The Basics of Newspaper Digitization
OhioDIG Fall Meeting | November 14, 2013

Jenni Salamon, Project Coordinator, NDNP-OH
Kevin Latta, Quality Control Technician, NDNP-OH
Agenda

- Background
- Content Selection
- Technical Specifications
- Project Workflow
- Hosting Content
Background
Newspaper Digitization at OHS

- 2008 – Program started with National Digital Newspaper Program (NDNP)
- 2010 – Expanded to Ohio Memory
- Utilizes NDNP digitization standards
National Digital Newspaper Program

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

Goals
- Preservation
- Increased access
- Develop newspaper digitization standards
- Build state partners’ capacity to cultivate local programs
NDNP Standards

- Allow for high quality images and keyword searching
- Interoperable between multiple digital library hosting systems
- Sustainable, long-term access
NDNP in Ohio

- **Phase 1: 2008 – 2010**
  - 14 papers from 11 counties
  - 1880-1922
  - Almost 110,000 pages

- **Phase 2: 2010 – 2012**
  - 26 papers from 21 counties (3 overlap)
  - Focus on Civil War era (1845-1894)
  - Over 100,000 pages

- **Phase 3: 2012 – 2014**
  - 32 titles from 25 counties
  - 1836-1921
  - German - Tägliches Cincinnatier Volksblatt (1914-1918)
  - Additional 100,000 pages
Content Selection
Tips for Selection

 Choose what is important to your institution/users

 What fits into your collecting plans, your institutional purpose?

 Copyright
   Newspaper
     Published before 1923? In the public domain!
     Published 1923 or after? Need permission from current or last publisher!
  
   Microfilm
     Who created microfilm? May need to ask permission to digitize!
     Best to work from second generation negatives whenever possible
Hardcopy vs. Microfilm

**Hardcopy**
- More costly
- More time-consuming
- Better image quality (probably)

**Microfilm**
- Less expensive
- More efficient
- Preserve hardcopy
- Quality issues
  - Physical condition
  - Microfilming process

*Some issues can be corrected during digitization process, but some can’t!*
NDNP Digitization Standards

http://www.loc.gov/ndnp/guidelines/
NDNP Technical Specifications

- Digital Deliverables

  - Image files – three formats
    - Grayscale, uncompressed TIFF 6.0 images
    - Compressed JPEG2000 images
    - PDF with Hidden Text

  - XML files
    - Administrative, structural, descriptive and technical metadata
    - Optical Character Recognition (OCR) for all pages
NDNP Technical Specifications

- **Scanning**
  - 300-400 DPI
  - Deskew
  - Crop
  - Capture targets

- **OCR**
  - Conform to ALTO (Analyzed Layout and Text Object) XML scheme
  - Bounding box coordinate data
ALTO XML File
Structural and Technical Metadata

- METS (Metadata Encoding and Transmission Standard) format preservation metadata

- Structural metadata

- 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
<?xml version="1.0" encoding="UTF-8" ?>

  http://www.loc.gov/mods/v3 .../schema/ndnpmods-3-0.xsd http://www.loc.gov/ndnp ..schema/ndnp.xsd
  http://www.loc.gov/standards/premis ..schema/object.xsd http://www.loc.gov/mix/
  - <metsHdr CREATEDATE="2006-10-23T14:02:03" RECORDSTATUS=""/>
    - <agent ROLE="CREATOR" TYPE="ORGANIZATION">
        <name>University of California, Riverside</name>
    </agent>
  </metsHdr>
  - <dmdSec ID="techTargetModsBib1">
    - <mdWrap MDTYPE="MODS" LABEL="Tech Target metadata">
      - <xmlData>
        <mods>
          <titleInfo>
            <title>ISO 16067 Calibrated Target</title>
          </titleInfo>
          <note type="agencyResponsibleForReproduction" displayLabel="University of California, Riverside">CU-Riv</note>
        </mods>
      </xmlData>
    </mdWrap>
  </dmdSec>
  - <dmdSec ID="techTargetModsBib2">
    - <mdWrap MDTYPE="MODS" LABEL="Tech Target metadata">
      - <xmlData>
        <mods>
          <titleInfo>
            <title>ISO 16067 Calibrated Target</title>
          </titleInfo>
        </mods>
      </xmlData>
    </mdWrap>
  </dmdSec>
</mets>
Embedding Metadata

- TIFF Tags
- JP2 XML Box
- PDF/A XMP

**TIFF Tag #:** 269
**TIFF Tag Name:** DocumentName
**TIFF Type:** ASCII
**Notes:** Use Microfilm reel # (barcode). If not converted from microfilm, use normalized LCCN.

**TIFF Tag #:** 42016
**TIFF Tag Name:** ImageUniqueID
**Z39.87 #:** 6.2.1
**Z39.87 Name:** UniqueImageID
**TIFF Type:** ASCII
Metadata Collection: Reel-Level

- Pre-Digitization Inspection

<table>
<thead>
<tr>
<th>Title</th>
<th>LCCN</th>
<th>Reel Number</th>
<th>Start Date</th>
<th>End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star (Cincinnati, Ohio)</td>
<td>sn87056710</td>
<td>00296027091</td>
<td>1875-01-02</td>
<td>1875-10-15</td>
</tr>
<tr>
<td>The Cincinnati daily star.</td>
<td>sn85025759</td>
<td>00296027091</td>
<td>1875-01-02</td>
<td>1875-10-15</td>
</tr>
<tr>
<td>The Cincinnati daily star.</td>
<td>sn85025759</td>
<td>00296027108</td>
<td>1875-10-16</td>
<td>1876-07-31</td>
</tr>
<tr>
<td>The Cincinnati daily star.</td>
<td>sn85025759</td>
<td>0029602711A</td>
<td>1876-08-01</td>
<td>1877-04-30</td>
</tr>
<tr>
<td>The Cincinnati daily star.</td>
<td>sn85025759</td>
<td>00296027121</td>
<td>1877-05-01</td>
<td>1878-01-31</td>
</tr>
<tr>
<td>The Cincinnati daily star.</td>
<td>sn85025759</td>
<td>00296026931</td>
<td>1878-02-01</td>
<td>1878-11-15</td>
</tr>
<tr>
<td>The Cincinnati daily star.</td>
<td>sn85025759</td>
<td>00296026943</td>
<td>1878-11-16</td>
<td>1879-08-15</td>
</tr>
<tr>
<td>The Cincinnati daily star.</td>
<td>sn85025759</td>
<td>00296026955</td>
<td>1879-08-16</td>
<td>1880-04-15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date Created</th>
<th>Position</th>
<th>Reduction Ratio</th>
<th>Number of Resolution Targets</th>
<th>Loose Leaves</th>
<th>Bound Volume</th>
<th>Dimensions</th>
<th>Pages per issue</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>2b</td>
<td>17x</td>
<td>2</td>
<td>true</td>
<td>false</td>
<td>22x29 in.</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1980</td>
<td>2b</td>
<td>17x</td>
<td>2</td>
<td>true</td>
<td>false</td>
<td>22x29 in.</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1980</td>
<td>2b</td>
<td>17x</td>
<td>2</td>
<td>true</td>
<td>false</td>
<td>22x29 in.</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1980</td>
<td>2b</td>
<td>17x</td>
<td>2</td>
<td>true</td>
<td>false</td>
<td>22x29 in.</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1980</td>
<td>2b</td>
<td>17x</td>
<td>2</td>
<td>true</td>
<td>false</td>
<td>22x29 in.</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>1980</td>
<td>2b</td>
<td>17x</td>
<td>2</td>
<td>true</td>
<td>false</td>
<td>22x29 in.</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>AB</th>
<th>AC</th>
<th>AD</th>
<th>AE</th>
<th>AF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>120</td>
<td>0.82</td>
<td>0.80</td>
<td>0.84</td>
</tr>
<tr>
<td>1993</td>
<td>120</td>
<td>0.82</td>
<td>0.80</td>
<td>0.84</td>
</tr>
<tr>
<td>1993</td>
<td>120</td>
<td>0.30</td>
<td>0.74</td>
<td>0.80</td>
</tr>
<tr>
<td>1993</td>
<td>136</td>
<td>0.92</td>
<td>0.90</td>
<td>0.86</td>
</tr>
<tr>
<td>1993</td>
<td>120</td>
<td>0.94</td>
<td>0.96</td>
<td>0.98</td>
</tr>
<tr>
<td>1993</td>
<td>120</td>
<td>0.84</td>
<td>0.84</td>
<td>0.92</td>
</tr>
<tr>
<td>1993</td>
<td>120</td>
<td>0.92</td>
<td>0.84</td>
<td>0.88</td>
</tr>
<tr>
<td>1993</td>
<td>120</td>
<td>0.86</td>
<td>0.80</td>
<td>0.84</td>
</tr>
<tr>
<td>1993</td>
<td>132</td>
<td>0.94</td>
<td>0.90</td>
<td>0.94</td>
</tr>
<tr>
<td>1993</td>
<td>119</td>
<td>0.84</td>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td>1993</td>
<td>119</td>
<td>0.94</td>
<td>0.98</td>
<td>0.94</td>
</tr>
</tbody>
</table>
**Metadata Collection: Issue-Level**

- **Collation**

<table>
<thead>
<tr>
<th>LCCN</th>
<th>Title</th>
<th>Reel Number</th>
<th>Volume Number</th>
<th>Issue Date</th>
<th>Issue Date As Labeled</th>
<th>Issue Number</th>
<th>Edition Number</th>
<th>Edition Label</th>
<th>Section Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>sn85038189</td>
<td>Meigs County times. (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>1</td>
<td>1844-02-07</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sn85038189</td>
<td>Meigs County times. (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>1</td>
<td>1844-02-14</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sn85038189</td>
<td>Meigs County times. (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>1</td>
<td>1844-02-27</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sn85038189</td>
<td>Meigs County times. (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>1</td>
<td>1844-04-03</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sn85038189</td>
<td>Meigs County telegraph (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>3</td>
<td>1851-04-17</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sn85038189</td>
<td>Meigs County telegraph (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>3</td>
<td>1851-04-24</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sn85038189</td>
<td>Meigs County telegraph (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>3</td>
<td>1851-05-01</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sn85038189</td>
<td>Meigs County telegraph (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>3</td>
<td>1851-05-08</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sn85038189</td>
<td>Meigs County telegraph (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>3</td>
<td>1851-05-16</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sn85038189</td>
<td>Meigs County telegraph (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>3</td>
<td>1851-05-22</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sn85038189</td>
<td>Meigs County telegraph (Pomeroy, Ohio)</td>
<td>00296033010</td>
<td>3</td>
<td>1851-05-29</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Page Count Published</th>
<th>Page Count Filmed</th>
<th>Missing Issue</th>
<th>Missing Page Nums</th>
<th>Duplicate Page Nums</th>
<th>Blocked Page Count</th>
<th>Delivered Page Count</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td>1,3</td>
<td>2</td>
<td>4</td>
<td>rr=20, created 1978; deliver second copies pp. 1, 3; pages light throughout reel</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>deliver second copy p. 3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>deliver second copy p. 3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>deliver second copy p. 3</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td>1,2</td>
<td>2</td>
<td>4</td>
<td>deliver second copy p. 1; pp. 1-2 damaged</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
Cataloging Requirements

- Updated MARC record
- CONSER-compliant
- Accurate application of titles/LCCNs means proper:
  - Organization and delivery of content on ChronAm
  - Resource description
Newspaper Selection & Analysis

- Preliminary Selection
- Microfilm Review
- Final Selection
Film to Vendor

1. Duplicate Master Microfilm Reels
2. Microfilm Analysis
3. Scan Film
4. Send to Digitization Vendor
Digitization

Scan Film
400 dpi, grayscale, 8 bit per pixel TIFFs

Create derivative images
JP2, PDF

Generate OCR
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

Upload to Digital Repository
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
The First Thanksgiving Proclamation

By the President of the United States of America
A Proclamation

When we renew the solemnities which affect us in many other forms the present condition of the United States affords much matter of consolation, love.

Deeply impressed with this sentiment, I, George Washington, President of the United States, do recommend to all the people of this nation to set aside the first Thursday in November as a day of public thanksgiving and praying and on or after such day.

In testimony whereof I have caused the seal of the United States of America to be affixed to this present and that the same be published, signed by me.

Done at the City of Philadelphia the First day of February, one thousand seven hundred and ninety-six. Of the Independence of the United States of America the nineteenth.

G. Washington
Accessing Digitized Content

- How will users access the content?
  - Vendor solution
  - In-house hosting platform
  - Regional/statewide digital library

- Sustainability

- Interoperability
Chronicling America

Search Pages  Advanced Search  All Digitized Newspapers 1836-1922

Select State(s):
All states
Alabama
Arizona
California
Colorado
District of Columbia
Florida
Georgia

Or Select Newspaper(s):
All newspapers
Abbeville progress. (Abbeville, Vermilion Parish, La.)
Aberdeen herald. (Aberdeen, Chehalis County, W.T.)
The Abilene reflector. (Abilene, Kan.)
Abilene weekly reflector. (Abilene, Kan.)
The Abingdon Virginian. (Abingdon [Va.])
The Adeir County news. (Columbia, Ky.)
Adams County news. (Ritzville, Wash.)

Select Year(s)*
Newspaper pages are available for newspapers published between 1836-1922
from 1836 to 1922
Or Date Range
from to

Limit Search: only front page or Specific page

Language

Enter Search
...with any of the words:
...with all of the words:
...with the phrase:
...with the words: within 5 words of each other

Clear Search
Marietta daily leader., January 14, 1896, Image 1
About Marietta daily leader. (Marietta, Ohio) 1895-1906

Image provided by: Ohio Historical Society, Columbus, OH

# Chronicling America

## Browse Issues: Marietta daily leader.
Marietta, Ohio (1895-1906)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S M T W T F S</td>
<td>S M T W T F S</td>
<td>S M T W T F S</td>
<td>S M T W T F S</td>
</tr>
<tr>
<td>5</td>
<td>6 7 8 9 10 11</td>
<td>2 3 4 5 7 8</td>
<td>1</td>
<td>2 3 4 5 6 7</td>
</tr>
<tr>
<td>12</td>
<td>13 14 15 16 17</td>
<td>8 9 10 11 12 13</td>
<td>7 8 9 10 11 12</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>19</td>
<td>20 21 22 23 24</td>
<td>14 15 16 17 18</td>
<td>13 14 15 16 17</td>
<td>9 10 11 12 13</td>
</tr>
<tr>
<td>26</td>
<td>27 28 29 30 31</td>
<td>21 22 23 24 25</td>
<td>20 21 22 23 24</td>
<td>18 19 20 21 22</td>
</tr>
</tbody>
</table>

**Show all front pages**

Single edition: dates in **bold**

Multiple editions: dates in *bold italics*
ChronAm Software
Download and Info

https://github.com/libraryofcongress/chronam/
Ohio Memory (CONTENTdm)

The Amherst News Times has evolved much in the same way the community of Amherst has evolved over the years. The history of newspapers in Amherst goes back to the late 1880’s when the North Amherst Reporter, a two-page paper with world news headlines was printed and managed by Harvey Closs and his family.

Fred Avery purchased the paper in 1920 and renamed it the North Amherst Argus. When Avery sold the business back to Clock in 1960 the name reverted back to The North Amherst Reporter. The “North” was dropped from the title when North Amherst became Amherst in 1963.

Around 2014 a second newspaper, the Amherst News, was started by then Amherst Mayor J. Smith and his business partner Oscar Dunn. A year later they bought out the Amherst Reporter, renaming the paper by integrating the name Amherst News.
Ohio Memory (CONTENTdm)
### Object Description

<table>
<thead>
<tr>
<th>Title</th>
<th>Amherst News-Times, 1932-08-04</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place</td>
<td>Amherst (Ohio)</td>
</tr>
<tr>
<td>Date of Original</td>
<td>1932-08-04</td>
</tr>
<tr>
<td>Source</td>
<td>LCCN: sn84028333, Amherst News-Times, 1932-08-04 15 28</td>
</tr>
<tr>
<td>Submitting Institution</td>
<td>Amherst Public Library</td>
</tr>
<tr>
<td>Type</td>
<td>Text</td>
</tr>
<tr>
<td>Digitization Information</td>
<td>300dpi, 8 bit Grayscale, Model: SunRise 2000, Software: iArchives, Inc., 3.240</td>
</tr>
<tr>
<td>Format</td>
<td>newspapers</td>
</tr>
<tr>
<td>LCCN</td>
<td>sn84028333</td>
</tr>
</tbody>
</table>

### Description

<table>
<thead>
<tr>
<th>Title</th>
<th>page 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place</td>
<td>Amherst (Ohio)</td>
</tr>
<tr>
<td>Creator</td>
<td>Amherst News-Times, 1932-09-04</td>
</tr>
<tr>
<td>Source</td>
<td>Real number: 00000000014</td>
</tr>
<tr>
<td>Submitting Institution</td>
<td>Amherst Public Library</td>
</tr>
<tr>
<td>Type</td>
<td>Text</td>
</tr>
<tr>
<td>Digitization Information</td>
<td>300dpi, 8 bit Grayscale, Model: SunRise 2000, Software: iArchives, Inc., 3.240</td>
</tr>
<tr>
<td>File Size</td>
<td>3720.04KB</td>
</tr>
</tbody>
</table>

Full Text:

> v. the amherst news-times amherst, lore in county, ohio, thursday august 4, 1932 sandstone cantsowrheairworld y- of fiea well, folks, day for big picnic is nearly here, and it tip to people make most of an opportunity that presented them by business men, people's picnic, with men adding us hosts, peams type being this year - just suited conditions are facing at bills time... haven't money pent would be required give our children a good time park like crystal beach, but ill cascade there plenty amusements free, and other forms entertainment provided full everyone, peams picnics designed primarily year, any rate for children, and mothers who need get out where they can spend quiet away from their worries. type, co-operation or all men, attendance local people, suit success. http://www.revelations in made week community
Thank You!

- Jenni Salamon, Project Coordinator, NDNP-OH
  - jsalamon@ohiohistory.org
- Kevin Latta, Quality Control Technician, NDNP-OH
  - klatta@ohiohistory.org