From Microfilm to Digital Images: The National Digital Newspaper Program

ALAO TEDSIG 2012 Spring Workshop | May 11, 2012

Jenni Salamon, Project Coordinator, NDNP-OH
Phil Sager, Digital Projects Developer
Agenda

Background
NDNP Training
Technical Specifications & Cataloging
Requirements
NDNP-OH Project Workflow
NDNP and the End User
Background
National Digital Newspaper Program

Funded by the National Endowment for the Humanities
Managed by the Library of Congress

Goals

- Preserve historic newspapers
- Provide increased access to historic newspapers
  - Keyword searchable
  - Freely available
  - No need to visit library or archives
NDNP in Ohio

First grant cycle: July 1, 2008 – June 30, 2010
- 14 papers from 11 counties
- 1880-1922
- Almost 110,000 pages

Second grant cycle: July 1, 2010 – August 31, 2012
- 26 papers from 21 counties (3 overlap from previous cycle)
- Focus on Civil War era (1845-1894)
- Over 100,000 pages
NDNP Training
Training

NDNP Annual Awardee Meetings
- Introduces project to new awardees
- Overview changes to project
- Returning awardees talk about their projects

NDNP Awardee Wiki and Listserv
- Documentation on all aspects of project
- Tips for pre-digitization and post-digitization processing and workflow
- Links to relevant resources
Training

meta | morphosis

- Film-to-digital institute hosted by the University of Kentucky
- Not required, but encouraged
- Covers everything from microfilm basics to technical specifications to vendor selection
- Solid background on all aspects of the project
Training

On-the-Job

- Open communication between all awardees
- Kick-off meeting at beginning of grant with digitization vendor
- Working closely with vendor to ensure project success
- Creation of internal workflow and instructional documents for NDNP-OH staff
- Updating processes as necessary given staff resources, time and experience
Technical Specifications and Cataloging Requirements
Technical Specifications

Digital Deliverables
- Scanned 300-400 dpi
- Three formats
  - Grayscale, uncompressed TIFF 6.0 images
  - Compressed JPEG2000 images
  - PDF with Hidden Text
- Accompanying structural and technical metadata
- OCR for all pages
Technical Specifications

Scanning
- Deskew images with a skew of greater than 3 degrees
- Crop to visible edge of page
- Capture grayscale preservation microfilm targets

OCR
- Conform to ALTO XML scheme
  - ALTO (Analyzed Layout and Text Object) is an XML Schema that details technical metadata for describing the layout and content of physical text resources
- Bounding box coordinate data
  - Each column is sectioned and coordinates are used to place words
For More Information about ALTO...

International metadata standard
LC is governing body
www.loc.gov/standards/alto
Structural and Technical Metadata

METS (Metadata Encoding and Transmission Standard) format preservation metadata

Structural metadata
- Relate pages to title, date and edition
- Sequence pages within issue or section
- Identify image and OCR files

Technical metadata to support functions of digital repository

Other metadata standards used include: MIX, PREMIS & MODS
Metadata Dictionary

Defines metadata elements and formats
General, title, page, reel level
Defines fields for metadata and collation files
Gather information from:

- MARC records on Chronicling America
- Microfilm itself
Embedding Metadata

TIFF Tags

JP2 XML Box

PDF/A XMP

<table>
<thead>
<tr>
<th>TIFF Tag #</th>
<th>TIFF Tag Name</th>
<th>TIFF Type</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>269</td>
<td>DocumentName</td>
<td>ASCII</td>
<td>Use Microfilm reel # (barcode). If not converted from microfilm, use normalized LCCN.</td>
</tr>
<tr>
<td>42016</td>
<td>ImageUniqueID</td>
<td>ASCII</td>
<td></td>
</tr>
<tr>
<td>6.2.1</td>
<td>UniqueImageID</td>
<td>ASCII</td>
<td></td>
</tr>
<tr>
<td>6.2.1</td>
<td>UniqueImageID</td>
<td>ASCII</td>
<td></td>
</tr>
</tbody>
</table>
Metadata Collection – Reel-Level

Pre-Digitization Inspection

- Record data values for reels, issues, pages
  - Start and end dates
  - Page dimensions
  - Density readings
  - Resolution
  - Guide to contents
  - Reduction ratio
  - Etc.
- Vendor-created spreadsheet
Metadata Collection – Issue-Level

Collation

- Review use copies of selected microfilm reels
- Check for missing issues/pages, duplicate issues/pages and other abnormalities
- Vendor-created spreadsheet
Cataloging Requirements

Updated MARC record
CONSER-compliant
Accurate titles/LCCNs are extremely important!

- Digitized content “hangs” on preexisting catalog records on Chronicling America
- Ensures that each digitized issue is in the right place
Controlled Vocabulary

A. Lincoln, his last 24 hours / by W. Emerson Reck.
Author: Reck, W. Emerson (Waldo Emerson), 1903-
Description: vii, 232 p. : ill. ; 24 cm.
Notes: Includes index.
Presidents -- United States -- Biography.
ISBN: 0899502164 (alk. paper)
OCLC Number: ocm15109485

Not for NDNP!

Titles, LCCNs and issue dates are controlled
Text in newspaper is full-text searchable
  ▪ Includes every word printed on page (and even some that aren’t!)

No metadata field for relevant subject headings, article summaries or descriptions
Must rely on OCR engine to find information
NDNP-OH Project Workflow
Newspaper Selection and Analysis

Preliminary Selection (Advisory Board)

Microfilm Analysis (Project Staff)

Final Selection (Advisory Board)

No Digitization

NO

YES

NO

No Digitization
Newspaper Selection

Advisory board
- Librarians, educators, curators, archivists, historians, etc.

Selection criteria
- Historical significance
- Geographical significance
- Chronological coverage

At least 1 paper from each OHIOHSHA region (10,000 pages)
Represent as many Ohio voices as possible
Film to Vendor

Collect data on master film
Master negative duplicated
Collect data on duplicate film
Upload metadata spreadsheets to vendor-developed portal
Forward silver-halide film to digitization vendor
Digitization

Scan Film (Digitization Vendor)

400 dpi grayscale, 8 bit per pixel

Create Image Derivatives (Digitization Vendor)

generate OCR (Digitization Vendor)

- TIFF 6.0, JPEG2000, PDF
- OCR text file, ALTO XML
iArchives Digitization Workflow

Film/Paper Scanning → Split, De-Skew, Crop → Image Processing → Image Metadata → Article Metadata → OCR Framework → Post Process → Indexing, Hosting

Shared Storage (NAS) → Workflow Manager → DB

KEY:
- Automatic process [image processing, OCR, ...]
- Manual process [image + article metadata]
- Quality Control
Validation and Inspection

Verify and Review
Quality of Product
(Project Staff)

Release deliverables to the Library of Congress
(Project Coordinator)
Validation and Inspection

Check directory and file structure
Validation and Inspection

Adds technical metadata about images through validation
Perform verification using LC’s Digital Viewer and Validator (DVV)
Spot check integrity of images, metadata and OCR
Repository Ingest
Chronicling America
Repository Ingest – Detailed View

NDNP Workflow

- Hard Drive
- Validation & Verification
- QC
- Archival Storage
- Content Transfer Services
- Visible Inventory
NDNP and the End User

Chronicling America
Ohio Memory (CONTENTdm)
Chronicling America

NDNP’s final product
Hosted by the Library of Congress
Contains over 4 million images from 1836-1922
Over 700 papers
28 states (incl. D.C.)
Updated on rolling basis
Chronicling America

Basic and Advanced Search Options
Chronicling America – Viewing An Image

Navigation Bar
View Options (Text, PDF)
Download

Zoom
Toggle Full Screen
Clip Image (Print)

Persistent Link: http://chroniclingamerica.loc.gov/lcnr/sn/0765213/1896-01-14/ed-1/seg-1/
For More Information and to Download the LC Newspaper Viewer....

http://sourceforge.net/apps/trac/loc-ndnp/
Ohio Memory

Amherst News-Times

- Partnership with Amherst Public Library
- Use NDNP technical specifications
CONTENTdm FlexLoader

Utility allows loading of NDNP-spec content into CONTENTdm-hosted digital library

Advantages:

- Allows you to crosswalk a subset of NDNP elements to CONTENTdm fields.
- Ingests NDNP-formatted data as is, including all word highlighting (which is the product of the ALTO files).

Disadvantages:

- Availability of elements available for crosswalking is statically determined (e.g. right now MIX data largely unmappable).
- Doesn’t automatically read in image information like Project Client does.
- Assume it will take output from any vendor capable of producing valid NDNP output, but it should be tested.
CONTENTdm FlexLoader
CONTENTdm FlexLoader
Ohio Memory – Viewing an Image

View Options (Image, Text)
Zoom/Rotate
Print
Thumbnails
Most of this information is not viewable on *Chronicling America* image-viewing screen!
Thank You!

Questions? Comments?

- Jenni Salamon, Project Coordinator, NDNP-OH
  - jsalamon@ohiohistory.org
- Phil Sager, Digital Projects Developer
  - psager@ohiohistory.org