IFLA International Newspaper Conference


11-13 April 2012
BnF, Paris

With the support of:
Sustainability in the U.S. National Digital Newspaper Program

U.S. NATIONAL ENDOWMENT FOR THE HUMANITIES and LIBRARY OF CONGRESS

Presented by Sue Kellerman, Pennsylvania State University and NDNP Award Manager
IFLA International Newspaper Conference
13 April 2012
Working with U.S. Newspapers

• Newspapers = fundamentals of U.S. history

• Many types of users, high demand for access

• No single U.S. collection – 140,000 titles published since 1690 (collected across the country)

Newspaper format challenges

• Physical characteristics
  • Large, brittle, acid paper, poor ink, light damage

• Content characteristics
  • Many subjects on a page, small text, hard to identify parts
U.S. National Digital Newspaper Program, 2004-

GOALS:

- To enhance access to historic American newspapers from every state and territory

- To apply emerging technologies to the products of the United States Newspaper Program, 1982-2011 (inventory, description, microfilm)
  - 140,000 titles cataloged, 900,000 holdings, more than 75 million pages filmed

- To develop best practices for the digitization of historic newspapers (through shared community)

- Provide free and open access to content
U.S. National Digital Newspaper Program, 2004 -

- **Timeline**
  - 2004= agency-level agreements
  - 2005=1st awards (6 recipients in 2005)
  - 2007= launch of Chronicling America web site
  - 2012+ … 28 states and counting, >5 million pages…

- **Program Components**
  - Over time, NEH grants **2-year awards (up to $350k) to state projects**, to select and digitize historic newspapers, primarily from microfilm, for full-text access (100,000 pages per award).
  - Program provides **uniform** selection guidelines, focused on **historic significance AND appropriateness for conversion (film quality)**
  - LC creates and **hosts Chronicling America Web site** to provide freely accessible search and discovery for digitized papers and descriptive newspaper records (records created by USNP).
  - State projects **repurpose NDNP contributions for local purposes**, as desired.
PARTICIPANTS AS OF 2011 AWARDS:
28 institutions | >5.6 million pages by 2013 | 1836-1922

7–28–11
U.S. National Digital Newspaper Program and Chronicle America: Historic American Newspapers

Available Now:

- >4.8 million pages
- 1836-1922
- 700+ titles from 25 states and Washington, DC
- http://chroniclingamerica.loc.gov/

- Awards made, 2005-2011
  - 5.6 million pages currently funded
  - Next awards announced – Sept 2012

Coming Soon (Summer 2012):

- Content from states who joined in 2011 (Indiana, North Dakota, West Virginia)
- More pages!
## Technical Sustainability – Objects and Management

<table>
<thead>
<tr>
<th>Objectives:</th>
<th>Aggregate, serve, sustain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principles:</td>
<td>Open, modular, certain to change, able to evolve</td>
</tr>
<tr>
<td></td>
<td>Support OAIS workflows (ingest, archive, disseminate, manage)</td>
</tr>
<tr>
<td>Realities:</td>
<td>Distributed producers, finite funding, public funds, expect change (keep doors open)</td>
</tr>
</tbody>
</table>

### Resulting High-Level Requirements:
- High quality objects appropriate for reuse and for preservation,
- Technical consistency across objects,
- Open formats and sustainable practices,
- Scalable data environment,
- Reliable data management tools,
- Efficient delivery for access
Digital Objects

Digital objects = open, appropriate to use, achievable

- Newspaper Page = TIFF, JPEG2000, PDF, ALTO XML for OCR
- Issue = METS XML
- Reel = METS XML

Roles:
- TIFF = archive (grayscale, 300-400 dpi)
- JPEG2000 = Production (compressed, tiled for pan and zoom, derivatives on the fly, and high-resolution download)
- PDF = Portability (full page, searchable text)
- ALTO XML = search (keywords, locational information for words)
- Issue = object structure; includes preservation and technical metadata
- Reel = provenance, technical metadata
Digital Object Tools – PMT Image

- Preservation Microfilm Target (for Imaging)
  - Standardized measure of imaging quality
  - Compare imaging capabilities
  - Ongoing quality control
  - (produced in cooperation with Image Science Associates – imagescienceassociates.com)

- Includes:
  - Tonal measures (exposure) – grayscale boxes (OECF),
  - Resolution measures – wedge, checkerboard
  - Sharpness measures – slant edge (SFR),
  - Spatial distortion – fiducials (corners)
  - Noise measures
Digital Object Tools – Digital Viewer and Validator

- Ensure consistent, reliable data
- Conformance to specifications
- Help with quality assurance
- Used at LC, Awardees, Vendors
- Command-line or Visual Interface

- Automated Analysis
  - JHOVE-based
  - >100 characteristics
  - Validates all formats to specification
  - Extended Java: adds “digital signatures” (checksum), PREMIS, MIX to XML
  - Verifies checksums on demand

- Subjective Quality Review
  - Visual interface for quality assurance
  - Compare metadata to image
  - Display embedded file metadata
  - Confirm relationships
Enhancing Access - Data in Action

Access to 5 million+ pages
- Search by place, time, keyword
- Title, Date, Edition, Section, Page
- Visual search results (Thumbnail view with hit-highlights)
- Pan and Zoom
- Full-screen view
- Download, Clip/Print
- Share by Email and Social Media

Access to US Newspaper Directory
- Search by place, time, keyword, format, subject, etc. (CONSER/WorldCat data)
- Keyword search – e.g., “http” (external Web site links) or “times”
- Bibliographic descriptions
- Library holdings
- Newspaper histories for digitized titles
Enhancing Access – Behind the Scenes

- Open access, Free, Sustainable, Scalable
- Web Browser, Machine and API* access to data
- Open Source Software Base
  - Apache HTTPD Web Server
  - Django Web Publishing Framework
  - JQuery JavaScript Library
  - MySQL
  - SOLR/Lucene Search Server
  - Python Libraries

- Published open-source as “Library of Congress Newspaper Viewer” on SourceForge.net - http://sourceforge.net/projects/loc-ndnp/?source=directory
- Semantic Web and RDF
  - Persistent URL, Embedded Dublin Core, MODS

- Multiple interfaces to data supports multiple uses of data – browse/search, harvest, mine, visualization, etc.

*API (Application Programming Interface) – see http://chroniclingamerica.loc.gov/about/api/
Enhancing Access - Beyond the Web Site

- **Subscribe** to Weekly update/highlights, Recent Additions (by RSS*/email)
  - Chronicling America Home Page – **Subscribe** button

- **Share** the wealth – send to email, Facebook, Twitter and more…
  - Chronicling America Home Page – **Share** button

- **Topics** in *Chronicling America* ([http://www.loc.gov/rr/news/topics/topics.html](http://www.loc.gov/rr/news/topics/topics.html))
  - 100+ subject-specific guides from the Newspaper and Current Periodicals Reading Room for exploring the history in *Chronicling America*, including sample links to articles!

- **Illustrated Newspapers in Flickr**
  - Illustrated supplements from *the New-York Tribune*, 1900-1909

- **Awardees** (state-specific) **Blogs and topics** pages
  - University of Kentucky - [http://kyndnp.blogspot.com/](http://kyndnp.blogspot.com/)
  - University of South Carolina - [http://library.sc.edu/blogs/newspaper/](http://library.sc.edu/blogs/newspaper/)
  - University of Vermont - [http://vtdnp.wordpress.com/](http://vtdnp.wordpress.com/)
  - And others… (see [http://www.loc.gov/ndnp/awards/](http://www.loc.gov/ndnp/awards/))

- **Data Visualization** (example) – Stanford University/Bill Lane Center for American West
  - [http://www.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us_newspapers](http://www.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us_newspapers)
Sustainability – Content/Data Management

Principles: People, process, technology changes
Asset must remain available for use
Transparency and automation required

Repository Workflows and Services:
- Use of BagIt for data management
  - Draft IETF specification, incorporates manifest and checksum/hash values
  - In use by NDIIPP partners
- LC Content Transfer Services (CTS)
  - Automated, subjective checkpoints, malware, verification, periodic spotchecks
  - Workflow tracking, preservation metadata created and managed
  - Web-based access control layer, transparency
  - Storage information management
  - Reporting and statistics
- Data Storage Policies
Sustainability – Organizational Infrastructure

- NEH and LC Agreements
  - Formal Agreements
  - Shared program development
  - Cost sharing
  - Stakeholders and Roles – NEH Division of Preservation and Access, LC Office of Strategic Initiatives and Serial and Government Publications Division
  - Communication

- LC Stakeholders
  - Program
    - Includes curatorial and technical stakeholders
    - Guide program development and outreach
  - Technical and curatorial groups
    - Data preservation architecture and repository development
    - Software development and open data modeling
    - Digital library standards and practices
    - Content type curators
    - Integration with other digital collections

- Lessons learned shared both within NDNP and beyond
Summary

Enhancing Access to Newspapers in a Sustainable Digital Library

- Uniform body of content from multiple producers and locations
- Provide basic access to many types of users
- Develop shared community of practice
- Implement sustainable (and scalable) infrastructure to manage large data sets for ongoing access and transparency
History’s “Rough Draft” ...or “Everything Old is New Again”

The Washington Herald (Washington, DC), 14 February 1912
Thank you!

- NDNP Public Web
  [http://www.loc.gov/ndnp/](http://www.loc.gov/ndnp/)

- NDNP Web Service
  *Chronicling America: Historic American Newspapers*
  [http://chroniclingamerica.loc.gov](http://chroniclingamerica.loc.gov)

- Contact us at [ndnptech@loc.gov](mailto:ndnptech@loc.gov)