A Testbed for Collection-Item Metadata Relationships

Richard J. Urban
University of Illinois, USA
rjurban@illinois.edu

Karen M. Wickett
University of Illinois, USA
wickett2@illinois.edu

Allen H. Renear
University of Illinois, USA
renear@illinois.edu

Abstract

This paper describes the development of a testbed for formalized categories of collection-item metadata relationships. Because these categories are characterized using logic-based formal languages and are most naturally analyzed by exploring inferences and logical relationships, the testbed is based on contemporary semantic web architecture. We describe the design and development of the testbed, discussing some challenges that we overcame in the process of translating OAI-PMH XML records into DCAM-compliant *descriptions sets* that could be represented in RDF.

Keywords: metadata; cultural heritage; RDF; OAI-PMH; collections

1. Introduction

The IMLS Digital Collections and Content (IMLS DCC) project was initially conceived as a gateway to more than 200 digital collections funded by Institute of Museum and Library Services National Leadership Grants. Over five years, the development team created infrastructure to support a centralized collection registry and a repository of item-level Dublin Core records harvested using the Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH). Beginning in October 2007, IMLS DCC began a new phase of the project that expanded the scope to include digital collections related to United States history. This new repository, known as *Opening History* (http://imlsdcc.grainger.uiuc.edu/history), includes more than 700 collection-level records and over one million item-level records. The new grant included support for a project to carry out research on collection/item metadata relationships.

1.1. Collection/Item Metadata Relationships (CIMR)

Previous research indicated that collection-level metadata can provide important contexts for understanding information quality and for building robust search and retrieval services (Shreeves et al. 2005; Foulonneau & Cole, 2005). The Collection/Item Metadata Relationships (CIMR) research group is concerned with identifying and describing relationships between metadata that describe collections and metadata about items that are members of those collections, in order to support search, browsing and management of large-scale aggregations (Renear et al., 2008). In the first year of the project, we developed formal definitions for three categories of relationships between item-level and collection-level metadata:

- **Attribute-value propagation:** whenever a collection has some value for an attribute, every item that is a member of that collection has the same value for the same attribute.
- **Value propagation:** whenever a collection has some value for some attribute, every item in the collection has that value for some other attribute.
- **Value Constraint:** when a collection-level attribute/value pair implies that values within some item-level attributes must be within a particular range.

These categories describe high-level patterns that could be instantiated by any metadata vocabulary. To assess their applicability to real metadata, we are identifying Dublin Core properties that are likely to display propagation or constraint behavior and examining the distribution of values between collection and item descriptions. The CIMR categories are oriented towards a rules-based approach to digital library inferencing. Successful implementation of this approach requires that rules be tested against real-world metadata repositories. In order to test rules and to evaluate and refine our categories, we have developed a testbed environment that supports exploration of semantic descriptions derived from IMLS DCC records. The results of our research are yet to come, however the process of constructing a testing environment has already been unexpectedly illuminating.

2. A Semantic Web Architecture for IMLS DCC Metadata

The IMLS Digital Collections and Content project uses a variety of reliable and effective digital library tools and standards. Item-level OAI-PMH XML records are harvested, preprocessed, and loaded into a relational database (Cole & Shreeves, 2004). While these familiar technologies provide a robust infrastructure for the public search and retrieval service, we concluded that this infrastructure would not be an optimal environment for testing CIMR rules. Because CIMR rules are developed in first order logic and refer to facts implied by the metadata, it is natural to use the logic-based approach provided by contemporary semantic web knowledge representation languages, such as RDF, OWL, SWRL, and SPARQL. Even where tests could be carried out with SQL queries against a relational database, the translation from logical expressions into a relational language complicates analysis and coordination with our emerging theory of collection/item metadata, particularly as we anticipate exploring the use of additional constraints, from metadata schemas, or conjectured independently. Apart from the advantage for studying collection/item relationships, this approach also better aligns our contributions with emerging trends towards semantic web architectures for networked metadata. These approaches can allow us to augment existing metadata records using logic-based inference rules - rules that we are deriving from our broader CIMR categories during the current testing phase (Wickett et al., 2009).

In addition, approaches that result in bespoke stylesheets, scripts and database models cannot readily be shared with those in the larger Dublin Core community interested in logic-based approaches. Consequently, we selected available metadata processing and semantic web toolkits for the foundation of the testbed suite, including: Open-RDF Sesame triple store (http://www.openrdf.org/), the MIT SIMILE Project's OAI2RDF utility (http://simile.mit.edu/wiki/Gadget) and the California Digital Library's Date Normalization Utility (CDL-DNU - https://confluence.ucop.edu/display/Curation/Date+Normalization)

2.1. Selecting Candidate Collections/Item Metadata

We selected 33 collections with associated item-level metadata for the CIMR testbed based on several factors:[1]

1. We did not consider item-level properties that did not have clear relationships to corresponding collection-level properties. For example, *dc:title* for a collection does not share an obvious relationship with *dc:title* of the items (other than both being titles).
2. Several DC-CAP properties explicitly suggest relationships to properties of the items contained in the collection. For example, *cld:itemFormat* is defined as "The media type, physical or digital, of one or more items within the collection" (Dublin Core Collection Description Task Group, 2007). Ideal metadata for the testbed included item-level metadata with both *dc:type* and *dc:format* properties.

[1] Of the 700 collection-level descriptions in the *Opening History* repository, approximately 300 have item-level metadata. Our selection represents ~10% of available collections with item-level records.

3. Research on the information-seeking behaviors of the project's target audience - academic historians and amateur scholars - has demonstrated the importance of temporal and spatial metadata (Case, 1991). These kinds of properties were also important in our initial development of the CIMR categories (Renear et al., 2008). Item-level metadata that included *dc:date* and *dc:coverage* information was given the highest priority in our selection criteria.

To select collections, we first made a gross characterization of item descriptions according to the frequency of attributes within a given collection. We then processed collections that exhibited the desirable features identified above using SIMILE's Gadget metadata exploration utility, which allowed us to take a closer look at the frequency of *values* that appeared in each collection.

3. From OAI-PMH XML to RDF: Some Problems

Construction of the CIMR testbed involved mapping the OAI-PMH XML record to a RDF representation, requiring us to explore the application of the DCMI Abstract Model to the IMLS DCC aggregation (Powell et al., 2007). There are a number of challenges to migrating OAI-PMH XML, which does not follow current DCAM-compliant serialization patterns (Powell, 2009), into RDF. We describe a few of the particularly interesting practical issues below.

3.1. Identifying the Described Resource

The first challenge we encountered was how to appropriately identify the *described resources*. The default SIMILE OAI2RDF stylesheet generated RDF graphs that used the *OAI Identifier* as the subject URI for *DCAM descriptions*. Although this seems reasonable at first glance (Haslhofer (2008) uses a similar convention) it is technically in error and can cause problems later in a linked data environment. The OAI Identifier *does not identify* the resource which is the subject of the triples, rather it identifies the *OAI Item*, which is defined as: "… a container that stores or dynamically generates metadata about a single resource in multiple formats" (Open Archives Initiative, 2008).

The OAI Identifier is an appropriate subject for attributions signaled by OAI XML elements such as "`<setSpec>2771</setSpec>`", as these are not used to make assertions about the described resource but to make assertions about the OAI Item, a metadata container. However the OAI Item is *not* the resource described by XML elements such as "`<oai_dc:format>glass plate negative </oai_dc:format>`" This problem is a classic example of the difficulty in moving from a data description language with a loosely defined semantics (in this case OAI-PMH) to a logic-based language with a precisely defined semantics (RDF). (Renear et al., 2002)

It quickly became clear that it would be difficult to reliably and systematically select an identifier for the described resource from within the OAI metadata records. In some cases URIs were used as identifiers, but there was no way to confirm that these referenced the intended resource (and in some cases it was clear that they did not), or to make principled selections when several URIs were given. Literal values (e.g. local identifiers such as call numbers, accession numbers, etc.) had the same problems as URIs. It was obvious that any effort to generate described resource URIs from these unpromising materials would be time-consuming to develop, and error-prone in any case. Our solution was to create a new URI to identify the described resource that the OAI Item metadata described. This strategy does assume that each OAI Item contains metadata about "one, and only one," resource. This approach is consistent with the DCMI 1:1 principle and the basis for overloading strategies used elsewhere (Haslhofer 2008). For convenience, we added a "CIMR:" namespace prefix to the existing OAI Identifiers to form a URI for CIMR resources.

3.2. Connecting Collections with their Items

An important assumption in CIMR's agenda is that collection-level metadata stands in some relationship to item-level metadata. When we examined the metadata available from the IMLS

DCC OAI services, a collection membership property was not clearly expressed in any of the records. While the underlying database includes a primary/foreign key relationship between collections and their items, this property is not included in the shared OAI metadata (although metadata may reference a source repository or website). A harvester interrogating the collection-level metadata would not be aware of the associated item-level IMLS DCC OAI service and vice versa. Fortunately the IMLS DCC item-level OAI data provider included an undocumented feature. Each of the collections was available as an item-level OAI set identified by the same value used in collection-level OAI Items. We used the *setSpec* value to construct a URI that explicitly picked out the corresponding collection.

While we could have used *dc:relation* or *dcterms:isPartOf* to express collection membership, the current usage of this property introduces semantic ambiguities. In some records *isPartOf* does represent a relationship to a collection, however in many records *isPartOf* indicates other kinds of parthood, such as the relationship between a page and a book. At the collection-level *hasPart* (the symmetric property of *isPartof*) is restricted to indicating "sub-collections" - not items. For the CIMR testbed, we defined a specific property *cimr:isGatheredInto* (as a sub-property of *dcterms:relation*) for item-level records that includes the collection URI. The *isGatheredInto* property is based on the Dublin Core Collections Application Profile (DC-CAP) data model, which states that items are "gathered into" collections. (Dublin Core Collection Description Task Group, 2007). The addition of the *isGatheredInto* property creates a complete collection/item RDF graph for metadata included in the CIMR testbed.

3.3. Dealing with Dates

A well-known problem for large-scale metadata aggregations, including IMLS DCC, is the diversity found in date values (Dushay & Hillman, 2003; Shreeves et al. 2005). A call to the California Digital Library's Date Normalization Utility (CDL-DNU) is incorporated into the OAI2RDF stylesheet and CDL-DNU normalized values are injected into our RDF graphs. These values are kinds of dates, but they are distinguished from the dates that are native to the record by defining additional CIMR properties (as sub-properties of *dcterms:date* and *dcterms:temporal*). e.g.:

- **cimr:date.normalized** (a literal value that maybe a date range, in the case of original "circa" dates, the CDL tool expresses them as a range such as1905-1915)
- **cimr:date.min** (a typed literal (gYear) for the minimum year in a range)
- **cimr:date.max** (a typed literal (gYear) for the maximum year in a range)

Expressing, in RDF, concepts such as date ranges and "circa" dates is challenging (Davis, 2009). Although the CDL tool could derive typed literal values from source metadata, date ranges conforming to the W3CDTF are still untyped literal values. To express the two values that make up a range, a blank node was introduced to contain the minimum and maximum date values, each as a typed literal. Original and normalized values are still available via direct query to recommended Dublin Core terms.

4. Preliminary Testing & Future Research

The first round of rule testing has touched on each of the three main categories from the CIMR framework: date attributes are a natural source to test value constraint rules; type and format attributes (e.g. *cld:itemType* and *dc:type*) provide test cases for value propagation rules; and *dc:language* attributes, which appear at both the collection and item levels, supply a potential case of attribute value-propagation. Future work will present the result of these examinations. Preliminary testing has suggested that the most accurate rules with respect to the metadata may have a different logical structure than initially conjectured. Generalization and specialization relationships between values seem to play an important role in how relationships obtain across descriptions. Future work will also consider how the use of controlled vocabularies in different

descriptive environments and the mapping of metadata into OAI-PMH influence the appearance of these metadata relationships.

Acknowledgements

This research was supported by a 2007 IMLS National Leadership Research and Demonstration grant LG- 06-07-0020-07 hosted by the GSLIS Center for Informatics Research in Science and Scholarship (CIRSS), Carole L. Palmer, Principal Investigator. Implementation of the CIMR testbed would not have been possible without assistance from GSLIS research programmers Larry Jackson and Amit Kumar.

References

Case, D. O. (1991). The collection and use of information by some American historians: a study of motives and methods. *The Library Quarterly*, *61*(1), 61–82.

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

Davis, I. (2009). Representing Time in RDF. Retrieved from: http://blog.iandavis.com/2009/08/time-in-rdf-1

DCMI Usage Board. (2008). DCMI Metadata Terms. Retrieved from http://dublincore.org/documents/dcmi-terms/

Dublin Core Collection Description Task Group. (2007). Dublin Core Collections Application Profile. Retrieved from http://dublincore.org/groups/collections/collection-application-profile/

Dushay, N., & Hillmann, D. I. (2003). Analyzing metadata for effective use and re-use. In Proceedings, Dublin Core Metadata Conference, DC-2003.

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 Fifth ACM/IEEE Joint Conference on Digital Libraries [Denver, June 7–11]. New York, Association for Computing Machinery (2005): 32-41.

Haslhofer, B., & Schandl, B. (2008). The OAI2LOD Server: Exposing OAI-PMH metadata as linked data. In International Workshop on Linked Data on the Web (LDOW2008.

Open Archives Initiative. (2008). Open Archives Initiative Protocol for Metadata Harvesting. Retrieved from: http://www.openarchives.org/OAI/openarchivesprotocol.html

Powell, A., Nilsson, M., Naeve, A., Johnson, P., & Baker, T. (2007). DCMI Abstract Model. Retrieved from http://dublincore.org/documents/abstract-model/

Powell, A. (2009). Helpful Dublin Core RDF Usage Patterns. Retrieved from: http://efoundations.typepad.com/efoundations/2009/10/a-couple-of-useful-dublin-core-rdf-usage-patterns.html

Renear, A.H., Dubin, D., Sperberg-McQueen, C.M. and Huitfeldt, C. (2002). Towards a semantics for XML markup. In R. Furuta, J. I. Maletic, and E. Munson, editors, *Proceedings of the 2002 ACM Symposium on Document Engineering*, pages 119-126, McLean, VA, November 2002. Association for Computing Machinery.

Renear, A.H., Wickett, K.M., Urban, R.J., Dubin, D., Shreeves, S.L. (2008). Collection/Item Metadata Relationships. In Proceedings of the International Conference on Dublin Core and Metadata Applications, Berlin, Germany, September 22-26, 2008.

Shreeves, S.L., Knutson, E.M., Stvilia, B., Palmer, C.L., Twidale, M.B., & Cole, T.W. (2005). Is ‘quality' metadata ‘shareable' metadata? The implications of local metadata practice on federated collections. In H.A. Thompson (Ed.) Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, April 7-10 2005, Minneapolis, MN . Chicago, IL: Association of College and Research Libraries. p.223-237.

Summers, E., Isaac, A., Redding, C., & Krech, D. (2008). LCSH, SKOS and Linked Data. In Metadata for semantic and social applications: proceedings of the International Conference on Dublin Core and Metadata Applications: Berlin, 22-26 September 2008: DC 2008: Berlin, Germany (p. 25).

Wickett, K.M., Urban, R.J., Zheng, W., Renear, A.H. (2009). A testbed approach for metadata inference rule development. Workshop On Integrating Digital Library Content with Computational Tools and Services. ACM/IEEE Joint Conference on Digital Libraries (JCDL 2009), June 2009, Austin, TX. Retrieved from: http://nema.lis.uiuc.edu/sgdlwiki/files/wickett_abstract.pdf