 12% of *Description* metadata elements in The European Library, 6% in American Memory, and 4% in Opening History. Indications of uniqueness of the content of a digital collection were found in 11% of *Description* metadata elements in American Memory, 5% in Opening History, and 3% in The European Library. Indications of digital collection's comprehensiveness were found in 6% of *Description* metadata elements in American Memory, 4% in Opening History, and 3% in The European Library. Language of items in a digital collection was mentioned in almost 6% of *Description* metadata elements overall (8% in American Memory, 4% in Opening History, and 4% in The European Library). Between 1% and 6% of *Description* metadata elements (6% in American Memory, 4% in Opening History, and 1% in Opening History) acknowledged funding sources – public or corporate – that helped build the digital collection. Information about the intended audience of a digital collection was found in 7% of *Description* metadata elements 7% in Opening History and 2% in American Memory, while no such indications were found in the sample of collection metadata records from The European Library. Finally, the information about copyright and frequency of additions to the digital collection was included the least often across the three aggregations. Copyright information was found in only 3% of *Description* metadata elements in American Memory, 2% in The European Library, and 1% in Opening History). Indications of the frequency of additions to a digital collection were found in only 2% of *Description* metadata elements in The European Library and 1% in Opening History, while no such indications were found in the sample of collection metadata records from the American Memory.

TABLE 2: Representative examples of non-subject-specific collection properties in *Description*

Properties	Examples
collection size	"hundreds of personal letters, diaries, photos, and maps" "more than 70,000 volumes of digitized texts, 80,000 still images, and 30 hours of sound recordings"
collection title	"The 1936 Gainesville Tornado: Disaster and Recovery" "Warsaw in Words and Images"
collection development	"a sample of the photographic archives" "a selection of framed items from the collections of the ... Library" "effort has been made to offer a balanced number of items for each inaugural event" "to inventory and to describe the decoration of the manuscripts held in the Bibliothèque Nationale de France" "titles published between 1850 and 1950 were selected and ranked by teams of scholars" "to stimulate the documentation and preservation of ethnic materials and foster a greater interest in the history and cultures of the peoples of the region"
item creators	"Among the authors represented are Frederick Douglass, Booker T. Washington, Ida B. Wells-Barnett, Benjamin W. Arnett, Alexander Crummel, and Emanuel Love" "monasteries of Mount Athos: Chilandar, Vatoped, Simonopetra and Kutlumush"
provenance	"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" "documents belonging to the collection of the Army Museum" "selected from various Library of Congress holdings"
hosting/contributing/ institutions	"Archives Department provides access to the digitized Roman Catholic Church registers of birth, marriage and death (1599-1907). The Art Museum presents digital images" "project brings Tufts, and the Virginia Center for Digital History together with the University to build a digital repository"
funding sources	"digitized as the result of an Illinois State Library FY98 Educate and Automate grant" "funded by Reuters America, Inc., and The Reuters Foundation" "funds provided by the Institute of Museum and Library Services, under the federal Library Services and Technology Act" "made possible by a major gift from Citigroup Foundation" "made possible through the generous support of the AT&T Foundation"

TABLE 2 (Cont.): Representative examples of non-subject-specific collection properties in *Description*

Properties	Examples
navigation and functionality	"accessed by the scanned county photomosaic or line indexes" "accessible by date of issue or by keyword searching" "allows the user to browse the highlights thematically or by number" "arranged chronologically by Japanese periods" "grouped by county" "may be searched or browsed in a variety of ways, including by keyword, subject, creator, title, and date" "organized according to seven major categories" "overall organization of the database is by tribe" "the indexes for all categories are searched simultaneously"
uniqueness	"rare historic published monographs and serials" "rare and unique library and archival resources" "sources that are rare, unusual, out-of-print, or difficult, if not impossible, to access" "unique historical treasures from ... archives, libraries, museums, and other repositories"
importance	"an archive of unparalleled importance" "collection of the most important and influential 19th and early 20th century American cookbooks" "important books, government documents, manuscripts, maps, musical scores, plays, films, and recordings" "materials are significant in their place within the fabric of American history and culture" "the most outstanding representatives of Yiddish literature"
comprehensiveness	"a rich diversity of materials" "a comprehensive and integrated collection of sources and resources on the history and topography of London" "almost complete collection of Norwegian printed newspapers" "one of the most ambitious and comprehensive effort to date to deliver educational content on the Civil Rights Movement" "the most comprehensive library of manuscripts" "such a large body of materials presents a full spectrum of representation and opinion"
language	"English- and Yiddish-language playscripts" "entirely printed in Latin" "European, Slavic, Middle Eastern, and English- and Spanish-language folk music" "many of the publications are in Vietnamese"
audience	"Alabama residents and students, researchers, and the general public in other states and countries" "middle and high school students" "schoolchildren, genealogists, historians, authors, producers, and special interest groups" "those studying political reorganization in Georgia and the growth of Atlanta as well as the Civil Rights Movement, the Cold War, the Vietnam conflict, Middle East tensions, and Watergate"
copyright	"historical sheet music registered for copyright" "materials are royalty-free and available free of charge" "materials with expired copyrights" "restricted to items that are not covered by copyright protection"
frequency of additions	"annual growth is ca. 700 publications" "regular additions to the collection are expected" "some 10,000 volumes per year"

Differences, sometimes significant, in the frequency of occurrence of certain collection properties in the collection-level *Description* metadata elements were observed among the three digital libraries. Overall, 13 out of 19 collection properties were found more often in American Memory than in the two other digital libraries, with the most pronounced difference in uniqueness (2.14 times more compared to the aggregation with the second highest rate of occurrence of this collection property), size (1.97 times more compared to the digital library with the second highest rate of occurrence), and hosting/contributing/participating institutions (1.65 times more compared to the digital library with the second highest rate of occurrence). Geographic coverage, navigation and functionality, and audience were the three collection properties found more often in Opening History; the most significant difference was observed in the case of audience (3.5 times more

compared to the digital library with the second highest rate of occurrence of this collection property). Two collection properties – importance and frequency of additions – occurred significantly more often in The European Library *Description* elements (2.07 times and 2.0 times more compared to the digital library with the second highest rate of occurrence). Indication of language(s) of items a digital collection were found equally often in Opening History and The European Library and less often in American Memory.

Although more research is needed into digital library developers' decisions around collection-level *Description* element, it is obvious that the differences identified above might be explained by the specifics of the policies followed, the tools used in describing digital collections in the three digital libraries, and the collection development approaches. For example, the fact that only free-text *Description* is displayed to the end-user in American Memory might be influencing the decisions on how rich *Description* metadata element values should be in this digital library which results in longer and richer *Description* values. More consistent indication of uniqueness and comprehensiveness of a digital collection in the *Description* may be due to American Memory's collection development policy, which emphasizes digitizing collections of unique materials and great educational value (Arms, 1996). Wider encoding of geographic coverage information in Opening History *Description* metadata element might be due to the focus on local history in Opening History collection development policy (Opening History, 2009).

Comparison with existing best practice recommendations for the content of *Description* metadata element (Table 3) makes it clear that while meeting most of the best practice recommendations, collection-level *Description* metadata elements in Opening History, American Memory, and The European Library also routinely include 7 additional kinds of information about digital collections that are not covered by any of available recommendations: comprehensiveness, copyright, frequency of additions, funding sources, hosting/contributing/participating institutions, size, and title. Encoding these additional collection properties in *Description* metadata elements might be considered an emerging best practice that has yet to be reflected in the best practice documents.

TABLE 3: *Description* element: mapping the findings to existing best practice guidelines

	Existing guidelines regarding information to be included in free-text *Description*					**Findings of this study**
	Item-level metadata guidelines that are applicable to collections			**Collection-level metadata guidelines**		
META GUIDLINES	***EAD [Scope Content element]***	***CCO/CDWA***	***OSU Knowledge Bank Metadata Application Profile***	***National Union Catalog of Manuscript Collections*** http://www.loc.gov/coll/nucmc/lcforms.html	***OLAC Summary Notes for Catalog Records***	**collection properties encoded in *Description* elements**
	SUBJECT-SPECIFIC INFORMATION					
	Form and arrangement of materials	N/A	N/A	Types of materials included in the collection	Specific types and forms of materials present	Object types/Genres
	Significant organizations, individuals, subjects represented	Subject	N/A	Topics with which the materials in the collection deal	Significant people and topics covered	Topical coverage
	Places represented	N/A	N/A	Geographical areas, with which the materials in the collection deal	Significant places covered	Geographic coverage
	Events represented	N/A	N/A	Associated dates, events, and historical periods dealt with by the materials in the collection	Significant events covered, span of dates covered by the collection	Temporal coverage

TABLE 3 (Cont.): *Description* element: mapping the findings to existing best practice guidelines

	NON-SUBJECT-SPECIFIC INFORMATION					
META GUIDELINES (Cont.)	N/A	N/A	Provenance, history of the work	N/A	History of the work	Provenance
	Strengths	Significance	N/A	N/A	N/A	Importance
	N/A	N/A	N/A	N/A	Unique characteristics of the collection	Uniqueness
	Significant organizations, individuals represented	N/A	N/A	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	N/A	Item creator(s)
	N/A	Function	N/A	N/A	Reason and function of the collection	Collection development
	N/A	N/A	Nature of the language of the resource	N/A	N/A	Language
	N/A	N/A	N/A	N/A	Audience	Audience
	N/A	N/A	N/A	N/A	User interaction	Navigation and functionality
EMERGING PRACTICE	N/A	N/A	N/A	N/A	N/A	—Comprehen-siveness —Copyright —Frequency of additions —Funding sources —Hosting / participating / contributing institution —Size —Title

Although more research is needed into digital library developers' decisions around collection-level *Description* element, it is obvious that the differences identified above might be explained by the specifics of the policies followed, the tools used in describing digital collections in the three digital libraries, and the collection development approaches. For example, the fact that only free-text *Description* is displayed to the end-user in American Memory might be influencing the decisions on how rich *Description* metadata element values should be in this digital library which results in longer and richer *Description* values. More consistent indication of uniqueness and comprehensiveness of a digital collection in the *Description* may be due to American Memory's collection development policy, which emphasizes digitizing collections of unique materials and great educational value (Arms, 1996). Wider encoding of geographic coverage information in Opening History *Description* metadata element might be due to the focus on local history in Opening History collection development policy (Opening History, 2009).

Comparison with existing best practice recommendations for the content of *Description* metadata element (Table 3) makes it clear that while meeting most of the best practice recommendations, collection-level *Description* metadata elements in Opening History, American Memory, and The European Library also routinely include 7 additional kinds of information about digital collections that are not covered by any of available recommendations: comprehensiveness, copyright, frequency of additions, funding sources, hosting/contributing/participating institutions, size, and title. Encoding these additional collection

properties in *Description* metadata elements might be considered an emerging best practice that has yet to be reflected in the best practice documents.

4. Conclusions

Best practice recommendations for creating rich collection-level subject metadata are needed. These guidelines can be incorporated in the *Framework of Guidance for Building Good Digital Collections* NISO Recommended Practice document (NISO Framework Working Group, 2007) or *IFLA Guidelines for Digital Libraries* that are currently under development. The findings of this study with respect to the emerging best practices in application of free-text collection-level subject metadata could be instrumental in developing these recommendations.

This exploratory research focused on free-text collection metadata practices in national- and international-level digital libraries of one type – aggregations of cultural heritage digital collections that are created for humanities and social sciences scholars, educators, and enthusiasts. Additionally, a comparative study of controlled-vocabulary subject metadata in the same three large-scale cultural heritage digital libraries and relations between the values encoded in free-text and controlled-vocabulary metadata is currently underway. The task of developing best practice guidelines warrants analysis of metadata in digital libraries that have a different subject focus (e.g., science and technology as in the United States National Science Digital Library) and scale (e.g., state-level digital libraries such as Texas Heritage Online or regional-level digital libraries such as Mountain West Digital Library). A combination of multiple obtrusive and unobtrusive research methods (e.g., content analysis, transaction log analysis, survey, interview, observation) in a larger study will allow researchers not only to compare patterns of application of collection-level subject metadata in a representative sample of digital libraries of varying subject focus and scale, but also:

- to understand how decisions about collection-level subject metadata (e.g., regarding the subject metadata elements to be used, the suggested length of subject metadata element values, the collection properties to be represented in subject metadata element values, the controlled vocabularies, etc.) are made,
- to observe patterns of user interactions with digital libraries and user engagement with collection-level metadata, and
- to determine how collection-level subject metadata assists end-users in their information seeking in digital libraries.

Acknowledgements

I wish to thank the research and implementation team of the IMLS NLG Research and Demonstration grant-funded Digital Collections and Content project at the University of Illinois at Urbana-Champaign (particularly Dr. Carole L. Palmer, Dr. Allen Renear, Amy Jackson, and Myung-Ja Han) for providing Opening History collection metadata for analysis, help with data coding, and valuable feedback. I also wish to acknowledge contributions of Sally Chambers and Christa Maher, who provided The European Library and American Memory metadata for analysis and answered my questions about free-text collection-level metadata in these digital libraries.

References

Arms, Caroline R. (1996). Historical collections for the National Digital Library: Lessons and challenges at the Library of Congress. *D-Lib Magazine*, April 1996; May 1996. Retrieved April 30, 2011, from http://www.dlib.org/dlib/april96/loc/04c-arms.html ; http://www.dlib.org/dlib/may96/loc/05c-arms.html.

Baca, Murtha, and Patricia Harpring (Eds.), (2009). *Categories for the Description of Works of Art (CDWA)*. Retrieved April 30, 2011, from http://www.getty.edu/research/publications/electronic_publications/cdwa/index.html.

Baca, Murtha, Patricia Harpring, Elisa Lanzi, Linda McRae, and Ann Baird Whiteside. (2006). *Cataloging Cultural Objects: A Guide to Describing Cultural Works and their Images.* Chicago: American Library Association.

Bearman, David. (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*, (pp. 140-149).

Dublin Core Metadata Initiative. (2007). *Dublin Core Collections Application Profile*. Retrieved April 30, 2011, from http://dublincore.org/groups/collections/collection-application-profile.

Encoded Archival Description. (2002). Retrieved April 30, 2011, from http://www.loc.gov/ead/.

Greenberg, Jane. (2003). Metadata and the World Wide Web. In *Encyclopedia of Library and Information Science,* pp. 1876-1888. New York: Marcel Dekker.

Hillmann, Diane I. (2008). Metadata quality: from evaluation to augmentation. *Cataloging & Classification Quarterly*, *46*(1), 65-80. Retrieved August 12, 2011, from http://hdl.handle.net/1813/7899.

Hillmann, Diane I. (2005). *Using Dublin Core: The Elements*. Dublin Core Metadata Initiative. Retrieved April 30, 2011, from http://dublincore.org/documents/usageguide/elements.shtml.

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

Miller, Paul. (2000, Sept.). Collected wisdom: some cross-domain issues of collection-level description. *D-Lib Magazine, 6.* Retrieved April 30, 2011, from http://www.dlib.org/dlib/september00/miller/09miller.html.

National Union Catalog of Manuscript Collections. (2011). *Online Data Sheet for Participating Institutions*. Retrieved April 30, 2011, from http://www.loc.gov/coll/nucmc/lcforms.html.

NISO Framework Working Group. (2007). *A Framework of Guidance for Building Good Digital Collections*. (3rd ed.). Bethesda, MD: National Information Standards Organization. Retrieved April 30, 2011, from http://www.niso.org/publications/rp/framework3.pdf.

OLAC Cataloging Policy Committee, Summary/Abstracts Task Force. (2002). *Summary Notes for Catalog Records*. Retrieved April 30, 2011, from http://www.olacinc.org/drupal/?q=node/21.

Ohio State University Libraries. (2006). *OSU Knowledge Bank Metadata Application Profile for Digital Video*. Retrieved April 30, 2011, from http://library.osu.edu/staff/techservices/KBAppProfileDV.php.

Opening History. (2009). *Opening History (OH) Aggregation Collection Development Policy*. Retrieved April 30, 2011, from http://imlsdcc.grainger.uiuc.edu/docs/CollectionDevelopmentPolicy.pdf.

Palmer, Carole L., Oksana L. Zavalina, and Katrina Fenlon. (2010). Beyond size and search: building contextual mass in digital aggregations for scholarly use. *Proceedings of the American Society for Information Science and Technology, 47*, 1, 1-10.

Xie, Hong (Iris). (2006). Evaluation of digital libraries: Criteria and problems from users' perspectives. *Library & Information Science Research, 28*(3), 433-452.

Xie, Hong (Iris). (2008). Users' evaluation of digital libraries (DLs): their uses, their criteria, and their assessment. *Information Processing & Management, 44*(3), 1346-1373.

Zavalina, Oksana L., Carole L. Palmer, Amy S. Jackson, and Myung-Ja Han. (2008). Evaluating descriptive richness in collection-level metadata. *Journal of Library Metadata*, *8*(4), 263-292.