Panel: Aspects of Sustainability in Digital Humanities

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

Digital Humanities 2008
June 26, 2008, Oulu Finland

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.
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, Myung-Ja Han, Amy Jackson, Mark Newton , Sarah L. Shreeves, Michael Twidale, Oksana Zavalina

Slides: Renear. Last Change: June 26, 2008. *Not for circulation.*

Introduction

- Collections are designed to support particular research and scholarship.
- Toward that end *collections*, as well as their items are carefully described.
  - … purpose, subject, method of selection, spatial/temporal coverage, completeness, representativeness, summary statistical features, etc.
- This enables collections to function as more than aggregates of items.

  As intended by their creators and curators.

  (Curral, Moss & Stuart 2005; Heaney, 2000; Lagoze, et al. 2006; Palmer 2004)
- *Collection-level metadata* is thus critical to the distinctive intellectual and cultural role of collections as more than a set of individual objects.
- Unfortunately….

  collection-level metadata is poorly understood and accommodated, diminishing the value, and sustainability of digital collections.

Origins of our focus on this problem

These issues were initially raised during and IMLS project:
Digital Collections and Content (DCC) [PI: Carole L. Palmer]
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign,
Funded by the Institute for Museum and Library Services 2003-2007.
(Palmer & Knutson, 2004; Foulonneau et al. 2005; Palmer, et al. 2006)

Deliverables…

– a *collection metadata* schema
(based on RSLP CD and concurrent work on *DC Collection Application Profile*) .
– a *collection registry* for all 202 IMLS digital collections.
– an *item-level metadata repository,* harvesting 76 collections (OAI-PMH).
– an *experimental portal* for federated/metasearch. [imlsdcc.grainger.uiuc.ed].]

Amongst our research findings…

Users need collection-level information, for enhancing search and understanding

— *but this is difficult to provide.*

The New Project: DCC/CIMR

In 2007 the DCC receives a new three year IMLS grant (PI: C. Palmer).

A major deliverable:

show how a formal description of collection-level/item-level metadata relationships can help registry users *locate* and *use* digital items across multiple collections.

CIMR (Collection/Item Metadata Relationships), consists of three phases:

- Developing a logic-based framework of collection-level/item-level metadata relationships with associated inference rules.
- Conducting empirical studies to see if the conjectured framework matches the understanding and behavior of metadata specification designers, metadata creators, and collection registry users.
- Implementing pilot applications to support searching, browsing, and navigation; including RDF/OWL bindings for inference rules.

Initially focusing on the *Dublin Core Collections Application Profile* (DCCAP).

Examples of collection/item metadata relationships

Category	**Example Element**
Attribute/Value Propagation:	*marcrel:OWN*
Value Propagation:	*cld:itemType* */ dc:type*
Value Constraint: *cld:dateItemsCreated*	*/ dc:created* / temporally_within

Attribute/Value Propagation: *marcrel:OWN*

Consider the DCCAP metadata element *marcrel:OWN*…

We say that metadata elements with this behavior a/v-propagate.

An informal definition might be:

an attribute *a/v-propagates* =df

if a collection has some value for the attribute then each item in the collection has the same value for that (same) attribute.

Or, in first order logic:

An attribute A *a/v-propagates* =df

$$\forall x \forall y \forall z\ [(\mathtt{IsGatheredInto}(x,y)\ \&\ \mathtt{A}(y,z)) \supset \mathtt{A}(x,z)\ ]$$

*IsGatheredInto, from the DCMI DCCAP, represents the item/collection relationship

We assume: $\forall x \forall y\ [\mathtt{IsGatheredInto}(x,y) \supset (\mathtt{Member}(x)\ \&\ \mathtt{Collection}(y))$.

Value Propagation: *cld:itemType / dc:type*

Consider the DCMI CD AP metadata element *cld:itemType*, (refined: homogeneous collections, no repetition).

cld:itemType does **not** a/v-propagate

However…

We call this *value propagation,* or *v-propagation*

Informal definition of *v-propagation*:

an attribute *v-propagates* =df if a collection has some value for the attribute then each item in the collection has that value for some other attribute.

[Notice that a/v-propagation is a special case of v-propagation: an attribute a/v-propagates precisely when it v-propagates to itself.]

Or, in first order logic:

An attribute A v-propagates to an attribute B =df

$$\forall x \forall y \forall z\ [(\mathrm{IsGatheredInto}(x,y)\ \&\ \mathbf{A}(y,z)) \supset \mathbf{B}(x,z)\ ]$$

Value Constraints:
cld:dateItemsCreated / *dc:created* / temporally_within

Consider the DCMI collection-level attribute *cld:dateItemsCreated*

cld:dateItemsCreated does not a/v propagate

nor does it v-propagate to *dc:created*

However…

The date values for dc:created on items may not fall outside the date range given as the value for *cld:dateItemsCreated*

We refer to these cases as value constraints (or *v-constraints*)

Informally:

an attribute **A** *v-constrains* an attribute **B** with respect to constraint **C** =df if a collection has the value *z* for **A** and an item in the collection has the value *w* for **B**, then *w* is related to *z* by **C**.

In FOL:

an attribute A *v-constrains* an attribute **B** with respect to a constraint **C** =df

$$\forall x \forall y \forall z \forall w\ [(\mathrm{IsGatheredInto}(x,y)\ \&\ \mathbf{A}(y,z)\ \&\ \mathbf{B}(x,w)) \supset \mathbf{C}(w,z)]$$

NB: Propagation is *not* “inheritance”

In short,

IsGatheredInto is neither *subclass* nor *instance*

[Our use of “propagation” follows Brachman (1991)]

Sustainability … how the framework can help.

Ensuring that information will be as valuable as possible, to multiple audiences, for multiple purposes, via multiple tools, and over time

…and doing this as inexpensively, which is to say as mechanically, as possible.

When formal specifications and tools based on them are in place, collection/item metadata relationships will integrated directly into management and use.

In the mean time, sensitivity to the issues raised here can still improve matters through documentation and policies, and by informing system design.

Preliminary Guidance for Practitioners

For **metadata standards developers**:

" Metadata specifications should explicitly document the relationships between collection-level metadata and item-level metadata.

> [Though currently we do not have an adequate understanding (let alone an agreed framework) for such documentation.]

For **systems designers**:

" Information recorded at the collection level should be, when possible, 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 *conceptualize* it; relational tables may remain in normal forms.]

" Information that does not propagate without loss must, to the fullest extent possible, be made evident and available to users working with multiple collections.

> [We have no idea how to do this systematically]

For **collection managers**:

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

" Representations based on propagation must be repeated as new objects are added or removed, and new information about objects and collections becomes available.

Research…

Topics ahead

Some questions emerge…

- how many relationship categories are there?
- which metadata attributes fall into which categories?
- how expressive a logic is needed to express propagation rules?
  - how much of first order logic?
  - what extensions? (modal, default, …?)
  - what are the consequences for computational tractability?
- when does propagation convert information without loss?
- what about propagation from items to collections?

One result: Finishing the job requires modal logic (*at least*)

E.g., our current definition of a/v propagates

…

An attribute A *a/v-propagates* =df

I. a) □ yz□□ [Collection(y) & A(y,z)] &

b) b xz□□ [Member(x) & ~A(x,z)] &

c) c xyz□□□ [A(x,z) & ~A(y,z)] &

II. I xyz∀∀∀ [(IsGatheredInto(x,y) & A(y,z)) ⊃A(x,z)].

See: The return of the trivial: Formalizing collection/item metadata relationships. Renear, A.H., Wickett, K.M., Urban, R.J., and Dubin, D. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM Press, New York 2008.

And most importantly: Non-Reducible Collection Attributes

There are “mission-critical” collection-level attributes that both

(i) resist any conversion and

(ii) clearly result in loss of important information if discarded.

Examples are metadata indicating that a collection

- is complete or incomplete, large or small
- is representative of a period or style
- is developed according to some systematic method
- is heterogeneous with respect to genre or type of object, etc.
- is representative (in some respect) of a domain
- was developed according to some particular method
- was designed for some particular purpose
- has certain summary statistical features …. *and so on*.

Such information is tightly tied to the distinctive role for which a collection is intended to play in the support of research and scholarship.

If is lost or inaccessible, the collection cannot be useful, as a collection, in the way originally intended by its creators

[Last Slide]

In a slogan…

Not only sustaining collections *in sensu diviso*,

…. but sustaining collections *in sensu composito*

Questions?

For additional background: Collection/Item Metadata Relationships, A. H. Renear, K. Wickett, R. Urban, D. Dubin. Forthcoming in the *Proceedings of the DCMI Dublin Core Conference*, Berlin Germany, September 2008.

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).
- Renear, A. H., K. Wickett, R. Urban, D. Dubin. (Forthcoming). In: *Proceedings of the DCMI Dublin Core Conference*, Berlin Germany, September 2008.
- Renear, A.H., Wickett, K.M., Urban, R.J., and Dubin, D. (2008). The return of the trivial: Formalizing collection/item metadata relationships. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*. ACM Press, New York 2008.
- 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