Collection/Item Metadata Relationships

Allen H. Renear, Karen M. Wickett, Richard J. Urban, David Dubin, Sarah L. Shreeves
Center for Research in Information and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

DC 2008: International Conference on Dublin Core and Metadata Applications
September 24, 2008; Berlin, Germany

Why collection-level metadata is important

- Collections are designed to support research and scholarship.

- Toward this end collection descriptions indicate such things as:
  - *purpose*
  - *subject*
  - *method of selection*
  - *spatial/temporal coverage*
  - *completeness*
  - *representativeness*
  - *summary statistical features*

    …etc.

- These descriptions enable collections to function as more than simply aggregates of items,
  - as intended by their creators and curators
  - as required by their users

But unfortunately….

Collection-level metadata is poorly understood and accommodated

Most retrieval systems flatten the world, ignoring collection context

Retrieval systems that do use metadata use only item-level metadata

Even simple discovery is impeded:

If the *owner* of a collection is indicated only at the collection-level, then retrieval accessing only item-level metadata…

— cannot usefully process queries constrained by owner

— cannot display the owner of item in the result set

Origins of our focus on this problem: DCC

IMLS Digital Collections and Content

University of Illinois at Urbana-Champaign
Grainger Library &
Graduate School of Library and Information Science
Funded by IMLS, 2003-2007
Timothy Cole, Principal Investigator
Carole L. Palmer, Sarah L. Shreeves, Michael B. Twidale, Co-Investigators

Deliverables…

- a *collection metadata* schema
  Based on RSLP CD and concurrent work on *DC Collection Application Profile*.
- a *collection-level metadata registry*
  for 202 IMLS digital collections.
- an *item-level metadata repository*
  76 collections harvested using OAI-PMH.
- an *experimental portal* for searching aggregated metadata.
  [http://imlsdcc.grainger.uiuc.edu]

Among the research findings:

Users need collection-level information, for discovery and understanding

(Palmer & Knutson, 2004;
Foulonneau et al. 2005;
Palmer, et al. 2006)

But what information?

And how to provide it?

So we included this problem in our next IMLS proposal…

Climax Miners, Leadville, CO. Courtesy Colorado School of Mines

The new project

In 2007 the DCC received a new three year IMLS grant

Carole L. Palmer, Principal Investigator
Timothy Cole, Allen H. Renear, Michael B. Twidale, Co-Investigators

A major deliverable:

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

CIMR: Collection/Item Metadata Relationships

Three phases:

1) Develop a logic-based framework of collection/item metadata relationships and inference rules.
2) Conduct empirical studies to see if the framework matches the behavior of metadata specification designers, metadata creators, and registry users.
3) Implement pilot applications to support searching, browsing, and navigation; including RDF/OWL formulations and inference rules.

Our initial focus is on the *Dublin Core Collections Application Profile (DCCAP).*

Where we are now

Phase 1:

Develop a logic-based framework of collection/item metadata relationships and inference rules.

The next few slides…
three simple examples of collection/item metadata relationships

Attribute/Value Propagation: *marcrel:OWN*

Consider the DCCAP metadata element **marcrel:OWN**…

Plausibly: whoever owns a collection owns each of its items

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

Informal definition

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 attribute.

Or, in first order logic:

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

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

[IsGatheredInto(x,y) is adapted from from the DCMI DCCAP.]

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

Consider the DCCAP metadata element **cld:itemType**.*

*a refinement, assuming homogeneous collections and no repetition of elements.

cld:itemType* does not a/v-propagate…

However,

if a collection has a value for cld:itemType* then
each of its items has the same value for *dc:type.*

We call this *v-propagation.*

Informal definition

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.

Or, in first order logic:

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

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

Value Constraints: *cld:dateItemsCreated / dcterms:created*

cld:dateItemsCreated* does *not* a/v propagate

nor does it v-propagate to dcterms:created

However,

if a collection has a temporal range for cld:dateItemsCreated*, then its items may not have values for dcterms:created that fall outside that range.

this is a *constraint*: the value of dcterms:created must be *temporally-within* the range given by cld:dateItemsCreated*

Informal Definition

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 first order logic:

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

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

How will the framework help?

- *Metadata specification developers* use the framework to classify metadata elements in their specifications.
- *Metadata librarians* use these classifications to confirm their understanding of the metadata elements they are assigning.
- *Software architects* use these classifications to guide the configuration of inferencing features in retrieval systems.

What is missing?

A completed shared framework

... a project for the community

University of Washington Libraries, Special Collections Division. PH Coll 548

Prior work? Of course.

- Relationships such as those just described have been studied elsewhere — which is a good thing.
- However as far as we know no one has focused on the *IsGatheredInto* relationship.

Some research questions

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

One result: Finishing the job requires modal logic

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

I. a) $\Diamond\, \exists y \exists z\, [\mathrm{Collection}(y) \;\&\; \mathrm{A}(y,z)] \;\&$

b) $\Diamond\, \exists x \exists z\, [\mathrm{Member}(x) \;\&\; {\sim}\mathrm{A}(x,z)] \;\&$

c) $\Diamond\, \exists x \exists y \exists z\, [\mathrm{A}(x,z) \;\&\; {\sim}\mathrm{A}(y,z)] \;\&$

II. $\Box\, \forall x \forall y \forall z\, [(\mathrm{IsGatheredInto}(x,y) \;\&\; \mathrm{A}(y,z)\,) \supset \mathrm{A}(x,z)\,].$

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.

Most importantly: Non-Reducible Collection Attributes

- Some vital collection-level attributes resist conversion to item-level attributes

- Examples are metadata indicating that a collection

  -- is complete or incomplete
  -- is representative (in some respect)
  -- is heterogeneous with respect to genre or type of object, etc.
  -- was developed according to some particular method
  -- was designed for some particular purpose
  -- has certain summary statistical features

  …. *and so on*.

- These are tightly tied to the distinctive role a collection is intended to play in the support of research and scholarship.

- If this information is inaccessible, the collection cannot be useful, as a collection, in the way originally intended by its creators.

Questions?

We are just getting started and welcome comments and advice.

Acknowledgements

- This research is supported by:
  The Institute of Museum and Library Services,
  a federal agency that fosters innovation, leadership, and a lifetime of learning.
  National Leadership Grant for Research & Demonstration
  Carole L. Palmer, Principal Investigator
- Hosted by the:
  Center for Informatics Research in Science and Scholarship
  Graduate School of Library & Information Science
  University of Illinois at Urbana-Champaign
- Project documentation: http://imlsdcc.grainger.uiuc.edu
- We have benefited from many discussions with other DCC/CIMR project members and with participants in the IMLS DCC Metadata Roundtable, including: Thomas Dousa, Myung-Ja Han, Amy Jackson, Mark Newton, Oksana Zavalina, Wu Zheng.

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

INSTITUTE *of* Museum and Library SERVICES

References

Arms, W.Y. Dushay, N., Fulker, D. & Lagoze, C. (2003). A case study in metadata harvesting: the NSDL. *Library Hi Tech*, 21(2), pp. 228–237.

Brachman, R. J. (1983). What ISA is and isn't: An analysis of taxonomic links in Semantic Networks. *IEEE Computer*, 16 (10), pp. 30-6.

Brachman R. J. et al. (1991). "Living With Classic: When and how to use a KL-ONE-like language", in *Principles of Semantic Networks: Explorations in the Representation of Knowledge,* ed. John F. Sowa, Morgan Kaufman, pp. 401-456.

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*. California Digial Library. http://www.cdlib.org/.

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, June 20-21, Washington, DC. Digital Library Federation.

DCMI. (2007). Dublin Core Collections Application Profile. http://dublincore.org/ Retrieved April 13, 2008,

Dushay, N. & Hillmann, D.I. (2003). Analyzing metadata for effective use and re–use. *DC–2003: Proceedings of the International DCMI Metadata Conference and Workshop*, [United States]: Dublin Core Metadata Initiative, pp. 161–170.

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

Gasser, L. & Stvilia, B. (2001). *A new framework for information quality*. Technical report ISRN UIUCLIS--2001/1+AMAS. Champaign, Ill.: University of Illinois at Urbana Champaign.

Guarino, N. & Welty, C. (2004). An overview of OntoClean. S. Staab and R. Studer, eds, *The Handbook on Ontologies*. Springer.

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

Hutt, A. & Riley, J. (2005). Semantics and Syntax of Dublin Core Usage in Open Archives Initiative Data Providers of Cultural Heritage Materials. *Proceedings of the 5th ACM/IEEE–CS Joint Conference on Digital Libraries, Denver, Colo. (June 7–11 June)*. New York: ACM Press, pp. 262–270.

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

Lalmas, M. (1998). Logical models in information retrieval. *Information Processing and Management*. 34, 1.

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

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

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

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

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

Renear, A. H., Urban, R., Wickett, K., Palmer, C.L., & Dubin, D. (2008a). Sustaining collection value: Managing collection/item metadata relationships. Proceedings of the Digital Humanities conference, 25-29 June 2008, Oulu, Finland

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

Sebastiani, F. (1998). On the role of logic in information retrieval, *Information Processing and Management* 34, 1.

Shreeves, S., Knutson, E., Stilva, B., et al. (2005). Is 'Quality' Metadata, 'Shareable' Metadata? The Implications of local metadata practices for 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. pp. 223-237.

Stvilia, B., Gasser, L., Twidale, M., Shreeves, S.L. & Cole, T.W. (2004). Metadata quality for federated collections. *Proceedings of ICIQ04—9th International Conference on Information Quality*. Cambridge, MA: 111–25.

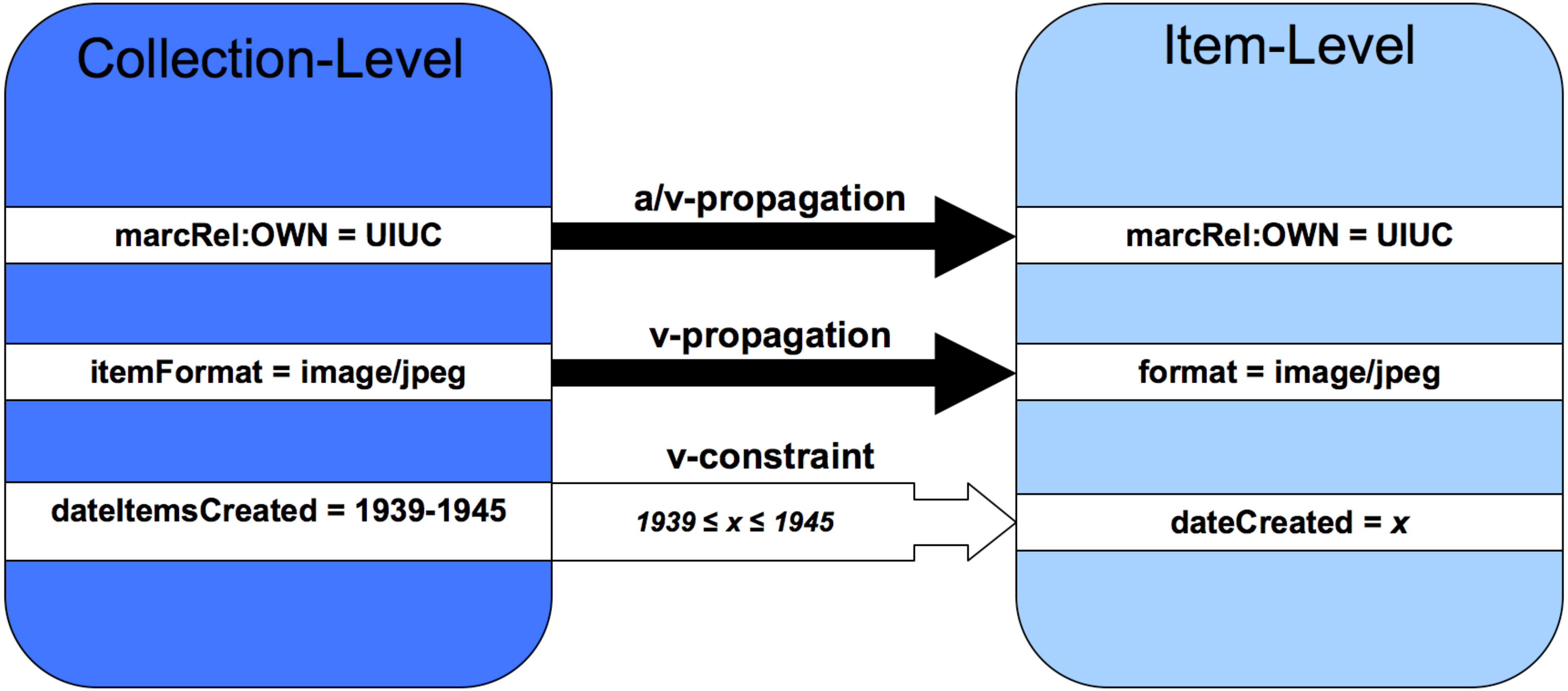
van Rijsbergen, C. J. (1986). A non-classical logic for information retrieval. *The Computer Journal* 29,6.

Warner, S., Bekaert, J., Lagoze, C., Lin, X., Payette, S., & Van de Sompel, H. (2007). Pathways: Augmenting interoperability across scholarly repositories. *International Journal on Digital Libraries*.

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.

Woods, W. (1975). What's in a link: Foundations for semantic networks. *Representation and Understanding*, D. Bobrow and A. Collins (eds.), Academic Press.

Examples of collection/item metadata relationships



Collection-Level

marcRel:OWN = UIUC

itemFormat = image/jpeg

dateItemsCreated = 1939-1945

a/v-propagation

v-propagation

v-constraint

1939 ≤ x ≤ 1945

Item-Level

marcRel:OWN = UIUC

format = image/jpeg

dateCreated = *x*

But unfortunately….

Collection-level metadata is poorly understood and accommodated

Cabinet Photograph of Lincoln Home Parlor. Courtesy Lincoln Home Historic Site.

NB: Propagation is *not* “inheritance”

IsGatheredInto

is neither

subclassOf

nor

instanceOf

Our use of “propagation” follows Brachman (1991)

Table Showing Contraband Items. Colorado State Penitentiary.
Courtesy Cañon City Public Library