Assessing Descriptive Substance in Free-Text Collection-Level Metadata

Oksana Zavalina zavalina@uiuc.edu
Carole L. Palmer clpalmer@uiuc.edu
Amy S. Jackson amyjacks@uiuc.edu
Myung-Ja Han mhan3@uiuc.edu
University of Illinois at Urbana-Champaign, USA

Abstract

Collection-level metadata has the potential to provide important information about the features and purpose of individual collections. This paper reports on a content analysis of collection records in an aggregation of cultural heritage collections. The findings show that the free-text *Description* field often provides more accurate and complete representation of subjects and object types than the specified fields. Properties such as importance, uniqueness, comprehensiveness, provenance, and creator are articulated, as well as other vital contextual information about the intentions of a collector and the value of a collection, as a whole, for scholarly users. The results demonstrate that the semantically rich free-text *Description* field is essential to understanding the context of collections in large aggregations and can serve as a source of data for enhancing and customizing controlled vocabularies.

Keywords: descriptive metadata; collection-level metadata; Dublin Core Collection Application Profile; federated digital collections; IMLS Digital Collections and Content project.

1. Introduction and Background

It has long been recognized that contextual metadata is important for facilitating access to documents in archival collections (e.g., Bearman, 1992). More recently, digital collections have come to be understood as information seeking contexts (Allen & Sutton, 1993; Lee, 2000). As digital collections are aggregated into larger meta-collections, and grow in size and complexity, the need for a coherent contextual framework increases. Collection-level metadata can provide the necessary relational and contextual framework (Macgregor, 2003; Miller, 2000) through "unitary"[1] and "analytic"[2] descriptive approaches (Heaney, 2000).

Cultural heritage institutions have purposefully conceptualized and developed their digital collections in many ways, as "displays", "tours", "tools", "lessons", and to provide a record of cultural events (Palmer et al., 2006). However, in a large digital federation or aggregation, the purpose of the original, deliberately built collections becomes difficult to discern. Collection-level metadata has the potential to provide important information about features of a parent collection and why it might be of value to users. But the qualitative aspects of collections are difficult to describe in a systematic way, as they may embody a good deal of intellectual intent and tend to be highly complex and mutable.

This paper reports on the current phase of the Digital Collections and Content (DCC) project that is investigating how to represent collection context for scholarly use of large-scale, heterogeneous digital aggregations. The DCC provides integrated access to over 200 digital collections funded by the Institute of Museum and Library Services (IMLS), National Leadership Grant program, through a centralized collection registry and metadata repository. The DCC collection metadata schema used for the registry was adapted from a preliminary version of the Dublin Core Collection Description Application Profile (DC CDAP) and the UKOLN RSLP schema (Heaney, 2000). The information used to encode collection registry records is gathered directly from resource developers through a survey, with complementary information taken from collection websites and the descriptive text provided in the grant proposals submitted to IMLS. Once the initial record has been created, it is sent to the local collection administrator for review

[1] Defined as: "consists only of information about the collection as a whole."
[2] Defined as: "consists of information about the individual items within [a collection] and their content."

and editing. Needed updates, changes, and additions of information and links to related collections are made through the DCC collection record edit interface. The DCC project coordinator is responsible for final review and release of all collection records made accessible through the public interface.

Previous DCC reports have discussed the various ways that resource developers conceive of collections, the attributes they find most important in describing collections, and the different "cultures of description" evident among libraries, museums, archives, and historical societies (Knutson, Palmer, & Twidale, 2003; Palmer & Knutson, 2004). In addition, preliminary DCC usability studies suggested that collection and subcollection metadata help users ascertain features like uniqueness, authority, and representativeness of objects retrieved and can lessen confusion experienced searching large-scale federations (Foulonneau et al., 2005; Twidale & Urban, 2005). The analysis presented here builds on previous DCC work[3] to extend our understanding of the role of collection metadata and provide an empirical foundation for our ongoing analysis of item-level and collection-level metadata relationships (Renear et al., forthcoming).

2. Methods

The objectives of the study were to identify the range of substantive and purposeful information about collections available within the DCC Collection Registry, determine patterns of representation, and assess the adequacy of the DCC collection-level metadata schema[4] for representing the richness and diversity of collections in the aggregation. The results presented here are based on a systematic, manual analysis of 202 collection-level records. The free-text in the *Description* field was both qualitatively and quantitatively analyzed to identify types of information provided about a digital collection and the degree of agreement between information provided in the free-text *Description* field and relevant information found in other free-text and controlled vocabulary fields. Hereafter, we use the term "collection properties" to refer to the types of information identified in the collection records.[5]

3. Findings

Table 1 lists the properties found only in the *Description* field of the DCC collections record. The properties are subdivided into three groups. The first consists of three properties that are special claims about collections: Importance (e.g., "collection of the most important and influential 19th and early 20th century American cookbooks"), Uniqueness (e.g., "unique historical treasures from … archives, libraries, museums, and other repositories"), and Comprehensiveness (e.g., "a comprehensive and integrated collection of sources and resources on the history and topography of London"). These properties are of particular interest as the kind of self-assessed value commonly used to distinguish special collections. Although not prominent enough to include in the table, a related property, "Strength", appeared in three records.[6]

The second group contains two other common descriptive properties also not delineated in the DCC collection metadata schema: Creator of items in the collection (e.g., "The Museum Extension Projects of Pennsylvania, New Jersey, Connecticut, Illinois, and Kansas crafted most of the items currently in the collection") and Provenance (e.g., "in December 2002, the … Library acquired the Humphrey Winterton Collection of East African photographs"). Item Creator[7] and Provenance elements might serve an even greater number of DCC collections than those currently

[3] Described in detail in our five-year report http://imlsdcc.grainger.uiuc.edu/docs/FinalReport_ResearchMethods.pdf

[4] Available at: http://imlsdcc.grainger.uiuc.edu/CDschema_elements.asp

[5] No predefined list of categories was used for analysis. The categories emerged from coding performed by two coders who are authors on this paper. A test of intercoder reliability showed 80.4% agreement in assigning the codes to specific cases.

[6] See Johnston (2003) for discussion on inclusion of a *Strength* element in the Dublin Core Collection Description Application Profile.

[7] The DCC collection description metadata schema currently uses dc:creator element in a limited way to indicate a grant project responsible for creation of the digital collection, but does not include creators of items and collections.

exploiting the *Description* field for these purposes. There are DCC collections related to single or multiple authors that could benefit from more formal representation of item creators. In this case, a new element would need to be specified, since the existing DC CDAP *Collector* element is designed to cover creator of the collection not creator of items in the digital collection. Also, a large number of the collections come from museums, and a smaller but substantial group from historical societies and archives. These institutions are likely to have conventions for documenting chain of custody. Here, the DC CDAP *Custodial History* element is a good model, since it covers the kind of provenance information found in our free-text metadata.

The third group contains Subject and Object. Formal elements do exist for these properties, but the analysis shows that the *Description* field provides extensive additional coverage (e.g., "broad range of topics, including ranching, mining, land grants, anti-Chinese movements, crime on the border, and governmental issues"; "souvenirs of all kinds, including plates, cups, vases, trays, bottles, sewing boxes and games").

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

TABLE 1. Collection properties unique to *Description* field.

Collection Property	***Number of collections***	***%***
Subjects	181	91.4
Object types	149	75.3
Collection development policy	102	52.0
Collection title	103	52.0
Size	53	26.8
Audience	34	17.0
Navigation and functionality	32	16.2
Participating/contributing institutions	30	15.2
Funding sources	10	5.1

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

Table 2 shows nine collection properties represented but not unique to the free-text *Description* field. The subject information in the *Description* field ranges from specific statements to subject keywords scattered throughout the text. In most cases (66.7%), the *Description* field provides more accurate and specific coverage than the fields intended for subject indexing*: Subjects*, *GEM Subjects*, *Geographic Coverage,* and *Time Period*. Fifty percent of the *Description* fields include indications of temporal coverage, ranging from specific dates and date ranges (e.g., 19th century) to known historical periods (e.g., World War I, California Golden Rush). Sixty percent of *Description* fields include indications of geographic coverage of varying granularity (e.g., "Austro-Hungarian Empire"; "Mayan city of Uxmal in Yucatan, Mexico and a Native American Mississippian site, Angel Mounds U.S.A.").

The *Description* field often lists additional, or more specific, types of objects than covered by the formal element, *Objects Represented*. Broad terms, such as "physical artifacts", are common, as are more specific terms, such as "lanterns, torches, banners". Formats and genres are also frequently specified, as with "leaflets", "songbooks", and "political cartoons". Object types and formats are sometimes conflated, even within the same sentence, in the *Description* field, as well

as in *Objects Represented*. This lack of disambiguation between type and format is a known metadata quality problem in digital object description (see, for example, Jackson et al., 2008).

Over half of the *Description* fields contain evidence of collection development policies (e.g., "titles published between 1850 and 1950 were selected and ranked by teams of scholars for their great historical importance"). Some identify other locally accessible materials or plans for future collection development, a potentially significant aspect of collector intentionality: "it is planned to provide access to a complimentary collection of Richmond related Civil War period resources"; "lesson plans, activities and photo essays designed by teacher advisors and educational consultants will be added in the future". Others explicitly state a purpose: "support global efforts to conserve, study, and appreciate the diversity of palms".

While duplicative of the *Title* field, many titles found in the *Description* field (either full title or part of title) provide concise statements with subject-specific information, as well as information on the object types in a collection. Collection size statements in the *Description* field range from quantitative specifications (e.g., "209 cartoons, 12 Christmas cards, and 3 facsimiles of cartoons") to general orientations (e.g., "hundreds of personal letters, diaries, photos, and maps"). In 28% of the cases, the *Description* field is the only source of this important information. In 30% of the collection records the size data in the *Description* and *Size* fields do not match; these discrepancies seem to reflect, sometimes clearly, the difference between projected and actual size of the digital collection (e.g., "When finished, the collection guide will consist of well over 100,000 online stereoviews" in the *Description* field and "38254 Stereographic Photoprints" in the *Size* field).

Audience information, found in 17% of *Description* fields (e.g., "Alabama residents and students, researchers, and the general public"), often complements and clarifies controlled vocabulary values in the *Audience* field. For example, in a record where the *Audience* field lists "General public, K-12 students, undergraduate students, K-12 teachers and administrators, Scholars/researchers/graduate students", the *Description* field specifies "anthropologists, art historians, cultural studies scholars, historians, political scientists and sociologists".

Some aspects of navigation or functionality represented in the *Description* field are also found in the formal *Interaction with Collection* field of the same record (e.g., "accessible by date of issue or by keyword searching" in *Description* and "search, browse" in *Interaction with Collection*). In most cases, information in the two fields is complementary.

Institutions participating in the digitization project and contributing items to digitize (e.g., "project brings … together with the University to build a digital repository") and funding sources that helped support digital collections (e.g., "funds provided by the Institute of Museum and Library Services, under the federal Library Services and Technology Act") are also often acknowledged in *Description* fields.

4. Discussion and Conclusions

Our findings identify the various kinds of substantive descriptive information provided in the free-text *Description* element, much of which clearly enriches the collection-level records and provides important scholarly context for the collections within the DCC. There is consistent representation of subjects and object types that is more accurate in coverage and offers more detail than that represented in the other fields specified for those purposes. Moreover, "special claims" about a collection's importance, uniqueness, or comprehensiveness are not represented in any other way within the record and add vital qualitative and contextual information about the intentions of collectors and the role the collection plays in the larger universe of related content. Provenance and Item Creator properties are not accommodated in the current DCC collection metadata schema, but were strongly represented within the *Description* field. All of these data represent distinguishing features potentially of interest to scholarly and other research audiences.

Based on these findings, the first activity slated for collection record enhancement in the DCC is to align the DCC collection description schema with the DC CDAP, which was released after

development of the DCC schema. The *Custodial History* field will accommodate some of the key information currently found only in the *Description* field. A newly defined field for creators of items in a collection and a specified field for special claims about collections are also under consideration. Moreover, the *Description* field is clearly a semantically-rich source from which to mine terms to develop a customized controlled vocabulary for use in the DCC and similar aggregations of cultural heritage digital materials. The research team is exploring how to enhance the current controlled vocabulary with frequently used terms and concepts used in the *Description* field. This terminology would be more representative of the language used by collection creators to explain the purpose and value of their content and would provide a more accurate record of the materials included in cultural heritage collections. The next step in our study of free-text collection-level metadata is a comparative analysis of collection records from sources other than the DCC, produced by libraries, museums, and archives. A broader understanding of the use of the *Description* field in various organizational contexts will be particularly meaningful as we continue to explore the general relationship between content and context and the ways in which collection-level description can complement item-level description.

Acknowledgements

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

References

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

Bearman, David. (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 held 30 April to 3 May 1990 at the National Archives of Canada*, 140-149.

Foulonneau, Muriel, Timothy W. Cole, Thomas G. Habing, and Sarah L.Shreeves. (2005). Using collection descriptions to enhance an aggregation of harvested item-level metadata. *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, 32-41.

Heaney, Michael. (2000). *An Analytical Model of Collections and Their Catalogues*. Retrieved April 12, 2008, from http://www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf.

Jackson, Amy S., Myung-Ja Han, Kurt Groetsch, Megan Mustafoff, and Timothy W. Cole. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata, 8 (1)*.

Johnston, Pete. (2003). *Report from Meeting of DC CD WG at DC-2003.* Retrieved April 12, 2008, from http://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0310&L=DC-COLLECTIONS&D=0&I=-3&P=59.

Knutson, Ellen M., Carole L. Palmer, and Michael Twidale (2003). Tracking metadata use for digital collections. *Proceedings of the International DCMI Metadata Conference and Workshop*, 243-244.

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

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

Miller, Paul. (2000, September). Collected wisdom: some cross-domain issues of collection-level description. *D-Lib Magazine, 6(9)*. Retrieved June 14, 2008, from http://www.dlib.org/dlib/september00/miller/09miller.html.

Palmer, Carole L., and Ellen M. Knutson. (2004). Metadata practices and implications for federated collections. *Proceedings of the 67th ASIS&T Annual Meeting,* 456-462.

Palmer, Carole L., Ellen M. Knutson, Michael Twidale, and Oksana Zavalina. (2006). Collection definition in federated digital resource development. *Proceedings of the 69th ASIS&T Annual Meeting*, 161-162.

Renear, Allen H., Richard J. Urban, Karen M. Wickett, Carole L. Palmer, and David Dubin. (forthcoming). Sustaining collection value: Managing collection/item metadata relationships. *Proceedings of the Digital Humanities Conference*, 25-29 June 2008, Oulu, Finland.

Twidale, Michael, and Richard J. Urban. (2005). *Usability Analysis of the IMLS Digital Collection Registry*. Retrieved June 14, 2008, from http://imlsdcc.grainger.uiuc.edu/3YearReport/docs/UsabilityReport1.pdf.