Batch Processing With Drush

How The University of Wyoming managed a large digital backlog with islandora_book_batch

Wednesday October 15, 2014  By Amanda Lehman
Background: A Digital Backlog

• About 4TB of digitized images
  • Book pages,
  • Photos,
  • Etc.
• Batch uploading increased UW’s digital holdings in Islandora by 85% in approximately 5 months
  • From 460 GB to 3200 GB
  • About 45,000 book page and image objects
Emmett D. Chisum Special Collections

The Grace Raymond Hebard Collection is considered to be the most comprehensive collection of published materials about Wyoming history, culture, and natural history. It is the library of record for University of Wyoming publications. This collection, holding more than 50,000 cataloged titles, helps to document the history and culture of Wyoming and the region by collecting, preserving, and providing access to unique, rare, and valuable books, pamphlets, ephemera, maps, paintings, and other documents. Visit the Special Collections Web Site for more information.

Art In The Libraries

The University of Wyoming Libraries are a strong supporter of art in the Libraries. We do this by purchasing student art to our permanent collection, purchasing art by professional artists, and featuring temporary exhibits by both student and professional local artists. In 2000, the University of Wyoming Libraries established the UW Libraries Purchase Award for student art. The Libraries purchase student art each year and the tradition continues today. The student artist adds a lot of ambiance to the public spaces of the library, and at the same time, students exhibit and give their work exposure to patrons of the library. This page features images and locations of the art in the building. Enjoy! Contact Jennifer Mayer, Fine Arts Librarian, mayer@uwyo.edu.
While waiting for the migration of existing I6 objects, we had the chance to play with different modes of ingest:

- HTTP
- Drush Batches

There are still many issues and lots of things that don’t work at all.
What to do:
Local Processing (before the Drush)

• Walk-through Tutorial on the web (link at the end of the presentation)
• Basic process:
  1. Organize your files into a digestible format for islandora_batch or book_batch modules.
  2. Upload to FTP
  3. Run Drush Scripts for ‘preprocessing’ into Batch Queue, and for Ingest.
1. Organizing files: For islandora_batch - Identical file names
1. Organizing files: For book_batch – Folder Structure

BatchFolder/
my_first_cool_book/
   Page1/
      OBJ.tif
      MODS.xml [page]
      OCR.txt
   Page2/
      OBJ.tif
      DC.xml
      MODS.xml [book]
      PDF.pdf
my_2nd_cool_book/
   Page1/...etc.
Tools for preparing books

- Copying Page Metadata (if you want it)
  - Batch file
- Batch renaming files
  - Advanced Renamer on PC or Automator/Renamer on Mac
- Moving Files into Folders
  - Windows .bat file
- SSH or other File Transfer Protocol
Gather/Rename Files

- Objects
- Metadata/etc.

Batch into Folders

- Simple .bat file that moves and renames.

Add Book Datastreams

- MODS
- PDF
- Etc.

@echo off
for %%a in (*.tif*) do (
    if  %%~a NEQ %0 (md "%%~na" 2>nul
    move "%%a" "%%~na\OBJ.tif"
)
)

for %%a in (*.xml*) do (move "%%a" "%%~na\MODS.xml"

Name | Date modified | Type
---|--------------|-----
001.tif | 8/13/2012 9:49 AM | TIF File
001.txt | 9/25/2014 11:20 AM | TXT File
001.xml | 9/19/2014 9:36 AM | XML File
002.tif | 8/13/2012 9:49 AM | TIF File
002.xml | 9/19/2014 9:36 AM | XML File
003.tif | 8/13/2012 9:49 AM | TIF File
003.xml | 9/19/2014 9:36 AM | XML File
004.tif | 8/13/2012 9:49 AM | TIF File
004.xml | 9/19/2014 9:36 AM | XML File
005.tif | 8/13/2012 9:49 AM | TIF File
005.xml | 9/19/2014 9:36 AM | XML File
filestofolder_OBJ/MODS.bat | 9/23/2014 11:17 AM | Windows Batch
2. Uploading to FTP

- Talk to your system admin to get accounts set up, etc.
- Transfer the appropriate folder(s) into your ingest space.
3. Islandora batch & book_batch Commands

- **Preprocessing** - Different Commands and content models:
  - `islandora_batch_scan_preprocess` (large image, pdf, etc)
  - `islandora_book_batch_preprocess` (--content_models=islandora:bookCModel)
  - Check your queue with the right view: http://yoursite/admin/reports/islandora_batch_queue

- **Same ingest command:**
  - `islandora_batch_ingest`

- **To begin:** drush --help or drush #command --help
  - Options like --create_pdf, etc.
Example Book Batch Preprocess

- `username@server:~$ drush --root=/PathtoDrupal/drupal7 --uri=http://yoursite.coalliance.org --user=admin or username islandora_book_batch_preprocess --namespace=wyu --parent=wyu:25864 --content_models=islandora:bookCModel --type=directory --target=/home/amandarl/ingest/25864books`
Sample Image Batch Preprocess

- `username@server :~$ drush --root=/PathtoDrupal/drupal7 --user=username --uri=http://uwyo.coalliance.org islandora_batch_scan_preprocess --namespace=wyu --content_models=islandora:sp_large_image_cmodel --parent=wyu:61177 --type=directory --target=/home/amandarl/ingest/61177Art`
Check: Islandora Batch Queue/Sets

Enable: Islandora batch queue, Batch sets at /admin/structure/views

View Report: /admin/reports/islandora_batch_queue

- Shared servers/solr? I limit my time/objects to about 12 hrs (overnight) or 1000 objects.
Same Command to Ingest

- `username@server:~$ drush --root=/PathtoDrupal/drupal7 --uri=http://yoursite.coalliance.org --user=admin or username islandora_batch_ingest 2>&1 | tee logfile9-23_BookBatch.txt`

- `-v[erbose] or -d[ebug] if you want to see more details about the processing, memory, derivatives, etc.`
Troubleshooting

- **Metadata Errors** - “tag mismatch, line 1”, etc.
  - Fix the xml file, re-upload, restart ingest!
  - You may have to redo derivatives.
- **Empty string as source, invalid foreach()**
  - Files may be in wrong structure, or too big (before it broke, I found a zip limit of about 4GB).
- **Books: each page an object instead of in a book?**
  - Folder structure too shallow - Folder of Books of Pages…
  - [Helpful Thread](#)
Questions or Comments?

- Thanks to Ed Fugikawa and Robin Dean at the Alliance Digital Repository

- Here is the tutorial document: https://docs.google.com/document/d/1HRaOYulcjHM8UXUDp_4RdH7EXtlIMGfDezcZY7Ymi_w/edit?usp=sharing