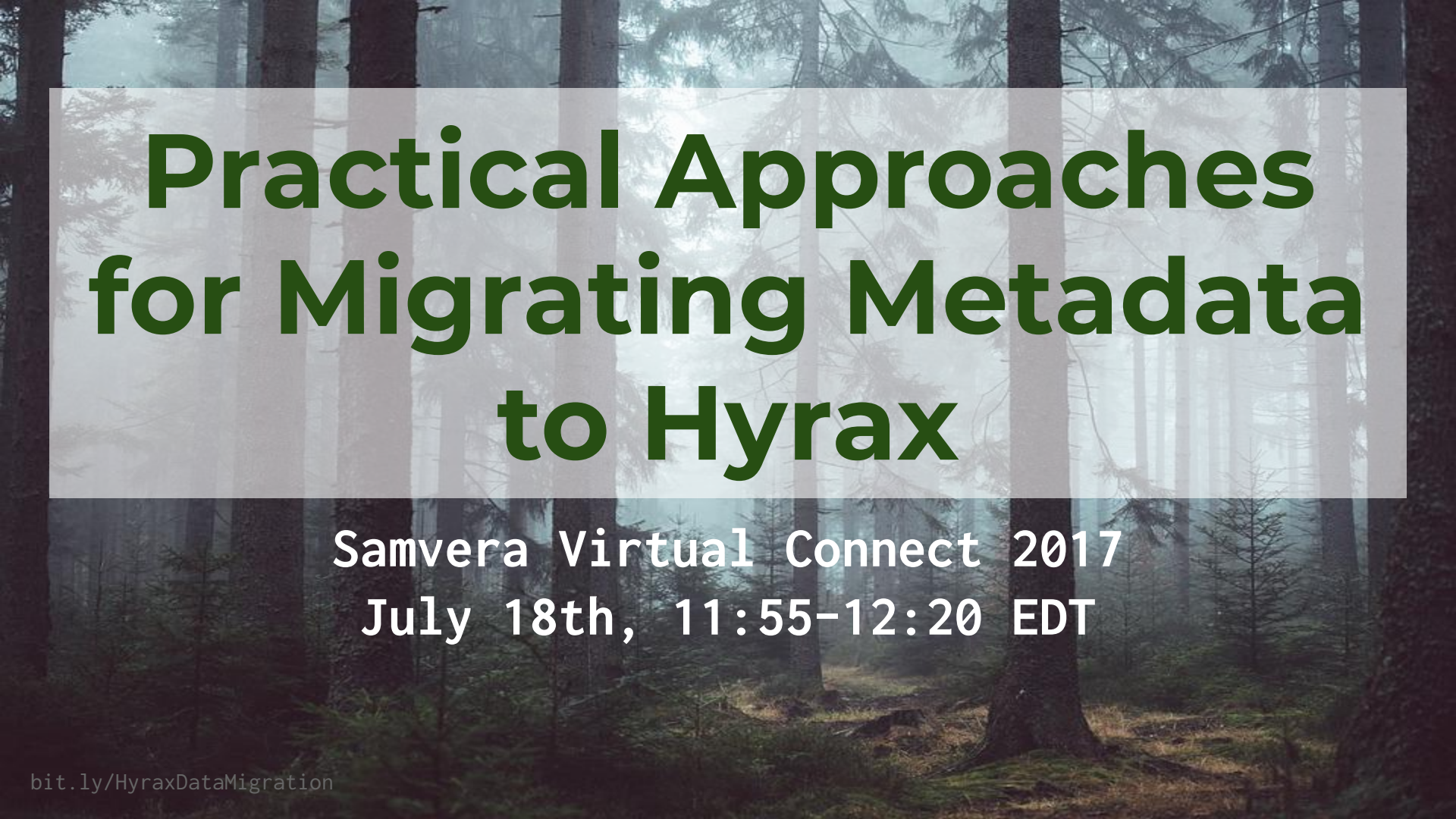


Practical Approaches for Migrating Metadata to Hyrax

Samvera Virtual Connect 2017
July 18th, 11:55-12:20 EDT

bit.ly/HyraxDataMigration


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

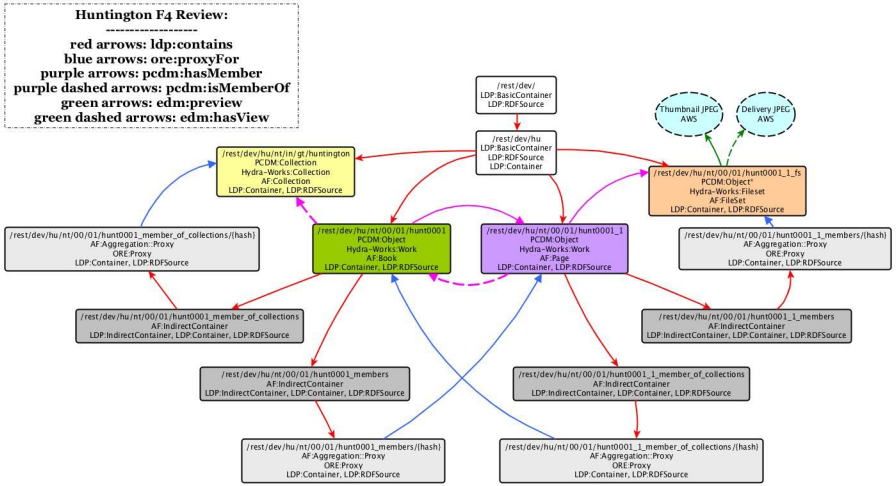
bit.ly/HyraxDataMigration

Status Quo of Hyrax Metadata

bit.ly/HyraxDataMigration

Object Models & Application Profiles

Object Model



Application Profile

PCDM:Collection > HydraWorks:Collection : Digital Collection

This is the collection resource representing Huntington Digital Collection.

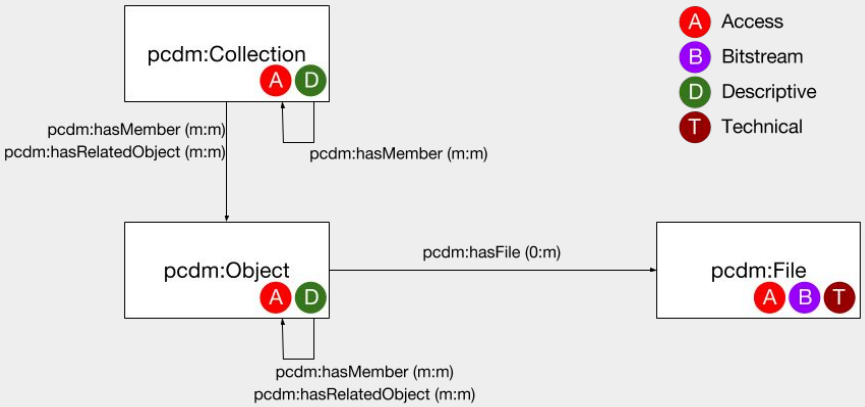
Descriptive Profile

predicate	value	notes
dcterms:title	"Huntington Free Library Native American Collection"^^xsd:string	Need language typing
dcterms:abstract	"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."^^xs:string	Need language typing
dcterms:date	"2010"^^http://id.loc.gov/datatypes/edtf	Need date (data type) typing. Make sure is not dcterms:created
dcterms:identifier	"6790930"	identifier typing to be added in phase 2 of migration.
		To be removed

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
Works: http://pcdm.org/2016/02/16/works

bit.ly/HyraxDataMigration

PCDM:Collections MAP

Default Usage in Hyrax

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:
http://samvera.github.io/customize-metadata-model.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

HW:FileSet MAP

Default Usage in Hyrax

“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

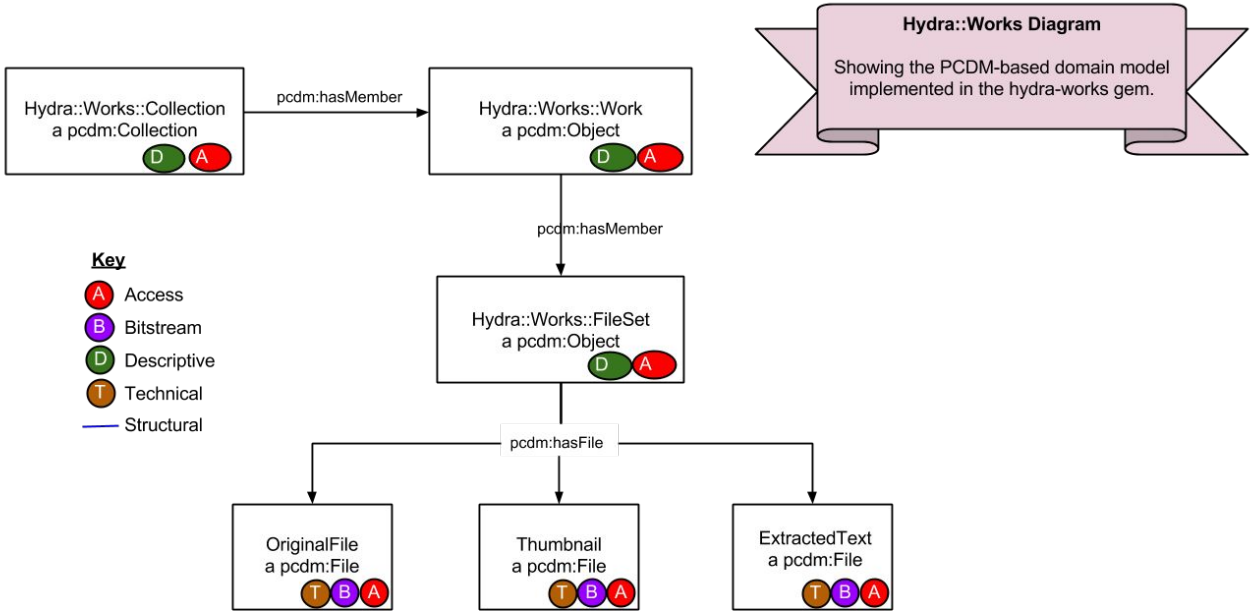
“A File is a sequence of binary data”

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

Property	Predicate	Multiple
label	ActiveFedora::RDF::Fcrepo::Model.downloadFilename	**FALSE**
relative_path	::RDF::URI.new('http://scholarsphere.psu.edu/ns#relativePath')	**FALSE**
import_url	::RDF::URI.new('http://scholarsphere.psu.edu/ns#importUrl')	**FALSE**
resource_type	::RDF::Vocab::DC.type	TRUE
creator	::RDF::Vocab::DC11.creator	TRUE
contributor	::RDF::Vocab::DC11.contributor	TRUE
description	::RDF::Vocab::DC11.description	TRUE
keyword	::RDF::Vocab::DC11.relation	TRUE
license	::RDF::Vocab::DC.rights	TRUE
rights_statement	::RDF::Vocab::EDM.rights	TRUE
publisher	::RDF::Vocab::DC11.publisher	TRUE
date_created	::RDF::Vocab::DC.created	TRUE
subject	::RDF::Vocab::DC11.subject	TRUE
language	::RDF::Vocab::DC11.language	TRUE
identifier	::RDF::Vocab::DC.identifier	TRUE
based_near	::RDF::Vocab::FOAF.based_near	TRUE
related_url	::RDF::RDFS.seeAlso	TRUE
bibliographic_citation	::RDF::Vocab::DC.bibliographicCitation	TRUE
source	::RDF::Vocab::DC.source	TRUE

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
  http://samvera.github.io/customize-metadata-controlled-vocabulary.html

Underlying Model -

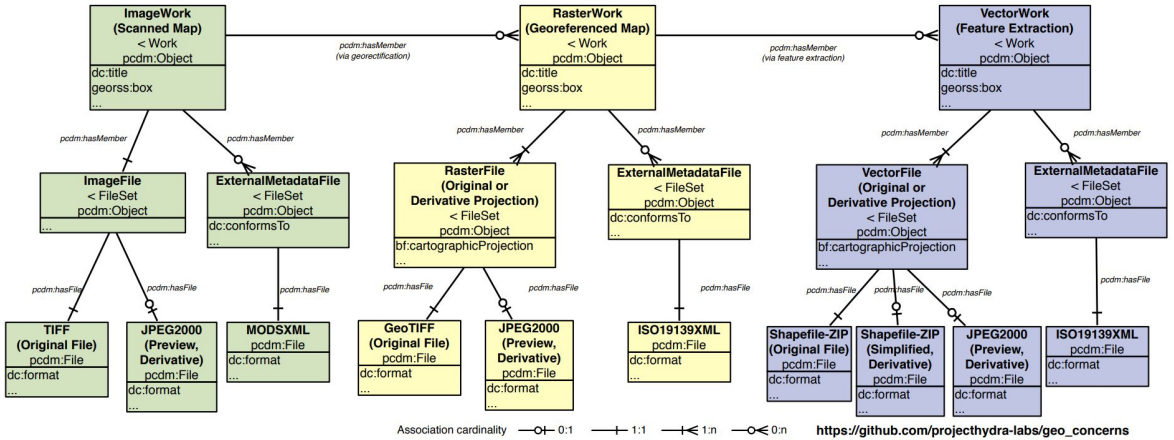
- Resource Description Format (RDF)
  https://www.w3.org/TR/rdf11-concepts/

```ruby
# Generated via
#  `rails generate hyrax:work GenericWork`
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|
    index.as :stored_searchable
  end

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

  property :department, predicate: ::RDF::URI.new("http://lib.my.edu/departments"), multiple: false do |i
ndex|
    index.as :stored_searchable, :facetable
  end
end
```

bit.ly/HyraxDataMigration

Build, Reuse & Share MAPs

- Share both structure and application profiles.
- Spread development and maintenance labor around.
- Reusing saves work!



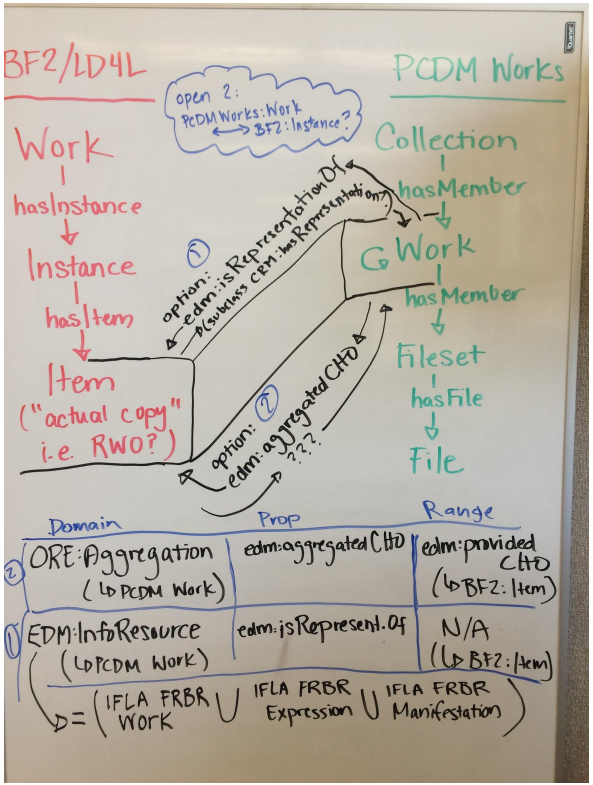
bit.ly/HyraxDataMigration

Where to Start in Migration Planning

bit.ly/HyraxDataMigration

Questions to Work Out Pre-Migration

- Classes of Objects Determined? Types of Objects?
- Relationships between Object Classes? **Nested Objects? “Complex” Objects?**
- 1st Class Resources? 2nd Class? Context Classes?
- External Data? **Entity Resolution? Look-Ups? New URIs? Identifiers?**
- 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**



bit.ly/HyraxDataMigration

Performing Assessment

Do Your Metadata Homework, Using Whatever Tools Work For You.

bit.ly/HyraxDataMigration

```
   1 b'Fact sheet and poster'
   1 b'Fact sheet and posters'
3875 b'Image'
   1 b'Instruction Manual'
  45 b'Interview'
 125 b'Learning Object'
   2 b'Literature Review'
   3 b'Musical Score'
 275 b'Newsletter'
1270 b'Other'
 243 b'Pamphlets'
5224 b'Paper or Project'
  20 b'Paper or project'
6990 b'Periodical'
   3 b'Poster'
  37 b'Preprint'
 759 b'Presentation'
   1 b'Proceedings'
  59 b'Project Plan'
 338 b'Project Report'
 119 b'Report'
  31 b'Software'
 426 b'Sound'
 229 b'Syllabus'
3330 b'Technical Report'
  19 b'Technical report'
   1 b'Video/Moving Image and Poster'
 721 b'Video/Moving Image'
   9 b'Video/Moving Object'
  33 b'Website'
   6 b'Working Paper'
 535 b'article'
 600 b'book chapter'
  39 b'book'
   1 b'case study'
   1 b'conference proceedings'
   3 b'dataset'
6158 b'dissertation or thesis'
   4 b'fact sheet'
   1 b'final report'
  13 b'learning object'
   1 b'literature review'
3873 b'newsletter'
  21 b'other'
   4 b'paper or project'
   4 b'periodical'
   1 b'preprint'
  73 b'presentation'
   1 b'procedure manual'
   1 b'project research'
 932 b'report'
  53 b'syllabus'
   9 b'technical report'
  11 b'video/moving image'
   1 b'vidoe/moving image'
┌─<sul.cmharlow @ li-dl-2628-db6b in ~/C/s/g/c/metadataQA>
└─< (master)* >──» exit
┌─<sul.cmharlow @ li-dl-2628-db6b in ~/C/s/g/c/metadataQA>
└─< (master)* >──» ttygif myrecording
```

Clarify & Iterate on Expectations

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

bit.ly/HyraxDataMigration

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 current maps to the dlxs 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:P60376** [curator] = string (phase 2 removal)
- **rdau:P60376** [curator] = Agent URI (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** = URL
- **dc:subject** = 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?

bit.ly/HyraxDataMigration

Full Representation

CIDOC-CRM FRBR-OO BIBFRAME

EDM VRACore-RDF MODSRDF

DPLA Schema

DC Profiles

Common Abstractions

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

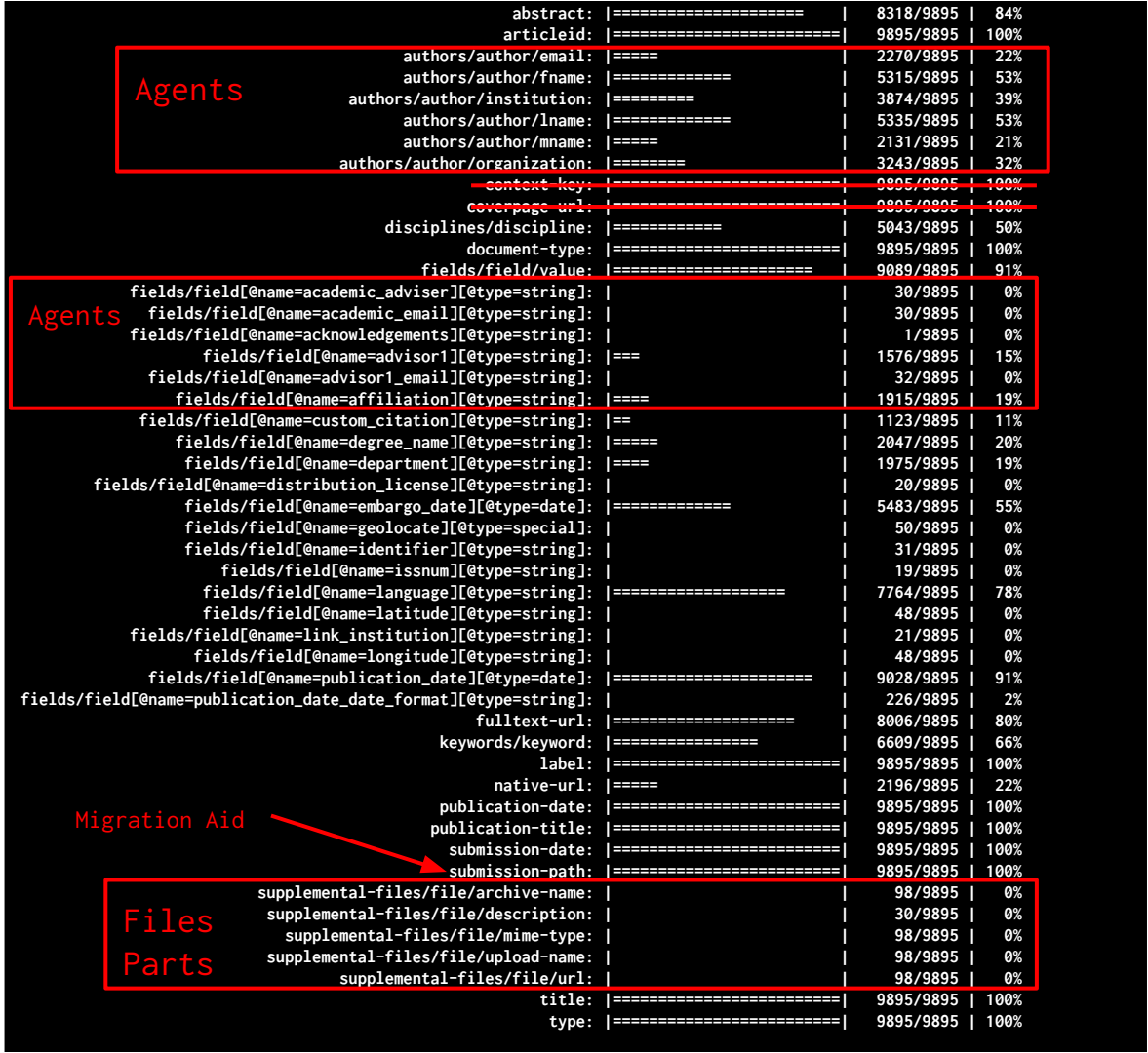
```
   1 b'Fact sheet and poster'
   1 b'Fact sheet and posters'
3875 b'Image'
   1 b'Instruction Manual'
  45 b'Interview'
 125 b'Learning Object'
   2 b'Literature Review'
   3 b'Musical Score'
 275 b'Newsletter'
1270 b'Other'
 243 b'Pamphlets'
5224 b'Paper or Project'
  20 b'Paper or project'
6990 b'Periodical'
   3 b'Poster'
  37 b'Preprint'
 759 b'Presentation'
   1 b'Proceedings'
  59 b'Project Plan'
 338 b'Project Report'
 119 b'Report'
  31 b'Software'
 426 b'Sound'
 229 b'Syllabus'
3330 b'Technical Report'
  19 b'Technical report'
   1 b'Video/Moving Image and Poster'
 721 b'Video/Moving Image'
   9 b'Video/Moving Object'
  33 b'Website'
   6 b'Working Paper'
 535 b'article'
 600 b'book chapter'
  39 b'book'
   1 b'case study'
   1 b'conference proceedings'
   3 b'dataset'
6158 b'dissertation or thesis'
   4 b'fact sheet'
   1 b'final report'
  13 b'learning object'
   1 b'literature review'
3873 b'newsletter'
  21 b'other'
   4 b'paper or project'
   4 b'periodical'
   1 b'preprint'
  73 b'presentation'
   1 b'procedure manual'
   1 b'project research'
 932 b'report'
  53 b'syllabus'
   9 b'technical report'
  11 b'video/moving image'
   1 b'vidoe/moving image'
┌─<sul.cmharlow @ li-dl-2628-db6b in ~/C/s/g/c/metadataQA>
└─< (master)* >──» exit
┌─<sul.cmharlow @ li-dl-2628-db6b in ~/C/s/g/c/metadataQA>
└─< (master)* >──» ttygif myrecording
```

bit.ly/HyraxDataMigration

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

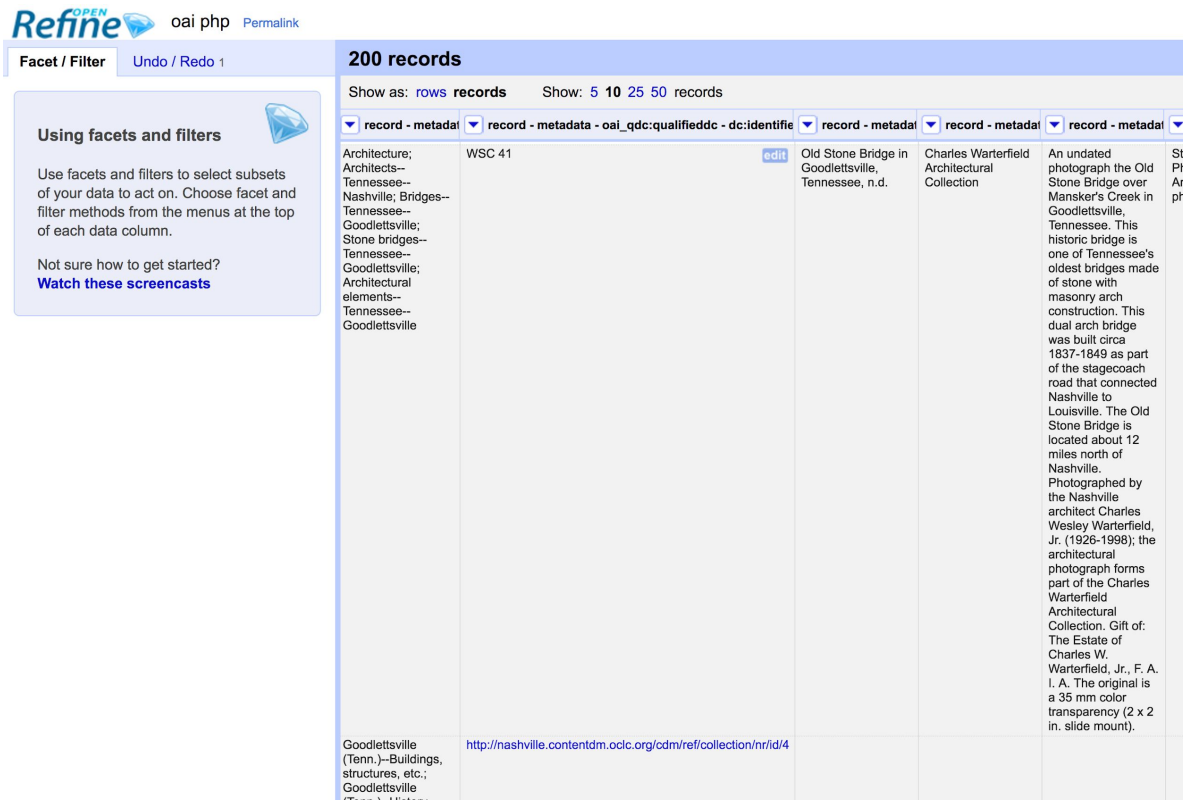
bit.ly/HyraxDataMigration



abstract:	|========== |	8318/9895	84%
articleid:	|==========|	9895/9895	100%
authors/author/email:	|===== |	2270/9895	22%
authors/author/fname:	|============= |	5315/9895	53%
authors/author/institution:	|========= |	3874/9895	39%
authors/author/lname:	|============= |	5335/9895	53%
authors/author/mname:	|===== |	2131/9895	21%
authors/author/organization:	|======== |	3243/9895	32%
context-key:	|==========|	9895/9895	100%
coverpage-url:	|==========|	9895/9895	100%
disciplines/discipline:	|============ |	5043/9895	50%
document-type:	|==========|	9895/9895	100%
fields/field/value:	|========== |	9089/9895	91%
fields/field[@name=academic_adviser][@type=string]:	| |	30/9895	0%
fields/field[@name=academic_email][@type=string]:	| |	30/9895	0%
fields/field[@name=acknowledgements][@type=string]:	| |	1/9895	0%
fields/field[@name=advisor1][@type=string]:	|=== |	1576/9895	15%
fields/field[@name=advisor1_email][@type=string]:	| |	32/9895	0%
fields/field[@name=affiliation][@type=string]:	|==== |	1915/9895	19%
fields/field[@name=custom_citation][@type=string]:	|== |	1123/9895	11%
fields/field[@name=degree_name][@type=string]:	|===== |	2047/9895	20%
fields/field[@name=department][@type=string]:	|==== |	1975/9895	19%
fields/field[@name=distribution_license][@type=string]:	| |	20/9895	0%
fields/field[@name=embargo_date][@type=date]:	|============= |	5483/9895	55%
fields/field[@name=geolocate][@type=special]:	| |	50/9895	0%
fields/field[@name=identifier][@type=string]:	| |	31/9895	0%
fields/field[@name=issnum][@type=string]:	| |	19/9895	0%
fields/field[@name=language][@type=string]:	|========== |	7764/9895	78%
fields/field[@name=latitude][@type=string]:	| |	48/9895	0%
fields/field[@name=link_institution][@type=string]:	| |	21/9895	0%
fields/field[@name=longitude][@type=string]:	| |	48/9895	0%
fields/field[@name=publication_date][@type=date]:	|========== |	9028/9895	91%
fields/field[@name=publication_date_date_format][@type=string]:	| |	226/9895	2%
fulltext-url:	|========== |	8006/9895	80%
keywords/keyword:	|========= |	6609/9895	66%
label:	|==========|	9895/9895	100%
native-url:	|===== |	2196/9895	22%
publication-date:	|==========|	9895/9895	100%
publication-title:	|==========|	9895/9895	100%
submission-date:	|==========|	9895/9895	100%
submission-path:	|==========|	9895/9895	100%
supplemental-files/file/archive-name:	| |	98/9895	0%
supplemental-files/file/description:	| |	30/9895	0%
supplemental-files/file/mime-type:	| |	98/9895	0%
supplemental-files/file/upload-name:	| |	98/9895	0%
supplemental-files/file/url:	| |	98/9895	0%
title:	|==========|	9895/9895	100%
type:	|==========|	9895/9895	100%

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


You’re Not Alone in this Process
bit.ly/HyraxDataMigration

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

bit.ly/HyraxDataMigration

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

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

bit.ly/HyraxDataMigration


Thank you!
http://bit.ly/HyraxDataMigration