

A Digital Preservation Repository for Duke University Libraries

Jim Coble

Digital Repository Developer

jim.coble@duke.edu

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



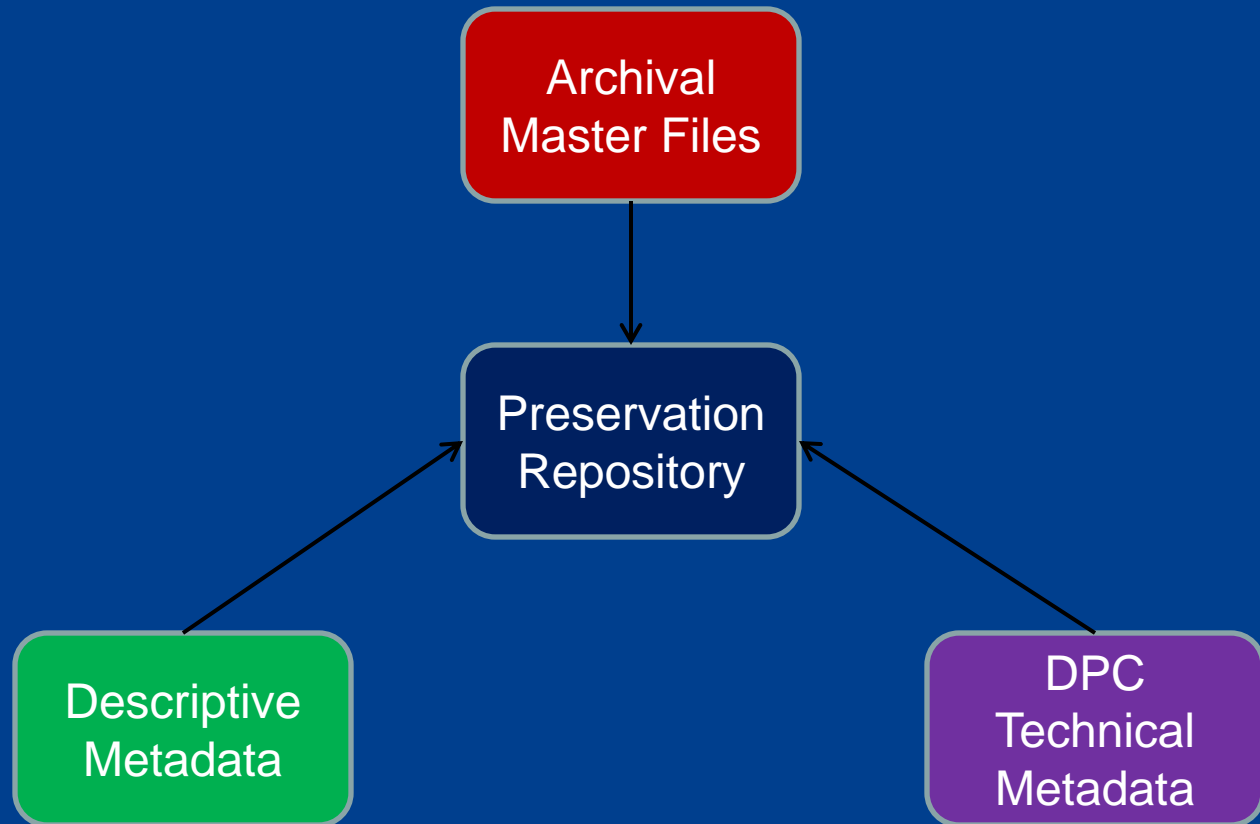
Digital Collections Program

- 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/>

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

Initial Project Goal



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:

```
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: kwlph010010010  
  - identifier: kwlph010010020  
  - identifier: kwlph010010030  
  - identifier: kwlph010010040
```

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 ...

```
...
<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

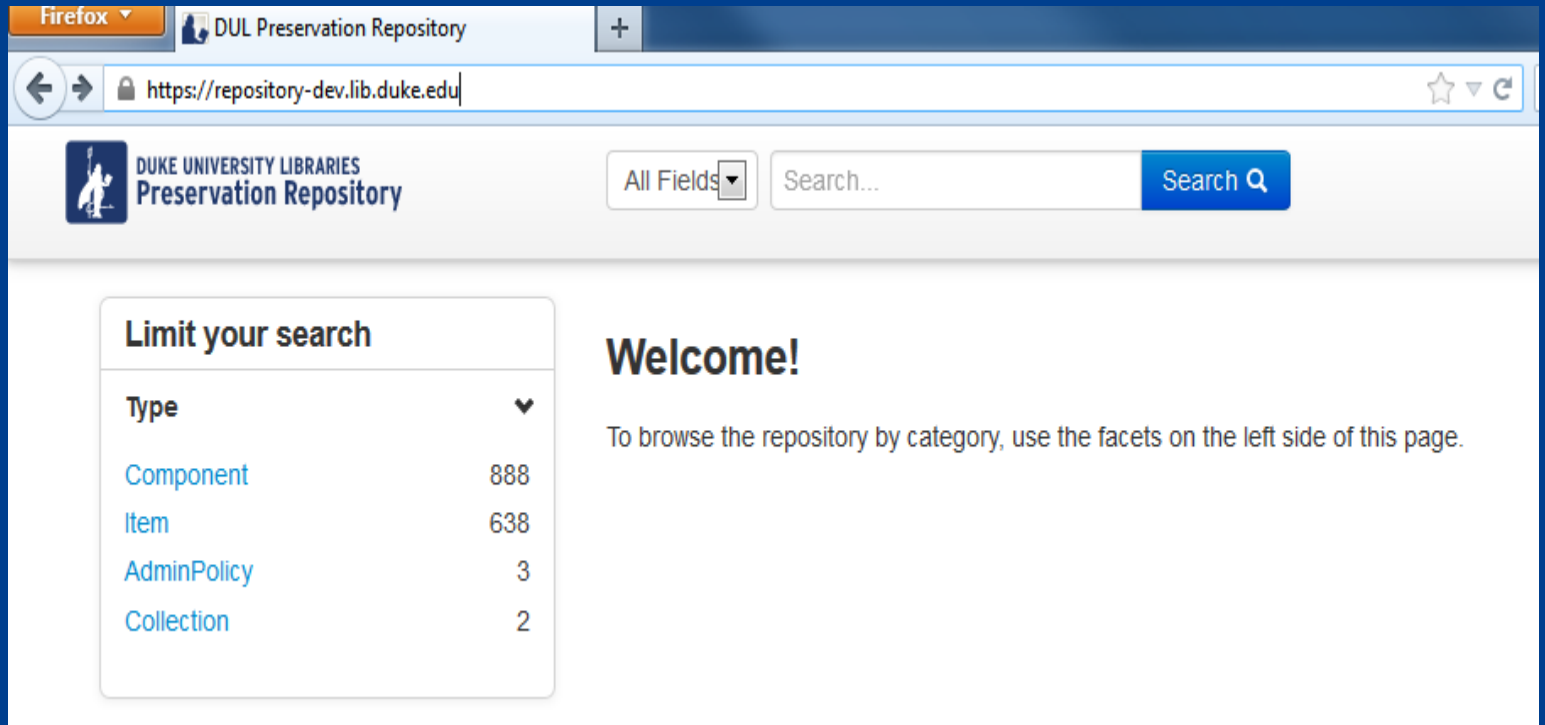
- Export sets can be named and stored for re-use
- By default, zip file is also stored
- Staff member can delete the zip file (to save space) and re-generate it as needed from the export set record
- When no longer needed, export set record can be deleted



Screenshot Walk-Through

Open Repositories 2013

Repository Home Page



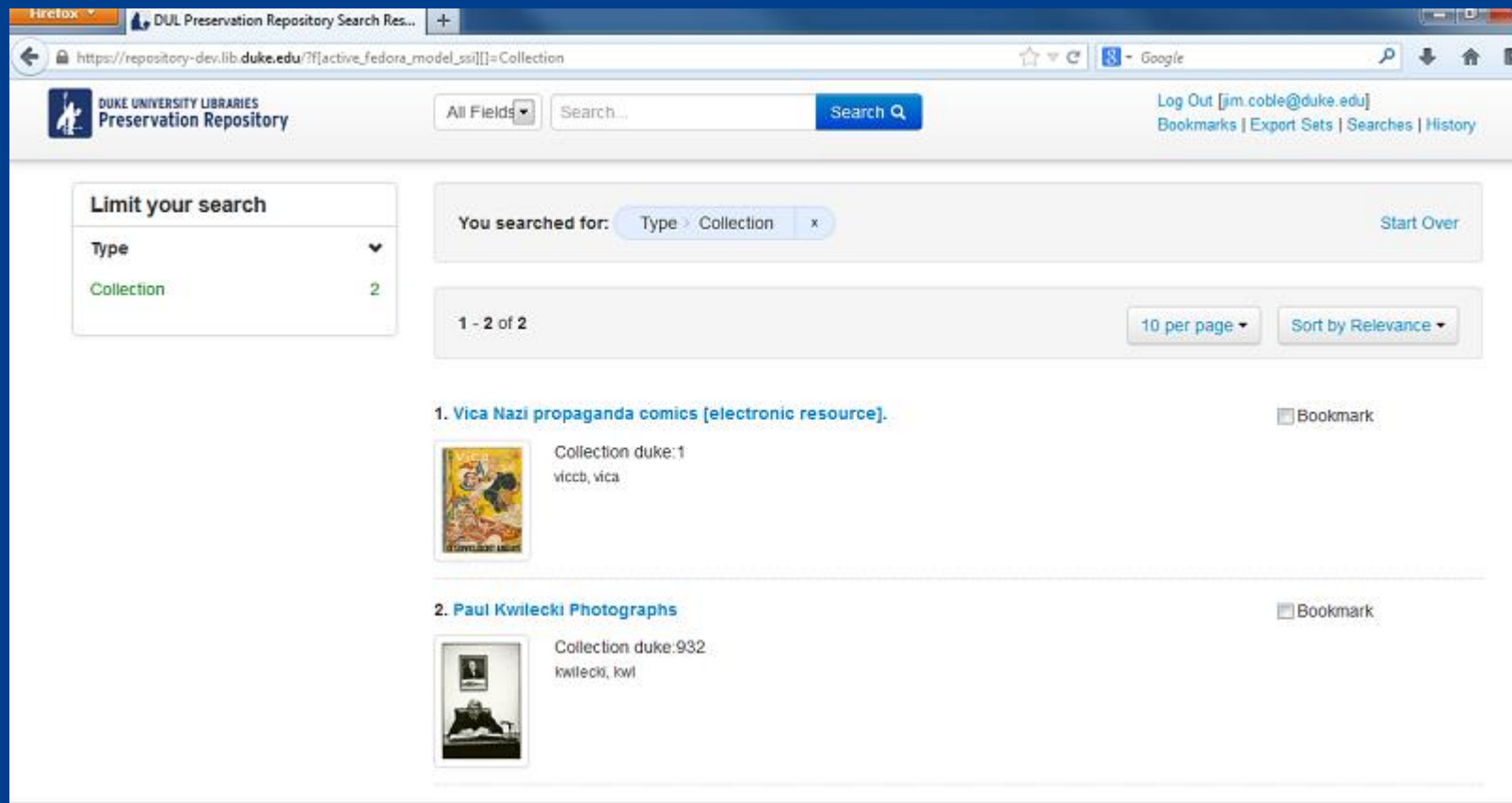
The screenshot shows a Firefox browser window with the address bar displaying `https://repository-dev.lib.duke.edu`. The page header includes the Duke University Libraries logo and a search bar with a dropdown menu set to "All Fields" and a "Search" button. Below the header, there is a "Limit your search" section with a table of facets and a "Welcome!" section with a brief instruction.

Limit your search	
Type	
Component	888
Item	638
AdminPolicy	3
Collection	2

Welcome!

To browse the repository by category, use the facets on the left side of this page.

Collection Index



Firefox | DUL Preservation Repository Search Res... | +

https://repository-dev.lib.duke.edu/?f[active_fedora_model_ssi][]=Collection

DUKE UNIVERSITY LIBRARIES
Preservation Repository

All Fields Search... Search

Log Out [jim.coble@duke.edu]
Bookmarks | Export Sets | Searches | History



Limit your search

Type ▾

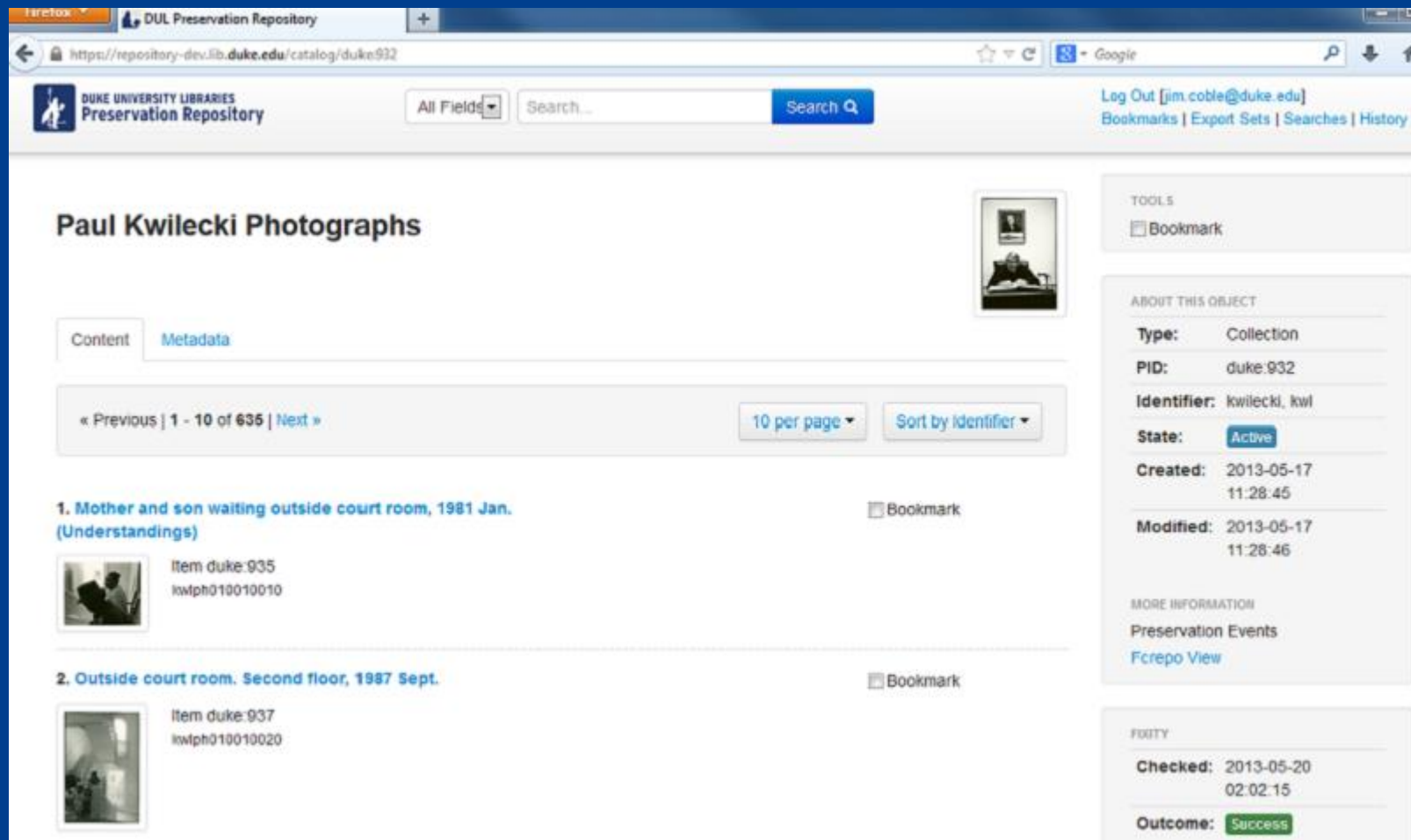
Collection 2

You searched for: Type > Collection x Start Over

1 - 2 of 2 10 per page ▾ Sort by Relevance ▾

- Vica Nazi propaganda comics [electronic resource].** Bookmark
 Collection duke:1
viccb, vica
- Paul Kwilecki Photographs** Bookmark
 Collection duke:932
kwilecki, kwl

Collection Content: Items



The screenshot displays the 'DUL Preservation Repository' interface. The main heading is 'Paul Kwilecki Photographs'. Below the heading, there are tabs for 'Content' and 'Metadata'. A navigation bar shows '« Previous | 1 - 10 of 635 | Next »', '10 per page', and 'Sort by Identifier'. Two items are listed:

- 1. Mother and son waiting outside court room, 1981 Jan. (Understandings)**
Item duke:935
kwiph010010010
- 2. Outside court room. Second floor, 1987 Sept.**
Item duke:937
kwiph010010020

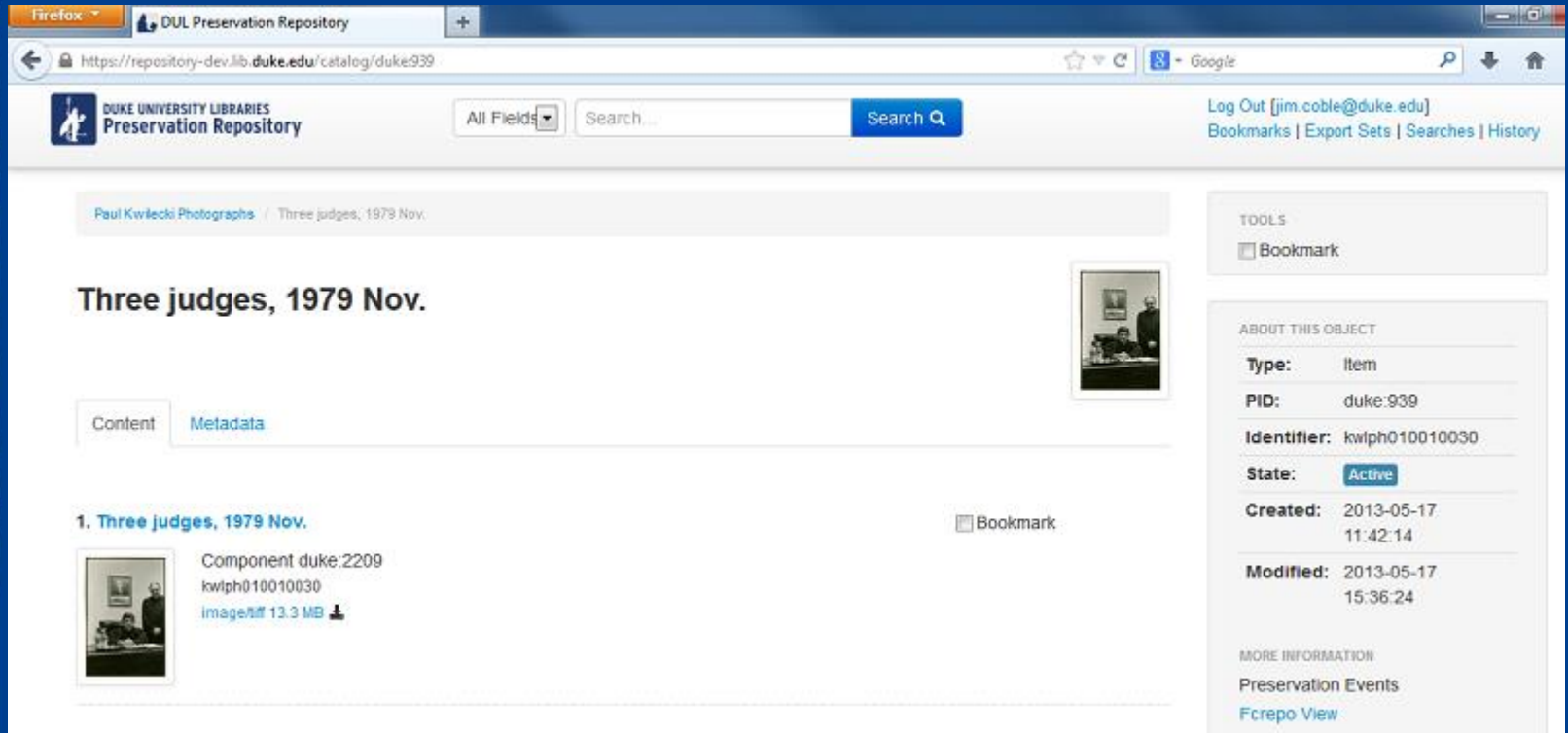
Each item has a small thumbnail image and a 'Bookmark' button. On the right side, there is a 'TOOLS' section with a 'Bookmark' button, and an 'ABOUT THIS OBJECT' section with the following details:

- Type:** Collection
- PID:** duke:932
- Identifier:** kwilecki, kwl
- State:** Active
- Created:** 2013-05-17 11:28:45
- Modified:** 2013-05-17 11:28:45

Below this is a 'MORE INFORMATION' section with 'Preservation Events' and 'Fcrepo View' links. At the bottom right, a 'FOOTY' section shows:

- Checked:** 2013-05-20 02:02:15
- Outcome:** Success

Item Contents: Components



Firefox | + | **DUL Preservation Repository** | +

https://repository-dev.lib.duke.edu/catalog/duke939

DUKE UNIVERSITY LIBRARIES
Preservation Repository

All Fields Search Search


Log Out [jim.coble@duke.edu]
Bookmarks | Export Sets | Searches | History

Paul Kwilecki Photographs / Three judges, 1979 Nov.

Three judges, 1979 Nov.

Content Metadata

1. **Three judges, 1979 Nov.** Bookmark

 Component duke:2209
kwlph010010030
image/tiff 13.3 MB

TOOLS
 Bookmark

ABOUT THIS OBJECT

Type:	Item
PID:	duke:939
Identifier:	kwlph010010030
State:	Active
Created:	2013-05-17 11:42:14
Modified:	2013-05-17 15:36:24

MORE INFORMATION
Preservation Events
[Fcrepo View](#)

Item Metadata

Firefox | DUL Preservation Repository | +

https://repository-dev.lib.duke.edu/catalog/duke:939

DUKE UNIVERSITY LIBRARIES Preservation Repository | All Fields | Search... | Search Q | Log Out [jim.coble@duke.edu] | Bookmarks | Export Sets | Searches | History

Paul Kwilecki Photographs / Three judges, 1979 Nov.

Three judges, 1979 Nov.

Content | Metadata

Download XML | Fcrepo View

Title	Three judges, 1979 Nov.
Identifier	kwlph010010030
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-

TOOLS
 Bookmark

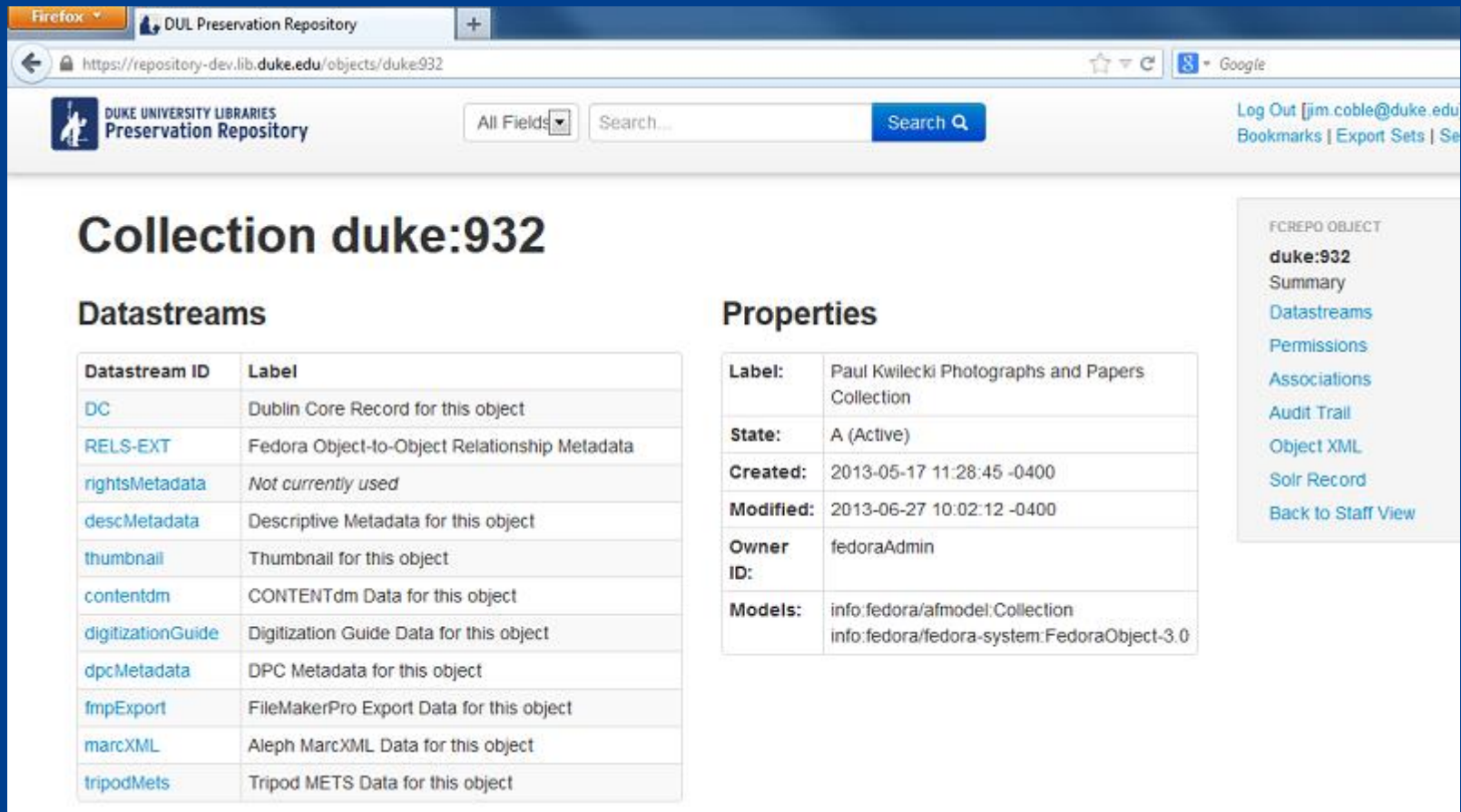
ABOUT THIS OBJECT

Type:	Item
PID:	duke:939
Identifier:	kwlph010010030
State:	Active
Created:	2013-05-17 11:42:14
Modified:	2013-05-17 15:36:24

MORE INFORMATION
[Preservation Events](#)
[Fcrepo View](#)

FIXITY

Collection FCRepo View



Firefox | DUL Preservation Repository | + | https://repository-dev.lib.duke.edu/objects/duke:932 | Google

DUKE UNIVERSITY LIBRARIES Preservation Repository | All Fields | Search... | Search Q | Log Out [jim.coble@duke.edu] | Bookmarks | Export Sets | Se

Collection duke:932

Datastreams

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

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
Models:	info:fedora/afmodel:Collection info:fedora/fedora-system:FedoraObject-3.0

FCREPO OBJECT

- duke:932**
- [Summary](#)
- [Datastreams](#)
- [Permissions](#)
- [Associations](#)
- [Audit Trail](#)
- [Object XML](#)
- [Solr Record](#)
- [Back to Staff View](#)

Creating Export Set

The screenshot shows a web browser window with the URL <https://repository-dev.lib.duke.edu/bookmarks>. The page title is "Bookmarks". At the top, there is a search bar with "All Fields" selected and a "Search" button. To the right, there are links for "Log Out [jim.coble@duke.edu]", "Bookmarks", "Export Sets", "Searches", and "History". Below the search bar, there are two buttons: "Clear Bookmarks" and "Create Export Set", with the latter circled in red. The main content area displays a list of three bookmarks:

- 1. Three men talking in front of court house, 1970 (Understandings)** In Bookmarks
Component duke:2211
kwlph010010040
image 15.6 MB
- 2. Outside court room. Second floor, 1987 Sept.** In Bookmarks
Component duke:2207
kwlph010010020
image 23.6 MB
- 3. Three judges, 1979 Nov.** In Bookmarks
Component duke:2209
kwlph010010030
image 13.3 MB

Creating Export Set

Firefox | DUL Preservation Repository | +

https://repository-dev.lib.duke.edu/export_sets/new

DUKE UNIVERSITY LIBRARIES
Preservation Repository

Search... Search Q

New Export Set

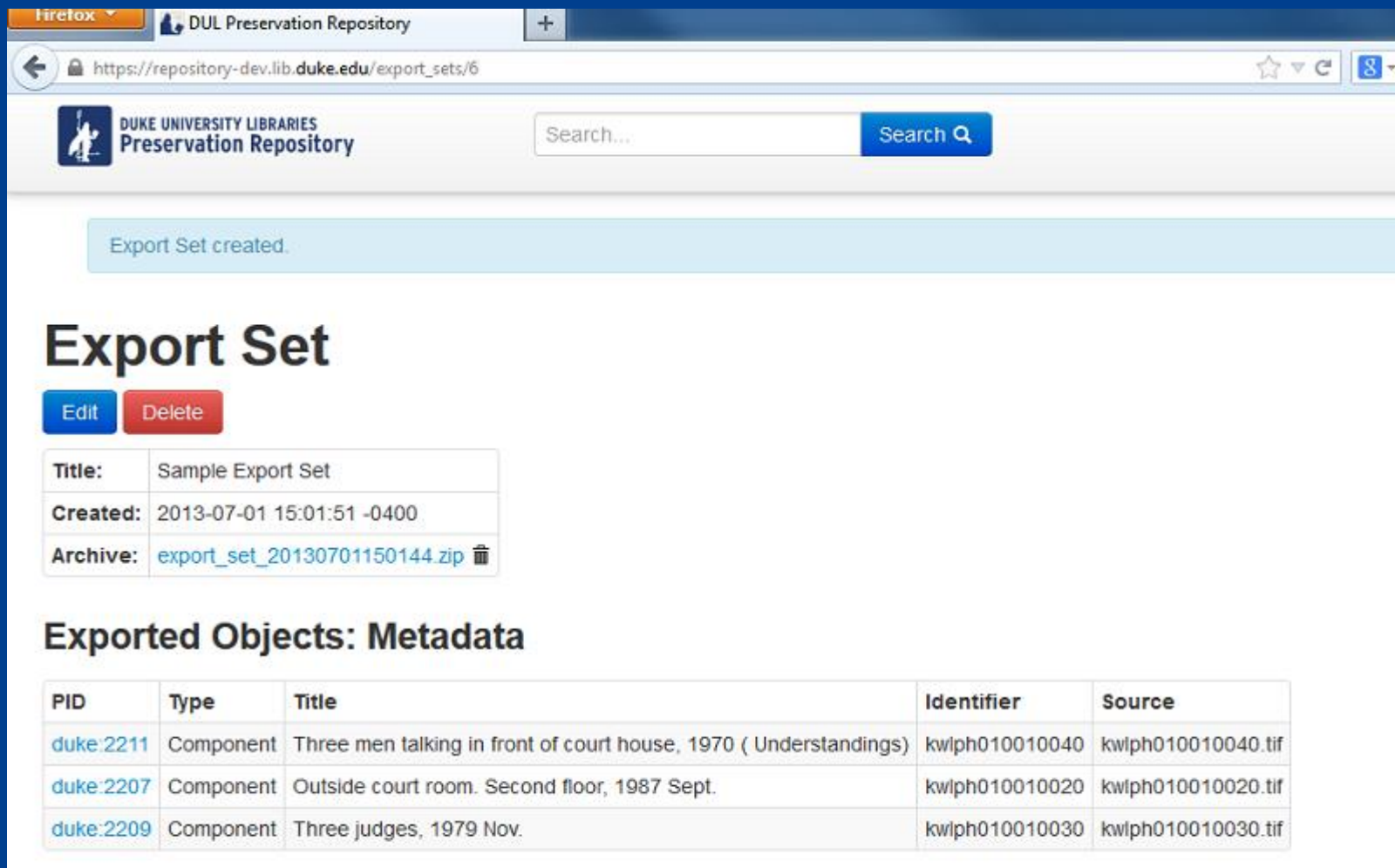
Title

Select content-bearing objects from your bookmarks to export:

#	PID	Type	Title	Identifier	Source
<input checked="" type="checkbox"/>	duke:2211	Component	Three men talking in front of court house, 1970 (Understandings)	kwlph010010040	kwlph010010040.tif
<input checked="" type="checkbox"/>	duke:2207	Component	Outside court room. Second floor, 1987 Sept.	kwlph010010020	kwlph010010020.tif
<input checked="" type="checkbox"/>	duke:2209	Component	Three judges, 1979 Nov.	kwlph010010030	kwlph010010030.tif

Create Export set

Export Set Created



Firefox | DUL Preservation Repository | +

https://repository-dev.lib.duke.edu/export_sets/6


DUKE UNIVERSITY LIBRARIES
Preservation Repository

Search... Search

Export Set created.

Export Set

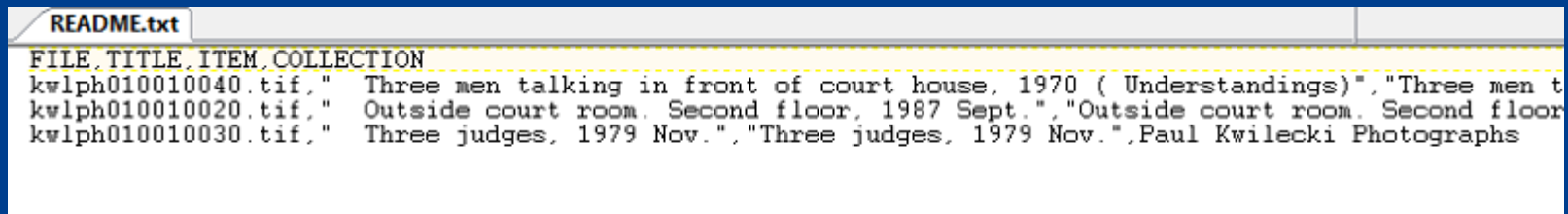
Edit Delete

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

Exported Objects: Metadata

PID	Type	Title	Identifier	Source
duke:2211	Component	Three men talking in front of court house, 1970 (Understandings)	kwlph010010040	kwlph010010040.tif
duke:2207	Component	Outside court room. Second floor, 1987 Sept.	kwlph010010020	kwlph010010020.tif
duke:2209	Component	Three judges, 1979 Nov.	kwlph010010030	kwlph010010030.tif

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

Questions?

Jim Coble

jim.coble@duke.edu

Digital Repository Developer

Duke University Libraries

Project

<https://github.com/duke-libraries/dul-hydra>