Sustaining Collection Value: Managing Collection/Item Metadata Relationships

Allen H. Renear, Richard Urban, Karen Wickett, Carole L. Palmer, David Dubin
Center for Research in Information and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
[March 18, 2008. Accepted for DH2008.]

Abstract

Many aspects of managing collection/item metadata relationships are critical to sustaining collection value over time. Metadata at the collection-level not only provides context for finding, understanding, and using the items in the collection, but is often essential to the particular research and scholarly activities the collection is designed to support. Contemporary retrieval systems, which search across collections, usually ignore collection level metadata. Alternative approaches, informed by collection-level information, will require an understanding of the various kinds of relationships that can obtain between collection-level and item-level metadata. This paper outlines the problem and describes a project that is developing a logic-based framework for classifying collection-level/item-level metadata relationships. This framework will support (i) metadata specification developers defining metadata elements, (ii) metadata librarians describing objects, and (iii) system designers implementing systems that help users take advantage of collection-level metadata.

Introduction

Collections of texts, images, artifacts, and other cultural objects are usually designed to support particular research and scholarly activities. Toward that end collections themselves, as well as the items in the collections, are carefully developed and described. These descriptions indicate such things as the purpose of the collection, its subject, the method of selection, size, nature of contents, coverage, completeness, representativeness, and a wide range of summary characteristics, such as statistical features. This information enables collections to function not just as aggregates of individual data items but as independent entities that are in some sense more than the sum of their parts, as intended by their creators and curators (Curral, Moss & Stuart, 2005; Heaney, 2000; Lagoze, et al. 2006; Palmer, 2004). Collection-level metadata, which represents this information in computer processable form, is thus critical to the distinctive intellectual and cultural role of collections as something more than a set of individual objects.

Unfortunately, collection-level metadata is often unavailable or ignored by retrieval and browsing systems, with a corresponding loss in the ability of users to find, understand, and use items in collections (Lee, 2000, 2003, 2005; Wendler, 2004). Preventing this loss of information is particularly difficult, and particularly important, for "metasearch", where item-level descriptions are retrieved from a number of different collections simultaneously, as is the case in the increasingly distributed search environment (Christenson & Tennant, 2005; Dempsey, 2005; DLF, 2005; Foulonneau, et al., 2005; Lagoze, et al., 2006; Warner, et al., 2006).

The now familiar example of this challenge is the "'on a horse' problem", where a collection with the collection-level subject "Theodore Roosevelt" has a photograph with the item-level annotation "on a horse" (Wendler, 2004). Item-level access across multiple collections (as is provided not only by popular Internet search engines, but also specialized federating systems, such as OAI portals) will not allow the user to effectively use a query with keywords "Roosevelt" and "horse" to find this item, or, if the item is retrieved using item-level metadata alone, to use collection-level information to identify the person on the horse as Roosevelt.

The problem is more complicated and consequential than the example suggests and the lack of a systematic understanding of the nature of the logical relationships between collection-level metadata and item-level metadata is an obstacle to the development of remedies. This understanding is what is required not only to guide the development of context-aware search and exploitation, but to support management and curation policies as well. The problem is also timely: even as recent research continues to confirm the key role that collection context plays in the scholarly use of information resources (Brockman, et al., 2001; Palmer, 2004), the Internet has made the context-free searching of multiple collections routine.

In what follows we describe our plans to develop a framework for classifying and formalizing collection-level/item-level metadata relationships. This undertaking is part of a larger project, recently funded by US Institute for Museum and Library Services (IMLS), to develop tools for improved retrieval and exploitation across multiple collections.[1]

Varieties of Collection/Item Metadata Relationships

In some cases the relationship between collection-level metadata and item-level metadata attributes appears similar to non-defeasible inheritance. For instance, consider the Dublin Core Collections Application Profile element *marcrel:OWN*, adapted from the MARC cataloguing record standard. It is plausible that within many legal and institutional contexts whoever owns a collection owns each of the items in the collection, and so if a collection has a value for the *marcrel:OWN* attribute then each member of the collection will have the same value for *marcrel:OWN*. (For the purpose of our example it doesn't matter whether or not this is actually true of *marcrel:OWN*, only that some attributes are sometimes used by metadata librarians with an understanding of this sort, while others, such as *dc:identifier*, are not).

In other cases the collection-level/item-level metadata relationship is almost but not quite this simple. Consider the collection-level attribute *myCollection:itemType*, intended to characterize the type of objects in a collection, with values such as "image", "text", "software", etc. (we assume heterogeneous collections).[2] Unlike the preceding case we cannot conclude that if a collection has the value "image" for *myCollection:itemType* then the items in that collection also have the value "image" for that same attribute. This is because an item which is an image is not itself a *collection* of images and therefore cannot have a non-null value for *myCollection:itemType.*

However, while the rule for propagating the information represented by *myCollection:itemType* from collections to items is not simple propagation of attribute and value, it is nevertheless simple enough: if a collection has a value, say "image", for *myCollection:itemType*, then the items in the collection have the same value, "image" for a corresponding attribute, say, *myItem:type,* which indicates the type of item (cf. the Dublin Core metadata element *dc:type*). The attribute *myItem:type* has the same domain of values as *myCollection:itemType,* but a different semantics.

When two metadata attributes are related as *myCollection:itemType* and *myItem:type* we might say the first can be v-converted to the other. Roughly: a collection-level attribute **A** v-converts to an item-level attribute **B** if and only if whenever a collection has the value *z* for **A**, every item in the collection has the

---

[1] Information about the IMLS Digital Collections and Content project can be found at: http://imlsdcc.grainger.uiuc.edu/about.asp.

[2] In our examples we will use imaginary metadata attributes. The namespace prefix "*myCollection:*" indicates collection-level attributes and the prefix "*myItem:*" indicates item-level attributes. No assumptions should be made about the semantics of these attributes other than what is stipulated for illustration. The current example, *myCollection:itemType,* does intentionally allude to *cld:itemType* in the Dublin Core Collections Application Profile, and "image", "text", "software", are from the DCMI Type Vocabulary; but our use of *myCollection:itemType* differs from *cld:itemType* in entailing that *all* of the items the collection are of the indicated type.

value *z* for **B**. This is the simplest sort of convertibility—the attribute changes, but the value remains the same. Other sorts of conversion will be more complex. We note that the sort of propagation exemplified by *marcrel:OWN* is a special case of v-convertibility: *marcrel:OWN* v-converts to itself.

This analysis suggests a number of broader issues for collection curators. Obviously the conversion of collection-level metadata to item-level metadata, when possible, can improve discovery and exploitation, especially in item-focused searching across multiple collections. But can we even in the simplest case be confident of conversion without loss of information? For example, it may be that in some cases an "image" value for *myCollection:itemType* conveys more information than the simple fact that each item in the collection has "image" value for *myItem:type*.

Moreover there are important collection-level attributes that both (i) resist any conversion and (ii) clearly result in loss of important information if discarded. Intriguingly these attributes turn out to be carrying information that is very tightly tied to the distinctive role the collection is intended to play in the support of research and scholarship. Obvious examples are metadata indicating that a collection was developed according to some particular method, designed for some particular purpose, used in some way by some person or persons in the past, representative (in some respect) of a domain, had certain summary statistical features, and so on. This is precisely the kind of information that makes a collection valuable to researchers, and if it is lost or inaccessible, the collection cannot be useful, as a collection, in the way originally intended by its creators.

The DCC/CIMR Project

These issues were initially raised during an IMLS Digital Collections and Content (DCC) project, begun at the University of Illinois at Urbana-Champaign in 2003. That project developed a collection-level metadata schema[3] based on the RSLP[4] and Dublin Core Metadata Initiative (DCMI) and created a collection registry for all the digital collections funded through the Institute of Museum and Library Services National Leadership Grant (NLG) since 1998, with some Library Services and Technology Act (LSTA) funded collections included since 2006[5]. The registry currently contains records for 202 collections. An item-level metadata repository was also developed, which so far has harvested 76 collections using the OAI-PMH protocol[6].

Our research initially focused on overcoming the technical challenges of aggregating large heterogeneous collections of item-level records and gathering collections descriptions from contributors. We conducted studies on how content contributors conceived of the roles of collection descriptions in digital environments (Palmer & Knutson, 2004; Palmer, et al. 2006), and conducted preliminary usability work. These studies and related work on the CIC Metadata Portal[7], suggest that while the boundaries around digital collections are often blurry, many features of collections are important for helping users navigate and exploit large federated repositories, and that collection and item-level descriptions should work in concert to benefit certain kinds of user queries (Foulonneau, et al., 2005).

In 2007 we received a new three year IMLS grant to continue the development of the registry and to explore how a formal description of collection-level/item-level metadata relationships could help registry users locate and use digital items. This latter activity, CIMR, (Collection/Item Metadata Relationships),

---

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

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

[5] http://www.imls.gov

[6] http://www.openarchives.org/OAI/openarchivesprotocol.html

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

consists of three overlapping phases. The first phase is developing a logic-based framework of collection-level/item-level metadata relationships that classifies metadata into varieties of convertibility with associated rules for propagating information between collection and item levels and supporting further inferencing. Next we will conduct empirical studies to see if our conjectured taxonomy matches the understanding and behavior of metadata librarians, metadata specification designers, and registry users. Finally we will design and implement pilot applications using the relationship rules to support searching, browsing, and navigation of the DCC Registry. These applications will include non-convertible and convertible collection-level/item-level metadata relationships.

One outcome of this project will be a proposed specification for a metadata classification code that will allow metadata specification designers to indicate the collection-level/item-level metadata relationships intended by their specification. Such a specification will in turn guide metadata librarians in assigning metadata and metadata systems designers in designing systems that can mobilize collection level metadata to provide improved searching, browsing, understanding, and use by end users. We will also draft and make electronically available RDF/OWL bindings for the relationship categories and inference rules.

Preliminary Guidance for Practitioners

A large part of the problem of sustainability is ensuring that information will be valuable, and as valuable as possible, to multiple audiences, for multiple purposes, via multiple tools, and over time. Although we have only just begun this project, already some preliminary general recommendations can be made to the different stakeholders in collection management. Note that tasks such as propagation must be repeated not only as new objects are added or removed but, as new information about objects and collections becomes available.

For metadata standards developers:

1. Metadata standards should explicitly document the relationships between collection-level metadata and item-level metadata. Currently we have neither the understanding nor the formal mechanisms for such documentation but they should be available soon.

For systems designers:

2. Information in convertible collection-level metadata should be propagated to items in order to make contextual information fully available to users, especially users working across multiple collections. [This is not a recommendation for how to manage information internally, but for how to represent it to the user; relational tables may remain in normal forms.]

3. Information in item-level metadata should, were appropriate, be propagated to collection level metadata.

4. Information in non-convertible collection-level metadata must, to the fullest extent possible, be made evident and available to users.

For collection managers...

5. Information in non-convertible metadata must be a focus of data curation activities if collections are to retain and improve their usefulness over time.

When formal specifications and tools based on them are in place, relationships between metadata at the collection and item levels will integrated more directly into management and use. In the mean time, attention and sensitivity to the issues we raise here can still improve matters through documentation and policies, and by informing system design.

Acknowledgements

This research is supported by a 2007 IMLS NLG Research & Demonstration grant hosted by the GSLIS Center for Informatics Research in Science and Scholarship (CIRSS). Project documentation is available at http://imlsdcc.grainger.uiuc.edu/about.asp#documentation. We have benefited considerably from discussions with other DCC/CIMR project members and with participants in the IMLS DCC Metadata Roundtable, including: Timothy W. Cole, Thomas Dousa, Dave Dubin, Myung-Ja Han, Amy Jackson, Mark Newton , Carole L. Palmer, Sarah L. Shreeves, Michael Twidale, Oksana Zavalina

References

Brockman, W. et al. 2001. Scholarly Work in the Humanities and the Evolving Information Environment. Washington, DC: Digital Library Federation/Council on Library and Information Resources.

Christenson, H., & Tennant, R. 2005. Integrating Information Resources: Principles, Technologies, and Approaches. http://www.cdlib.org/inside/projects/metasearch/nsdl/nsdl_report2.pdf

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

Dempsey, L. 2005. From metasearch to distributed information environments. Lorcan Dempsey's Weblog (October 9, 2005). http://orweblog.oclc.org/archives/000827.html

DLF. 2005. The Distributed Library: OAI for Digital Library Aggregation: OAI Scholars Advisory Panel Meeting, June 20-21, Washington, DC. http://www.diglib.org/architectures/oai/imls2004/OAISAP05.htm

Foulonneau, M., Cole, T. W., Habing, T. G., & Shreeves, S. L. 2005. Using collection descriptions to enhance an aggregation of harvested item-level metadata. In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM Press, New York, NY, 32-41.

Heaney, M. 2000. An Analytic Model of Collections and Their Catalogues, UK Office for Library and Information Science.

Lagoze, C. et al. 2006. Metadata aggregation and "automated digital libraries": A retrospective on the NSDL experience. In Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM Press, New York.

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

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

Lee, H. 2000. What is a collection? JASIS, 51 (12), 1106-1113.

Palmer, C. 2004. Thematic research collections. In S. Schreibman, R.Siemens, & J. Unsworth (Eds.). Companion to Digital Humanities. Oxford: Blackwell, pp. 348-365.

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

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

Warner, S., Bekaert, J., Lagoze, C., Lin, X., Payette, S., & Van de Sompel, H. 2006. Pathways: Augmenting interoperability across scholarly repositories. Accepted for International Journal on Digital Libraries special issue on Digital Libraries and eScience.

Wendler, R. 2004. The eye of the beholder: Challenges of image description and access at Harvard. In Hillmann, D. I. and Westbrooks, E. L., eds., Metadata in Practice. American Library Association, Chicago, IL, pp. 51-6