Exploring the richness of collection-level subject metadata in three large-scale digital libraries

Oksana L. Zavalina

Department of Library and Information Sciences,
College of Information,
University of North Texas,
1155 Union Circle 311068 Denton, TX 76203-5017, USA
Email: Oksana.Zavalina@unt.edu

Abstract: A growing number of digital libraries worldwide are now generating collection-level metadata to describe entire digital collections as integral wholes. This paper reports results of the study that used an in-depth comparative content analysis to assess free-text and controlled-vocabulary collection-level subject metadata in three large-scale digital libraries in the European Union and the USA. As observed by this study, the emerging best practices in creating collection-level subject metadata include: (a) describing collection's subject matter with mutually complementary values in controlled-vocabulary and free-text metadata; (b) encoding a variety of collection properties in free-text metadata, including both subject properties (topical, geographic and temporal coverage, and types/genres of objects) and non-subject 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 collection or contribute to it, funding sources, item creators, importance, uniqueness and comprehensiveness of a digital collection.

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

Reference to this paper should be made as follows: Zavalina, O.L. (XXXX) 'Exploring the richness of collection-level subject metadata in three large-scale digital libraries', *Int. J. Metadata, Semantics and Ontologies*, Vol. X, No. Y, pp.xxx–xxx.

Biographical notes: Oksana L. Zavalina obtained her Library and Information Science degrees from the Kyiv National University of Culture and Arts (Ukraine) and the University of Illinois (USA). She worked as a bibliographer at the National Parliamentary Library of Ukraine, as a library director at the Kyiv School of Economics, as a data analyst and a cataloguer at the University of Illinois. In 2010, she joined the faculty of the College of Information at the University of North Texas. Her research interests focus on the role of collection-level metadata in semantic access to information contained in the digital libraries.

This paper is a revised and expanded version of the paper entitled 'Free-text collection-level subject metadata in large-scale digital libraries' presented at 'the International Conference on Dublin Core and Metadata Applications in the Hague', September 2011, Netherlands.

1 Introduction and background

Cultural heritage institutions, with support of funding agencies, have invested intensively in digitisation projects. Thousands of digital collections produced by numerous digitisation projects worldwide have substantially broadened user access to materials of cultural, historical and educational value. Digital aggregations now bring together hundreds, sometimes thousands, 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, 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; Xie, 2008). Metadata – "structured data about an object that supports functions associated with the designated object" (Greenberg, 2003, p.1876) – is used to organise information in digital libraries 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*. The latter is further subdivided into geographical and temporal coverage.

Metadata that describes collection as an integral whole (as opposed to individual items) has a long history. It has been recognised in 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. However, virtually no research to date has evaluated and compared the collection-level metadata in digital aggregations.

In discussions of metadata, the terms 'richness', 'detailed description', 'level of description' or 'quality' of metadata seem to be used interchangeably (e.g. Arms, 1996; Duval et al., 2002). While a variety of metadata quality criteria have been suggested, the three most widely accepted criteria are metadata accuracy, consistency and completeness (Park, 2009; Park and Tosaka, 2010).

Metadata accuracy is measured as the degree to which the metadata values match characteristics of the described object (e.g. Stvilia et al., 2007). For example, if the object was originally published in 1912 and digitised in 2012, which of the two dates is included as a data value in *Date Created* metadata element? If the object is about Georgian Republic, a country in Eastern Europe, commonly known as Georgia, does the *Geographic Coverage* metadata element contain the correct value or does it mislead the user into thinking that the object deals with Georgia State in the USA?

Metadata consistency is another important metadata quality criterion which is further subdivided into semantic and structural consistency (Park, 2009). Semantic consistency refers to an extent to which the same values or elements are used for representing similar concepts (Bruce and Hillmann, 2004). For example, does the *Language* metadata element in the same digital library contain the value 'Deutsch' in some records and the value 'German' in others? Does the name of the journal in which the article was published appear consistently in *Relation* metadata element, or do some records use *Source* element for this same information? Do the values in *Geographic* Coverage element represent only the places that the information object is about or do the places of publication/production of information object also appear as data values in this element? Structural consistency is evaluated as a degree to which the same structure is followed in representing information in certain metadata elements (Bruce and Hillmann, 2004). For example, do all the records in the digital library encode dates in the same format, or do they utilise a variety of data entry formats (including YYYY-MM-DD, DD-MM-YYYY, DD-MM-YY, MM-DD-YYYY and MM-DD-YY)?

Metadata completeness – the third most important metadata quality criterion (Park, 2009) – is evaluated as an extent to which objects are described using all applicable metadata elements to their full access capacity. Some of the assessment criteria used to evaluate metadata completeness (Moen et al., 1998) include the number of metadata elements per record, practice of presenting blank (i.e. non-populated but displayed) metadata elements, utilisation and selected characteristics of mandatory and optional elements.

Large-scale digital libraries that aggregate metadata from different sources inevitably face problems with metadata quality, and thus evaluation of metadata gains more and more importance (Hillmann, 2008). Yet almost no research to date has attempted to evaluate collection-level metadata. Zavalina et al.'s (2008) study started addressing this research gap by assessing collection-level metadata in the Digital Collections and Content registry of IMLS-funded digital collections. However, because that study focused on a single-digital library, generalisabilty of its results is limited. More recently, Zavalina's (2011a) study examined consistency of application of controlled-vocabulary collection-level subject metadata elements in several digital libraries and the role this metadata played in information retrieval in one of them. The study reported in this paper complements Zavalina's (2011a) study, and the current paper is substantial revision and expanded version of Zavalina, 2011b. In addition to detailed comparative analysis of free-text collection-level metadata in the three large-scale digital libraries in the USA and the European Union, this paper presents results of a comparative analysis of the values in free-text *Description* and four controlled-vocabulary subject metadata fields in the same three large-scale 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 metadata element in collection-level metadata records.

Three large-scale digital libraries were selected for analysis: The European Library (see http://www.theeuropeanlibrary.org) that aggregates cultural heritage digital collections created by the national libraries in the European Union and neighbouring 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. All three of these digital libraries collect cultural heritage materials, including resources covering various aspects of the US history (Opening History and American Memory) and the history of the member states of the European Union (The European Library). Owing to the very similar nature of the two US-based digital libraries, certain digital collections are included in both of them.

At the time of this paper submission, these three digital libraries aggregated over 1700 digital collections: 1450 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 analysed. 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 (33.1% of the population of 1450). This sample size allows for generalisations 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 data values. The descriptive statistics indicators were measured for the sample of collection-level 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 data value length (absolute, average, median; range, variance and standard deviation) was measured. The correlation coefficient (Pearson's r) between the collection-level description element data value length and the number of collection properties encoded in it was calculated for each of the three digital libraries.

The preliminary list of coding categories used in the 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 et al., 2008) and had included 14 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-level 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; and 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 totalling 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 six metadata records – two 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 data values in metadata records describing physical collections of manuscripts (National Union Catalog of Manuscript Collections, 2011) 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) (Visual Resources Association, 2006), Categories for the *Description* of Works of Art (CDWA) (Baca and Harpring, 2009), Encoded Archival Description (2002),[2] and OSU Knowledge Bank Metadata Application Profile for Digital Video (Ohio State University Libraries, 2006).[3]

A subset of collection-level metadata records in the sample was further analysed: 39 records from American Memory, 33 records from Opening History, and 27 records from The European Library. The resulting 99 collection-level metadata records were closely examined, qualitatively and quantitatively, to determine how the data values in different collection-level subject metadata elements within a record relate to each other (e.g. one-way or two-way complementarity, redundancy, etc.).

3 Findings and discussion

3.1 Collection properties in free-text collection-level subject metadata

Each of the following 19 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 19 collection properties were found in collection-level 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 six 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 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 data 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.

Table 1 *Description* metadata element data 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*	*Average*	*Median*	*Variance*	*St. Dev.*	*Range*	*Average*	*Median*	*Variance*	*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 Library	7–181	39	27	1014	32	1–8	4.63	4	2.39	1.54	.57562

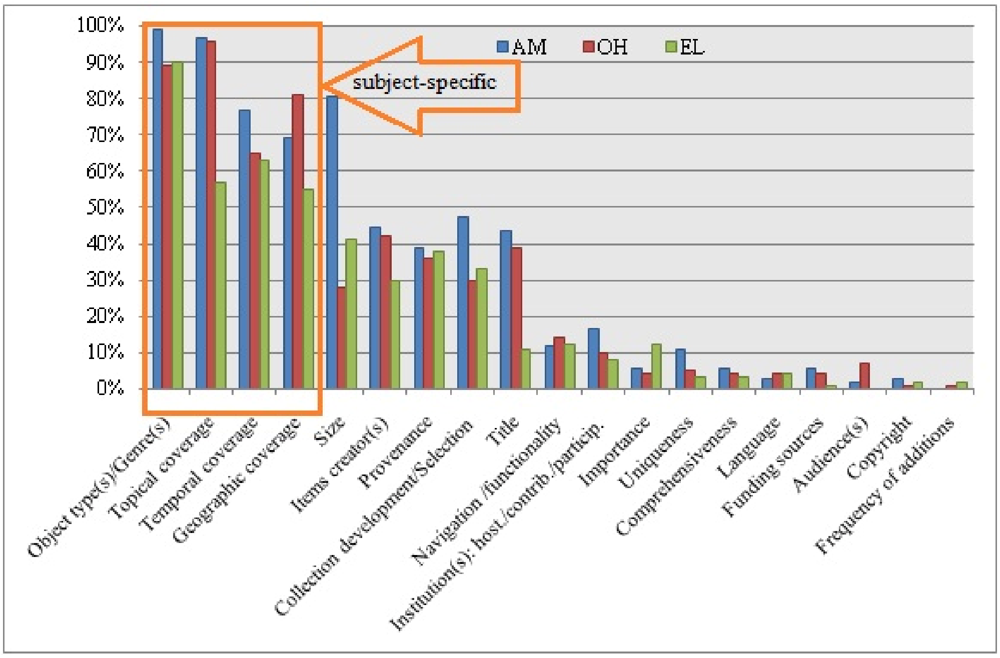
In each of the three digital libraries under investigation, the value length of free-text *Description* was found to have a medium positive correlation with the number of collection properties encoded in this metadata element (Table 1). The highest Pearson's r value (.60913) was recorded in the American Memory which had the highest median value length of the *Description* metadata element. This finding suggests that the longer *Description* metadata element data 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.

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 and/or genre information was included in *Description* metadata elements the most often. Object type terms, such as 'physical artefacts', '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. 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. 'decolonisation', 'life as a soldier', 'American discovery', 'drafting and ratification of Constitution', etc.)

Figure 1 Distribution of collection properties in *Description* (% of records) (see online version for colours)



Temporal and geographic coverage of a digital collection were the third and fourth most widely represented collection properties in *Description* metadata elements. Temporal coverage 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'). Some representative examples of geographic coverage information include 'Austro-Hungarian Empire', 'Dutch Indies'.

In addition to the subject-specific information (i.e. object type/genre, topical, temporal and geographic coverage), free-text *Description* metadata elements were found to include a variety of other collection properties: collection size, collection title, collection development policy, item creators, provenance, hosting and/or contributing institutions, navigation and functionality, funding sources, uniqueness, importance, comprehensiveness, language, audience, copyright and frequency of additions. Table 2 includes representative examples of these collection properties.

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 digitised 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/ participating institutions	'Archives Department provides access to the digitised 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'
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' 'organised according to seven major categories' 'overall organisation of the database is by tribe' 'the indexes for all categories are searched simultaneously'
Funding sources	'digitised 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'
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'

Table 2 Representative examples of non-subject-specific collection properties in *Description* (continued)

Properties	*Examples*
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 reorganisation 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 (Table 3). 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 digital library 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 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 data 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 emphasises digitising 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 of this study's findings with five existing best practice recommendations for the content of *Description* metadata element in either collection-level or item-level metadata made it clear that collection-level description metadata elements in Opening History, American Memory and The European Library meet the guidelines for high quality metadata. National Union Catalog of Manuscript Collections (2011) suggests that collection-level metadata creators for manuscript collections provide in the *Description* element information about types of materials included in the collection; topics with which the materials in the collection deal; geographical areas with which the materials in the collection deal; associated dates, events and historical periods dealt with by the materials in the collection; names, dates and biographical identification of persons and names of corporate bodies significant (by quality and/or quantity of material) to the collection, and specific phases of career/activity of the major person/body responsible. Summary Notes for Catalog Records (OLAC Cataloging Policy Committee, 2002) recommend inclusion of the information about specific types and forms of materials present, significant people and topics covered, significant places covered, significant events covered, span of dates covered by the collection, history of the work, unique characteristics of the collection, reason and function of the collection, audience and user interaction. Encoded Archival Description (2002) recommends inclusion of such characteristics as form and arrangement of materials; significant subjects represented; places represented; events represented; significant organisations and individuals represented; and collection strengths. CCO (Visual Resources Association, 2006) and CDWA (Baca and Harpring, 2009) suggest recording information about subject, significance and function in item-level free-text *Description* element. OSU Knowledge Bank Metadata Application Profile for Digital Video (Ohio State University Libraries, 2006) recommends inclusion of provenance and history of the work, as well as the nature of the language of the resource.

As demonstrated by the data collected and analysed in this study, all of the best practice recommendations for *Description* metadata element identified above have been met by the free-text collection-level description element in the three large-scale digital libraries whose metadata was analysed in this study. In addition, collection-level description metadata elements in Opening History, American Memory, and The European Library include seven 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.

3.2 Complementarity between free-text and controlled-vocabulary collection-level subject metadata

A significant proportion of collection-level metadata records in the sample included cases of one-way complementarity, when information in one collection-level subject metadata element complemented information in one or more other metadata elements, by providing additional details absent elsewhere. The highest occurrence of one-way complementarity between collection-level subject metadata elements was observed in Opening History. In 76% of collection-level metadata records analysed in this study it was the free-text *Description* metadata element that complemented information found in one or more of the controlled-vocabulary subject metadata elements: *Subjects, Geographic Coverage*, *Temporal Coverage* and *Objects*.

As seen in Figure 2 and Table 4, free-text *Description* metadata element data values most often complemented topical information found in the *Subjects* element. Representative examples include: ‘Spanish cartographer, … history, urbanism, public works and agriculture from a strictly geographic point of view’ in *Description* vs. ‘900 History and geography, 911 Historical geography’ in *Subjects*; ‘interior design, … homes of US presidents’ in *Description*, with these topics not mentioned in *Subjects*; ‘early developments in the National Park, … landscape and park facilities’ in *Description* vs. ‘Great Basin, Social studies, State history’ in *Subjects*.

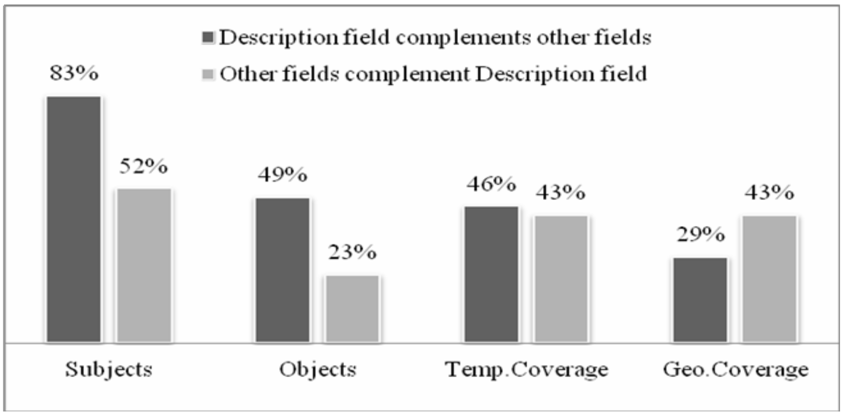
Table 3 Comparative frequencies of occurrence of collection properties in *Description*

Collection property	*American Memory (% of metadata records)*	*Opening History (% of metadata records)*	*The European Library (% of metadata records)*
Audience	2	7	–
Collection title	44	39	11
Collection development	48	30	33
Collection size	81	28	41
Comprehensiveness	6	4	3
Copyright	3	1	2
Frequency of additions	–	1	2
Funding sources	6	4	1
Geographical coverage	69	81	55
Hosting/contributing/ participating institutions	17	10	8
Importance	6	4	12
Item creators	45	42	30
Language	8	4	4
Navigation and functionality	12	14	11
Object type/genre	99	89	90
Provenance	39	38	36
Temporal coverage	77	65	63
Topical coverage	97	96	57
Uniqueness	11	5	3

Table 4 Comparative frequencies of occurrence of one-way complementarity

Free-text Description metadata element complements information in controlled-vocabulary metadata element	*American Memory (% of metadata records)*	*Opening History (% of metadata records)*	*The European Library (% of metadata records)*
Geographic Coverage	19	39	33
Objects	70	30	44
Subjects	86	76	70
Temporal Coverage	51	67	15
Multiple controlled-vocabulary metadata elements	11	45	10
Controlled-vocabulary metadata element complements information in free-text Description metadata element	*American Memory (% of metadata records)*	*Opening History (% of metadata records)*	*The European Library (% of metadata records)*
Geographic Coverage	24	55	56
Objects	14	52	–
Subjects	30	70	60
Temporal Coverage	3	67	72

Figure 2 Complementarity of collection-level subject metadata (% of records overall)



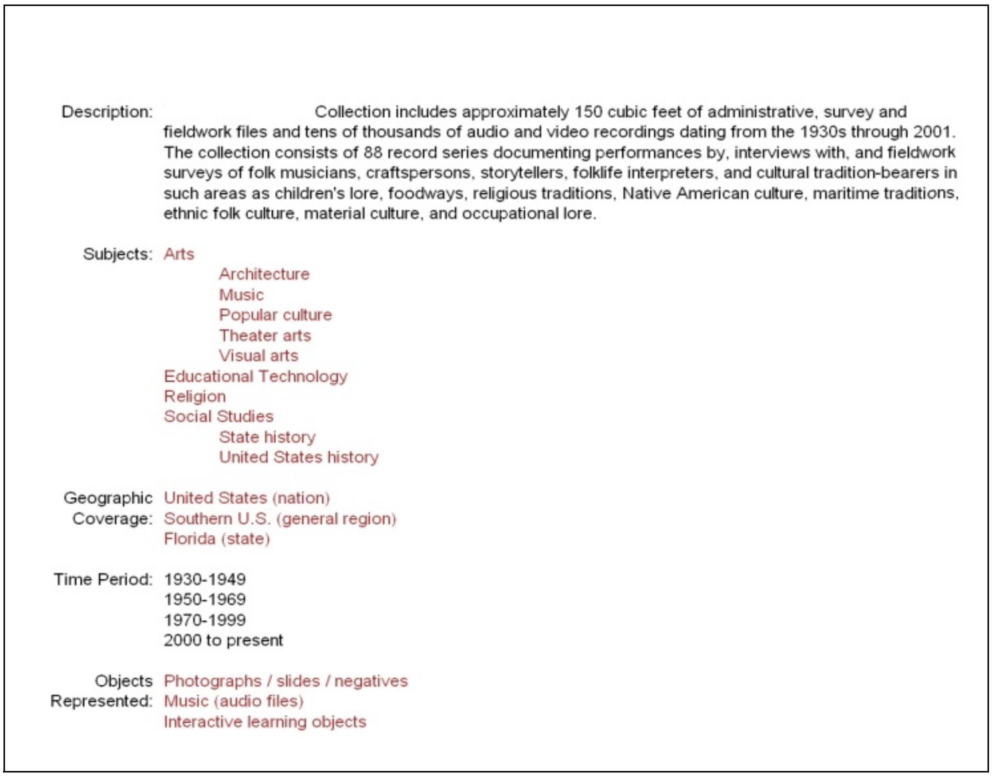
Objects metadata element was the second most often complemented by object-, type- or genre-specific information in *Description* field. Representative examples included: ‘uniform books, ego documents, photographs and sketches’ in *Description* vs. ‘images’ in *Objects*; ‘digital pre-print originals and online publications’ in *Description* while *Objects* field was missing; ‘historical photographs, … portraits, … aerial shots’ in *Description* vs. ‘photographs/slides/negatives’ in *Objects*; ‘rare books, government documents, manuscripts, maps, musical scores, plays, films and recordings’ in *Description* vs. ‘software, multimedia’ in *Objects*.

Temporal Coverage metadata element was also often complemented by *Description*. Representative examples included: ‘16th century, 17th century, 18th century, 19th century, 20th century’ in *Temporal Coverage* vs. ‘Since the Eighty Years’ War’ in *Description*; ‘from 1895 to 1920s’ in *Description* vs. ‘1850 to 1899, 1900 to 1929’ in *Temporal Coverage* field.

Geographic Coverage was complemented by *Description* metadata element less often overall. Representative examples included: ‘Hispanic America … Spanish territories in America and Oceania’ in *Description* vs. ‘Hispanic America’ in *Geographic Coverage*; ‘Hungary or the Central European region’ in *Description* vs. machine-readable ‘hu’ in *Geographic Coverage*; ‘American states, the District of Columbia, and London, England’ in *Description* vs. ‘USA’ in *Geographic Coverage*; ‘Baja California, Mexico in an area south-east of Ensenada’ vs. ‘Mexico (nation)’ in *Geographic Coverage*.

The cases of free-text *Description* metadata element complementing several controlled-vocabulary subject metadata elements in the same collection-level metadata record were observed. As shown in Figure 3, *Description* includes keywords that complement both *Subjects* and *Objects* with topical information (‘foodways, religious traditions, Native American culture, maritime traditions, ethnic folk culture, material culture’), genre information (‘children’s lore’, ‘occupational lore’, ‘performances’, ‘interviews’, ‘surveys’), and occupational subject information (‘musicians, craft persons, storytellers, folklife interpreters’), while also specifying the dates encoded in *Temporal Coverage* field. In fact, in almost a quarter (22%) of collection-level metadata records in the sample *Description* field complemented two or more controlled-vocabulary subject metadata fields.

Figure 3 An example of multiple complementarities and two-way complementarity



Description: Collection includes approximately 150 cubic feet of administrative, survey and fieldwork files and tens of thousands of audio and video recordings dating from the 1930s through 2001. The collection consists of 88 record series documenting performances by, interviews with, and fieldwork surveys of folk musicians, craftspersons, storytellers, folklife interpreters, and cultural tradition-bearers in such areas as children's lore, foodways, religious traditions, Native American culture, maritime traditions, ethnic folk culture, material culture, and occupational lore.

Subjects: Arts
Architecture
Music
Popular culture
Theater arts
Visual arts
Educational Technology
Religion
Social Studies
State history
United States history

Geographic Coverage: United States (nation)
Southern U.S. (general region)
Florida (state)

Time Period: 1930-1949
1950-1969
1970-1999
2000 to present

Objects Represented: Photographs / slides / negatives
Music (audio files)
Interactive learning objects

Information in controlled-vocabulary subject metadata elements also complemented information encoded in free-text *Description* (Figure 2 and Table 4). For example, in this same collection-level metadata record (Figure 3), *Geographic Coverage* provided spatial information absent in *Description* – 'the USA (nation), Southern USA (general region), Florida (state)' – while *Subjects* listed additional topics (e.g. 'Architecture') not covered by *Description*.

The *Subjects* metadata element was found to complement *Description* the most often. Representative examples included: '860 Spanish and Portuguese literatures' in *Subjects* when this topic was not mentioned at all in *Description*; 'Tennessee Valley Authority, African Americans, forestry' in *Subjects* when these topics were not mentioned at all in *Description*; 15 specific subject strings (e.g. 'North Carolina – African-Americans, North Carolina – Agriculture, North Carolina – Economics and Business' in *Subjects* vs. much broader topical and spatial coverage in *Description* ('North Carolina, … story of the Tar Heel State'). Interestingly, this kind of one-way complementarity was observed twice more often in Opening History and The European Library than in the American Memory. This significant difference is most probably due to the fact that American Memory free-text *Description* metadata element data values are the most detailed among the three digital libraries, as discussed in Section 3.1. Such a high level of detail, arguably necessitated by the fact that only the free-text *Description* element data values are displayed to the end user in American Memory, makes it hard to complement the values found in free-text *Description* elements in this digital library.

The *Temporal Coverage* metadata element was found to complement *Description* in almost a half of collection metadata records overall. Representative examples included: '1400–1699, 1700–1799, 1800–1849, 1850–1899, 1900–1929, 1930–1949, 1950–1969, 1970–1999, 2000 to present, Pre-1400' in *Temporal Coverage* when no time information was provided in *Description*; '1783–1789' in *Temporal Coverage* when no time information was provided in *Description*; '1200–1900' in *Temporal Coverage* vs. 'European age of chivalry' in *Description*. This kind of one-way complementarity was observed drastically less often in the American Memory (3% of records) than in the two other digital libraries (67% and 72% of records). While it was common for the American Memory's controlled-vocabulary collection-level subject metadata elements to complement free-text *Description* element data values less often than in the other two digital libraries, such a dramatic 22- to 24-fold difference in the case with *Temporal Coverage* controlled-vocabulary element is likely due to the comparatively low level of application of this element in American Memory collection-level metadata records: only 67% of American Memory records in the sample included *Temporal Coverage* element.

The *Geographic Coverage* metadata element was found to complement *Description* much more often than the *Description* complemented *Geographic Coverage*. Representative examples included: 'Poland, Lithuania, Ukraine, Belarus' in *Geographic Coverage* vs. 'Poland' in *Description*; 'Germany' in *Geographic Coverage* when no geographic information was provided at all in *Description*; 'Europe, Italy, Great Britain' in *Geographic Coverage* vs. 'the USA and abroad' in *Description*; 'the USA (nation), Midwest USA (general region), Illinois (state), Randolph (county), Knox (county)' in *Geographic Coverage* vs. 'Randolph County, Illinois' in *Description*. Again, this kind of one-way complementarity was observed almost 50% less often in the American Memory than in the two other digital libraries.

The *Objects* metadata element data values also often complemented information in *Description* in two digital libraries – Opening History and American Memory – and again, significantly less often in American Memory. No cases of *Objects* metadata element data values complementing the values in *Description* were observed in The European Library, which can be explained by inconsistent application of *Objects* metadata element in this digital library: in 59% of collection-level metadata records in The European Library sample the *Objects* metadata element was blank or missing, while in the remaining 41% this field contained a broad single-word term (e.g. 'images', 'maps'). Representative examples of the *Objects* metadata element data values complementing *Description* included: 'Film transparencies – Colour, Cityscape photographs' in *Objects* vs. 'photographs' in *Description*; 'Gelatin silver prints, Safety film negatives, Nitrate negatives' in *Objects* vs. 'original negatives and photographic prints' in *Description*; 'books and pamphlets, photographs/slides/negatives, newspapers, posters and broadsides, periodicals, prints and drawings' in *Objects* vs. 'manuscripts, photographs, ephemera and published materials' in *Description*.

In addition, one-way complementarity between controlled-vocabulary metadata elements was also observed. In particular, subject headings' geographical subdivisions (such as in 'Japanese Americans–California–Manzanar') and temporal qualifiers (as in 'World War, 1914–1918') in *Subjects* metadata element included information that complemented *Temporal Coverage* and *Geographic Coverage* values. For example, in Opening History, *Subjects* complemented *Geographic Coverage* in 12% of collection-level metadata records and *Temporal Coverage* in 18% of the records in the sample.

The cases of two-way complementarity between the two collection-level subject metadata elements were less numerous than cases of one-way complementarity (Table 5). No cases of two-way complementarity were observed between the two or more controlled-vocabulary subject metadata elements. Two-way complementarity between the free-text (*Description*) and controlled-vocabulary subject metadata element, in contrast, occurred in 40% of collection-level metadata records overall.

Two-way complementarity was widespread in Opening History (79% of records overall), but occurred less often in The European Library (41%) and significantly less often in American Memory (8%). Most often two-way complemenarity was observed between *Description* and *Subjects* elements (29% on average across the three digital libraries; the least often in American Memory). Two-way complementarity between *Description* and *Temporal Coverage* was observed only in Opening History. Two-way complementarity between *Description* and *Geographic Coverage* was observed in two digital libraries: Opening History and The European Library (11% of the records on average). The least two-way complementarity was observed between *Description* and *Objects* metadata element (7% on average across the three digital libraries; significantly more often in Opening History than in the other two digital libraries). Representative examples of two-way complementarity included:

- 'letters' in *Description* vs. 'autograph albums' in *Subjects* (taken together, the values in two fields provide more comprehensive genre information)
- 'dance instruction manuals, anti-dance manuals, histories, treatises on etiquette' in *Description* vs. 'Ballroom dancing – USA' in *Subjects* (*Subjects* information specifies *Description* information from 'dance' to 'ballroom dancing' and adds geographic coverage information, while *Description* adds information on specific aspects of dancing – 'etiquette' – and genre of materials in collection not covered by any other metadata field in this record).
- 'towns of Coal City, Braidwood and Wilmington' in *Description* vs. 'Illinois (state), Grundy (county)' in *Geographic Coverage* (state and county information in *Geographic Coverage* and town information in *Description* complement each other for a more specific geographic representation).
- 'contemporary, … European age of chivalry, … prior to 1900' in *Description* vs. '1200–1900' in *Temporal Coverage* (while *Temporal Coverage* specifies the lower limit of the 'prior to 1900' range of years – '1200' – and provides the time frame for 'European age of chivalry', *Description* introduces another – 'contemporary' – time period not covered by *Temporal Coverage*).
- 'newspaper photographs' in *Description* vs. 'photographs/ slides/negatives, archival finding aids' in *Objects* (*Description* specifies genre information in *Objects* from general 'photographs' to 'newspaper photographs, while *Objects* adds another genre not mentioned in *Description* – 'archival finding aids').

Among the digital libraries examined in this study, only The European Library had a noticeable proportion (19%) of redundancy between the values in different collection-level subject metadata elements. Very little redundancy was observed in the Opening History and American Memory collection-level metadata records. Examples of redundancy include restating of identical geographic information (e.g. 'Estonia', 'the Netherlands', 'Ljubljana' in both *Description* and *Geographic Coverage* metadata element), temporal information (e.g. '1763' in both *Description* and *Temporal* Coverage), and genre information (e.g. 'photographs' in both *Description* and *Subjects*).

Table 5 Comparative frequencies of occurrence of two-way complementarity

	American Memory (% of metadata records)	*Opening History (% of metadata records)*	*The European Library (% of metadata records)*
Overall	8	79	41
Description and Subjects	5	58	30
Description and Temporal Coverage	–	39	–
Description and Geographic Coverage	–	24	11
Description and Objects	3	18	–

4 Conclusions

Duval et al. (2002) point that richness of metadata descriptions should be determined by local policies and best practices designated by the agency creating the metadata. The study reported in this paper and a related study (Zavalina, 2011a) collectively sought to define and evaluate the specific instance of metadata richness – the richness of collection-level subject metadata. The following major indicators of the richness of collection-level subject metadata were identified and considered:

- consistency of application of subject metadata elements in collection-level metadata records (Zavalina, 2011a)
- variety of collection properties represented in the free-text collection-level subject metadata (this study)
- complementarity between the values in different collection-level subject metadata elements (this study).

Results of this study indicate that encoding of mutually complementary subject-specific information in free-text and controlled-vocabulary collection-level metadata elements is already a common practice among some of the large-scale digital libraries, and possibly is recognised by collection-level metadata creators as a benchmark in crafting rich collection-level metadata in digital libraries.

Despite the differences observed in these three digital libraries, most of their free-text *Description* collection-level metadata elements were found to provide rich description of digital collections, covering a variety of collection properties and complementing information encoded in controlled-vocabulary collection-level subject metadata elements. The emerging best practices in collection-level description observed in this study suggest enriching *Description* metadata element data values by encoding a variety of additional, non-subject collection characteristics. These properties include title, size, collection development policy, copyright information, provenance, intended audience, navigation and functionality, language of items in collection, frequency of additions, participating or contributing institutions, funding sources, collection strengths (importance, uniqueness and comprehensiveness) and creators of items in collection.

The findings presented in this paper demonstrate the high level of mutual complementarity between free-text and controlled-vocabulary subject metadata elements in collection-level metadata in large-scale digital libraries that aggregate cultural heritage digital collections. Quite predictably, the free-text *Description* metadata element, due to its natural language values and higher length, often complemented information in controlled-vocabulary subject metadata element. However, it was also observed in this study that controlled-vocabulary subject metadata elements, especially *Geographic Coverage*, complemented information encoded in *Description* quite often. Although most newly created digital libraries limit their collection-level metadata to *Title* and free-text *Description* elements, this empirical study results demonstrate that providing more detailed collection-level metadata, including both free-text and controlled-vocabulary subject metadata elements, improves subject access.

Best practice recommendations for creating rich collection-level subject metadata are needed. These guidelines can be incorporated into the *Framework of Guidance for Building Good Digital Collections* NISO Recommended Practice document (NISO Framework Working Group, 2007) or International Federation of Library Associations and Institutions (IFLA) *Guidelines for Digital Libraries* that are currently under development (IFLA Working Group on Guidelines for Digital Libraries, 2012). The findings of this study with respect to the emerging best practices in application of free-text and controlled-vocabulary collection-level subject metadata could be instrumental in developing these recommendations.

This exploratory research focused on collection-level subject 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. 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, and 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 data values, the collection properties to be represented in subject metadata element data 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

The author of this paper wishes to thank the research and implementation team of the Digital Collections and Content project at the University of Illinois at Urbana-Champaign funded by the United States Institute for Museum and Library Services through the National Leadership Grant (particularly Dr. Carole L. Palmer, Dr. Allen Renear, Amy Jackson and Myung-Ja Han) for providing Opening History collection-level metadata for analysis, help with data coding, and valuable feedback. The author also wishes to acknowledge contributions of Sally Chambers and Christa Maher, who provided The European Library and American Memory collection-level metadata for analysis and answered the questions about subject metadata in these digital libraries.

References

Arms, C.R. (1996) ‘Historical collections for the national digital library: lessons and challenges at the library of congress’, *D-Lib Magazine*. Available online at: http://www.dlib.org/dlib/april96/loc/04c-arms.html; http://archive.ifla.org/VII/s12/mom/appendx3.htm (accessed on 1 March 2012).

Baca, M. and Harpring, P. (Eds) (2009) *Categories for the Description of Works of Art (CDWA)*, Getty Research Institute, Santa Monica.

Visual Resources Association (2006) *Cataloging Cultural Objects: A Guide to Describing Cultural Works and their Images*, American Library Association, Chicago.

Bearman, D. (1992), ‘Contexts of creation and dissemination as approaches to documents that move and speak’, *Documents that Move and Speak: Audiovisual Archives in the New Information Age: Proceedings of a Symposium*, National Archives of Canada, Ottawa, pp.140–149.

Bruce, T.R. and Hillmann, D.I. (2004) ‘The continuum of metadata quality: defining, expressing, exploiting’, in Hillman, D. and Westbrook, L. (Eds): *Metadata in Practice*, American Library Association, Chicago, pp.238–256.

Dublin Core Metadata Initiative (2007) *Dublin Core Collections Application Profile*. Available online at: http://dublincore.org/groups/collections/collection-application-profile (accessed on 1 March 2012).

Duval, E., Hodgins, W., Sutton, S. and Weibel, S.L. (2002) ‘Metadata principles and practicalities’, *D-Lib Magazine*, Vol. 8, No. 4. Available online at: http://www.dlib.org/dlib/april02/weibel/04weibel.html (accessed on 1 March 2012).

Encoded Archival Description (2002) Available online at: http://www.loc.gov/ead/ (accessed on 1 March 2012).

Greenberg, J. (2003) ‘Metadata and the world wide web’, in Kent, A. et al. (Eds): *Encyclopedia of Library and Information Science*, Marcel Dekker, New York, pp.1876–1888.

Hillmann, D.I. (2005) *Using Dublin Core: The Elements*. Available online at: http://dublincore.org/documents/usageguide/elements.shtml (accessed on 1 March 2012).

Hillmann, D.I. (2008) ‘Metadata quality: from evaluation to augmentation’, *Cataloging & Classification Quarterly*, Vol. 46, No. 1, pp.65–80.

IFLA Working Group on Guidelines for Digital Libraries (2012) Available online at: http://www.ifla.org/en/digital-libraries/guidelines (accessed on 1 March 2012).

Macgregor, G. (2003) ‘Collection-level descriptions: metadata of the future?’, *Library Review*, Vol. 52 No. 6, pp.247–250.

Miller, P. (2000) ‘Collected wisdom: some cross-domain issues of collection-level description’, *D-Lib Magazine*. Available online at: http://www.dlib.org/dlib/september00/miller/09miller.html (accessed on 1 March 2012).

Moen, W.E., Stewart, E.L. and McClure, C.R. (1998) *The Role of Content Analysis in Evaluating Metadata for the U.S. Government Information Locator Service (GILS): Results from an Exploratory Study*. Available online at: http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm (accessed on 1 March 2012).

National Union Catalog of Manuscript Collections (2011). *Online Data Sheet for Participating Institutions*. Available online at: http://www.loc.gov/coll/nucmc/lcforms.html (accessed on 1 March 2012).

NISO Framework Working Group (2007) *A Framework of Guidance for Building Good Digital Collections*, 3rd ed., National Information Standards Organization, Bethesda, Maryland.

Ohio State University Libraries (2006) *Knowledge Bank Metadata Application Profile for Digital Video*. Available online at: http://library.osu.edu/staff/techservices/KBAppProfileDV.php (accessed on 1 March 2012).

OLAC Cataloging Policy Committee (2002) *Summary Notes for Catalog Records*. Available online at: http://www.olacinc.org/drupal/?q=node/21 (accessed on 1 March 2012).

Opening History (2009) *Opening History (OH) Aggregation Collection Development Policy*. Available online at: http://imlsdcc.grainger.uiuc.edu/docs/CollectionDevelopmentPolicy.pdf (accessed on 1 March 2012).

Park, J. (2009) ‘Metadata quality in digital repositories: a survey of the current state of the art’, *Cataloging & Classification Quarterly*, Vol. 47, No. 3, pp.213–228.

Park, J. and Tosaka, Y. (2010) ‘Metadata quality control in digital repositories and collections: criteria, semantics, and mechanisms’, *Cataloging & Classification Quarterly*, Vol. 48, No. 8, pp.696–715.

Stvilia, B., Gasser, L., Twidale, M.B. and Smith, L.C. (2007) ‘A framework for information quality assessment’, *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 12, pp.1720–1733.

Xie, H. (2006) ‘Evaluation of digital libraries: criteria and problems from users’ perspectives’, *Library & Information Science Research*, Vol. 28, No. 3, pp.433–452.

Xie, H. (2008) 'Users' evaluation of digital libraries (DLs): their uses, their criteria, and their assessment', *Information Processing & Management*, Vol. 44, No. 3, pp.1346–1373.

Zavalina, O.L. (2011a) 'Contextual metadata in digital aggregations: application of collection-level subject metadata and its role in user interactions and information retrieval', *Journal of Library Metadata*, Vol. 11, Nos. 3/4, pp.104–128.

Zavalina, O.L. (2011b) 'Free-text collection-level subject metadata in large-scale digital libraries: a comparative content analysis', *Proceedings of the International Conference on Dublin Core and Metadata Applications*, DCMI, The Hague, the Netherlands, pp.147–157.

Zavalina, O.L, Palmer, C.L., Jackson, A.S. and Han, M-J. (2008) 'Evaluating descriptive richness in collection-level metadata', *Journal of Library Metadata*, Vol. 8, No. 4, pp.263–292.

Notes

[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 catalogues 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 (Hillmann, 2005) 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, p.4.3).