Practical Approaches for Migrating Metadata to Hyrax

Samvera Virtual Connect 2017
July 18th, 11:55-12:20 EDT
Link to the slides, notes, links: http://bit.ly/HyraxDataMigration
Today’s Facilitators

Christina

Data Ops Repository Specialist, Stanford Univ. Libraries,
@cm_harlow

Tom

Ph.D. Student at University of Wash. iSchool; DCE, @no_reply
This Session’s Goals: Help You...

- Understand the Hyrax Metadata “Status Quo”
- Know Where to Start with Metadata Migration
- Point Out Questions, Priorities, & “Gotchas”
- Point You to Helpful Resources & Communities
Status Quo of Hyrax Metadata
Object Models & Application Profiles

Object Model

Application Profile

PCDM:Collection > HydraWorks:Collection: Digital Collection

This is the collection resource representing Huntington Digital Collection.

Descriptive Profile

<table>
<thead>
<tr>
<th>predicate</th>
<th>value</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>dcterms:title</td>
<td>&quot;Huntington Free Library Native American Collection&quot;</td>
<td>Need language typing</td>
</tr>
<tr>
<td>dcterms:abstract</td>
<td>&quot;One of the largest collections of books and manuscripts of its kind, the Huntington collection contains extensive materials documenting the history, culture, languages, and arts of the native tribes of both North and South America. Contemporary politics and human rights issues are also important components of the collection. Full text of a selection of 91 books from the Huntington Free Library Native American Collection representing the various genres in the collection.&quot;</td>
<td>Need language typing</td>
</tr>
<tr>
<td>dcterms:date</td>
<td>&quot;2010&quot;<a href="http://id.loc.gov/datatypes/edtf">http://id.loc.gov/datatypes/edtf</a></td>
<td>Need date (data type) typing. Make sure is not dcterms:created</td>
</tr>
<tr>
<td>dcterms:identifier</td>
<td>&quot;6790930&quot;</td>
<td>Identifier typing to be added in phase 2 of migration.</td>
</tr>
</tbody>
</table>

bit.ly/HyraxDataMigration
PCDM & Hydra Works

Default Usage in Hyrax

- PCDM is a low-level object model for repository data.
- Hydra Works extends PCDM by breaking Objects into Works and FileSets.

Why shared models?

PCDM: [http://pcdm.org/2016/04/18/models](http://pcdm.org/2016/04/18/models)
Works: [http://pcdm.org/2016/02/16/works](http://pcdm.org/2016/02/16/works)
Collection’s members are Objects and/or other Collections.

**Hyrax::CoreMetadata**
- Title (dct:title)
- Depositor (marcrel:dpt)
- Date Uploaded (dct:dateSubmitted)
- Date Modified (dct:modified)

**Hyrax::BasicMetadata**
Creator, Contributor, Rights, Description, &c.

See:
[1.html#basic-metadata](1.html#basic-metadata)
HW:Work
MAP

Default Usage in Hyrax

‘A work or intellectual entity, such as a book, film, dissertation, etc.’

A hw:Work is a pcdm:Object.

Use Works for repository objects. I.e. books, but also pages (if you handle pages).

Core & Basic Metadata
A group of related Files

A `hw:FileSet` is a `pcdm:Object`.

Use FileSets for groupings of files that aren’t repository objects; e.g. a scanned text and related OCR output.

**Core & Basic** Metadata

Use FileSet metadata for information about the group.

FileSets may be members of more than one Object.
PCDM:File MAP

Default Usage in Hyrax

“An File is a sequence of binary data”

Files are members of one Object.

File Metadata:
- Label (rdfs:label)
- File Name (ebu:filename)
- File Size (ebu:filesize)
- Date Created (ebu:dateCreated)
- Date Modified (ebu:dateModified)
- Byte Order (sweetjpl:byteOrder)
- File Hash (premis:hasMessageDigest)
Works Model

https://github.com/duraspace/pcdm/wiki/Diagrams#hydra
Hyrax Basic Metadata

- Basic Metadata is intended to capture general use cases.
- Can be relatively easily changed (or not used at all) for Works.
- See: [http://samvera.github.io/customize-metadata-model.html#basic-metadata](http://samvera.github.io/customize-metadata-model.html#basic-metadata)
Adding a Field

- Select a predicate
- Select an accessor name
- Single- or Multi-value?
- Controlled Vocabulary?
- Required? Other validations?
- Ordered? (are you really modeling order, or something else?)
- Data types (e.g. dates, integers) and language tags are handled natively.

**Documentation -**

- Defining Metadata
  http://samvera.github.io/customize-metadata-model.html
- Controlled Vocabularies

**Underlying Model -**

- Resource Description Format (RDF)
  https://www.w3.org/TR/rdf11-concepts/
class GenericWork < ActiveFedora::Base
  include ::Hyrax::WorkBehavior
  include ::Hyrax::BasicMetadata
  # Change this to restrict which works can be added as a child.
  # self.valid_child_concerns = []
  validates :title, presence: { message: 'Your work must have a title.' }

  property :contact_email, predicate: ::RDF::Vocab::VCARD.hasEmail, multiple: false do |
    index
      as :stored_searchable
    end

  property :contact_phone, predicate: ::RDF::Vocab::VCARD.hasTelephone do |
    index
      as :stored_searchable
    end

  property :department, predicate: ::RDF::URI.new("http://lib.my.edu/departments"), multiple: false do |
    index
      as :stored_searchable, :facetable
    end
end
Build, Reuse & Share MAPs

- Share both structure and application profiles.
- Spread development and maintenance labor around.
- Reusing saves work!
Where to Start in Migration Planning
Questions to Work Out Pre-Migration

- Classes of Objects Described? Types of Objects?
- Relationships between Object Classes? Nested Objects? “Complex” Objects?
- 1st Class Resources? 2nd Class? Context Classes?
- Discovery Access Points & Facets?
- Who Adds What Metadata?
- Metadata Current State? Fix What Errors When?
- Separate out Migration & Enhancements And…

DO NOT AIM FOR PERFECT
Performing Assessment

Do Your Metadata Homework, Using Whatever Tools Work For You.
Clarify & Iterate on Expectations

Assess, Prioritize Functionalities, Normalize, Migrate, Normalize, Enhance

The RDF relationships expected for this class, symmetry of those properties not assumed (hence A -> B and B -> A are both given)

Structure of Digital Objects in our Fedora 4/Hydra stack using PCDM:

PCDM:Collection > HydraWorks:Collection : Digital Collection

This is the digital collection that currently maps to the dixs identifier sets (i.e. ‘hunt’, ‘bol’, etc.). This is required. Can be repeated/nested as needed.

Descriptive Metadata Profile:

- dcterms:abstract = string
- dcterms:alternative = string
  - rdau: investment curriculum = string (phase 2 removal)
- rdau: investment [curator] = string (phase 2 addition)
- dcterms:date = EDTF literal
- dcterms:identifier = string
- dc:language = string (phase 2 removal)
- dc:publisher = string (phase 2 removal)
- dcterms:publisher = Agent URI (phase 2 addition)
- dcterms:relation = URI
  - dcterms:relation = string (phase 2 removal)
- dcterms:subject = Concept URI (phase 2 addition)
- dcterms:title = literal

Structural Profile:

If no secondary PCDM:Collection for Set:

- Digital Collection -PCDM:hasMember-> Intellectual Work
- Digital Collection <-PCDM:isMemberOf- Intellectual Work

If there is a secondary PCDM:Collection for Set:

- Digital Collection -PCDM:hasMember-> Set
- Digital Collection <-PCDM:isMemberOf- Set
“Graceful Degradation”

What is the shared, underlying logic that your metadata can revert to?

Tip of the Hat to Tom Baker for The Phrase “Graceful Degradation”
Some Rough / Common Groups of Recommendations
DSpace

- OAI-PMH DIMS Feed is Most Complete
- Beware Namespaces
- Complex Fields via Elements & Qualifiers
- Break Out Object Versions, File Info, Identifiers for Parts / Filesets / Objects
- Beware Language Tags
**Bepress (Digital Commons)**

- OAI-PMH Metadata-Export Is Most Complete
- Check Completeness of Data Dumps
- Some Regular Fields
- Authors Broken Into Multi-Parts
- Local Fields Beware
- Supplemental Files => Filesets, Related Objects, Parts

<table>
<thead>
<tr>
<th>Field Description</th>
<th>Complete</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>abstract</td>
<td>8318/9895</td>
<td>84%</td>
</tr>
<tr>
<td>articleid</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>authors/author/email</td>
<td>2270/9895</td>
<td>22%</td>
</tr>
<tr>
<td>authors/author/fname</td>
<td>5315/9895</td>
<td>53%</td>
</tr>
<tr>
<td>authors/author/institution</td>
<td>3874/9895</td>
<td>39%</td>
</tr>
<tr>
<td>authors/author/organization</td>
<td>3874/9895</td>
<td>39%</td>
</tr>
<tr>
<td>authors/author/iname</td>
<td>3353/9895</td>
<td>34%</td>
</tr>
<tr>
<td>authors/author/mname</td>
<td>2131/9895</td>
<td>21%</td>
</tr>
<tr>
<td>authors/author/banner</td>
<td>3243/9895</td>
<td>32%</td>
</tr>
<tr>
<td>context-key</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>coverpage-url</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>disciplines/discipline</td>
<td>5043/9895</td>
<td>50%</td>
</tr>
<tr>
<td>document-type</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>fields/field/value</td>
<td>9089/9895</td>
<td>91%</td>
</tr>
<tr>
<td>fields/field[@name=academic_adviser][@type=string]</td>
<td>30/9895</td>
<td>0%</td>
</tr>
<tr>
<td>fields/field[@name=academic_email][@type=string]</td>
<td>30/9895</td>
<td>0%</td>
</tr>
<tr>
<td>fields/field[@name=acknowledgements][@type=string]</td>
<td>1/9895</td>
<td>0%</td>
</tr>
<tr>
<td>fields/field[@name=advisor][@type=string]</td>
<td>1576/9895</td>
<td>15%</td>
</tr>
<tr>
<td>fields/field[@name=advisor_email][@type=string]</td>
<td>32/9895</td>
<td>0%</td>
</tr>
<tr>
<td>fields/field[@name=affiliation][@type=string]</td>
<td>1915/9895</td>
<td>19%</td>
</tr>
<tr>
<td>fields/field[@name=custom_citation][@type=string]</td>
<td>1123/9895</td>
<td>11%</td>
</tr>
<tr>
<td>fields/field[@name=degree_name][@type=string]</td>
<td>2047/9895</td>
<td>20%</td>
</tr>
<tr>
<td>fields/field[@name=department][@type=string]</td>
<td>1975/9895</td>
<td>19%</td>
</tr>
<tr>
<td>fields/field[@name=distribution_license][@type=string]</td>
<td>5482/9895</td>
<td>55%</td>
</tr>
<tr>
<td>fields/field[@name=embargo_date][@type=date]</td>
<td>50/9895</td>
<td>0%</td>
</tr>
<tr>
<td>fields/field[@name=embargo_date_date_format][@type=string]</td>
<td>31/9895</td>
<td>0%</td>
</tr>
<tr>
<td>fields/field[@name=issnum][@type=string]</td>
<td>19/9895</td>
<td>0%</td>
</tr>
<tr>
<td>fields/field[@name=language][@type=string]</td>
<td>7764/9895</td>
<td>78%</td>
</tr>
<tr>
<td>fields/field[@name=latitude][@type=string]</td>
<td>48/9895</td>
<td>0%</td>
</tr>
<tr>
<td>fields/field[@name=link_institution][@type=string]</td>
<td>21/9895</td>
<td>0%</td>
</tr>
<tr>
<td>fields/field[@name=location][@type=string]</td>
<td>48/9895</td>
<td>0%</td>
</tr>
<tr>
<td>fields/field[@name=publication_date][@type=date]</td>
<td>9005/9895</td>
<td>91%</td>
</tr>
<tr>
<td>fields/field[@name=publication_date_date_format][@type=date]</td>
<td>226/9895</td>
<td>2%</td>
</tr>
<tr>
<td>fulltext-url</td>
<td>8006/9895</td>
<td>80%</td>
</tr>
<tr>
<td>keywords/keyword</td>
<td>6609/9895</td>
<td>66%</td>
</tr>
<tr>
<td>keywords/keyword</td>
<td>6609/9895</td>
<td>66%</td>
</tr>
<tr>
<td>label</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>label</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>label</td>
<td>8006/9895</td>
<td>80%</td>
</tr>
<tr>
<td>keywords/keyword</td>
<td>6609/9895</td>
<td>66%</td>
</tr>
<tr>
<td>label</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>label</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>label</td>
<td>8006/9895</td>
<td>80%</td>
</tr>
<tr>
<td>keywords/keyword</td>
<td>6609/9895</td>
<td>66%</td>
</tr>
<tr>
<td>label</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>label</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>label</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>label</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>label</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>label</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>native-url</td>
<td>2196/9895</td>
<td>22%</td>
</tr>
<tr>
<td>publication-date</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>publication-date</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>submission-date</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>submission-path</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>supplemental-files/file/archive-name</td>
<td>98/9895</td>
<td>0%</td>
</tr>
<tr>
<td>supplemental-files/file/description</td>
<td>30/9895</td>
<td>0%</td>
</tr>
<tr>
<td>supplemental-files/file/mime-type</td>
<td>98/9895</td>
<td>0%</td>
</tr>
<tr>
<td>supplemental-files/file/upload-name</td>
<td>98/9895</td>
<td>0%</td>
</tr>
<tr>
<td>supplemental-files/file/url</td>
<td>98/9895</td>
<td>0%</td>
</tr>
<tr>
<td>title</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
<tr>
<td>type</td>
<td>9895/9895</td>
<td>100%</td>
</tr>
</tbody>
</table>
ContentDM

- Check TSV Export First for Most Complete
- Watch Out For Page-Level Objects Mapped to PCDM Objects / “Parts”
- Split Concatenated Fields
- Beware Local Fields

bit.ly/HyraxDataMigration
Hydra & F3

- Overlapping Datastreams to PCDM Objects
- RELS-EXT != PCDM Relationships
- Keep that MODS or XML Metadata in Hyrax As Needed In Your Transition
- Fedora Migrate ... ?
  https://github.com/samvera-labs/fedora-migrate
- FOXML & Fedora File System-based Migration ... ?
  https://github.com/fcrepo4-exts/migration-utils
- Consider Administrative Data Placement for Files / Filesets

bit.ly/HyraxDataMigration
You’re Not Alone in this Process
Helpful Resources & Groups

- **Samvera Community Groups**
  - [DSpace Migration Group](#)
  - [Documentation Group](#)
  - [Metadata Group](#)
  - [Hyku Metadata Guidance](#)

- **Outside of Samvera...**
  - Tools: [OpenRefine](#), [Catmandu](#), ...
  - [PCDM Community](#)
  - Our Fedora 4 Peers: [Islandora](#)
  - [DLF Metadata Assessment Group](#)

- **Peer Groups...**
  - [Hyrax Renegade Metadata Docs](#)
  - [Shared Cultural Heritage MAPs](#)
  - [DPLA MAP v.4](#)
Helpful Resources & Groups

- **Samvera Community Groups**
  - DSpace Migration Group
  - Documentation Group
  - Metadata Group
  - Hyku Metadata Guidance

- **Outside of Samvera...**
  - Tools: [OpenRefine](https://openrefine.org), [Catmandu](https://catmandu-project.org), ...
  - PCDM Community
  - Our Fedora 4 Peers: [Islandora](https://www.islandora.org/)
  - [DLF Metadata Assessment Group](https://www.dlf.org/resources/dlf-metadata-assessment-group)

- **Peer Groups...**
  - [Hyrax Renegade Metadata Docs](https://github.com/HyraxRenegade)
  - [Shared Cultural Heritage MAPs](https://www.dpl.org/dpla/metadata-asset-profiles)
  - [DPLA MAP v.4](https://dpla.metadata-assessment-profiles.org)

And Please Remember to Share Back Your Experiences & Lessons as Well!

bit.ly/HyraxDataMigration
Thank you!