Web-based workflow software to support book digitization and dissemination

The Mounting Books project

books.northwestern.edu

CNI Spring Task Force Meeting, April 7, 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
Components mentioned

**Mounting Books**
the one-year project funded by the Andrew W. Mellon Foundation

**Book Workflow Interface** or **BWI**
the back-end workflow system that manages the scanning and publication process (code-named "Crabcake")

**Book Viewer**
the public book viewing software in use at the public site, **Northwestern Books** (books.northwestern.edu)

**The repository**
the Library's Fedora-based digital repository
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

In Suez bestand früher eine sehr drückende Hafennorm, infolge welcher die Schiffe, die in den Hafen einliefen, nur in demselben Turmus befrachtet werden durften, als sie angekommen waren; das anstehende eingeladene Schiff musste also warten, bis alle seine Vorgänger befacht waren. Es ward hiermit nicht nur alle Concurrrency ausgeschlossen, sondern auch der freien Bewegung des Verkehrs eine hemmende Fessel angelegt. Erst vor wenigen Jahren ward von der ägyptischen Regierung diesem Unfug der Fassh (so hiess dieses Rotationsystem) 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 versetzen, gewinnt der Hafen sehr an Lebhaftigkeit. Im Gefolge der Dampfschiffahrt finden sich natürlich auch europäische Segelschiffe, meist englischer Flagge, in Suez ein, die mit Kohlen befrachtet aus England kommen und dann gewöhnlich von Deschediah oder
27) Ein grosser Schiffs-Dampf

gegenwärtig der ägyptischen

wischen Suez und Dsched-

Dampfer der Einrichtung.
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.

Secondly, 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 (29.8%) from those who indicated they used the bus more often than last year (43.9%). This conclusion is predicated on the assumption that the difference indicates those persons who were bus patrons in both 1972 and 1973.

Thirdly, 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 (69.0%) and are between the ages of 35-66 years (53.1%) who do not have an automobile available for the trip (81.6%) and who use the bus for transportation to work (56.7%).

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 1972 were identical to those used in 1973 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 (63.4%) wished greater hour service while patrons in 1973 reaffirmed that desire (74.7%). Frequency of service is also a major concern of the bus patrons. In 1972, 36.6% would use the bus more often if service was more frequent while in 1973, 39.4% responded affirmatively.

In 1972 the primary criticism of bus service was that buses did not run on schedule (28.6%). The shift in concern appears significant in that frequency of service is 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.

Secondly, 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 (29.8%) from those who indicated they used the bus more often than last year (43.9%). This conclusion is predicated on the assumption that the difference indicates those persons who were bus patrons in both 1972 and 1973.

Thirdly, 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 (69.0%) and are between the ages of 35-66 years (53.1%) who do not have an automobile available for the trip (81.6%) and who use the bus for transportation to work (56.7%).

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 1972 were identical to those used in 1973 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 (63.4%) wished greater hour service while patrons in 1973 reaffirmed that desire (74.7%). Frequency of service is also a major concern of the bus patrons. In 1972, 36.6% would use the bus more often if service was more frequent while in 1973, 39.4% responded affirmatively.

In 1972 the primary criticism of bus service was that buses did not run on schedule (28.6%). The shift in concern appears significant in that frequency of service is a problem declined by only 2.4% while an inability to maintain schedules doubled (13.8% to 28.2%).
Kirtas @ Northwestern University Library

2007  Shift away from Brittle reformatting on Kirtas
New Digital Collections department formed
Applied to Andrew W. Mellon Foundation

2008  Collection/Patron/Brittle, quarterly queues
New focus on digital-only delivery

Mellon-funded Mounting Books project

2009  Book Workflow Interface (BWI)
books.northwestern.edu
Simple workflow

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

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

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
The first steps in the workflow:
Starting a new book job
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># Imgs</th>
<th>Project(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Job</td>
<td>005323948</td>
<td>Rural science for standards III &amp; IV : new syllabus / by G.P. Chidga.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Scan</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Open Job</td>
<td>005328580</td>
<td>The fugitive and other literary works in prose and poetry / of G.M. Woodward.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005328509</td>
<td>Conforto de gli affetti condannati a’ morte / del dottor collegiato Gio Battista Gorgiari. Con Indice delle cose più 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>2</td>
</tr>
<tr>
<td>Open Job</td>
<td>005328447</td>
<td>La crise des transports : illusions &amp; réalités ?</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313912</td>
<td>Proceedings ... annual convention, [electronic resource]</td>
<td>2</td>
<td>normal</td>
<td>n/a</td>
<td>Scan</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313912</td>
<td>Proceedings ... annual convention</td>
<td>6</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313763</td>
<td>Los antoño de ribas / cuiudo pe?; la felibresa do?a Cauldon</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313520</td>
<td>Informacie?n de estado</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313543</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>Book Builder</td>
<td>2</td>
<td>258</td>
<td>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313540</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>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313347</td>
<td>ha-iShkol. [electronic resource]</td>
<td>1</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313347</td>
<td>ha-iShkol</td>
<td>3</td>
<td>normal</td>
<td>n/a</td>
<td>Touchup</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313247</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>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313033</td>
<td>Knives of diamonds : being tales of mine and veld / Illustrated by E. F. Shorie and T. A. Ayton Symington.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Book Builder</td>
<td>3</td>
<td>290</td>
<td>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005313034</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>Touchup</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Open Job</td>
<td>005312938</td>
<td>Knives of diamonds : being tales of mine and veld / Illustrated by E. F. Shorie and T. A. Ayton Symington.</td>
<td>n/a</td>
<td>normal</td>
<td>n/a</td>
<td>Batch</td>
<td>2</td>
<td>n/a</td>
<td>2</td>
</tr>
</tbody>
</table>

Start a new book job
# Book job details

## Book Details

<table>
<thead>
<tr>
<th>Status</th>
<th>Book Details</th>
<th>Book Job Diagram</th>
<th>Process Vars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bib #:</td>
<td>0005313543</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Barcode:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OCLC:</td>
<td>(OCLC)00721661</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volume:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Title:</td>
<td>Chicago traction: a history, legislators</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author:</td>
<td>Author: Norton, Samuel Wilber.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Date:</td>
<td>1907.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Location:</td>
<td>TRANSPORTATION Library</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Call Number:</td>
<td>HE 4491.C4 1888</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Schedule

- 2009 Winter

## Projects

- Collection
- Patron
  - Brittle

## Priority

- Normal

## Job Type

- Kirtas
  - Vendor

## Visibility

- Public
  - Private

## Has Foldouts

- No

## Requester's contact information

- Name: Brittle
- Email: preservation@northwestern.edu

## Close job

- Requester notified date: 
- Notification method: 
- Notified by: 
- Book routed date: 
- Confirmed receipt by: 
- Confirmation date: 

## Copyright

### Access

- World
- U.S. only
- Northwestern only
- Custom (select all that apply):
  - Time-limited link
  - Northwestern group(s)
  - Northwestern user(s)

### Copyright status

- Copyrighted
- Public domain
- Unknown

## Notes and description of copyright analysis

- 

## Book Structure

- Root

---

Suspend this book job | End this book job | Delete this book job
Scan, Move files, and Batch processes
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.
Ingesting the files into Fedora, associating rescans
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

- Vendor book structure generator
  - yes → Book Builder
  - no → import foldouts
- done → Purge Rescanned Pids
- done → Book Builder
- done → Yes Foldout Pending Flag Set
- no → return to book build
- yes → Does Job Modify Existing Book
- done → Book Builder
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.
Book Builder demonstration
Foldouts
(images entered scanned separately from the Kirtas machine)
Book approval

- Generate a Handle that links to book in public interface
- Update Voyager 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):

```
Book Object: Datastreams
  - Dublin Core
  - MODS
  - Relationship Data

Page Object: Datastreams
  - Dublin Core
  - MODS
  - Relationship Data
  - image1.jpg

Page Object: Datastreams
  - Dublin Core
  - MODS
  - Relationship Data
  - image2.jpg

Page Object: Datastreams
  - Dublin Core
  - MODS
  - Relationship Data
  - image3.jpg
```
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>MIMETYPE</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>
</tbody>
</table>
# Page Image 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 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>MIX</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
where we are now...
future growth...

**Workflow**

- Book Building
- OCR & derivatives
- Repository
- OPAC/public viewer
- Outsourced digitization
- Annotation
- ebook distribution
- Portico/HathiTrust
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 1

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

OpenSky Solutions
James Chartrand

Project funded by the Andrew W. Mellon Foundation

Thank you