Creating a Reproducible Metadata Transformation Pipeline Using Technology Best Practices

Cara Key
*Louisiana State University*, carakey@lsu.edu

Mike Waugh
*Louisiana State University*, mwaugh2@lsu.edu

Follow this and additional works at: [https://digital.sandiego.edu/symposium](https://digital.sandiego.edu/symposium)

Part of the Cataloging and Metadata Commons, Databases and Information Systems Commons, and the Software Engineering Commons


This 45-minute concurrent session is brought to you for free and open access by Digital USD. It has been accepted for inclusion in Digital Initiatives Symposium by an authorized administrator of Digital USD. For more information, please contact digital@sandiego.edu.
Creating a Reproducible Metadata Transformation Pipeline Using Technology

Best Practices

Presenter 1 Title
Metadata Librarian

Presenter 2 Title
Technology Initiatives Development and Management Librarian

Session Type
45-minute concurrent session

Abstract
Over the course of two years, a team of librarians and programmers from LSU Libraries migrated the 186 collections of the Louisiana Digital Library from OCLC's CONTENTdm platform over to the open-source Islandora platform.

Early in the process, the team understood the value of creating a reproducible metadata transformation pipeline, because there were so many unknowns at the beginning of the process along with the certainty that mistakes would be made. This presentation will describe how the team used innovative and collaborative tools, such as Trello, Ansible, Vagrant, VirtualBox, git and GitHub to accomplish the task.

Location
Room G

Keywords
digital libraries, metadata transformation, automated workflows, project management, innovation, collaboration

Creative Commons License

This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 License
Building a Reproducible Metadata Pipeline Using Technology Best Practices

Cara Key, Metadata Librarian, LSU Libraries
Mike Waugh, Technology Development Librarian, LSU Libraries
1. Background
2. Problems
3. Solutions
   a. Building the metadata pipeline
   b. Visualizing the pipeline
   c. Work in action

Image: Pipeline Construction Use of Equipment
Harold Cosini
Standard Oil (NJ) Collection, Photographic Archives, Archives and Special Collections, University of Louisville
Background
About LSU and LSU Libraries

- Louisiana State University
  - Flagship campus located in Baton Rouge, LA
  - Public Research University
  - 27,656 Students
- LSU Libraries
  - ARL, ASERL
  - LOUIS consortium
  - Louisiana Digital Consortium (LDC)
Development Team

- Louisiana Digital Library primary project for new development team
- 4 programmers + 4 librarians
- New hires
- Agile project management

Image: Burlesque circus parade participants in Baton Rouge
Andrew D. Lytle Collection, Mss. 893, 1254, Louisiana and Lower Mississippi Valley Collections, LSU Libraries, Baton Rouge, La.
About the Louisiana Digital Consortium

- 18 institutions + more joining
  - Academic libraries
  - Public libraries
  - Museums
  - Archives
  - Historical organizations

- Developed and managed by LSU Libraries
- Participating institutions responsible for the management of their own collections and metadata
Louisiana Digital Library

- Louisiana Digital Library
  - Historical and cultural materials related to Louisiana
    - Photographs
    - Maps
    - Manuscripts
    - Books
    - Oral Histories
  - 186 collections +
  - First collections added in 1993
  - Migrated to CONTENTdm in 2002
Reasons for Choosing Islandora

- Leaving CONTENTdm
  - Improved support
  - Platform stability
  - Facelift

- Open Source
  - More flexibility
  - Platform for innovation

- Islandora
  - Familiarity
    - Drupal
    - Tulane used Islandora
Problems

Image: Alligator juvenile
Col. Joseph S. Tate Photograph Album, Mss. 4963
Louisiana and Lower Mississippi Valley Collections
LSU Libraries, Baton Rouge, La.
Metadata Conversion

- CONTENTdm records are Dublin Core (ish)
- Converting to MODS for Islandora
- Existing crosswalks are imprecise
- Existing crosswalks don’t handle local fields

Image: Pipeline Walkers
Harold Cosini. Louisiana images from the Standard Oil Collection. Standard Oil (NJ) Collection, Photographic Archives, Archives and Special Collections, University of Louisville: Louisville, KY
Inconsistent Content

- Collection metadata was not standardized
- Records had been created:
  - Over a long period of time...
  - By different staff...
  - At different institutions...
  - Using different interpretations of DC schema & local guidelines
- Many custom fields were present
  - Schema was extended to fit collection needs
- CONTENTdm export issues
  - Collections had different download settings
    - Archival quality vs derivative version vs no download
  - Unreliable indexing
    - Counts of items in collections would vary
Concurrent Platform Development

- The Islandora instance was being configured & customized at the same time as the metadata transformation
  - Display, utilities, architecture, settings, etc. could change
  - Metadata might be leveraged for new features in unexpected ways
  - Example: Updating MODS display XSLT
- Multiple people working on things at the same time
  - Communication was critical
- Challenges of planning a pipeline when key utilities don’t even exist
  - Example: Compound Batch module
Coordination with Content Owners

- LSU agreed to perform the migration for all LDL content, belonging to all consortium partner institutions
  - Ensuring partner needs are served
  - Encouraging participation
  - Handling complex communications
  - Managing expectations
  - General logistics

Image: Twelfth Night Revelers Tableau Design
The Historic New Orleans Collection: New Orleans, LA
TL;DR: Everything is new and in flux

- People
- Platform
- Schema
- Skills

We’re learning as we go…
Mistakes will be made!

Image: Breaking Ground for sewerage in Baton Rouge Louisiana in approximately 1901
Elk’s Souvenir of Baton Rouge. State Library of Louisiana Historic Photograph Collection
State Library of Louisiana: Baton Rouge, LA
Solutions
Stable Source Data

- The ‘Extract’ part of ETL
- Data freeze
  - Required buy-in from consortium partners
- CONTENTdm export script
  - Leverages CDM API
  - Pulls binaries and metadata
  - Keeps compound relationships
- Designated local staging area
  - Corrections & standardization
    - Extensive testing and checking
    - Fix data problems
    - Convert file types
  - Otherwise source data remain unchanged
Metadata Crosswalking

● Built a general template informed by standard DC to MODS crosswalks

● Customized per collection
  ○ Identifies collection problems
  ○ Handles different implementations of guidelines
  ○ Significant time investment

● Cleaned up using XSLT
  ○ Reusable as well as custom stylesheets
  ○ Example: Date formatting

● Preserved original description
  ○ Full DC record inside a MODS extension element
  ○ Display Label attributes preserve custom field names
Scripted Transformations

- Conversion of exported CONTENTdm content to Islandora ingest packages
- Early: Move-to-Islandora Kit (MIK)
- Later: cDM_to.mods
  - Converts metadata records from DC/JSON to MODS XML
  - Applies XSL transformations
  - Validates MODS records
  - Pairs MODS XML file with corresponding content file
  - Compares file counts with expected
  - Creates structure file for compound objects
  - Moves ingest package to staging area
  - Reports errors
  - Posts log message to Trello
Scripted Transformations

- As input, every collection has unique:
  - Source DC metadata files - extracted from CDM
  - Source content files - extracted from CDM and arranged in directory structure
  - DC to MODS mapping - customized from template
  - Configuration file (for MIK pipeline) - customized from template
  - List of XSLT stylesheets to be applied
- All of the above are preserved for each collection
Scripted Transformations

- As each collection goes through the pipeline, the mapping, configuration file*, and XSLTs are updated
  - Config used for MIK pipeline only; not used in cDM_to_mods
- The original items and metadata are untouched
- Metadata are reviewable prior to upload
- In case of problems at any point, the entire collection can be reprocessed
Git and GitHub

- **Version control**
  - Can see differences between versions of code
  - Can roll back bad changes
  - Can comment on changes

- **Allowed collaboration**
  - Code available to all members of team
  - Coordinated instances
    - Of conversion scripts
    - Of XSLT files
    - Of development environments
  - Branches allowed developers and librarians to experiment and write code without breaking others code

- **Caveat**
  - Learning curve for those new to git
Virtual Machines

- Virtual Machines
  - Virtual machine (VM) can run an Islandora stack (Linux web server, drupal, solr, etc.)
  - VM can be built on any platform
    - Laptops, Desktops and Servers (Production and Test servers)
    - Devs on Linux / Libs on Windows
  - VirtualBox
    - Free and widely supported
Virtual Machines

- Virtual machine management and configuration
  - Vagrant
    - Builds and manages the VM
    - Vagrant up
  - Ansible
    - In conjunction with Vagrant, customizes and configures the VM
    - Allows configuration to exist in code
      - Code is on GitHub repository
    - Example: which solution packs to install, and how to configure them

- Caveats
  - Can slow down a desktop/laptop
  - Builds can break
  - Can take a long time to build
  - Need sample data
Kanban board

- **For planning and task management**
  - Site development tasks
  - Scripts and automation

- **Columns indicate task status**
  - Backlog (To Do)
  - Work in progress
  - Blocked/Waiting
  - Done

- **Trello**
  - Assign cards and collaborate
  - Labels/categorize
  - Link cards, add attachments
  - Threaded comments
ETL Board
ETL Board

- Each collection is a Trello card
- Titled with collection name, namespace, institution
- Color coded labels (allowed for filtering)
- Threaded comments
  - Super important!
  - Entire history (comments and status changes) preserved
  - Allowed communication with collection owners
- Checklists
  - Automated using Trello API
ETL Board

- **Columns (representing status)**
  - Don’t migrate these
  - Starting point
  - Waiting on something
  - Has MIK Map and Config files
  - Needs XSL
  - Needs Troubleshooting
  - Whole Collection MODS verified; XML files in Approved.mods folder
  - Whole Collection packaged (Ready for upload)
  - In Process of Upload
  - Uploaded to Islandora, With Problems
  - Uploaded to Islandora
  - Verified correct with preliminary checks
  - Verified correct by collection owner (Done)
ETL Burndown chart

- Entered # of collections in each column in Google Sheet
  - Updated weekly
- Chart was visualization of the movement of collections to completion
- Method for team self-organization
  - Whole team can see progress
- Communicate project status to stakeholders
- Helped identify bottlenecks
  - Example: Collections were stalled at ingest for a period
Work in Action

Image: Cargo ship at dock
Charles L. Thompson Collection
Mss. 998, Louisiana and Lower Mississippi Valley Collections
LSU Libraries, Baton Rouge, La
Putting the pipeline’s reproducibility to the test

- Scripts were swapped
  - Example: MIK to cDM_to_mods
- Architecture was shifted
  - Example: Multisite to namespace
- Schema was revised
  - Example: Replacing @displayLabels
- And yes, mistakes were made
  - Example: Date bugs 1 & 2
Correct [date] bug in convert_to.mods and affected MODS records

Description

Square brackets are being stripped out of the date value, if it's an otherwise standard date - such as [1750] or [1750-05-17] (but not if it's something like [unknown] or [ca. 1750]).

There is existing XSLT to convert bracketed values using attribute 'qualifier="inferred"'.

To Do
Practical Example: Date Bug 1

- Correct code @garrettastrong
- Reprocess collections not yet uploaded
- Migrate islandora_crud module into production server
- Replace MODS datastream in production
  Add an item...

Replace MODS datastream in production
  Add an item...

Reprocess Collections Not Yet Uploaded
  Add an item...

- AAW: America at War (TAHIL)
- PSL: Historic Photographs of Southwest Louisiana (McNeese)
- lapur: Louisiana Purchase Bicentennial Collection (lsu)
- p15140coll26: Edna Tiny Tarbutton Collection (ulm)
- wi-boy: Wright Boyd Family Papers at LSU Libraries (lsu)
- APC: Antebellum Period (TAHI)
- LPC: Louisiana Purchase and Louisiana Colonial History (tahil)
- p16313coll1: Andrew Augustus Gunby Photograph Collection (ulm)
- lsu-p16313coll57: 4, 150, 11, 128, 115, 82, 165, 30, 10, 10, 21, 148, 163, 160, 83, 92, 156, 22, 162, Cpd/108, Cpd/189, Cpd/201...
  new MODS approved — collection in islandora is reportedly incomplete before doing crud, which crud will not remedy — moving this collection to chainsaw/re ingest

Add an Item

Add Comment

Write a comment...
Retrospective

- Culture shock
  - “Libs versus Devs”
  - Value of partnership
- Navigating Agile
  - “There sure are a lot of meetings”
  - Fostering communication
- Using Git, Vagrant, & tech tools as a n00b
  - Yes, there is a learning curve
  - ...But it doesn’t require a CS degree
Thanks!