Web-based workflow software to support book digitization and dissemination

The Mounting Books project

books.northwestern.edu

Open Repositories 2009 Meeting, May 19, 2009
What we will cover today

- History of Northwestern's Kirtas program
  - Scope of local book digitization
  - Workflow
- Goals of the Mounting Books project
- Details of the workflow system
- About Fedora architecture for Northwestern Books
- Future directions
Kirtas @ Northwestern University Library

2005  Acquisition and initial testing (in Preservation)

2006  Production of print facsimiles of Brittle books
Problem

We had a book scanner.

We weren't doing anything with the scans other than printing them, binding, returning to shelf.

We didn't have a system in place to publish them to the Library's Fedora repository.
The book digitization workflow
Production / Staffing

Kirtas:
   APT 1200 rated @ 1200 pages/hour (20 pages/minute)
   Assumes no setup time, 1200 page book
   Produces raw images

Real-world:
   APT 1200 produces 6-900 pages/hour
   Each student processes 1-200 pages/hour

Currently 6 students: 2 scanning, 4 processing

Before Book Workflow Interface software:
   120 books/quarter
Tonnen Tragfähigkeit. 


In Suez bestand früher eine sehr drückende Hafenordnung, infolge welcher die Schiffe, die in den Hafen einliefen, nur in demselben Turm bebruchten werden durften, als die angekommen waren; das zuletzt eingelaufene Schiff musste also warten, bis alle seine Vor- gänger befürchten waren. Es ward hiermit nicht nur alle Concurrenz ausgeschlossen, sondern auch der freien Bewegung des Verkehrs eine hemmende Fessel angelegt. Erst vor wenigen Jahren ward von der ägyptischen Regierung dieses Unfug der Fazeh zu haben dieses Rotationsysstem ein Ende gemacht. Hierdurch hat sich der Schiffsverkehr im Hafen von Suez in den letzten Jahren bedeutend gehoben.

Durch die zahlreichen Dampfschiffe, die jetzt die Linien von Suez nach Bombay, Kalkutta und Australien regelmäßig verkehren, gewinnt der Hafen sehr an Lebhaftigkeit. Im Gefolge der Dampfschifffahrt finden sich natürlich auch europäische Segelschiffe, meist englischer Flagge, in Suez ein, die mit Kohlen bebraucht aus England kommen und dann gewöhnlich von Descheddah oder
Ein großer Schachbrett ist daher die Einrichtung der Dampfer der ägyptischen Marine gegenwärtig vier Dampfer zwischen Suez und Dschedh.
Patron/Collection Image Processing

ANALYSIS OF 1973 "ON BOARD" BUS SURVEY

In comparing the results obtained in 1972 and 1973, three major conclusions seem apparent:

First, there appears to be a 30% turnover in bus users as evidenced by the response to question #1 (Did you respond to the bus survey last year? 70.2% yes; 29.8% no). This conclusion is based on the high degree of correlation between 1973 responses and those obtained in 1972.

Second, there appears to be greater bus usage by approximately 10% of the bus patrons. This conclusion is derived by subtracting those who did not respond to last year's survey (70.2%) from those who indicated they used the bus more often than last year (43.0%). This conclusion is predicated on the assumption that the difference indicates those persons who were bus patrons in both 1972 and 1973.

Third, the responses indicate increasing utilization of bus service for those persons exceeding the median age of the local population (26.9 years). This conclusion is derived by comparing the relative percent of the total population by age group as enumerated by the 1970 census, and the percent of bus patrons by age group as a function of the total bus patron population respectively.

Characteristics of Bus Riders

The typical bus patron, responding to the 1973 study, has similar characteristics to those of 1972. The overwhelming majority are women (89.0%) and are between the ages of 35-64 years (51.4%) who do not have an automobile available for the trip (63.6%) and who use the bus for transportation to work (46.7%).

Desirable Attributes for Bus Service

Several questions were asked to determine the consistency of responses and to solicit various aspects of those attributes considered desirable in bus service. Eleven of the questions utilized in 1973 were identical to those used in 1972 to provide a basis for comparison.

Two of the primary concerns in 1973 remain the same as those expressed in 1972. Most patrons in 1972 (64.6%) wished half hour service while patrons in 1973 reaffirmed that desire (74.7%). Frequency of service is also a major concern of the bus patron. In 1972, 36.6% would use the bus more often if service was more frequent while in 1973, 29.4% responded affirmatively.

In 1972 the primary criticism of bus service was that buses did not run often enough (29.1%). In 1973 the major concern has shifted to bus service not being on schedule (38.2%). The shift in concern appears significant in that frequency of service as a problem declined by only 2.4% while an inability to maintain schedules doubled (13.8% to 28.2%).

ANALYSIS OF 1973 "ON BOARD" BUS SURVEY

In comparing the results obtained in 1972 and 1973, three major conclusions seem apparent:

First, there appears to be a 30% turnover in bus users as evidenced by the response to question #1 (Did you respond to the bus survey last year? 70.2% yes; 29.8% no). This conclusion is based on the high degree of correlation between 1973 responses and those obtained in 1972.

Second, there appears to be greater bus usage by approximately 10% of the bus patrons. This conclusion is derived by subtracting those who did not respond to last year's survey (70.2%) from those who indicated they used the bus more often than last year (43.0%). This conclusion is predicated on the assumption that the difference indicates those persons who were bus patrons in both 1972 and 1973.

Third, the responses indicate increasing utilization of bus service for those persons exceeding the median age of the local population (26.9 years). This conclusion is derived by comparing the relative percent of the total population by age group as enumerated by the 1970 census, and the percent of bus patrons by age group as a function of the total bus patron population respectively.

Characteristics of Bus Riders

The typical bus patron, responding to the 1973 study, has similar characteristics to those of 1972. The overwhelming majority are women (89.0%) and are between the ages of 35-64 years (51.4%) who do not have an automobile available for the trip (63.6%) and who use the bus for transportation to work (46.7%).

Desirable Attributes for Bus Service

Several questions were asked to determine the consistency of responses and to solicit various aspects of those attributes considered desirable in bus service. Eleven of the questions utilized in 1973 were identical to those used in 1972 to provide a basis for comparison.

Two of the primary concerns in 1973 remain the same as those expressed in 1972. Most patrons in 1972 (64.6%) wished half hour service while patrons in 1973 reaffirmed that desire (74.7%). Frequency of service is also a major concern of the bus patron. In 1972, 36.6% would use the bus more often if service was more frequent while in 1973, 29.4% responded affirmatively.

In 1972 the primary criticism of bus service was that buses did not run often enough (29.1%). In 1973 the major concern has shifted to bus service not being on schedule (38.2%). The shift in concern appears significant in that frequency of service as a problem declined by only 2.4% while an inability to maintain schedules doubled (13.8% to 28.2%).
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>Book Workflow Interface (BWI). books.northwestern.edu</td>
</tr>
</tbody>
</table>
Simple workflow

Scan → Identify edges → Batch-process to crop, remove background → Clean up and quality control
Workflow reality

Problem

Send for rescans

Scan

Identify edges

Batch-process to crop, remove background

Clean up, insert foldouts and rescans, quality control

Foldouts removed

Foldouts scanned separately
Workflow reality over time

- Pages missing from original: request from ILL
- Foldouts removed
- Scan
- Identify edges
- Batch-process to crop, remove background
- Foldouts scanned separately
- Send for rescans
- Problem with one or few pages
- Clean up, insert foldouts and rescans, quality control
- Crash during insert or page rearrangement, book file corrupted, recovery not possible, start over

Tracking database and shared file server
Workflow reality over time

Pages missing from original: request from ILL

Send for rescans

Problem with one or few pages

Crash during insert or page rearrangement, book file corrupted, recovery not possible, start over

Foldouts removed

Scan

Identify edges

Batch-process to crop, remove background

Clean up, insert foldouts and rescans, quality control

Foldouts scanned separately

OCR, JPEG2000, PDF, technical metadata, submission to the repository, public viewing software

Tracking database and shared file server
Scope of the Mounting Books project

• Streamline and extend the book scanning workflow

• Publish digitized books online, link from the library catalog

• Share as many of our software components as possible
Demonstration

Finding books in Voyager
The BWI workflow diagram
Generated by the jBPM. Shows the steps taken to process a book
(view larger image)
# List of in-process book jobs

<table>
<thead>
<tr>
<th>Action</th>
<th>Bib #</th>
<th>Title</th>
<th>Vol /</th>
<th>Priority</th>
<th>Assigned</th>
<th>Status</th>
<th>Schedule</th>
<th># Images</th>
<th>Project(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Job</td>
<td>0005328948</td>
<td>Rural science for standards III &amp; IV. new syllabus / by G.P. Chadka.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Scan</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005328580</td>
<td>The fugitive and other literary works in prose and poetry / by G.M. Woodward.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005328526</td>
<td>Conforto de gli afflitti condannati a morte / del dottor collegiato Gio. Battista Gargiaria. Con l'indice delle cose piu' singolari e notabili.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005328447</td>
<td>La crise des transports : illusions &amp; re?ulite?.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313912</td>
<td>Proceedings ... annual convention. [electronic resource]</td>
<td>9</td>
<td>normal</td>
<td>n/a</td>
<td>Scan [Error]</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313912</td>
<td>Proceedings ... annual convention</td>
<td>8</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313753</td>
<td>Le amoure de ribas / culdo pe?r la Felibresse de?u Caulouin.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313570</td>
<td>Informaci?n de estado</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313543</td>
<td>Chicago traction : a history, legislative and political / by Samuel Wilber Norton.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313540</td>
<td>The new Slovakia / by R.W. Seton-Watson.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313347</td>
<td>ha-Eshkol [electronic resource]</td>
<td>1</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313347</td>
<td>ha-Eshkol</td>
<td>3</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313247</td>
<td>Panama : a personal record of forty-six years, 1861-1907 / by Tracy Robinson.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313053</td>
<td>Knaves of diamonds : being tales of mine and veld / Illustrated by E.F. Shere and T.A. Ayton Symington.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0005313053</td>
<td>Knaves of diamonds : being tales of mine and veld / Illustrated by E.F. Shere and T.A. Ayton Symington.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Open Job</td>
<td>0001329338</td>
<td>The social evil in Chicago : a study of existing conditions with recommendations</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Batch [Error]</td>
<td>3</td>
<td>n/a</td>
<td>3</td>
</tr>
</tbody>
</table>

Start a new book job
Status message:
book files moving

This book job is currently at the following state:
MoveProcessingFiles and the state is running.
Please refresh to see the most current status.

Suspend this book job

Add a note

Add note
The Queue Server

The Queue Server is notified of jobs that need to be processed, and queues them up to avoid overloading the server.

Each set of images has the following processes:
- Conversion to JPEG2000 using Aware software.
- MIX metadata generated with JHOVE
- Exif extraction from original camera images
- Three OCR derivatives are extracted (image plus text PDF, plain text, and XML) using ABBYY OCR software.
- The Queue Server also triggers FOXML generation and ingestion into Fedora.
The Book Builder

As part of the workflow, operators create and enhance structure for digitized books.

Features of the Book Builder:

- Operators can create chapters and other subdivisions.
- Each part of the structure (such as a chapter) can be given a number and a label.
- Automatic page numbering in descending or ascending order. Optionally, operators can apply their own custom page numbers.
- Application of informational tags indicating the presence of other elements such as figures and blank pages. These tags will be useful in the future for scholarly analysis and collection maintenance.
- Information about the structure is stored as METS.
Book Builder demonstration
Book approval

- Generate a Handle that links to book in public interface
- Update Voyager (Library catalog) with 856 URL to Handle
- Update the Search index
- Update PREMIS (future feature)
Fedora Datastreams and Disseminators

**Datastreams** are either pieces of content, or pointers to external content that are stored within one Fedora object. For example, a datastream may be a JPEG image, a Dublin Core XML file, or a redirect to a streaming movie on a server.

**Disseminators** are services that can provide a different view of the stored data.

For instance, a dissemination may include the ability to retrieve a different size of an image that is stored. The service takes the stored image, resizes it, and sends it to the user. The original image is not modified.
Atomistic vs. Compound Approach

Compound objects:
• Include many datastreams within one object
• Have information about the whole and its parts
• While the information is stored in one big object, it's harder to take advantage of Fedora's object model
• Objects can be as easily reused
Atomistic vs. Compound Approach

An example Compound object (a book object):

Datastreams

- Dublin Core
- MODS

- image1.jpg
- image2.jpg
- image3.jpg
- image4.jpg
- ...
- ...
- image2000.jpg
Atomistic vs. Compound Approach

Atomistic objects:
• Have a small number of datastreams per object
• Relationships are stored so the objects can be recognized as one entity.
• Many objects may be needed, but this takes best advantage of Fedora's object model.
• Best for object reuse; for example, all image objects in the repository can share the same dissemination services.
Atomistic vs. Compound Approach

Example Atomistic model (a book object and related page objects):
Each page object shares the same basic access model
(In fact, all of our Fedora image objects share the same dissemination access model)
## Page Imaged Book Model

<table>
<thead>
<tr>
<th>Datastream LABEL</th>
<th>ID</th>
<th>MIME TYPE</th>
<th>Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dublin Core</td>
<td>DC</td>
<td>text/xml</td>
<td>OAI_DC</td>
</tr>
<tr>
<td>MODS</td>
<td>MODS</td>
<td>text/xml</td>
<td>MODS</td>
</tr>
<tr>
<td>MARC</td>
<td>MARC</td>
<td>text/xml</td>
<td>MARC XML</td>
</tr>
<tr>
<td>XACML Policy file</td>
<td>POLICY</td>
<td>text/xml</td>
<td>XACML</td>
</tr>
<tr>
<td>Kirtas job tar file</td>
<td>KIRTAS_(1,2,3,...)</td>
<td>application/x-tar</td>
<td>KIRTAS JOB FILES</td>
</tr>
<tr>
<td>Tagged book Structure</td>
<td>BOOKSTRUCT</td>
<td>text/xml</td>
<td>METS</td>
</tr>
<tr>
<td>Relationships</td>
<td>RELS-EXT</td>
<td>text/xml</td>
<td>RELS-EXT</td>
</tr>
<tr>
<td>HTML viewer</td>
<td>HTML</td>
<td>text/html</td>
<td>XHTML</td>
</tr>
<tr>
<td>Datastream LABEL</td>
<td>ID</td>
<td>MIMETYPE</td>
<td>Schema</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------</td>
<td>----------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Dublin Core</td>
<td>DC</td>
<td>text/xml</td>
<td>OAI_DC</td>
</tr>
<tr>
<td>MODS record</td>
<td>MODS</td>
<td>text/xml</td>
<td>MODS</td>
</tr>
<tr>
<td>Archival image format</td>
<td>ARCHV-IMG</td>
<td>image/jpeg</td>
<td></td>
</tr>
<tr>
<td>Archival MIX</td>
<td>ARCHV-TECHMD</td>
<td>text/xml</td>
<td>MIX</td>
</tr>
<tr>
<td>Archival Exif Metadata</td>
<td>ARCHV-EXIF</td>
<td>text/xml</td>
<td>Exif as XML</td>
</tr>
<tr>
<td>Corrected Image format</td>
<td>PROC-IMG</td>
<td>image/jpeg</td>
<td></td>
</tr>
<tr>
<td>Corrected MIX</td>
<td>PROC-TECHMD</td>
<td>text/xml</td>
<td>MIX</td>
</tr>
<tr>
<td>Delivery image format</td>
<td>DELIV-IMG</td>
<td>image/jp2</td>
<td></td>
</tr>
<tr>
<td>Delivery MIX</td>
<td>DELIV-TECHMD</td>
<td>text/xml</td>
<td>MIX</td>
</tr>
<tr>
<td>Delivery metadata</td>
<td>DELIV-OPS</td>
<td>text/xml</td>
<td>SVG</td>
</tr>
<tr>
<td>Digital Provenance</td>
<td>PROVENANCE</td>
<td>text/xml</td>
<td>PREMIS</td>
</tr>
<tr>
<td>XACML Policy file</td>
<td>POLICY</td>
<td>text/xml</td>
<td>XACML</td>
</tr>
<tr>
<td>Relationships</td>
<td>RELS-EXT</td>
<td>text/xml</td>
<td>RELS-EXT</td>
</tr>
<tr>
<td>OCR in XML</td>
<td>OCR_XML</td>
<td>text/xml</td>
<td>Abbey OCR</td>
</tr>
<tr>
<td>OCR in PDF</td>
<td>OCR_PDF</td>
<td>application/pdf</td>
<td></td>
</tr>
<tr>
<td>OCR in plain text</td>
<td>OCR_TEXT</td>
<td>text/plain</td>
<td></td>
</tr>
<tr>
<td>HTML viewer page</td>
<td>HTML</td>
<td>text/html</td>
<td>XHTML</td>
</tr>
</tbody>
</table>
Web-enabled services

Web-enabled services allow us to more easily use software components for other projects. The web-enabled services used for this project include:

• Voyager (Integrated Library System) updater
• Handle persistent URL creator
• Queue server
• Yaz Proxy Z39.50 server
• SOLR search service
• Fedora digital repository
Other Technologies used in the project

- **METS** - used to store the book structure
- **Ext-JS** - JavaScript library, includes interface and AJAX
- **JSON** - Used to communicate between JavaScript and server-side Java
- **ABBYY** - OCR technology
- **Aware** - Generates and serves up derivatives from JPEG2000 files
- **JHove** - For generating MIX data from images
"we" =

Northwestern University Library
Administration, Bibliographic Services, Digital Collections, Library Technology, Preservation, Business & Finance, Special Collections and University Archives, Africana, Science and Engineering Library, Transportation, GovInfo, Music Library, Circulation ... and many more

Northwestern University Academic & Research Technologies
Flash viewer development

OpenSky Solutions
James Chartrand

Project funded by the Andrew W. Mellon Foundation

Thank you
where we are now...
future growth...
What next?

• PDF for download and printing, OCR tweaks
• Minor fixes to metadata display, OPAC linking
• Fine-tuning Book Builder navigation
• Northwestern Books beta launch early April
• Mounting Books software available June 30

then...
ebook support, print-on-demand, expanding access to Book Builder, METS/ALTO & TEI, integration with Portico, Google content/HathiTrust?