Experiences with Off-Production Server Ingests

Randy Fischer, Programmer
Gainesville, Florida
May 6th, 2016
Motivation

• The Florida Virtual Campus maintains an Islandora hosting service that currently serves 18 sites from a monolithic Islandora server.

• Ingests are ongoing, and because of the impact these ingests can have on the production services we have developed software to perform validation and derivative creation for packages on separate virtual hosts.
Architecture

FTP → Load Host → Production Server
FTP → Load Host

Backup Server
FTP Home Layout

/ftp/site/incoming/
/ftp/site/processing/
/ftp/site/warnings/
/ftp/site/errors/
Package Handling - Watcher

god:
Watcher Process

.../incoming/PKG-ID/
move to newly created directory

.../processing/aab/PKG-ID/
enter directory name in queue

Redis Queue
Package Handling - Watcher

god config:

God.watch do |w|
  w.uid = UID
  w.gid = GID
  w.env = { 'CONFIG_FILE' => "/usr/local/islandora/offline-ingest/config.yml" }
  w.name = "ftp-handler"
  w.log_cmd = "'/bin/logger -t '#{w.name}' -p local1.info"
  w.start = "'/usr/local/islandora/offline-ingest/tools/ftp-handler"
  w.keepalive
end
Package Handling - Ingestor

god config:

God.watch do |w|
  w.uid = UID
  w.gid = GID
  w.env = { 'CONFIG_FILE' => '/usr/local/islandora/offline-ingest/config.yml' }
  w.name = "ingest-handler-1"
  w.group = "ingest-handlers"
  w.log_cmd = "/bin/logger -t '#{w.name}' -p local1.info"
  w.start = "#/usr/local/islandora/offline-ingest/tools/ingest-handler digitool ftp"
  w.keepalive
end

more stanzas
Package Handling - Ingestor

**god:**

- **Ingest Process**
  - remove next directory name from queue
  - **fail:** move package to `/ftp/site/errors/`
  - **warn:** move package to `/ftp/site/warnings/`
  - **succeed:** delete package
  - write status info to a db

**Redis Queue**

- attempt ingest
Every affiliate site has a web site to report offline-ingest status

A simple sinatra data browser service: list, filter, drill down

write status info to db
### Ingest Report for fsu.digital.flvc.org: All 34645 Digital Objects Selected

<table>
<thead>
<tr>
<th>Date</th>
<th>Status</th>
<th>Package</th>
<th>Digitool</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-05-05 11:03:15</td>
<td>warning</td>
<td>FSU_WW2_16_0005 ⇒</td>
<td>Feldwebel Schaefer WWII 1940 Battle of France photograph album, 1941-1940</td>
<td></td>
</tr>
<tr>
<td>2016-05-05 07:56:12</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1980_v73n5 ⇒</td>
<td>Florida State University Bulletin Catalog Issue</td>
<td></td>
</tr>
<tr>
<td>2016-05-05 08:49:57</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1968_graduatebulletin ⇒</td>
<td>Florida State University Bulletin Graduate Issue</td>
<td></td>
</tr>
<tr>
<td>2016-05-05 21:30:37</td>
<td>warning</td>
<td>FSU_HPUA_catalog_lifeAtFSU ⇒</td>
<td>Life at FSU Bulletin</td>
<td></td>
</tr>
<tr>
<td>2016-05-04 20:05:11</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1980_v74n3 ⇒</td>
<td>Florida State University Bulletin Catalog Issue</td>
<td></td>
</tr>
<tr>
<td>2016-05-04 16:35:26</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1980_v73n5 ⇒</td>
<td>Florida State University Bulletin Catalog Issue</td>
<td></td>
</tr>
<tr>
<td>2016-05-04 17:47:08</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1980_v74n2 ⇒</td>
<td>Florida State University Schedule of Classes - Spring</td>
<td></td>
</tr>
<tr>
<td>2016-05-04 12:52:53</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1979_v72n5 ⇒</td>
<td>Florida State University Bulletin Catalog Issue</td>
<td></td>
</tr>
<tr>
<td>2016-05-04 11:50:18</td>
<td>success</td>
<td>FSU_HPUA_catalog_1979_winterschedule ⇒</td>
<td>Florida State University Schedule of Classes - Winter Quarter</td>
<td></td>
</tr>
<tr>
<td>2016-05-04 09:41:55</td>
<td>success</td>
<td>FSU_HPUA_catalog_1979_v73n1 ⇒</td>
<td>Florida State University Directory of Classes - Fall Quarter</td>
<td></td>
</tr>
<tr>
<td>2016-05-04 09:11:26</td>
<td>success</td>
<td>FSU_HPUA_catalog_1979_springschedule ⇒</td>
<td>Florida State University Directory of Classes - Spring Quarter</td>
<td></td>
</tr>
<tr>
<td>2016-05-04 09:33:03</td>
<td>error</td>
<td>FSU_HPUA_catalog_1979_v72n5 ⇒</td>
<td>Florida State University Bulletin Catalog Issue</td>
<td></td>
</tr>
<tr>
<td>2016-05-03 17:58:47</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1979_v72n4 ⇒</td>
<td>Florida State University Directory of Classes - Summer Quarter</td>
<td></td>
</tr>
<tr>
<td>2016-05-03 17:40:25</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1978_winterschedule ⇒</td>
<td>Florida State University Schedule of Classes - Winter Quarter</td>
<td></td>
</tr>
<tr>
<td>2016-05-03 17:58:46</td>
<td>error</td>
<td>FSU_HPUA_catalog_1979_spring schedule</td>
<td>Florida State University Tentative Course Listings - Summer Quarter</td>
<td></td>
</tr>
<tr>
<td>2016-05-03 17:32:26</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1978_summertentativelisting ⇒</td>
<td>Florida State University Schedule of Classes - Summer Quarter</td>
<td></td>
</tr>
<tr>
<td>2016-05-03 16:11:14</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1978_summerschedule ⇒</td>
<td>Florida State University Schedule of Classes - Summer Quarter</td>
<td></td>
</tr>
<tr>
<td>2016-05-03 15:31:54</td>
<td>warning</td>
<td>FSU_HPUA_catalog_1978_springschedule ⇒</td>
<td>Florida State University Schedule of Classes - Spring Quarter</td>
<td></td>
</tr>
</tbody>
</table>
Ingest Failure Report for Package FSU_HPUA_catalog_1979_v72n5 on fsu.digital.flvc.org

Title: Florida State University Bulletin Catalog Issue

Ingest Date: 2016-05-04 09:33:03
Processing Time: 0 seconds
Raw Ingest Size: 0.00 MB
Content Type: Book

Islandora Collections: Florida State University Bulletins (fsu:fsubulletins)

Package Errors
The Book Package FSU_HPUA_catalog_1979_v72n5 is missing the following 1 required file declared in the mets.xml file:
- FSU_HPUA_catalog_1979_v72n5_003%201.tif

Package Warnings
The Book Package FSU_HPUA_catalog_1979_v72n5 has the following 1 unexpected file that will not be processed:
- FSU_HPUA_catalog_1979_v72n5_003.tif
Ingest Success Report for Package FSU_HPUA_catalog_1980_v73n2 on fsu.digital.flvc.org

Title: Florida State University Directory of Classes - Winter Quarter
Islandora ID: fsu:305678

Ingest Date: 2016-05-04 14:23:44
Processing Time: 2 hours, 11 minutes, 40 seconds
Raw Ingest Size: 6334.07 MB


Purl: http://purl.flvce.org/fsu/id/FSU_HPUA_catalog_1980_v73n2
Islandora Collections: Florida State University Bulletins (fsu:fsubulletins)
Datastreams: Archived METS for future reference, Dublin Core Record, MODS Record, Relationships, Table of Contents, Thumbnail, XACML Policy Stream
what’s this package thing anyway?
The manifest.xml file

<?xml version="1.0" encoding="UTF-8"?>
<manifest xmlns="info:flvc/manifest/v1">
  <contentModel>islandora:sp_pdf</contentModel>
  <collection>fau:bio</collection>
  <submittingInstitution>FAU</submittingInstitution>
  <owningUser>Randy Fischer</owningUser>
  <owningInstitution>FAU</owningInstitution>
  <objectHistory source="digitool">
    admin_unit="FAU01",
    ingest_id="ing11657",
    creator="creator:OCABEZAS",
    creation_date="2011-12-14 10:20:49",
    modified_by="super:FAUDIG",
    modification_date="2013-06-14 08:50:43"
  </objectHistory>
</manifest>

selected elements saved in MODS stream
# The manifest.xml file

<table>
<thead>
<tr>
<th>ELEMENT</th>
<th>REQUIRED?</th>
<th>REPEATABLE?</th>
</tr>
</thead>
<tbody>
<tr>
<td>collection</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>contentModel</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>identifier</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>label</td>
<td></td>
<td></td>
</tr>
<tr>
<td>otherLogo</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>owningInstitution</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>owningUser</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>submittingInstitution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>embargo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pageProgression</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Packages: supported content models

- basic image
- large image
- pdf
- books
- newspapers
Packages: Books

.../incoming/PKG-ID/
.../incoming/PKG-ID/PKG-ID.xml
.../incoming/PKG-ID/manifest.xml
.../incoming/PKG-ID/mets.xml
.../incoming/PKG-ID/image-01
.../incoming/PKG-ID/image-02
.../incoming/PKG-ID/image-03
...usw
Code

github.com/FLVC/offline-ingest

- based on Hydra rubydora library (tuque in ruby)
- ingest daemon
- command line tools
Code

PackageClass (base class)

• BasicImagePackage
• LargeImagePackage
• PDFPackage

StructuredPageClass (base class)

• BookPackage
• NewspaperPackage
Advantages

- written in ruby
- handles failovers for image derivation (jp2)
- heavy lifting off main server (tesseract)
- reduces tedium of book, newspaper loading
- supports FLVC-specific data streams (e.g. TOC)
- load servers don’t need drupal, fedora installs
Disadvantages

- written in ruby
- multiple datastream uploads (reindexing)
- all or nothing ingest
- errors are opaque to end users
- I have to maintain it
Questions?