

## National Digital Newspaper Program (NDNP)

<http://loc.gov/ndnp/>

The National Digital Newspaper Program (NDNP) [1] is a partnership between the National Endowment for the Humanities (NEH) and the Library of Congress (LC) to create and maintain an Internet-based, freely-accessible, searchable database of U.S. newspaper information and select digitized pages. This database is known as Chronicing America [2].



Figure 1: NDNP Awardees 2005-2011

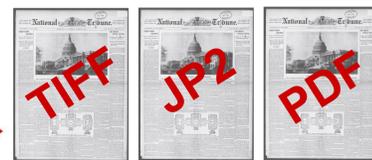
NEH provides 2-year grants for state institutions (awardees) to digitize 100,000 pages of microfilmed newspaper, published between 1836 and 1922.

To date, 28 states have contributed to the program.

## NDNP Digitization Specification



Figure 2: NDNP Microfilm to Digital Specification at a Glance



METS, MODS, ALTO, MARC, PREMIS, MIX



Figure 3: Digital Viewer and Validator Tool (DVV)

Awardees select titles, arrange for either in-house or vendor-based digitization to the NDNP specification [4], and complete quality assurance of the data using the LC provided Digital Viewer and Validator (DVV), a JSTOR/Harvard Validation Environment (JHOVE) based tool allowing viewing of NDNP data and validation of select technical aspects of the files [3].

## Acknowledgements

The National Digital Newspaper Program is funded by the National Endowment for the Humanities and supported by the Library of Congress. Chronicing America, including its API and the LC Newspaper Viewer software, has been developed by the Repository Development Center (RDC) at LC. Special acknowledgements include: Leslie Johnston and the staff of the RDC; Teri Sierra; Mark Sweeney; and Deborah Thomas.

## Chronicing America

<http://chronicingamerica.loc.gov/>

Chronicing America is the website providing access to the NDNP resources, which include:

- 140,114 newspaper records
- 385 title essays
- Almost 5 million digitized pages



Figure 4: From Chronicing America: Screen capture and RDF representation of Page 1, November 10, 1898 issue of the National Tribune newspaper. <http://chronicingamerica.loc.gov/locn/sn82016187/1898-11-10/ed-1/seq-1.rdf>

## Traffic

Opening Chronicing America's data to internet search engines and commercial users greatly increased the project's visibility and site traffic. In June 2009, the project reached a milestone of one million pages of content, and a site redesign added API access and crawler support. Site traffic doubled overall, and search referrals multiplied fifty-fold. A May 2011 redesign added social media sharing tools, increasing the direct linking capabilities of the site. Chronicing America has also been harvested by commercial users and integrated into proprietary databases. One subscription genealogy site now refers about 5% of total traffic. As detailed in Figures 6 and 7, use of the site continues to increase. From March 2011 to March 2012, total hits to the site increased over 300%. Genealogy sites and search engines that heavily use our API are also among the top referring domains.

## API

Using common Web protocols and linked data principles, the Application Programming Interface (API) for Chronicing America promotes a wide range of data use. Through the API, several views of the digitized content and metadata are publicly visible, with no restrictions to access. No API key is required [7].

## OpenSearch

Searching across the Chronicing America Newspaper Directory of over 140,000 newspaper MARC title records is possible using the OpenSearch protocol and the OpenSearch Description document describing Chronicing America's search engine.

## CORS and JSONP Support

Chronicing America has been constructed to integrate with third-party JavaScript applications, with support for both Cross-Origin Resource Sharing (CORS) and JavaScript Object Notation with padding (JSONP) responses.

## Sitemap Protocol

Chronicing America follows the Sitemap protocol [10], providing an index of URLs linked from the site's robots.txt file.

```
# Hi Robot, welcome to Chronicing America
User-agent: *
Disallow: /nothing---please-crawl-us--
Disallow: /beta
Sitemap: http://chronicingamerica.loc.gov/sitemap.xml
```

Figure 5: Chronicing America robots.txt file.

## Linked Data

Resource Description Framework (RDF) representations of the resources, using existing and new vocabulary terms, have been employed. A resource map of all newspapers is available for clients interested in harvesting NDNP objects. The concepts of newspaper titles and issues available in the data are described as Title (defined in DCMI Metadata Terms) and Issue (defined in the Bibliographic Ontology). The site also defines concepts not already defined in existing ontologies. This vocabulary includes elements suitable for newspaper information and the NDNP program, including these elements (Awardee, Batch, Page, number, section, sequence).

## Aggregations

The OAI-ORE specification allows LC to define aggregations of resources (pages) on the site that compose a single unit. For example, using the OAI-ORE vocabulary, the site can link Page resource to the JPEG2000, PDF, and OCR file counterparts. Batch, Title and Issue resources are related to Pages using the OAI-ORE vocabulary. One can "view HTML source" on Title, Issue, Page and Batch views to discover the referenced RDF/XML file in a link element as an OAI-ORE Resource Map.

More info about the NDNP API is available at: <http://chronicingamerica.loc.gov/about/api/>

	March 2011	March 2012
Site Hits		+ 323%
Robots		+ 132%
• Top Robot	Domestic search engine A	Domestic search engine A
• Robot #2	Commercial genealogy site A	Foreign search engine A
• Robot #3	Domestic search engine B	Commercial genealogy site A
Top Opensearch User	Commercial genealogy site A	Commercial genealogy site A
Unique Visitors		+ 91%
Site Referrals		+ 62 %
• Top Referring Domain	Domestic search engine C	Domestic search engine C
• Referrer #2	Domestic search engine B	Commercial genealogy site A
• Referrer #3	Commercial genealogy site A	Domestic search engine A

Figure 6: Chronicing America Traffic Statistics March 2011 and March 2012

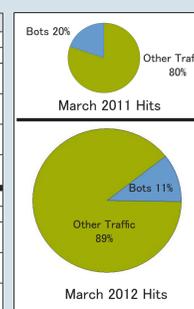


Figure 7: Chronicing America Hits Statistics March 2011 and March 2012

## Using NDNP Data

Chronicing America has welcomed academic research by providing API documentation on the site as well as joining the list of repositories for the NEH sponsored Digging into Data Challenge [12]. For a 2011 Challenge award, researchers from the Virginia Polytechnic Institute and the University of Toronto will mine the OCR text to examine public opinion and information dissemination during the 1918 Influenza Pandemic. In a visualization project, Stanford researchers used MARC records from the newspaper directory to map the growth of American newspapers [13] (Figure 7). Chronicing America also inspired a visualization project on the evolution of individual newspaper titles through mergers and acquisitions [14]. Linkypedia, a project developed by Ed Summers, shows how Chronicing America content has been reused or linked within Wikipedia [15]. Chronicing America's raw OCR text has been used in numerous other academic research endeavors.

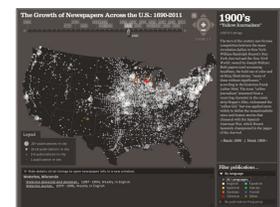


Figure 7: Stanford University, Data Visualization: Journalism's Voyage West.

## LC Newspaper Viewer (Open Source Software)

To further the goal of providing open access to historical newspapers, the Library of Congress released an open source version of the software underlying Chronicing America. Called the LC Newspaper Viewer, the application is architected using Apache Httpd, Django, MySQL, and Apache Solr and is available without restriction (from Sourceforge [16]) for use by other parties wishing to provide access to or view data produced to the NDNP specification.

Current public-facing use of the software includes the University of Oregon [17] (Figure 8), with internal use by several other awardees.



Figure 8: Historic Oregon Newspapers

## References and Links

- [1] Library of Congress. National Digital Newspaper Program. Accessed Jan. 30, 2012. <http://www.loc.gov/ndnp/>
- [2] Library of Congress. Chronicing America. Accessed Jan. 30, 2012. <http://chronicingamerica.loc.gov/>
- [3] Littman, J. 2006. A Technical Approach and Distributed Model for Validation of Digital Objects, D-Lib Magazine, 12, 5.
- [4] Murray, R. 2005. Toward a Metadata Standard for Digitized Historical Newspapers. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, 330-331.
- [5] The British Newspaper Archive. Accessed Jan. 30, 2012. <http://www.britishnewspaperarchive.co.uk/>
- [6] Klijn, E. 2008. The Current State-of-Art in Newspaper Digitization: A Market Perspective. D-Lib Magazine, 14, 1/2.
- [7] Library of Congress. Chronicing America. About the Site and API. Accessed Jan. 30, 2012. <http://chronicingamerica.loc.gov/about/api/>
- [8] Library of Congress. BagIt Specification. Accessed Jan. 30, 2012. <http://www.digitalpreservation.gov/documents/bagit-spec.pdf>
- [9] Library of Congress. Chronicing America. Data. Accessed Jan. 30, 2012. <http://chronicingamerica.loc.gov/data/>
- [10] Sitemaps.org. Accessed Jan. 30, 2012. <http://www.sitemaps.org/>
- [11] WorldCat Search API. Accessed Jan. 30, 2012. <http://www.worldcat.org/affiliate/tools?atype=wcapi>
- [12] Digging Into Data Challenge. Accessed Jan. 30, 2012. <http://www.diggingintodata.org>
- [13] Stanford University. Data Visualization: Journalism's Voyage West. Accessed Jan. 30, 2012. [http://www.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us\\_newspapers](http://www.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us_newspapers)
- [14] BeingNumerous. Genealogies of Old Newspapers. Accessed Jan. 30, 2012. <http://beingnumerous.com/blog/2010/05/genealogies-of-old-newspapers/>
- [15] Github. Linkypedia. Accessed Jan. 30, 2012. <https://github.com/edsu/linkypedia>
- [16] Sourceforge. LC Newspaper Viewer. Accessed Jan. 30, 2012. <http://sourceforge.net/apps/trac/loc-ndnp/>
- [17] University of Oregon. Historic Oregon Newspapers. Accessed Jan. 30, 2012. <http://oregonnews.uoregon.edu/>