Data Modeling 201

Building Models and Profiles with PCDM

bit.ly/C4LDataModeling201
Your Fearless Facilitators

Esmé Cowles
Hydra + Fedora + PCDM developer, Princeton, @escowles

Christina Harlow
Works with metadata somewhere in the world, @cm_harlow

Mark Matienzo
Collaboration & Interoperability Architect, Stanford @anarchivist

Steve Van Tuyl
Digital Repository Librarian at Oregon State, @badgerbouse

bit.ly/C4LDataModeling201
Link to Slides, Notes, Examples, Resources, and Other Workshop Materials
Communication Channels

- Alert a facilitator if you need help or have questions
- Code4Lib Slack: #c4l17-datamodeling channel
- Information about Code4Lib Slack: http://goo.gl/forms/p9Ayz93DgG

bit.ly/C4LDataModeling201
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:30-13:45</td>
<td>Introduction</td>
</tr>
<tr>
<td>13:45-14:15</td>
<td>Advanced Applied Data Modeling</td>
</tr>
<tr>
<td>14:15-14:35</td>
<td>State of PCDM &amp; Implementations</td>
</tr>
<tr>
<td>14:35-14:50</td>
<td>Break</td>
</tr>
<tr>
<td>14:50-15:00</td>
<td>Setup for breakout groups</td>
</tr>
<tr>
<td>15:00-16:10</td>
<td>Breakout groups: Data Modeling Hard Cases</td>
</tr>
<tr>
<td>16:10-16:30</td>
<td>Supporting &amp; Moving Work Forward</td>
</tr>
</tbody>
</table>

bit.ly/C4LDataModeling201
Our Expectations of You

- Follow the Code4Lib Code of Conduct
- Follow the Recurse Center Social Rules (a.k.a. "Hacker School Rules")
- Be ready to work on data models!
Code4Lib

Code of Conduct

http://2017.code4lib.org/conduct/

bit.ly/C4LDataModeling201
Recurse Center Social Rules
(a.k.a. Hacker School Rules)

https://www.recurse.com/manual#sub-sec-social-rules

- No feigning surprise
- No well-actually's
- No back-seat driving
- No subtle -isms
  - More info: https://www.recurse.com/blog/38-subtle-isms-at-hacker-school

bit.ly/C4LDataModeling201
Are you ready to data model?

bit.ly/C4LDataModeling201
Reminder

This is an informal workshop – ask questions and let facilitators know how we can help you

bit.ly/C4LDataModeling201
This workshop is an attempt to

- Create examples and models for digital objects using the Portland Common Data Model in a collaborative and hands-on fashion
- Solicit types of objects in need of data modeling
- Produce examples, documentation, model extensions and work that will be shared with the PCDM community

bit.ly/C4LDataModeling201
Our goals for this workshop

● Expand understanding, usage and examples of PCDM work explicitly, RDF modeling for repository resources generally
● Give participants hands-on experience modeling to take back to their day-to-day work
● Involve more people from the Code4Lib Community in PCDM development efforts

bit.ly/C4LDataModeling201
Now: your goals for this workshop?

- Why are you attending this workshop?
- What are your goals - immediate or long-term?
- What's your level of comfort and experience with data modeling?

We want to capture your goals, return to them throughout the workshop & going forward - feel free to add to responses to the shared notes.

bit.ly/C4LDataModeling201
Advanced Data Modeling

bit.ly/C4LDataModeling201
“Advanced” Data Modeling
What is a Model?

“When we want to make resources and their metadata available in a structured manner on the web, we first need to decide what characteristics of theirs are the most important to be represented. By doing so, we make an abstraction of the reality through the development of a model.”

- Linked Data for Libraries, Archives & Museums, p. 12

bit.ly/C4LDataModeling101
Open World Assumption

- Closed World Assumption (CWA) is the assumption that what is not known to be true must be false.
- Open World Assumption (OWA) is the opposite. In other words, it is the assumption that what is not known to be true is simply unknown.
- Our Global Knowledge Graph is OWA, i.e., incomplete.
- Where in Memory Institutions do we need OWA as opposed to CWA?
  - Patrons data
  - Collection Resources
  - Authorities
## RDF & RDFS

<table>
<thead>
<tr>
<th>Construct</th>
<th>Syntactic form</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Class</strong> (a class)</td>
<td>C rdf:type rdfs:Class</td>
<td>C (a resource) is an RDF class</td>
</tr>
<tr>
<td><strong>Property</strong> (a class)</td>
<td>P rdf:type rdfs:Property</td>
<td>P (a resource) is an RDF property</td>
</tr>
<tr>
<td><strong>type</strong> (a property)</td>
<td>I rdf:type C</td>
<td>I (a resource) is an instance of C (a class)</td>
</tr>
<tr>
<td><strong>subClassOf</strong> (a property)</td>
<td>C1 rdfs:subClassOf C2</td>
<td>C1 (a class) is a subclass of C2 (a class)</td>
</tr>
<tr>
<td><strong>subPropertyOf</strong> (a property)</td>
<td>P1 rdfs:subPropertyOf P2</td>
<td>P1 (a property) is a sub-property of P2 (a property)</td>
</tr>
<tr>
<td><strong>domain</strong> (a property)</td>
<td>P rdfs:domain C</td>
<td>domain of P (a property) is C (a class)</td>
</tr>
<tr>
<td><strong>range</strong> (a property)</td>
<td>P rdfs:range C</td>
<td>range of P (a property) is C (a class)</td>
</tr>
</tbody>
</table>

More on the RDF Data Model

- URI and IRI concepts
  - Used to reference resources unambiguously

- Literals
  - Describe data values with no clear identity like "100 km/h"
  - Literals may never be the origin of a node of an RDF graph
  - Edges may never be labeled with literals

- Blank nodes
  - Facilitate existential quantification for an individual with certain properties without naming it

- Lists
  - Container: adding new elements possible, ordered & unordered
  - Collections: ordered list; adding new elements impossible
Prince Example
RDF, RDFS, … OWL (Web Ontology Language)

- Discussed RDF & RDFS this morning
- OWL is based on description logics, a family of logics that are decidable fragments of first-order predicate logic.
- OWL includes
  - Individuals
  - Subclasses & Subproperties
  - Class Constructors
  - Property Chain Axioms
  - Property Characteristics (inverse, disjoint, symmetry, …)
  - Punning
LD4L Ontology Example
PROV

- PROV-O encodes PROV Data Model in OWL2
- Set of classes, properties, and restrictions that can be used to represent provenance information.
- Can also be specialized to create new classes and properties for modeling provenance information specific to different domain applications
PROV Core Data Model

https://www.w3.org/TR/prov-dm/
OAI-ORE (OAI Object Reuse and Exchange)

- Managed by Open Archives Initiative, the creators of OAI-PMH
- Started in 2006
- Generated to “Augment Interoperability”, i.e.:

“Develop, identify, and profile extensible standards and protocols to allow repositories, agents, and services to interoperate in the context of use and reuse of compound digital objects.”
OAI-ORE (OAI Object Reuse and Exchange)

https://www.openarchives.org/ore/1.0/datamodel
Tools for Data Modeling in RDF

- Protégé & Webprotégé
- TopBraid
- StarDog
- OntoStudio
- & diagramming interfaces like yEd with Graphoo

Querying RDF

- SPARQL - Getty SPARQL Endpoint for a demo
- LDPath
Conversions & Mappings

- RML / RDF Mapping Language and Convertor Engine
- Catmandu
- BF Convertor(s)

Validation & Expectations

- ShEx (Shapes Expressions)
  - ShEx Tester
- SHACL (Shapes Constraint Language)
Graph Stores & Triple Stores

- Fedora 4 (Graph Store)
- Cavendish
- Apache Cassandra
- Fuseki
- SDB/TDB
- Blazegraph
- Jena
- ...
State of PCDM & Implementations

bit.ly/C4LDataModeling201
Implementations

- Hydra
  - Hydra::PCDM, Hydra::Works, CurationConcerns & Sufia 7
  - CurationConcerns and Sufia are merging into Hyrax
  - Hyrax 1.0: migration path for Sufia 7.x
  - Hyrax 2.0: migration path for CC 1.x/2.x

- Islandora

- Others
Evolving data models

- Works extension
  - Used by Hydra implementation
  - But not really the current thinking

- FileSets
  - Also used by the Hydra implementation (and being built upon)
  - Not embraced by the rest of the community

- (Top)Range
  - From IIIF
  - Logical vs. physical structure
Evolving data models
Community

- PCDM Wiki
  - [https://github.com/duraspace/pcdm/wiki](https://github.com/duraspace/pcdm/wiki)
  - Profiles

- Mailing list
  - [http://groups.google.com/group/pcdm](http://groups.google.com/group/pcdm)

- Monthly calls
  - [https://github.com/duraspace/pcdm/wiki/PCDM-Community-Meetings](https://github.com/duraspace/pcdm/wiki/PCDM-Community-Meetings)

- Workshops
15 Minute Break
Reconvene at 2:40 PM
## Supporting Interoperability via “Profiles”

<table>
<thead>
<tr>
<th>Field Name</th>
<th>AF Model</th>
<th>Subject (can be iterated)</th>
<th>Domain</th>
<th>Predicate</th>
<th>Range</th>
<th>Obligation</th>
<th>&quot;Concept&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection Abstract</td>
<td>BasicMetadata</td>
<td>Digital Collection</td>
<td>HydraWorks:Collection</td>
<td>dcterm:abstract</td>
<td>literal</td>
<td>{0,1}</td>
<td>abstract</td>
</tr>
<tr>
<td>Collection Date</td>
<td>BasicMetadata</td>
<td>Digital Collection</td>
<td>HydraWorks:Collection</td>
<td>dcterm:date</td>
<td>literal (EDTF)</td>
<td>{0,1}</td>
<td>date</td>
</tr>
<tr>
<td>Collection Identifier</td>
<td>models/collection.rb</td>
<td>Digital Collection</td>
<td>HydraWorks:Collection</td>
<td>dcterm:identifier</td>
<td>literal</td>
<td>{1,n}</td>
<td>identifier</td>
</tr>
<tr>
<td>Collection Publisher</td>
<td>BasicMetadata</td>
<td>Digital Collection</td>
<td>HydraWorks:Collection</td>
<td>dc:publisher</td>
<td>literal</td>
<td>{0,n}</td>
<td>publisher</td>
</tr>
<tr>
<td>Collection Publisher URI</td>
<td>BasicMetadata</td>
<td>Digital Collection</td>
<td>HydraWorks:Collection</td>
<td>dcterm:identifier</td>
<td>URI &lt; dcterm:Agent</td>
<td>{0,n}</td>
<td>publisher_URI</td>
</tr>
<tr>
<td>Collection Related URL</td>
<td>models/collection.rb</td>
<td>Digital Collection</td>
<td>HydraWorks:Collection</td>
<td>dcterm:relation</td>
<td>URL</td>
<td>{0,n}</td>
<td>relatedURL</td>
</tr>
<tr>
<td>Collection Title</td>
<td>RequiredMetadata</td>
<td>Digital Collection</td>
<td>HydraWorks:Collection</td>
<td>dcterm:title</td>
<td>literal</td>
<td>{1,1}</td>
<td>title</td>
</tr>
<tr>
<td>Collection Subject</td>
<td>PROPOSED</td>
<td>Digital Collection</td>
<td>HydraWorks:Collection</td>
<td>dc:subject</td>
<td>literal</td>
<td>{0,n}</td>
<td>subject</td>
</tr>
<tr>
<td>Collection Subject URI</td>
<td>PROPOSED</td>
<td>Digital Collection</td>
<td>HydraWorks:Collection</td>
<td>dcterm:subject</td>
<td>URI</td>
<td>{0,n}</td>
<td>subject_URI</td>
</tr>
<tr>
<td>Collection Curator</td>
<td>PROPOSED</td>
<td>Digital Collection</td>
<td>HydraWorks:Collection</td>
<td>rdau:P60376</td>
<td>literal for now</td>
<td>{0,n}</td>
<td>curator</td>
</tr>
</tbody>
</table>
Breakouts
3 Parts to Our Breakouts

Break into 6-ish groups and...

Part 1 (20 minutes): Create a Model Profile for your group’s assigned object - Use papers, markers, Google Drive, etc.

Part 2 (20 minutes): Create Profiles/Models/Examples/Mappings for your group’s choice (see those provided [here](#))

Part 3 (20 minutes): Everyone Together! Compare and discuss models for the two assigned objects
Shared Objects

https://www.discogs.com/Dorothy-Collins-Experiment-Songs-From-Ballads-For-The-Age-Of-Science/release/1628463

http://sinaipalimpsests.org/about-project
Resources

● Definitions: http://pcdm.org/2016/04/18/models

● Domain Model: https://github.com/duraspace/pcdm/wiki#domain-model
Breakouts

2:50 - 3:10: Part 1, PCDM Profile for shared object
3:10 - 3:30: Part 2, PCDM Profile for Your Object Choice
3:30 - 3:50: Part 3, Pair & Share
Breakouts Recap

● What are the differences among models for shared objects?
● Is that okay that things differ?
● Where does this create confusion?
● What Communication Issues did you encounter reviewing another group’s work?
Building Out the PCDM Data Modeling Communities
Maintaining PCDM Momentum

- IRC: #pcdm on Freenode
- PCDM mailing list
- Notes & Shared Resources from Workshop - Keep Adding to this!

bit.ly/C4LDataModeling201
Broader or Related Communities Working on Modeling

- Hydra Metadata Interest Group
  - #metadata on project-hydra Slack
- Fedora 4 (& Fedora broadly)
- Islandora
  - CLAW / Islandora & Fedora 4 Architecture
  - Islandora Interest Groups (Includes Metadata)

bit.ly/C4LDataModeling201
Data Modeling Tools & Resources

- **Data Modeling Resources Going Forward**
  - Starter List of Tools
  - Some Links to Modeling Work in Cultural Heritage Institutions

- **Data Modeling Needs in PCDM & Related Communities**
  - Location for Open Discussions & Issues
  - [github.com/duraspace/pcdm/issues](https://github.com/duraspace/pcdm/issues)

Your ideas to continue support for data modeling in communities?

bit.ly/C4LDataModeling201
Thank you!

bit.ly/C4LDataModeling201