Workflows for Digital Curation and Preservation

Stacy Kowalczyk
PASIG Dublin 2012
October 17, 2012
Topics

• Goals
• A Very Brief Introduction to Workflow Systems
• Components for Curation
• Workflow Scenarios
• Future Work
Workflows for Curation

Goals

– Increase capacity and scalability of curation efforts
– Develop distributed curation processes
– Lower costs of curation activities
– Improve quality with systematic and repeatable processes
– Reduce human errors
Why Workflow Systems

- Repetitive and mundane activities simplified
- Facilitates and enforces best practices
- Enables efficient scheduling
- Machinery for coordinating the execution of services and linking together resources
- Facilitates outreach to researchers for direct deposit and automatic curation
Types of Workflow Systems

Kepler

Triana

Ptolemy II

Taverna

BPEL
Trident

- Open source project
- Based on Microsoft Workflow Foundation classes
- Supported by Microsoft Research and academic researchers
- Integrates with myExperiment
- Well accepted in the research community
  - well over 100 peer-reviewed and white papers were discovered from one scholarly aggregation service
- Graphical workflow design and execution interface
Trident Workflow Components

- Fixity
- Data Integrity
- Metadata Creation
- Format Normalization and Derivative Generation
- Persistent Identification
- Repository Integration
Fixity Components

- MD5 checksum generator

- MD5 checksum validator
Data Integrity Components

• JHOVE for format verification and validation

• Group validation (for object integrity)
Metadata Creation Components

• MIX data generator and validator

• METS data generator and validator
Format Components

• Format Conversions for normalization and derivative generation
  – .xlsx to .csv
  – .docx to .pdf
  – .ppt to .pdf
  – .tif to .jpg
  – Zipping on demand
  – Image (.tif or .jpg) to .pdf (single document and multipage)
Repository Component

• Ingest to DSpace via Sword

• DOI generator
Data Ingest Workflows

• Scenarios
  – Single part objects (individual images)
  – Multi-part objects (a book)
  – Multiple instantiations of a logical object (word, pdf and ppt of a research paper)
  – Multiple multi-part objects (a group of letters)
  – Research data products (multiple files of various types)
# Single Part Objects

<table>
<thead>
<tr>
<th>Date Scanned</th>
<th>DLP Slide ID</th>
<th>Image Width</th>
<th>Image Height</th>
<th>File size (MB)</th>
<th>Username</th>
<th>Hardware</th>
<th>Software</th>
<th>Slide #</th>
<th>Set</th>
<th>Box #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00055-01</td>
<td>3803</td>
<td>3740</td>
<td>40.7</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>Ghost of Christmas Future</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00056-01</td>
<td>3737</td>
<td>3684</td>
<td>39.4</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>Buy in Classroom</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00057-01</td>
<td>2396</td>
<td>2361</td>
<td>16.1</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>Man with Horse</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00058-01</td>
<td>2366</td>
<td>2381</td>
<td>16.1</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>Man with dog and bird</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00059-01</td>
<td>3755</td>
<td>3684</td>
<td>39.6</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>Buy your own cherries 1</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00060-01</td>
<td>3599</td>
<td>3684</td>
<td>39</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>Buy your own cherries 2</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00061-01</td>
<td>3561</td>
<td>3684</td>
<td>39.6</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>Buy your own cherries 3</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00062-01</td>
<td>2992</td>
<td>3703</td>
<td>30.6</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>&quot;From Greenland's icy mountains&quot;</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00063-01</td>
<td>3605</td>
<td>3646</td>
<td>37.6</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>A Kaffir's kraal, South Africa</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00064-01</td>
<td>3643</td>
<td>3571</td>
<td>37.2</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>A Native Village near the Zambesi</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00065-01</td>
<td>3079</td>
<td>3722</td>
<td>32.8</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>Transvaal: Swazi Boys</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00066-01</td>
<td>3505</td>
<td>31.17</td>
<td>32.1</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>Missionary travelling</td>
</tr>
<tr>
<td>12/01/2009</td>
<td>VAB7905-00067-01</td>
<td>3599</td>
<td>2872</td>
<td>30.4</td>
<td>heirober</td>
<td>B</td>
<td>Silverfast 6.6</td>
<td></td>
<td></td>
<td></td>
<td>Picking cotton</td>
</tr>
<tr>
<td>12/03/2009</td>
<td>VAB7905-00068-01</td>
<td>19403</td>
<td>3812</td>
<td>211.6</td>
<td>and</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Napoleon</td>
</tr>
<tr>
<td>12/03/2009</td>
<td>VAB7905-00069-01</td>
<td>19224</td>
<td>3530</td>
<td>194.2</td>
<td>and</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dancing</td>
</tr>
<tr>
<td>12/03/2009</td>
<td>VAB7905-00070-01</td>
<td>19060</td>
<td>3653</td>
<td>199.2</td>
<td>and</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>horse race</td>
</tr>
</tbody>
</table>
Single Part Objects Workflow

- Fixity Check
- Format Validation and Verification
- Image Quality Checks
- Derivative Generation
- Persistent Identification
- Create Tech Metadata
- Create Intellectual Metadata
- Create Object Metadata
- Deposit in Repository
Single Part Objects Workflow

- For each original image
  - MD5 checksum
  - JHOVE validation and verification report
  - ImageMagick report
  - MIX file
- For each derivative file
  - MD5 Checksum
  - DOI
- For each logical object
  - DC record
  - METS record
  - Sword package
Multi-part Object Workflow
Multi-part Object Workflow

• Comic Book
  – RIS
  – Set of .tif files
Multi-part Object Workflow

• For each individual image file
  – MD5 checksum
  – JHOVE validation and verification report
  – ImageMagick report
  – MIX file
• For each derivative file
  – MD5 Checksum
• For the whole object
  – DOI
  – DC record
  – METS record
• Sword Package
Multiple Instantiations of a Logical Object Workflow
Multiple Instantiations of a Logical Object Workflow

• Papers
  – Each logical object per subdirectory
  – RIS, word file and (perhaps) supplemental file

- Fixity Check
- Format Validation and Verification
- Format Normalization
- Derivative Generation
- Create Intellectual Metadata
- Persistent Identification
- Create Object Metadata
- Deposit in Repository
Multiple Instantiations of a Logical Object Workflow

• For each original object
  – MD5 Checksum
  – JHOVE report

• For each derivative object
  – MD5 Checksum
  – Output from normalization process
  – DOI for delivery object

• For the whole package
  – METS file
  – DC record
  – Sword Package
Multiple Multi-part Object Workflow
Multiple Multi-part Object Workflow

• Ball collection
  – RIS for collection and Inventory spreadsheet
  – Each logical object in separate subdirectory

Collection Integrity
Fixity Check
Format Validation and Verification
Image Quality Checks
Derivative Generation
Create Tech Metadata
Create Intellectual Metadata
Persistent Identification
Create Object Metadata
Create Collection Metadata
Deposit in Repository
Multiple Multi-part Object Workflow

- For each file
  - MD5 checksum
  - JHOVE report
  - MIX file
  - Scanning specifications
  - Derivative files

- For each logical object
  - Derivative object
  - DC record
  - METS file
  - DOIs

- For the whole collection
  - METS file
  - DC record
Research Data Products

<table>
<thead>
<tr>
<th>Name</th>
<th>Date Modified</th>
<th>Size</th>
<th>Kind</th>
</tr>
</thead>
<tbody>
<tr>
<td>forecast_20100430120000Z_run001</td>
<td>Jun 25, 2012 3:54 PM</td>
<td>--</td>
<td>Folder</td>
</tr>
<tr>
<td>forecast_20100502130000Z_run001</td>
<td>Jun 25, 2012 4:43 PM</td>
<td>--</td>
<td>Folder</td>
</tr>
<tr>
<td>forecast_20100502130000Z_run002</td>
<td>Jun 25, 2012 5:09 PM</td>
<td>--</td>
<td>Folder</td>
</tr>
</tbody>
</table>

```
<metadata>
  <idinfo>
    <citeinfo>
      <origin>Beth Plale, Keith Brewster, Craig Mattocks, Ashish Bhagale, Eran C Withana, Chathura Herath, Felix Terkhorn, Kavith</origin>
      <pubdate>20100728</pubdate>
      <title>forecast_20100430120000Z_run001</title>
      <geoform>raster digital data, NetCDF digital data, textual digital data</geoform>
      <onlinelink>http://dx.doi.org/10.5967/M0154FOX</onlinelink>
      <onlinelink>doi:10.5967/M0154FOX</onlinelink>
    </citeinfo>
  </idinfo>
  <description>
    <abstract>The Vortex2 project (http://www.vortex2.org/home/) supported 100 scientists using over 40 science support vehicles p</abstract>
    <purposes>This data was created to provide fine-grained, hourly forecasts for the Vortex2 scientists (see abstract for more det</purposes>
  </description>
</metadata>
```
Research Data Products

• Vortex
  – A subdirectory for each experiment

- Compress Data
- Fixity Check
- Persistent Identification
- Create Intellectual Metadata
- Create Object Metadata
- Deposit in Repository
Research Data Products

• Outputs
  – Zipped data file
  – MD5 Checksum
  – FGDC metadata record
  – Dublin Core record
  – METS record
  – Sword Package
Post Deposit Curation Workflow

• Scenarios
  – Fixity verification
  – Format normalization
  – New or additional derivative generation
  – Media migration
  – Persistent identifier updates
  – Metadata updates
Future Work

• Adding additional components
  – EAD from spreadsheet
  – MARC record support
  – Premis support

• Testing in the lab
  – Digital library scanning labs
  – Research labs
  – Integrating with a production repository
Acknowledgements

• This research was made possible through a generous grant by Microsoft Research

• And by the Data to Insight Center of Indiana University’s Pervasive Technology Institute

• Thanks to Kavitha Chandrashankar and Quan Zhou for their help with developing components, workflows, and documentation
Thank you

skowalcz@indiana.edu
http://d2i.indiana.edu