ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE

# Rule Categories for Collection/Item Metadata Relationships (CIMR)

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

ASIS&T 2010 Annual Meeting
October 25, 2010

CIRSS

INSTITUTE of Museum and Library SERVICES

Supported by a 2007 IMLS NLG Research & Demonstration grant

# Why collection-level metadata is important

- Collections are designed to support research and scholarship.
- Toward this end collection descriptions indicate:
  - *purpose*
  - *subject*
  - *method of selection*
  - *spatial/temporal coverage*
  - *completeness*
  - *representativeness*
  - *summary statistical features*, etc.
- Allowing collections to function as more than aggregates of items,
  - as intended by their creators and curators
  - as required by their users

# Unfortunately….

Collection-level metadata is poorly understood and accommodated

*For instance:*

Most retrieval systems ignore collection context

Retrieval systems that do use metadata only use 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 an item retrieved

The problems and limitations go beyond retrieval…

# CIMR Origins

A finding from the first IMLS DCC Project (2001-2007)

*Users need collection-level information, for discovery and understanding*

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

A deliverable for the second IMLS DCC Project (2007-2010)

"(C) Analyze relationships between collection-level metadata and item-level metadata … to better preserve context and enhance the functionality…".

# CIMR agenda

To improve our understanding of the semantics of collection and item metadata.

*By:*

- Providing a framework of rule categories for reasoning about collection-level and item-level descriptions.

- Exploring how to test the framework against available descriptions.

# Reasoning about ownership

marcrel:OWN

> "The person or organization that currently owns an item or collection."

We might expect that if someone owns a collection, then they own every item in that collection.

We can formalize this relationship with a rule:

$$\forall y \forall z((Collection(y) \,\&\, ownedBy(y,z)) \supset \forall x(isGatheredInto(x,y) \supset ownedBy(x,z)))$$

# Reasoning about type

*cld:itemType*

"the nature or genre of one or more items in the collection."

Unlike *marcrel:own*, which can be had by items and collections, *cld:itemType* can only be applied to collections.

The attribute at the item level is "the nature or genre of ... items", or *dc:type*.

Also, the rule will refer to *one or more* items instead of *every* item.

$$\forall y \forall z((Collection(y) \,\&\, itemType(y,z)) \supset \exists x(isGatheredInto(x,y) \,\&\, type(x,z)))$$

# Reasoning about date attributes

*cld:dateItemsCreated:*

"A range of dates over which the individual items within the collection were created."

We expect a rule linking *cld:dateItemsCreated* to an item-level date attribute like *dc:date*.

We also expect that date values will fall *within* the range indicated for the collection.

e.g. given "1850-1899" for a collection, we would expect items to have dates that fall within that range.

$$\forall y \forall z((Collection(y) \,\&\, dateItemsCreated(y,z)) \supset$$
$$\forall x(isGatheredInto(x,y) \supset \exists w(date(x,w) \,\&\, temporalWithin(w,z))))$$

# Rule Categories

We are interested in rules based on the *is gathered into* relationship between items and collections.

These are *propagation rules.* (*Propagation is not inheritance.*)

Determining the rules that govern metadata is an empirical matter, but rules can be classified according to their logical form.

This is what our framework does.

Framework structure:

- Top-level division: item-level quantification
- Further division: attribute conditions
- Further division: value conditions

# Item-Level Quantification

A *universal* propagation (UP) rule implies that something is true of *every* member of a collection.

Attributes $\mathrm{A}$ and $\mathrm{B}$ propagate universally iff

if a collection $y$ has the value $z$ for the attribute $\mathrm{A}$, then **every item** in the collection has some value $w$ for the attribute $\mathrm{B}$ such that $w$ is related to $z$ by the constraint $\mathrm{C}$.

An *existential* propagation (EP) rule implies that something is true of *at least one* member of a collection.

Attributes $\mathrm{A}$ and $\mathrm{B}$ propagate existentially iff

if a collection $y$ has the value for the attribute $\mathrm{A}$, then there is **some item** in the collection that has some value $w$ for the attribute $\mathrm{B}$ such that $w$ is related to $z$ by the constraint $\mathrm{C}$.

# Attribute Conditions

An *attribute propagation* (AP) rule connects an attribute at the collection level to the **same** attribute at the item level.

- e.g. *marcrel:OWN*

An *attribute differentiation* (AD) rule connects an attribute at the collection level to a **different** attribute at the item level.

- e.g. *cld:itemType* and *dc:type*

# Value Conditions

A *value propagation* (VP) rule implies that we will see the **same** value at the collection and item levels for the related attributes.

e.g. *cld:itemType* and *dc:type*

A *value constraint* (VC) rule implies that the values at the collection and item levels will be **related by a constraint**.

e.g. *cld:dateItemsCreated* and *dc:date*

values related by *temporal within* constraint.

# The categories

**Quantification Categories:**

UP: Attributes A, B propagate *universally* =df

$$\forall y \forall z((A(y,z) \,\&\, Collection(y)) \supset \forall x(isGatheredInto(x,y) \supset \exists w(B(x,w) \,\&\, C(x,z))))$$

EP: Attributes A, B propagate *existentially* =df

$$\forall y \forall z((A(y,z) \,\&\, Collection(y)) \supset \exists x(isGatheredInto(x,y) \,\&\, \exists w(B(x,w) \,\&\, C(x,z))))$$

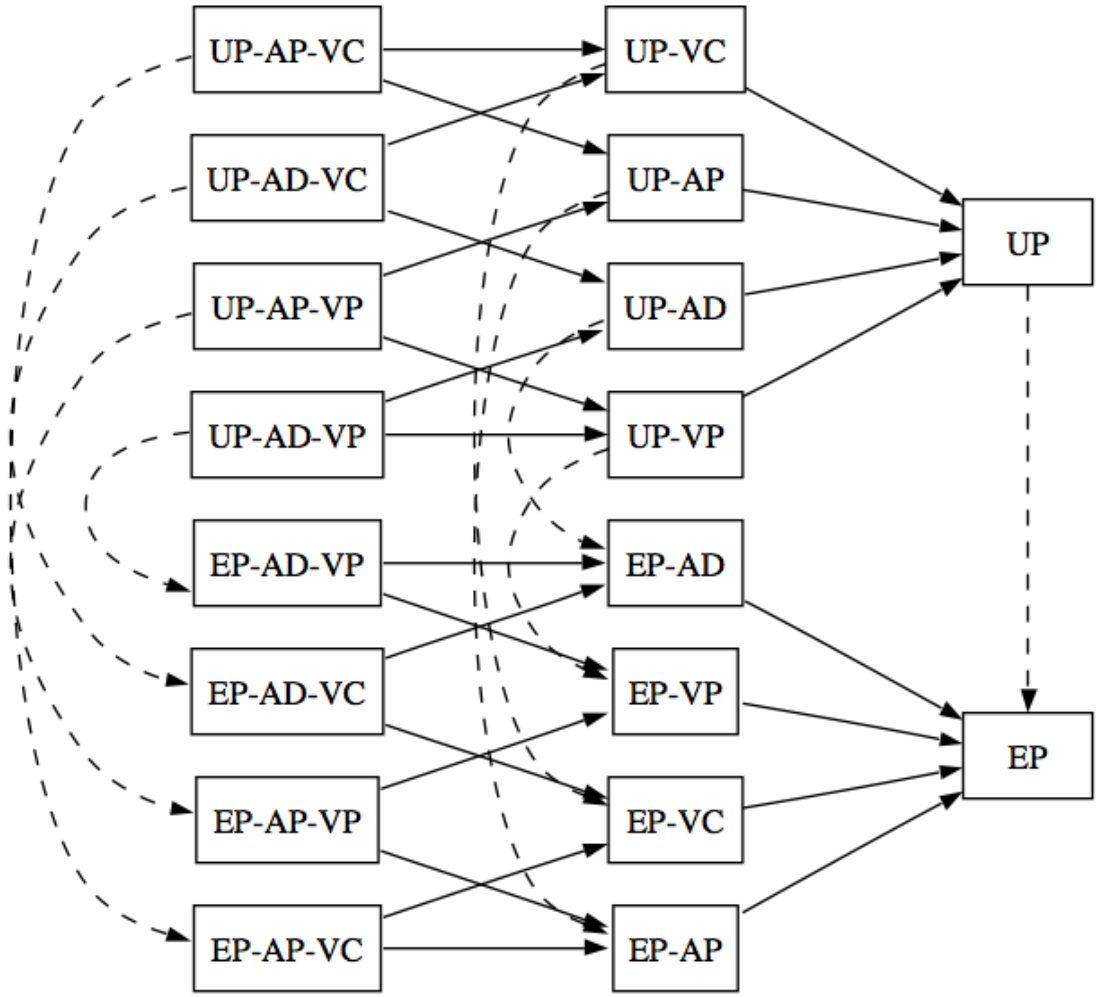
**Attribute Conditions:**

| | |
|---|---|
| A=B: | attribute propagation [AP] |
| ~(A=B): | attribute differentiation [AD] |

**Value Conditions:**

| | |
|---|---|
| C(x,y) ≡ x=y: | value propagation [VP] |
| ~[C(x,y) ≡ x=y]: | value constraint [VC] |

# Relationships between categories

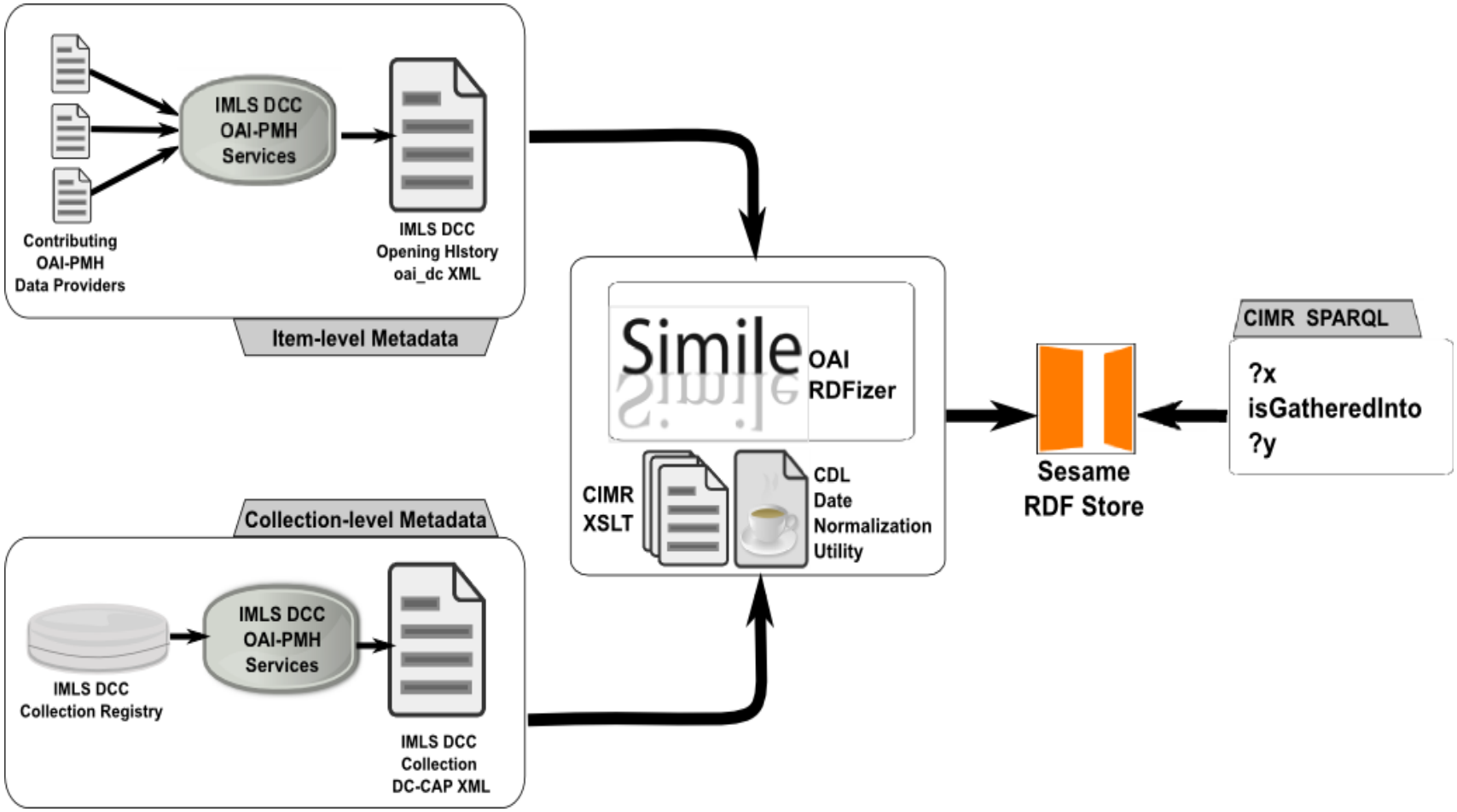



[dotted arrow]: implies, by universal instantiation existential generalization
(assuming no empty collections)
[solid arrow]: implies, by subtracting a specialization condition

# The CIMR testbed

To explore testing these rules against actual metadata, we built an RDF respository.

RDF was a natural choice since it is based in first order logic.



# Confirmation and refutation issues

We hoped to confirm or refute the conditional rules by searching the testbed for counterexamples.

However this is not as simple as it sounds, for several reasons.

*One important difficulty:*

- A counterexample to a conditional is a case where the antecedent is true and the consequent is false.
- What would this look like in an RDF repository?
  - lack of negation in RDF
  - open world assumption for semantic web
- Refutation therefore is only possible after adding additional constraints, drawn from an analysis of metadata schemas or commonsense knowledge.

  [e.g., rules implying nothing can have both TIFF and JPG as a value for dc:type]

# The categories

**Quantification Categories:**

UP: Attributes A, B propagate *universally* =df

$$\forall y \forall z((A(y,z) \,\&\, Collection(y)) \supset \forall x(isGatheredInto(x,y) \supset \exists w(B(x,w) \,\&\, C(x,z))))$$

EP: Attributes A, B propagate *existentially* =df

$$\forall y \forall z((A(y,z) \,\&\, Collection(y)) \supset \exists x(isGatheredInto(x,y) \,\&\, \exists w(B(x,w) \,\&\, C(x,z))))$$

**Attribute Conditions:**

| | |
|---|---|
| A=B: | attribute propagation [AP] |
| ~(A=B): | attribute differentiation [AD] |

**Value Conditions:**

| | |
|---|---|
| C(x,y) ≡ x=y: | value propagation [VP] |
| ~[C(x,y) ≡ x=y]: | value constraint [VC] |

# Generating candidate rules

Language rules -- *dc:language* and *dc:language*

- universal or existential quantification
- 2 candidate rules

Type rules -- *cld:itemType* and *dc:type*

- universal or existential quantification
- generalization, specialization, or equality between values
- combinations of value relationships
- 14 rules

Date rules – *dcterms:temporal* and *dc:date*

- universal or existential quantification
- temporal containment, comprehension, overlap, or equality between values
- combinations of value relationships
- 30 rules

# Rules confirmed

Language:

existential attribute/value propagation

$$\forall y \forall z((language(y,z) \,\&\, Collection(y)) \supset \exists x(isGatheredInto(x,y) \,\&\, language(x,z)))$$

Type

existential attribute differentiation with value constraint

$$\forall y \forall z((itemType(y,z) \,\&\, Collection(y)) \supset \exists x(isGatheredInto(x,y) \,\&\, \exists w(type(x,z) \,\&\, generalizes(w,z))))$$

Date

existential attribute differentiation with value constraint

$$\forall y \forall z((temporal(y,z) \,\&\, Collection(y)) \supset \exists x(isGatheredInto(x,y) \,\&\, \exists w(date(x,z) \,\&\, temporalWithin(w,z))))$$

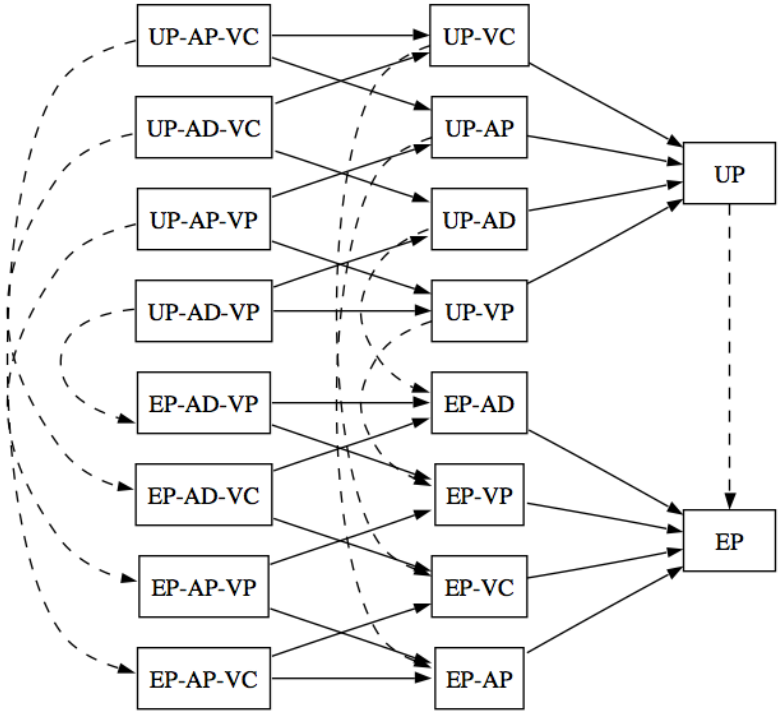
# Concluding remarks

A systematic understanding of collection/item metadata rules will provide support for improved functionality and information management

… and, by elucidating the nature of *isGatheredInto*, bring us a little closer to understanding what collections really are.

# Thank you

- Questions?



[dotted arrow]: implies, by universal instantiation existential generalization
(assuming no empty collections)
[solid arrow]: implies, by subtracting a specialization condition


- This research is supported by a 2007 IMLS NLG Research & Demonstration grant as part of the IMLS Digital Collections and Content project, Principal Investigator, Carole L. Palmer, Center for Informatics Research in Science and Scholarship (CIRSS). Project documentation is available at http://imlsdcc.grainger.uiuc.edu/about.asp.
- Thanks to Wu Zheng, Larry Jackson, Katrina Fenlon, Jacob Jett, Amit Kumar, Tim Cole, Thomas Dousa, Dave Dubin, Myung-Ja Han, Mark Newton, Sarah Shreeves, Michael Twidale, and Oksana Zavalina


INSTITUTE of Museum and Library SERVICES