Quality Review of Digital Newspapers: Lessons Learned at UNT

National Digital Newspaper Program
Awardee Conference
September 27, 2012
Digital Newspapers at UNT

• 2007: NDNP Award from NEH to digitize Texas newspapers for Chronicling America; establishment of the Texas Digital Newspaper Program (TDNP)

• 2009: Partnership with Oklahoma Historical Society – technical coordinator NDNP-OK award; supplemental award NDNP-TX

• 2010: Partnership with University of New Mexico – technical coordinator NDNP-NM award

• 2011: Supplemental awards NDNP-TX and NDNP-OK

• 2012: Launch of The Gateway to Oklahoma History; supplemental award NDNP-NM
Our Digital Newspaper Team

Ana Krahmer
Digital Newspaper Unit Supervisor

Ann Howington
NDNP-TX and TDNP

Andrew Weidner
NDNP-NM

Trista Barker
Microfilm Technician

Sarah Lynn Fisher
NDNP-OK and ODNP

2 Student Assistants
TDNP and ODNP

Cathy Hartman
Contracts and Grants

Mark Phillips
Technical Wizardry
NDNP Quality Review: Workflow

Pre-digitization:

1. Review microfilm
   - Note titles, volume and issue numbers, section and edition labels, number of pages for each date, and anomalies, comments on image quality

2. Communicate anomalies to vendor
   - Missing issues, missing/duplicate pages, misprinted dates/page numbers, incorrect page/section order
NDNP Quality Review: Workflow

Post-digitization:

1. Copy batch from external hard drive
2. Validate and/or verify batch
3. Check batch name in batch.xml and batch_1.xml
4. Using the DVV:
   - Start and end of each reel, check:
     - Compare issue dates in batch to reel collation
     - TIFF, JP2, PDF images; OCR (all pages); Headers
     - LCCN, Institution, Institution MARC Code, Title in Headers and Metadata
NDNP Quality Review: Workflow

Post-digitization, continued:

5. Review all TIFF images
   - Compare date, volume/issue numbers, edition and section labels in metadata to what is printed on the page
   - View thumbnails for each issue
   - OCR: Check 10% or one page per issue
     - Compare OCR text and accuracy percentage to image

6. Compile a list of corrections. Send to vendor or complete in-house.
NDNP Quality Review: Workflow

Post-digitization, continued:

7. QC Corrected batch
8. Validate and copy to external hard drive. Verify!
9. Compare number of PDF, JP2, and TIFF files. Numbers should match!
10. Compare number of issues in batch.xml to directory issues in the batch. Numbers should match!
Tip: Know Your Microfilm

* Careful review of microfilm and collation of data can reduce errors in your batch prior to quality review.
Collation Spreadsheet

<table>
<thead>
<tr>
<th>Title: The Daily Ardmoreite</th>
<th>Reel No: 00200296710</th>
<th>Year(s): 1912</th>
<th>Pages: 1200</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>January</strong></td>
<td><strong>May</strong></td>
<td><strong>September</strong></td>
<td><strong>October</strong></td>
</tr>
<tr>
<td>1 2 3 4 5 6 7</td>
<td>1 2 3 4 5 6 7</td>
<td>1 2 3 4 5 6 7</td>
<td>1 2 3 4 5 6 7</td>
</tr>
<tr>
<td>8 9 10 11 12 13 14</td>
<td>8 9 10 11 12 13 14</td>
<td>8 9 10 11 12 13 14</td>
<td>8 9 10 11 12 13 14</td>
</tr>
<tr>
<td>15 16 17 18 19 20 21</td>
<td>15 16 17 18 19 20 21</td>
<td>15 16 17 18 19 20 21</td>
<td>15 16 17 18 19 20 21</td>
</tr>
<tr>
<td>29 30 31</td>
<td>29 30 31</td>
<td>29 30 31</td>
<td>29 30 31</td>
</tr>
</tbody>
</table>

Page counts are automatically calculated using Excel formulas.
Collation Spreadsheet

Anomalies:

<table>
<thead>
<tr>
<th>Date</th>
<th>Vol</th>
<th>Issue No.</th>
<th>Duplicate Pages?</th>
<th>Missing Pages?</th>
<th>Section Labels</th>
<th>Page Count Published</th>
<th>Page Count Filmed</th>
<th>Blocked Page Count</th>
<th>Page Count Delivered</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1912-05-27</td>
<td>XVIII</td>
<td>209</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>No missing pages. Pages 2, 3, numbered 6, 7. Page 7 numbered 3.</td>
</tr>
<tr>
<td>1912-07-01</td>
<td>XVIII</td>
<td>239</td>
<td></td>
<td></td>
<td>26, 7, 11</td>
<td>12</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>Pages filmed: 1, 10, 9, 8, 5, 6, 3, 12. Reorder numerically with missing page targets.</td>
</tr>
<tr>
<td>1912-07-02</td>
<td>XVIII</td>
<td>240</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>8</td>
<td>0</td>
<td>Pages filmed: 1, 10, 9, 8, 5, 4, 3, 12. Reorder numerically with missing page targets.</td>
</tr>
<tr>
<td>1912-07-26</td>
<td>XVIII</td>
<td>261</td>
<td>6, 7</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>10</td>
<td>2</td>
<td>Use 2nd occurrence of these pages. First page 6, 7 filmed very light.</td>
</tr>
<tr>
<td>1912-07-30</td>
<td>XVIII</td>
<td>264</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>7</td>
<td>0</td>
<td>Missing issue - include target</td>
</tr>
<tr>
<td>1912-07-31</td>
<td>XVIII</td>
<td>265</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>7</td>
<td>0</td>
<td>Missing issue - include target</td>
</tr>
<tr>
<td>1912-09-18</td>
<td>XVIII</td>
<td>306</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>Missing issue - include target</td>
</tr>
</tbody>
</table>

Reel Information:

<table>
<thead>
<tr>
<th>Roll Information:</th>
<th>Roll Number/Etc.</th>
<th>Date Filed</th>
<th>OAI/ID</th>
<th>Title/Location</th>
<th>LCCN</th>
<th>OCLC</th>
<th>Start Date YYYY-MM-DD</th>
<th>End Date YYYY-MM-DD</th>
<th>Era</th>
<th>Date Filmed YYYY-MM-DD</th>
<th>Where Filmed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Film received 2/29/2012</td>
<td>TDA_19120425-19121015</td>
<td>2/29/2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reel Number/Etc.</td>
<td>OAI/ID</td>
<td>Date Filed</td>
<td>Title/Location</td>
<td>LCCN</td>
<td>OCLC</td>
<td>Start Date YYYY-MM-DD</td>
<td>End Date YYYY-MM-DD</td>
<td>Era</td>
<td>Date Filmed YYYY-MM-DD</td>
<td>Where Filmed</td>
<td></td>
</tr>
</tbody>
</table>
Tip: Allow Your Workflow to Evolve

* Incorporate checks for unanticipated errors into your workflow.

Example: OCR

Problem: Incomplete OCR files
Solution: 10% instead of random check exposed truncated OCR files!
Tip: Document Your Workflow

* Documentation prevents memory lapses and insures consistency between projects.

Digital Projects Unit Wiki

NDNP Digital Newspapers

- NDNP Microfilm Workflow
- NDNP Hard Drive Workflow
- Checking Invoices for NDNP
- NDNP Contact Information TExAS
- NDNP Contact Information OKLAHOMA
- NDNP Contact Information NEW MEXICO
- iArchives Correspondence
- NDNP Processing For Portal
- ftp from iArchives

Categories: Newspapers | NDNP

Newspapers
- Current Newspaper Projects Schedule
- Evaluating Materials for Microfilming
- NHPRC
- Working with non English language content in newspapers
- Sherman/Denison Newspapers
- Tulia Herald Project
- Microfilm Scanning Queue
- Microfilm Scanning Workflow
- NDNP_Processing_For_Portal
- Newspaper QC Workflow (scanned elsewhere)
- Newspaper Notes
- Born Digital Newspaper Processing
- TDNP Project Page
- TDNP Newspaper Project Checklist
- NDNP Projects
- TDNP Newspaper Workflows
- Problems/Solutions in Newspaper Processing
- Oklahoma Newspaper Projects
- Newspaper Group Microfilm Cabinet

This page was last modified on 12 October 2011, at 07:33. This page has been accessed 422 times.
A few more tips!

* Speed through the DVV using Automated Keyboard shortcuts

Free, open-source utility for Windows  
www.autohotkey.com

* Consult the NDNP wiki for sage advice, ask questions, learn from each other