

Case Study: Oregon Digital Migration from CONTENTdm to Hydra for digital collections

Hydra Connect 2016, Boston Public Library

Julia Simic & Linda Sato, University of Oregon Libraries

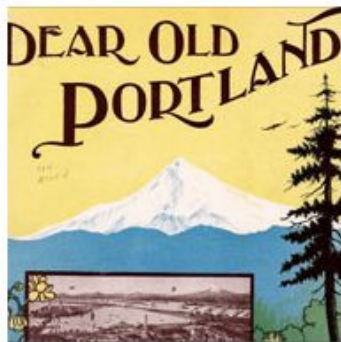
Ryan Wick & Margaret Mellinger, Oregon State University Libraries



Oregon Digital: Unique Digital Collections from OSU and UO Libraries

[Digital Home](#) [A-Z List](#) [Copyright & Use](#) [Order](#) [Help](#) [Contact](#)

DIGITALCOLLECTIONS



A-Z List of Collections

View a [list](#) of Digital Collections with short descriptions.



Take a Tour

View a [slide show](#) of images and documents in our many varied digital collections.



Featured Collection

OSU's [Braceros in Oregon Photograph Collection](#) documents the activities of Oregon's Bracero workers

SEARCH DIGITAL COLLECTIONS

[A-Z LIST OF COLLECTIONS](#)

Search CONTENTdm Collections:

 [Go](#)

Search

- UO Scholars' Bank
- Local & Regional Documents
- Renaissance Editions
- UO University Archives


 [Go](#)

Search OSU ScholarsArchive

 [Go](#)


The OSU and UO Libraries' Digital Collections are created to support the teaching and research mission of the Oregon University System. The collections are composed of unique digitized and born digital materials including photographs, journal articles, sheet music, manuscripts, ephemera, and more. The collections are managed by [UO Digital Library Initiatives](#) and [OSU Libraries Digital Access Services](#).

From CONTENTdm to Hydra

 LIBRARIES
University of Oregon

Oregon State
UNIVERSITY | Libraries

OREGON
DIGITAL

Search... 

Advanced Search

Login
Bookmarks
History

Limit your search

Creator >

Arranger >

Artist >

Author >

Composer >

Illustrator >

Interviewee >

Lyricist >




Photographer >

Topic >

OREGON DIGITAL

[Collections List](#)
View a list of Digital Collections.

The collections of Oregon Digital are primarily created to support the teaching and research missions of the [University of Oregon](#) and [Oregon State University](#). The collections are comprised of unique digitized and born-digital materials including photographs, articles, sheet music, manuscripts, ephemera, and more. Oregon Digital is collaboratively managed by the [Digital Scholarship Center](#) of the [University of Oregon Libraries](#) and the [Center for Digital Scholarship and Services](#) of Oregon State University Libraries.



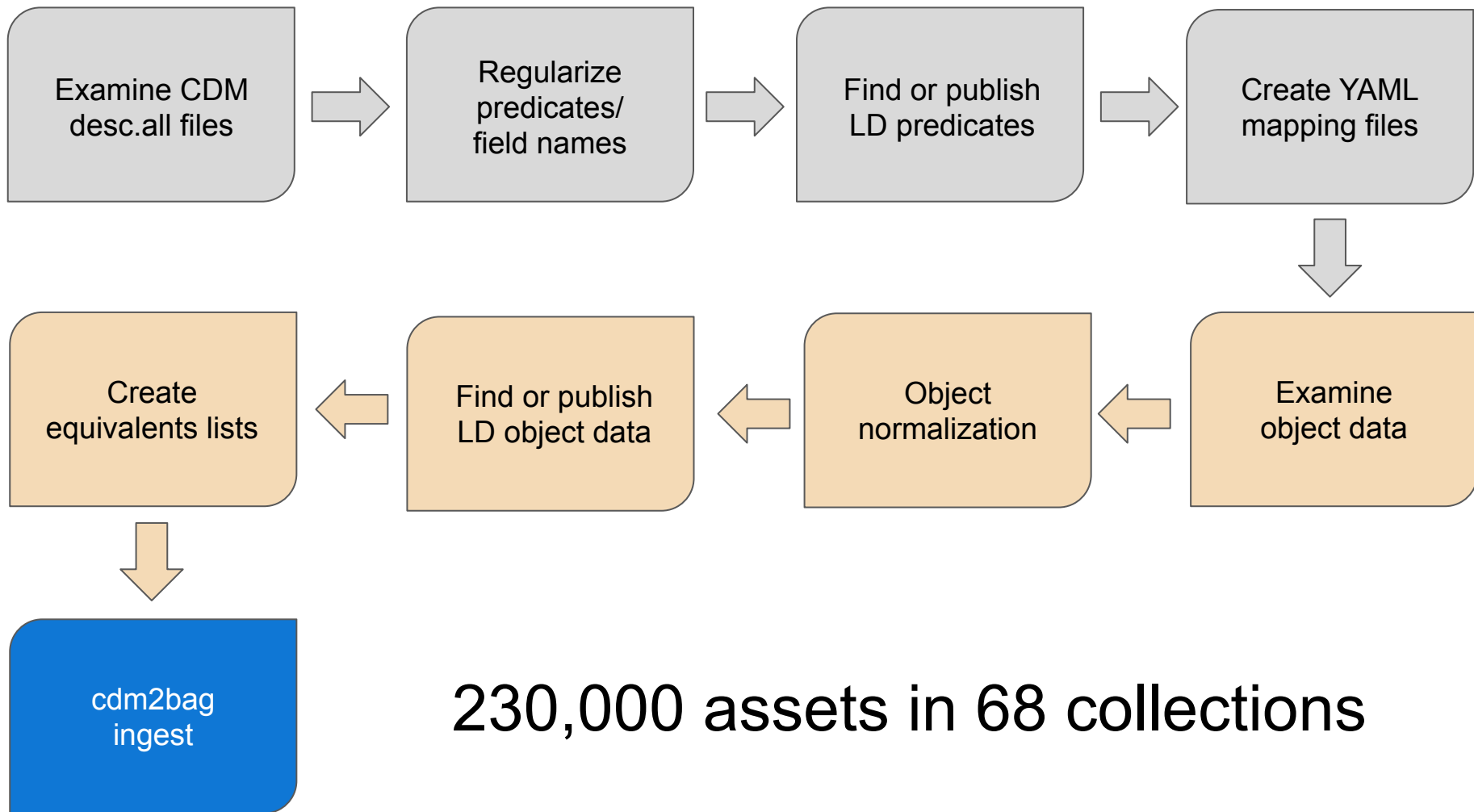
Hydra Prototype

- All descriptive metadata fields were in RDF, everything was a string
- Collection (set) landing pages and additional information pages
- Items could belong to multiple Collections (sets)
- Zoomable image viewer using IIP server and OpenSeadragon UI
- PDF/Document viewer using Internet Archive BookReader
- At the time, most in the community were not building repositories with many different item types or linked data.

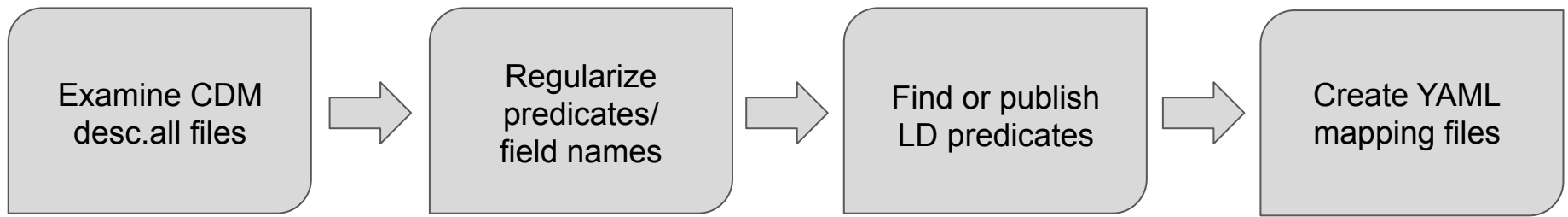
After the Prototype

- Decided to use deep RDF (URIs) where possible
- Needed to write RDF code that didn't exist elsewhere, such as **fetch** and **label** handling, some became ActiveTriples gem
- Set up a separate server for derivatives, ingest, fetch and index jobs
- Needed to have CONTENTdm item short URLs and viewer URLs redirected
 - CONTENTdm URL/ID stored with each new item, Rake task can export all of them and generate mapping files
 - Had some issues with map size until we used nginx
- Audio player - HTML5

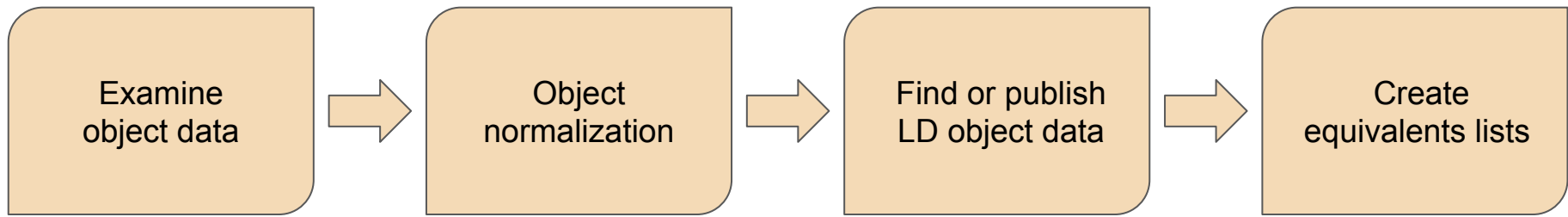
Migration Workflows



230,000 assets in 68 collections



- Exported CONTENTdm metadata by collection
- Examined field names and content to determine needed predicates
- Regularized and mapped field names to existing LD predicates
- Published new predicates in opaquenamespace.org
 - Originally JSON-LD files on [GitHub](https://github.com)
 - Now published using the [Controlled Vocabulary Manager](#)
- Deleted fields that were used only internally or were specific to CDM
- Determined if predicates should require string or URI objects within Oregon Digital
- Created [mapping files](#) (YAML)
- Documented in [Metadata Dictionary](#)



- Examined all Object data per collection
- Fixed typos, delimiter errors etc. using scripts and by hand
 - Non-UTF-8 diacritic problems
- Looked for usable LD vocabularies
 - Used a combination of scraping and hand gathering
- Defined predicates to validate only chosen vocabs
- Published new vocabularies in opaquenamespace.org (JSON-LD→CVM)
- Created lists of old entries and new URI values




cdm2bag ingest

- Cdm2bag Ruby script processed desc.all files, mapped fields and output bags
- Replace object data using a combination of mapping methods and lists
 - BagIt utilized to output one asset + metadata per bag
- Bad mapping, predicates or invalid objects stopped ingest
 - Problems with deep RDF predicates
 - Problems with unresolvable object URIs
- Some high resolution files were missing and/or never existed
- CDM full resolution links corrupted when assets were moved on servers and links were not updated
- Processing derivatives was very slow, large PDFs and TIFs
 - Considered pre-processing derivatives





Post-Migration Cleanup

Post Migration Cleanup

Groupings

Set	Historic Sheet Music Collection	
Exhibit	Women composers	
		

Administratives

Institution	University of Oregon Libraries	
Date Digitized	2008	
Replaces Url	https://oregondigital.org/u?sheetmusic,1852	
		


Raw statements

```
<http://purl.org/dc/dcmitype/Image> <http://www.w3.org/2000/01/rdf-schema#label> "Image"@en .
<http://purl.org/dc/dcmitype/Image> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/2000/01/rdf-schema#Class> .
<http://purl.org/NET/mediatypes/application/pdf> <http://www.w3.org/2000/01/rdf-schema#label> "application/pdf" .
<http://oregondigital.org/resource/oregondigital:02870v89h> <http://www.w3.org/2000/01/rdf-schema#label> "Text"@en .
<http://oregondigital.org/resource/oregondigital:02870v89h> <http://purl.org/dc/terms/creator> "Blake, Charlotte" .
<http://oregondigital.org/resource/oregondigital:02870v89h> <http://www.loc.gov/standards/mods/modsrdf/v1/physicalExtent> "8 p.; color" .
<http://oregondigital.org/resource/oregondigital:02870v89h> <http://www.loc.gov/standards/mods/modsrdf/v1/locationCopyShelfLocator> "Music ShColl
017160" .
<http://oregondigital.org/resource/oregondigital:02870v89h> <http://creativecommons.org/ns#license> <http://creativecommons.org/publicdomain
/mark/1.0/> .
<http://oregondigital.org/resource/oregondigital:02870v89h> <http://purl.org/dc/elements/1.1/rights> "This item is in the public domain.
Acknowledgement of the University of Oregon Libraries as a source is requested." .
<http://oregondigital.org/resource/oregondigital:02870v89h> <http://opaquenamespace.org/ns/full> "1853.pdf" .
```

Update Asset

Post Migration Cleanup

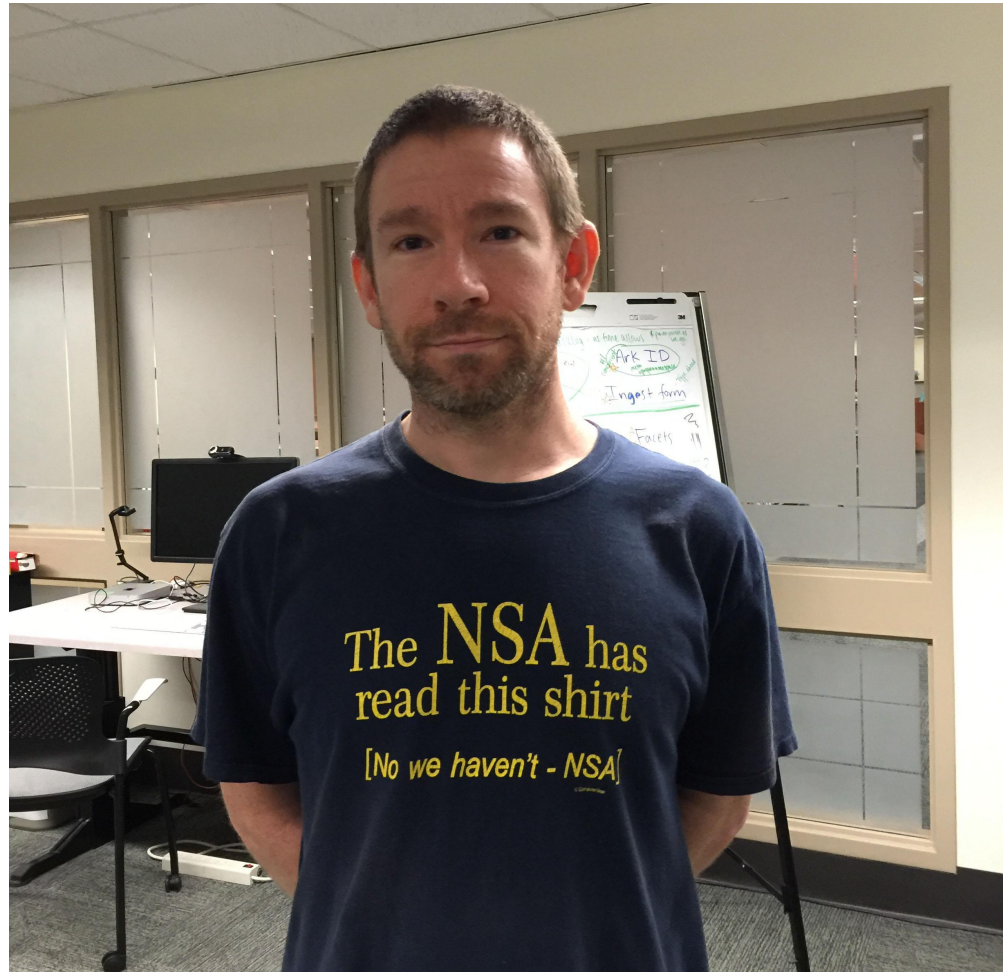
Descriptions

Description	Item from the OIMB 35mm slide collection	
Identification Verifi	Very Certain	
		

Subjects

LC Subject	http://opaquenamespace.org/ns/subject/itis_846496	
LC Subject	Wildflowers and Grasses	
Genus	http://opaquenamespace.org/ns/genus/itis_43473	
Phylum	http://opaquenamespace.org/ns/phylum/ubio_5952239	
Common Names	bog orchid or ladies tresses	
		

Bad Assets



Final Review: Image filename problems



SQKH_00002_KV



Final Review: Image problems

[Return to search](#)

Filters: Collection: herbarium; Tolerance: Images which are 75% different or worse (432 images awaiting review)

oregondigital:df65w978v - herbarium,2984

Title:

Silene andersonii Clokey

Description:

Identifiers:

OSC103809.tif

ContentDM



OregonDigital



[View on OregonDigital](#)

Compare notes: Failed pixels: 54683 out of 71360

The OD image is correct

The images are different, but it's not clear which is correct

The OD image is correct, but needs to be rotated or mirrored

The OD image needs cropping

The CDM image needs to be re-uploaded

Mountain West Digital Library

Correctly formatted dates

2015

2015-02

2015-02-26

any of the above as a range:

2015/2016

or as a series: [1801, 1926]

What we found:

ca. 1915

1877, c. 1885

c. 1900-1909

1913?

Decmeber 29, 1955

Summer 1957

1948, dismantled 1983

27-Aug-80

February 1975

1911-11 - 1912-08

1937 and later

1883, hospital opened

Kenneth Gunn

1925, 1928, completed 1948

Lessons Learned

- Metadata cleanup and review require significant resources, have metadata people closely involved with developers from the beginning
- Leadership needs to have all impacted staff onboard
- Ensure that original media files are named correctly and accessible
- Linked data encourages better quality but can be time consuming with poor metadata, tools have improved over years

Lessons Learned continued

- Being involved in the Hydra community is very important, share work and solve problems together
- Devote development resources to migration tooling and QA
- Compound objects are complicated

Resources

Metadata Dictionary:

<https://docs.google.com/spreadsheets/d/1nnJz49oqstQLF2PFBn5ZvWfWdF-8PVI5TFL1EMotQrA/edit?usp=sharing>

Oregon Digital Descriptive Metadata schema:

https://github.com/OregonDigital/oregondigital/blob/master/app/models/datastream/oregon_rdf.rb

Cdm2bag: <https://github.com/OregonDigital/cdm2bag>

- Collection field mapping: <https://github.com/OregonDigital/cdm2bag/blob/master/mapping.yml>

Csv2bag: <https://github.com/OregonDigital/csv2bag>

[OpaqueNamespace.org](https://opaquenamespace.org), powered by <https://github.com/OregonDigital/ControlledVocabularyManager/>

Oregon Digital Hydra Team, Past & Present

Oregon State University

Eviva Weinraub, Trey Pendragon,
Tom Johnson, Ryan Wick, Mike Eaton,
Brandon Straley, Greg Luis Ramírez,
Josh Gum, Ryan Ordway, Maura Valentino,
Brian Davis, Erin Clark, Michael Boock,
Margaret Mellinger, Chris Petersen,
Trevor Sandgathe, Hui Zhang, Susan McEvoy,
Helena Bales

University of Oregon

Karen Estlund, Sheila Rabun, David McCallum,
Julia Simic, Jeremy Echols, Duncan Barth,
Sarah Seymore, Linda Sato, Kate Jones

Questions?



oregondigital:df715547z

Title

Mika-in-a-bag

LC Subject

[American shorthair cat](#)

Photographer

[Rabun, Sheila](#)

Condition Of Source

adorable

Identifier

mika6

Type

[Image](#)

Format

[image/jpeg](#)

Has Version

Sheila's cat

Rights

Rights Reserved - Restricted Access