National Digital Newspaper Program: A Case Study in Sharing, Linking, and Using Data

Nathan Yarasavage
Library of Congress
101 Independence Ave SE
Washington, DC
202-707-2954
nyarasavage@loc.gov

Robin Butterhof
Library of Congress
101 Independence Ave SE
Washington, DC
202-707-8172
robu@loc.gov

Christopher Ehrman
Library of Congress
101 Independence Ave SE
Washington, DC
202-707-0223
cehr@loc.gov

ABSTRACT
This poster presents a case study describing how the National Digital Newspaper Program's (NDNP) metadata specification and public website, Chronicling America, have been designed to promote a wide range of data sharing. Through use of the website's extensive application programming interface (API) and open-source software counterpart, several institutions are benefiting from the publicly-funded program's data.

Categories and Subject Descriptors
H.3.5 [**Information Systems**]: Information Storage and Retrieval – On-line Information Services – Data sharing.
H.3.7 [**Information Systems**]: Information Storage and Retrieval - Digital Libraries - Collection, Standards.

General Terms
Documentation, Design, Standardization.

Keywords
Application Programming Interface, Digitization, Historical Newspapers, National Digital Newspaper Program

1. INTRODUCTION
The National Digital Newspaper Program (NDNP) 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. Chronicling America is the website providing access to these resources: 140,114 newspaper records, 385 title essays, and almost 5 million digitized pages at time of writing. NDNP is a distributed project modeled after the United States Newspaper Program (USNP), a national effort to inventory, catalog, and preserve on microfilm select at-risk original newsprint. The NDNP project builds upon the valuable bibliographic and microfilm assets created under the USNP. NEH provides 2-year grants for state institutions (awardees) to digitize 100,000 pages of microfilmed newsprint, published between 1836 and 1922. Post-1922 content is not considered in the public domain and is therefore outside the scope of this program. LC describes the digitization specifications and provides technical assistance to awardees. To date, 28 states have contributed to the program. Eventually data will be contributed from all states and territories. Awardees select titles, arrange for either in-house or vendor-based digitization, 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 [1][2][3].

2. NDNP METADATA SPECIFICATION
In 2005, the NDNP standard for metadata describing digitized newspapers was proposed by LC [4]. The NDNP metadata specification utilizes the Metadata Encoding and Transmission Standard (METS) to carry descriptive, administrative, and structural metadata about the newspaper's title, page, and reel information. Title (descriptive) information is mapped from existing Machine Readable Cataloging (MARC) metadata into the Metadata Object Description Schema (MODS) format. Optical Character Recognition (OCR) text is encoded with a METS extension schema known as Analyzed Layout and Text Object (ALTO) XML schema. The combination use of METS and ALTO, commonly referred to as METS/ALTO, has been widely adopted by many large-scale national newspaper digitization projects around the world including the Australian Newspapers Digitisation Program, the British Newspaper Archive, and the Bibliothèque nacionale de France [5][6]. During validation by the DVV, digital signatures and technical metadata describing the original source microfilm and digital assets are encoded in the METS files as PREservation Metadata: Implementation Strategies (PREMIS) and Metadata for Images in XML (MIX) metadata [1].

3. NDNP API
Using common Web protocols and linked data principles, the NDNP development team has designed an API for Chronicling America that promotes a wide range of potential data use. Through the API, several views of the digitized content and its metadata are publicly visible, with no restrictions to access [7]. Searching across the Chronicling America Newspaper Directory [2] of over 140,000 newspaper MARC title records is possible using the OpenSearch protocol and the OpenSearch Description document describing Chronicling America's search engine. Resource Description Framework (RDF) representations of the resources, using existing and new vocabulary terms has been employed. A resource map of all newspapers is available for clients interested in harvesting NDNP objects. All NDNP data submitted by awardees, except archival TIFF (Tagged Image File

Format) images, are packaged using the BagIt File Packaging Format [8] and made freely available to data harvesters via hypertext transfer protocol (http) [9]. Chronicling 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 [7]. Finally, Chronicling America follows the Sitemap protocol, providing an index of URLs linked from the site's robots.txt file [10].

4. USING NDNP DATA

While Chronicling America shares data freely through its API, the project also relies on outside data, namely, MARC records gathered via the WorldCat Search API [11]. These records comprise the bulk of the newspaper directory, a comprehensive listing of American newspapers published from 1690 to present. These records ensure up-to-date holdings information and represent a full catalog of U.S. newspapers. Chronicling 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]. Chronicling 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 Chronicling America content has been reused or linked within Wikipedia [15]. Opening Chronicling America's data to internet search providers (ISPs) 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, mostly from Google, multiplied fifty-fold. A May 2011 redesign added social media sharing tools, increasing the direct linking capabilities of the site. Chronicling America has also been harvested by commercial users and integrated into proprietary databases. One subscription genealogy site now refers about 5% of total traffic.

5. LC NEWSPAPER VIEWER

To further the goal of providing open access to historical newspapers, in 2010, the NDNP development team at LC released an open source version of the software underlying Chronicling America. Coined the LC Newspaper Viewer, the application is architected using Apache HTTPd, Django, MySQL, and Apache Solr and is available for download via Sourceforge. The LC Newspaper Viewer Software is available under a Berkeley Software Distribution (BSD) style license for use by individuals or organizations wishing to provide access to or view data produced to the NDNP specification. Customization of the software has enabled non-NDNP content creators to make use of it as well [16]. Current public-facing use of the software includes the University of Oregon [17]. Several other awardees report using the software locally at their institutions for internal quality review purposes.

6. ACKNOWLEDGMENTS

The NDNP is funded by the National Endowment for the Humanities and supported by LC. Chronicling 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.

7. REFERENCES

[1] Library of Congress. National Digital Newspaper Program. Accessed Jan. 30, 2012. http://www.loc.gov/ndnp/.

[2] Library of Congress. Chronicling America. Accessed Jan. 30, 2012. http://chroniclingamerica.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. Chronicling America. About the Site and API. Accessed Jan. 30, 2012. http://chroniclingamerica.loc.gov/about/api/.

[8] Library of Congress. BagIt Specification. Accessed Jan. 30, 2012. http://www.digitalpreservation.gov/documents/bagitspec.pdf.

[9] Library of Congress. Chronicling America. Data. Accessed Jan. 30, 2012. http://chroniclingamerica.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.

[14] BeingNumerous. Genealogies of Old newspapers. Accessed Jan. 30, 2012. http://beingnumero.us/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/.