A Digital Preservation Repository for Duke University Libraries

Jim Coble
Digital Repository Developer
jim.coble@duke.edu

Open Repositories 2013
Duke University

- Research university in Durham, NC, USA
- 14,500 students, graduate and undergraduate
- Duke University Libraries
  - Centrally administered library system
  - 240 staff
  - 6 million+ volumes
- Professional school libraries serving schools of Business, Law, Divinity, and Medicine
Initial Goal: Preservation Repository

- Focus: Preservation Infrastructure
  - Improve our processes around preservation of digital assets
  - Reduce initial complexity by ignoring discovery and access issues

- First Use Case: Digital Collections Program
  - Familiar with this content
  - Descriptive and technical metadata already exists
  - Separate discovery and access interface already exists
Digital Collections Program

- Digitized content, in-house and out-sourced
- 380,000 archival master files (~ 20 TB)
- Primarily still images, with some audio and video
- Locally developed public access interface
  - [http://library.duke.edu/digitalcollections/](http://library.duke.edu/digitalcollections/)
Current Scenario (Typical)

- Archival master files
  - Produced by library’s Digital Production Center (DPC)
  - Stored on filesystem
  - ACE-AM for periodic checksum validation
- Descriptive metadata
  - Produced by Cataloging and Metadata Services department
  - Maintained in CONTENTdm (or elsewhere)
- Technical metadata
  - Generated and maintained by DPC
  - Nothing ties these elements together except local knowledge and a DPC identifier

Open Repositories 2013
Initial Project Goal

Archival Master Files

Preservation Repository

Descriptive Metadata

DPC Technical Metadata

Open Repositories 2013
Technology

- Fedora Commons Repository
- Hydra Project Framework
  - Fedora (repository)
  - Solr (index)
  - Blacklight (discovery and access)
  - Hydra-Head (object creation / management)
Resources

- Experience on prior project (abandoned before production)
  - Fedora
  - Modeling digital collections content
- Two developers
  - Part-time, though proportion of time increased throughout this project
  - Web application development experience (Django/Python, Java servlets)
  - No prior Ruby or Rails experience
Timeline

- **Spring 2012**: Prototype using Fedora command line utilities and Django using “found time”
- **June 2012**: Project formally launched
- **July 2012**: OR 2012; growing interest in Hydra Project
- **October 2012**: HydraCamp at Penn State; Hydra-based development begins in earnest
- **February 2013**: Initial pilot completed
- **April 2013**: Duke becomes Hydra Partner
- **June 2013**: Production preservation repository launched with two collections ingested
Content Models

- Collection
  - Collection-level descriptive metadata
  - Aggregated metadata about items / components in some cases

- Item
  - Item-level descriptive metadata

- Component
  - Digital content file (e.g., TIFF image file)
  - Technical metadata

- Target
  - External digitization target image
  - Digital content file for target image
Additional Models

- **AdminPolicy**
  - Used in Hydra Framework to specify access rights
  - Individual objects are “governed by” a particular AdminPolicy

- **PreservationEvent**
  - Records PREMIS Event data for …
    - Ingest
    - Ingest validation
    - Periodic fixity checks
  - Associated with object to which it applies
Metadata Practices

- Collect metadata available at time of ingest
  - CONTENTdm
  - MarcXML from library catalog
  - Digitization Guide from DPC
  - etc
- Store collected metadata in its native formats in object datastreams
- Normalize one set of descriptive metadata into Qualified Dublin Core for indexing and display
Batch Ingest

- **Problem to solve**
  - 380,000 archival master files (~ 20 TB) spanning 8 years of digitization work
  - Some areas of relative consistency across the collections but also some divergences
- **Needed flexible batch ingest mechanism**
- **Solution**: Ingest “Manifest”
  - Enumerates the objects to be ingested in any given batch
  - Provides information about nature and location of content files, metadata, and related objects
Ingest Manifest

YAML File:

```yaml
basepath: /srv/fedora-working/ingest/KWL/component/
model: Component
adminpolicy: duke-apo:KwileckiCollection
label: Paul Kwilecki Photographs and Papers Image
parent:
  autoidlength: 14
  master: /srv/fedora-working/ingest/KWL/item/master/master.xml
metadata:
  - descmetadata
content:
  extension: .tif
  location: /nas/TUCASI_CIFS2/dpc-archive/Archived_NoAccess/na_KWL/ph/01/
  creator: DPC
checksum:
  location: sha256_na_KWL.xml
  source: dpc
objects:
  - identifier: kwlp010010010
  - identifier: kwlp010010020
  - identifier: kwlp010010030
  - identifier: kwlp010010040
```
Ingest Processor v1.0

- Reads manifest file
- Performs any needed pre-ingest steps
- Creates a repository object for each object in turn
  - Adds appropriate datastreams and relationships
  - Creates thumbnail image from uploaded digital content
  - Creates Ingestion PreservationEvent
- Validates each ingested object in turn
  - Compares repository object with manifest
  - Validates content file against external checksum if available
  - Creates Validation PreservationEvent and first Fixity Check PreservationEvent
Validation PreservationEvent

In PreservationEvent eventMetadata datastream ...

```xml
...<eventType>validation</eventType>
<eventDateTime>2013-06-04T14:52:17.960Z</eventDateTime>
<eventOutcomeInformation>
  <eventOutcome>success</eventOutcome>
  <eventOutcomeDetail>
    <eventOutcomeDetailNote>
      Identifier(s): kwlph010010020
      Verifying...PID found in master file...PASS
      Verifying...Component object found in repository...PASS
      Verifying...DC datastream present and not empty...PASS
      Verifying...RELS-EXT datastream present and not empty...PASS
      Verifying...descMetadata datastream present and not empty...PASS
      Verifying...content datastream present and not empty...PASS
      Verifying...DC datastream internal checksum...PASS
      Verifying...RELS-EXT datastream internal checksum...PASS
      Verifying...descMetadata datastream internal checksum...PASS
      Verifying...content datastream internal checksum...PASS
      Verifying...thumbnail datastream internal checksum...PASS
      Verifying...content datastream external checksum...PASS
      Verifying...child relationship to identifier kwlph010010020...PASS
      Object ingest...VALIDATES
    </eventOutcomeDetailNote>
  </eventOutcomeDetail>
</eventOutcomeInformation>
<linkingObjectIdentifier>
  <linkingObjectIdentifierType>object</linkingObjectIdentifierType>
  <linkingObjectIdentifierValue>info:fedora/duke:1275</linkingObjectIdentifierValue>
</linkingObjectIdentifier>
</event>
```
Export Sets

- Example service built on top of repository infrastructure
- Delivering archival master files to authorized patrons upon request
- Current process is manual
  - DPC staff locate master file(s) on filesystem
  - Possibly create a zip file
  - Place file(s) in pick-up location or copy onto CD, DVD, etc., for delivery
- Pre-Hydra prototype implementation was Django web app using Fedora REST API
Export Sets

- Built on bookmark functionality
  - Staff member searches for content-bearing objects of interest and bookmarks them
  - Export set can be created from bookmark list
- Content files are retrieved from the repository and bundled into a zip file
  - Staff member can download and deliver to patron
  - Zip file includes a README manifest listing the content files with basic metadata
Export Sets

- Export sets can be named and stored for re-use
- By default, zip file is also stored
- Staff member can delete the zip file (to save space) and re-generate it as needed from the export set record
- When no longer needed, export set record can be deleted
Screenshot
Walk-Through
Welcome!

To browse the repository by category, use the facets on the left side of this page.
Collection Index
Collection Content: Items

Paul Kwilecki Photographs

1. Mother and son waiting outside court room, 1981 Jan. (Understandings)
   Item duke:935
   kwph010010010

   Item duke:937
   kwph010010020

Open Repositories 2013
Item Contents: Components

Three judges, 1979 Nov.

1. Three judges, 1979 Nov.

Component duke:2209
kwph010010030
image/tiff 13.3 MB
Three judges, 1979 Nov.

Title: Three judges, 1979 Nov.
Identifier: kwph010010030
Description: Caption by photographer (Kwilecki): "On the wall a picture of Judge Crow (deceased). Standing, Judge Robert Culpeper. Seated, Judge Wallace Cato."
Date: 1979-11
Creator: Kwilecki, Paul, 1928-
Collection FCRepo View

Collection duke:932

Datastreams

<table>
<thead>
<tr>
<th>Datastream ID</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>Dublin Core Record for this object</td>
</tr>
<tr>
<td>RELS-EXT</td>
<td>Fedora Object-to-Object Relationship Metadata</td>
</tr>
<tr>
<td>rightsMetadata</td>
<td>Not currently used</td>
</tr>
<tr>
<td>descMetadata</td>
<td>Descriptive Metadata for this object</td>
</tr>
<tr>
<td>thumbnail</td>
<td>Thumbnail for this object</td>
</tr>
<tr>
<td>contentdm</td>
<td>CONTENTdm Data for this object</td>
</tr>
<tr>
<td>digitizationGuide</td>
<td>Digitization Guide Data for this object</td>
</tr>
<tr>
<td>dpcMetadata</td>
<td>DPC Metadata for this object</td>
</tr>
<tr>
<td>fmpExport</td>
<td>FileMakerPro Export Data for this object</td>
</tr>
<tr>
<td>marcXML</td>
<td>Aleph MarcXML Data for this object</td>
</tr>
<tr>
<td>tripodMets</td>
<td>Tripod METS Data for this object</td>
</tr>
</tbody>
</table>

Properties

| Label: | Paul Kwilecki Photographs and Papers Collection |
| State: | A (Active) |
| Created: | 2013-05-17 11:28:45 -0400 |
| Modified: | 2013-06-27 10:02:12 -0400 |
| Owner ID: | fedoraAdmin |

Open Repositories 2013
Creating Export Set

Bookmarks

1. Three men talking in front of court house, 1970 (Understandings)
   - Component duke:2211
   - kwiph010010040
   - image file: 15.6 MB

   - Component duke:2207
   - kwiph010010029
   - image file: 23.6 MB

3. Three judges, 1979 Nov.
   - Component duke:2209
   - kwiph010010030
   - image file: 13.3 MB

Open Repositories 2013
Creating Export Set

New Export Set

Title: Sample Export Set

Select content-bearing objects from your bookmarks to export:

<table>
<thead>
<tr>
<th>#</th>
<th>PID</th>
<th>Type</th>
<th>Title</th>
<th>Identifier</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>duke:2211</td>
<td>Component</td>
<td>Three men talking in front of court house, 1970 (Understanding)</td>
<td>kwiph010010040</td>
<td>kwiph010010040.tif</td>
</tr>
<tr>
<td>2</td>
<td>duke:2207</td>
<td>Component</td>
<td>Outside court room. Second floor, 1987 Sept.</td>
<td>kwiph010010020</td>
<td>kwiph010010020.tif</td>
</tr>
<tr>
<td>3</td>
<td>duke:2209</td>
<td>Component</td>
<td>Three judges, 1979 Nov.</td>
<td>kwiph010010030</td>
<td>kwiph010010030.tif</td>
</tr>
</tbody>
</table>

Create Export Set
Export Set Created

Export Set created.

Title: Sample Export Set
Created: 2013-07-01 15:01:51 -0400
Archive: export_set_20130701150144.zip

Exported Objects: Metadata

<table>
<thead>
<tr>
<th>PID</th>
<th>Type</th>
<th>Title</th>
<th>Identifier</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>duke:2211</td>
<td>Component</td>
<td>Three men talking in front of court house, 1970 ( Understandings)</td>
<td>kwiph010010040</td>
<td>kwiph010010040.tif</td>
</tr>
<tr>
<td>duke:2207</td>
<td>Component</td>
<td>Outside court room. Second floor, 1987 Sept.</td>
<td>kwiph010010020</td>
<td>kwiph010010020.tif</td>
</tr>
<tr>
<td>duke:2209</td>
<td>Component</td>
<td>Three judges, 1979 Nov.</td>
<td>kwiph010010030</td>
<td>kwiph010010030.tif</td>
</tr>
</tbody>
</table>
Export Set Zip File
Future Plans

 Version 1.1 – By September 2013
   Interface improvements
   Refactored batch ingest

 Future enhancements
   Ingest (batch and individual) performed by library staff
   Editing capability

 Future Use Cases
   Faculty scholarship, electronic theses and dissertations
   Electronic records and other born-digital content
   Datasets
   Image library for teaching / learning

Open Repositories 2013
Questions?

Jim Coble
jim.coble@duke.edu
Digital Repository Developer
Duke University Libraries

Project
https://github.com/duke-libraries/dul-hydra

Open Repositories 2013