Bulking Up: How Accepted Standards and Evolving Technology Advance Research in Chronicling America

2014 IFLA International Newspapers Conference
Salt Lake City, Utah, USA

Nathan Yarasavage, Deborah Thomas & Georgia Higley
Library of Congress
Chronicling America: Historic American Newspapers
http://chroniclingamerica.loc.gov/
World War II sapped rural newsrooms of manpower as young men enlisted in droves after the Pearl Harbor attacks in 1941. As in many other industries, women filled the gap in large numbers, and legal protections were passed for newspapers that needed to suspend publication during the War. Newspapers placed new emphasis on maps and information graphics to help explain the complex and far flung battlefields to the public. Back home, the War Relocation Authority permitted Japanese-Americans to publish their own newspapers in detention camps like Manzanar in California and Gila River in Arizona.

« Back 1930 | Next: 1950 »
- **1918-09-01**: Some discussions about war and Influenza cases in Washington
- **1918-09-22**: Outbreak of the flu in Washington - this is the most virulent epidemic period
- **1918-10-21**: Intense negotiations about the armistice
The Belmont chronicle and farmers mechanics and manufacturers advocate
How did we get from this...
... to this?!
GOALS:

- To enhance access to historic American newspapers from every state and territory
- To develop best practices for the digitization of historic newspapers (shared community)
- Free and open content, available to all
NDNP: Built for Sustainability

- Shared technical specifications
- Institutional cooperation
- Data integrity
- Data management
- Expect change
Shared technical specifications

NDNP Technical Guidelines for Applicants
(NDNP Tech Specs)

http://www.loc.gov/ndnp/guidelines/
Shared technical specifications

- Stable guidelines since 2005
- Changes limited to:
  - Clarifications to existing practice
  - Version updates (ALTO 2.0)
  - Expanding content scope
  - Simplifying some metadata requirements
    - Technical metadata from Microfilm collation now OPTIONAL
  - Section / Edition Labels now OPTIONAL

---

The National Digital Newspaper Program (NDNP)
Technical Guidelines for Applicants

Document Control
Overview of Technical Approach for 2012-14 NDNP Awards
Deliverables
Technical Details
Selection
Scanning and Master Image Format
OCR and Associated Information
Other Derivative Files
Metadata
Technical Validation of Digital Objects
Delivery of Digital Assets
Appendices
Appendix A: Digital Asset Metadata Elements - Dictionary
Appendix B: File Format Profiles and Specifications
Appendix C: XML Metadata Templates
Appendix D: Batch, File and Directory Structure on Delivery Media
Appendix E: Associated Online Resources

Document Control
Revision | Date | Summary of Change
--- | --- | ---

Overview of Technical Approach for 2012-14 NDNP Awards
The National Digital Newspaper Program is a long-term effort and the technical environment will change as the program continues. The National Endowment for the Humanities (NEH) and the Library of Congress (LC) have selected a technical approach to balance long-term objectives and shorter-term constraints. These include:

- convenient accessibility over the World Wide Web for the general public to the entire collection as it grows, through a consistent interface and using proven technology;
- page images of sufficient spatial and tonal resolution to support effective performance of OCR (optical character recognition) software and representation of printed halftones, given the limitations of microfilm, expecting that future improvements in OCR and image processing will be applied to the same images;
- the use of digital formats with a high probability of sustainability - in particular, using standard formats where possible and proprietary formats only where widely adopted;
- and attention to the cost of digital conversion and maintenance of the resulting assets.

The goal of the initial program phase is to enhance access to newspapers through a Web-accessible NDNP delivery application, open and freely available to the public. This delivery application will provide access to a substantive amount of newspaper content selected for its historical value with broad geographic coverage and consistent sustainable
Distributed Project / Shared Effort
NDNP Data

For every page:

- Archival Image: TIFF
- Production Image: JPEG 2000
- Printable Image: PDF
- OCR XML File: ALTO

For every issue and reel:

- METS XML File
  - Descriptive metadata
  - Structural metadata
  - Preservation metadata

For newspaper title:

- MARC record
  - Geographic metadata
  - Time Period
  - Subject metadata

http://www.loc.gov/ndnp/guidelines/
METS ALTO

- NDNP ALTO specification requires:
  - Column level text block zoning and
  - Coordinates to map or highlight text to image files.

- Chronicling America supports page level access with visual representation of search results.

NDNP / Chronicling America
Page Level Access

Zoom, pan, and clip tools are available
Chronam: Open Source software

https://github.com/LibraryofCongress/chronam
Chronicling America - API

Introduction
Chronicling America provides access to information about historic newspapers and select digitized newspaper pages. To encourage a wide range of potential uses, we designed several different views of the data we provide, all of which are publicly visible. Each uses common Web protocols, and access is not restricted in any way. You do not need to apply for a special key to use them. Together they make up an extensive application programming interface (API) which you can use to explore all of our data in many ways.

Details about these interfaces are below. In case you want to dive right in, though, we use HTML link conventions to advertise the availability of these views. If you are a software developer or researcher or anyone else who might be interested in programmatic access to the data in Chronicling America, we encourage you to look around the site, "view source" often, and follow where the different links take you to get started. When describing Chronicling America as the source of content, please use the URL and a Web site citation, such as "From the Library of Congress, Chronicling America: Historic American Newspapers site."

For more information about the open source Chronicling America software please see the LibraryOfCongress/chronicling America GitHub site. Also, please consider subscribing to the Chronicling America discussion lists if you want to discuss how to use or extend the software or data from its APIs.

The API
Jump to:
- **Search** the newspaper directory and digitized page contents using OpenSearch.
- **Auto Suggest** API for looking up newspaper titles.
- **Link** using our stable URL pattern for Chronicling America resources.
- **JSON** views of Chronicling America resources.
- **Linked Data** views of Chronicling America resources.
- **Bulk Data** for research and external services.
- **CORS and JSON** support for your JavaScript applications.

http://chroniclingamerica.loc.gov/about/api/
## OCR Data

**What is OCR?**

Optical character recognition (OCR) is a fully automated process that converts the visual image of numbers and letters into computer-readable numbers and letters. Computer software can then search the OCR-generated text for words, phrases, numbers, or other characters. However, OCR is not 100 percent accurate, and, particularly if the original item has extraneous markings on the page, unusual text styles, or very small fonts, the searchable text OCR generates will contain errors that cannot be corrected by automated means.

Although errors in the process are unavoidable, OCR is still a powerful tool for making text-based items accessible to searching. For example, important concept words often appear more than once within an article. Therefore, if OCR misreads one instance of a key word in a passage, but correctly reads the second instance, the passage will still be found in a full-text search.

To enable research and external services Chronicling America provides bulk access to its OCR data. The table below itemizes a list of files available for download. Each file will decompress into a directory structure that lets you easily map the OCR file to the URL identifier for that page. For example a file such as sm89000214/1903/05/01/ad-1/seq-1/ocr.txt maps to the URL: http://chroniclingamerica.loc.gov/record/sm89000214/1903-05-01/ad-1/seq-1/.

If you are interested in automated access to this data you may want to use the **Atom** and **JSON** versions of this table.

<table>
<thead>
<tr>
<th>Filename</th>
<th>Batch</th>
<th>Created</th>
<th>Size</th>
<th>SHA-1 Checksum</th>
</tr>
</thead>
<tbody>
<tr>
<td>batch wa indiianplum ver02.tar.gz</td>
<td>batch wa indiianplum ver02</td>
<td>2014-01-14T21:54:30-05:00</td>
<td>2.6 GB</td>
<td>b7131b222d6f2e22af04c77ea5471db782a94f63</td>
</tr>
<tr>
<td>batch wvu belgium ver01.tar.gz</td>
<td>batch wvu belgium ver01</td>
<td>2014-01-14T21:54:48-05:00</td>
<td>2.4 GB</td>
<td>819997e355e08cd32742891cd6051a52bf5664529</td>
</tr>
<tr>
<td>batch wvu denmark ver01.tar.gz</td>
<td>batch wvu denmark ver01</td>
<td>2014-01-14T21:44:51-05:00</td>
<td>2.1 GB</td>
<td>627345b3b9e31891d5664a024f808d2e0ff17</td>
</tr>
<tr>
<td>batch wvu eritrea ver01.tar.gz</td>
<td>batch wvu eritrea ver01</td>
<td>2014-01-14T21:31:30-05:00</td>
<td>2.1 GB</td>
<td>208fb744d5f563eb5b981d3e5b14e5bb3d60b75</td>
</tr>
<tr>
<td>batch vtu graniteville ver01.tar.gz</td>
<td>batch vtu graniteville ver01</td>
<td>2014-01-14T21:20:57-05:00</td>
<td>1.6 GB</td>
<td>6e3aff622b061b640e0624e164f3950d96f9b11</td>
</tr>
<tr>
<td>batch vtu burlington ver01.tar.gz</td>
<td>batch vtu burlington ver01</td>
<td>2014-01-14T21:16:31-05:00</td>
<td>1.7 GB</td>
<td>f41c1c6117d77c6b77600cb50da457b219d7d</td>
</tr>
<tr>
<td>batch wa fir ver01.tar.gz</td>
<td>batch wa fir ver01</td>
<td>2014-01-14T21:09:18-05:00</td>
<td>1.7 GB</td>
<td>c04d31a3d48151b07440dadcfc2d91a07213f75</td>
</tr>
<tr>
<td>batch wvu austria ver01.tar.gz</td>
<td>batch wvu austria ver01</td>
<td>2014-01-14T21:06:36-05:00</td>
<td>1.7 GB</td>
<td>edfb97412b2629a190d7c3d4008937a5b80d6</td>
</tr>
</tbody>
</table>

[http://chroniclingamerica.loc.gov/ocr/](http://chroniclingamerica.loc.gov/ocr/)
The Growth of US Newspapers, 1690-2011

1900’s “Yellow Journalism”

The turn of the century saw furious competition between the mass-circulation dailies in New York: William Randolph Hearst’s New York Journal and the New York World, owned by Joseph Pulitzer. Both papers used screaming headlines, the bold use of color and striking illustrations, “many of them without significance,” according to the historian Frank Luther Mott. The term “yellow Journalism” stemmed from a recurring character in the comic strip Hogan’s Alley, nicknamed the “yellow kid,” but was applied more widely to define the sensationalistic news and feature stories that climaxed with the Spanish–American War, which Hearst famously championed in the pages of the Journal.

< Back to 1880 | Next: 1909 >

Data sources: Chronicling America collection, Library of Congress; Urban population estimates from NASA

Data visualization by the Rural West Initiative, Bill Lane Center for the American West, Stanford University.

Credits: Dan Cheng, Kristy Clark, Yuanxia Ge, Geoff McGuire, Yingping Qin and Jason Wang

Notice something wrong? Please share your corrections and additions with us, and we’ll try to verify and include them.
Mapping Texts is a collaboration between scholars, staff and students at Stanford University and the University of North Texas. It is supported by the National Endowment for the Humanities.

http://mappingtexts.org/
“An Epidemiology of Information: Data Mining the 1918 Influenza Pandemic”

“The goal … is to develop methods for combining algorithmic techniques with the interpretive strength of traditional historical and rhetorical analysis in order to help researchers better understand reporting on the 1918 flu pandemic in American and Canadian newspapers.”

- Text mining
- Topic Modeling
- Tone Analysis

Professor E. Thomas Ewing, Principal Investigator and Project Director, Department of History, Virginia Tech

http://www.flu1918.lib.vt.edu/
“Infectious Texts” is sponsored by Northeastern University’s NULab for Texts, Maps, and Networks and generously funded by the National Endowment for the Humanities’ Office of Digital Humanities. The project team includes Professors Ryan Cordell, Elizabeth Maddock Dillon, and David Smith, as well as Ph.D. students Abby Mullen and Matthew Williamson.”
That 15,000 persons were probably killed and wounded in a typhoon that swept the Philippine Islands last Tuesday was reported yesterday in cable dispatches to the Bureau of Insular Affairs.

The typhoon swept the Visayas and is said to have practically destroyed Tacloban, the capital of Leyte, and to have wrought enormous damage and loss of life at Capiz, the capital of the province of Capiz.

Tacloban has a population of 12,000. Capiz has a population of over 20,000. Capiz is the terminal of the railroad from Ilolo. It is a most important sugar port.

Aid Rushed to Scene.

The first news of the catastrophe came in a dispatch from the governor general of the Philippines. No figures of the dead or injured were given, but it was stated that probably half the population of the two cities had been lost.

The governor general sent his dispatch on Thursday. He informed the department that he was rushing a shipload of food, clothing, and all available medical supplies to Tacloban. All telegraphic communication has been destroyed, and it
Beyond the *Chronicling America* Web Site

Regular update/highlights information by RSS* or email subscription
- [http://www.loc.gov/rss/ndnp/ndnp.xml](http://www.loc.gov/rss/ndnp/ndnp.xml); and **Recent Additions RSS feed** (newspaper titles added)
- Keep up with what’s in *Chronicling America*!

Open-source code for chronam application available on GitHub
- [https://github.com/LibraryOfCongress/chronam](https://github.com/LibraryOfCongress/chronam)
- chronam is the [Django](https://www.djangoproject.com) application that the Library of Congress uses to make its *Chronicling America* website (a core set of functionality for loading, modeling and indexing NDNP data)
- Check it out and take it for a spin!

Linked data and API** access in *Chronicling America*
- RDF/XML views available to Open Archives Initiative Object Reuse and Exchange (OAI-ORE)
- OpenSearch and Atom APIs for search queries, Bookmarkable URIs for all pages
- [http://chroniclingamerica.loc.gov/about/api/](http://chroniclingamerica.loc.gov/about/api/)

NDNP Extras! from the NDNP program site
- [http://www.loc.gov/ndnp/extras/](http://www.loc.gov/ndnp/extras/)
- Visualizations, tutorials, podcasts, teaching resources, state project blogs, how *Chronicling America* is used in other projects and more!
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)