Submitted on: 01.06.2017~~14.04.2017~~

2017 IFLA International News Media Conference
27-28 April 2017
Reykjavik, Iceland

Opening the Doors Wide: The US National Digital Newspaper Program, Open Data and the NEH Chronicling America Data Challenge

Deborah Thomas
Serial and Government Publications Division, Library of Congress, Washington, DC, USA.
E-mail address: deth@loc.gov

Leah Weinryb Grohsgal
Division of Preservation and Access, National Endowment for the Humanities, Washington, DC, USA.
E-mail address: lgrohsgal@neh.gov

Abstract:

What can you do with open access to millions of pages of historic newspapers? How do you prepare for this kind of access and what are the benefits? Chronicling America is an open access, searchable database of historic U.S. newspapers, produced by a long-term partnership between the US National Endowment for the Humanities (NEH) and the Library of Congress (LC). The program has generated millions of pages of digitized newspapers and descriptive information contributed by states and territories across the country, with open access and a well-documented API to explore it in a number of different ways. To spur use of the API and collection, in 2016, NEH hosted a contest to encourage researchers in thinking about creating interesting projects. The results demonstrate exciting possibilities for both creation of digital collections and reaching out to the research communities that use them.

The data available in Chronicling America and the mechanisms researchers and students can use to access it are described. The goals of the data challenge and snapshots of the six winning projects are included to provide a taste of the variety of research possible, focusing on important humanities themes and developing visualizations, maps, tools, and data mashups. Included are broader lessons for establishing connections between content holders and the research and educational community.

Keywords: historic newspapers, open data, digital humanities, research, digitization.

Fires and floods, elections, social issues, fashions, marriage and death reports, crime, lost and found, humor, song and scientific achievements – it can all be found in American newspapers published over the past three hundred years. This record of American culture – famously called "the rough draft of history" – is rich ground for the seeds of historic inquiry and research,

whether for the individual genealogist searching the family tree or the computational linguist mapping how texts traveled and repeated across the country. Regardless of the question, newspapers as data can play an important role in new kinds of historical research. However, to maximize the possible benefits, serving as many uses as possible, the data must be openly accessible, easily understood and standardized.

In 2004, the U.S. Library of Congress (LC) and the National Endowment for the Humanities (NEH) agreed to establish the National Digital Newspaper Program[i] (NDNP) with the intent to enhance access to historic American newspapers by supporting state-level digitization projects and aggregating the resulting digital issues into a sustainable dataset. Knowing the extraordinary quantity of newspapers published in the U.S. (approx. 154,000 titles since 1690), success for the program relied on determining achievable goals coupled with a long-term vision that would warrant the investment. To that end, the program developed policies and technical specifications that would support long-term growth and management of the program and the content and maximize available resources and potential uses. In particular, NDNP required that the data produced under the program be openly available and free for all to use.

There are several key elements to the program that have allowed for its success to date. Initially the chronological scope of the program was limited to the first decade of the twentieth century (1900-1910). The reasoning was two-fold – in 1900, all of the U.S. was settled in one form or another with published newspapers available, even though a given region was not yet a state. This made it possible for all states to participate; in addition, concentrating all digitization into a narrow time-period allowed us to demonstrate the value of aggregating selected but similar content (in terms of the history of journalism) from multiple states into one deep cross-searchable dataset. Each year, the chronological scope expanded to eventually reach its current range of 1690 (the publication of the earliest American newspaper) to 1964. (Under U.S. copyright law, unregistered publications published before 1964 are in the public domain.)

This gradual date expansion paralleled the planned gradual increase in participants in the program and geographic representation. In the first round of funding in 2005, NEH only made awards to six states to begin the program. These states were selected based on their experience and understanding of newspapers, experience with digitization, and understanding of digital library management needs. The intent of the selection was to ensure participants in the program were already experienced in digitization and understood the complexities of working with newspaper collections in terms of organization, description, and historical significance. The other aspect of the planned gradual increase in participants was more practical in terms of keeping the resources required of NEH and LC fairly steady over a lengthy period of time (approx. 20 years) in order to grow what would eventually be a substantial digital collection at a pace both institutions could support. Over time, NEH has awarded funding to 44 states and territories with a more limited number (approx. 25) actively producing data at one time.

Another basic concept of the program was to clearly define early on the roles of the participants; NEH provides funding to state awardees (one cultural heritage institution representing each state) and edits the essays that provide historical context for each title, LC provides technical expertise and supports the long-term access and management of the collection, and awardees select, digitize, and deliver the collections according to the technical specifications provided by the LC. Throughout the program, costs and resources are distributed and shared by all

participants. In addition, by engaging in this distribution of shared resources and requirements, the program has built a community of practice around newspaper digitization in the U.S., supporting digitization well-beyond the specific NEH/LC investment. Numerous awardees have continued digitization of their home-state newspapers beyond their participation in NDNP, making much more content available for research than the selected titles contributed for Chronicling America, [ii] the access point on LC's Web site to all NDNP-digitized newspapers.

A final, and perhaps, the most lasting component of the program, was to focus our technical specifications and implementation on open and sustainable data formats in order to plan for technical change over time and ensure the content received in the early years of the program would be, or could be made, compatible with content received in the later years of the program and beyond. Over time technology around the digitization of historical collections could be expected to change. Utilizing well-documented, common technical formats with clear and understandable metadata would allow for re-use, migration and/or transformation of the data to meet any number of as yet unknown future uses or needs.

Today the program includes participation from 44 states and territories. From Alaska to Maine, Puerto Rico to Hawaii, cultural heritage institutions from across the country have joined the program and contributed (or are in the processing of doing so) millions of historic digitized newspaper pages to this national collection. Newspapers from forty states and territories and the District of Columbia (represented by LC) are already available through Chronicling America's web site, approx. 12 million pages from 2200 titles. These U.S. newspapers are published primarily in English, but also French, German, Italian, and Spanish with more languages in process. In addition, awardees, curatorial experts in their own newspaper collections, have produced more than 1200 newspaper history essays on the historical value and context of each selected title. Also available from the Web site are searchable bibliographic and library holding records for the more than 154,000 U.S. newspapers published since 1690,[iii] to further enhance access to newspapers when open digitized content is not available.

For individual Web site users, Chronicling America provides a range of basic functionality that supports the in-depth use of historic newspapers. Limiting search to place, time and/or full-text keyword search (based on uncorrected OCR[iv] capture) provides access to newspaper pages with visual highlights to indicate where search matches occur. Page views include identity information for each page, with title, date and if included, edition; navigation to next and previous pages or dates, a view of the entire issue, downloadable page-level files, and the ability to pan across the image and magnify to eye-readable levels for reading.

However, beyond the features that support individual Web browsing, Chronicling America also supports access to all data through common Web protocols and formats, providing machine-level views of all data for harvesting and large-scale bulk download. As examples, researchers can harvest batched digitized page images as JPEG2000, PDF and/or METS-ALTO OCR, [v] or bulk OCR-only batches. Each newspaper page includes embedded Linked Data using a number of ontologies and supports JSON and RDF views. US Newspaper Directory bibliographic records are also available as MARCXML.[vi] The open API includes industry-standard endpoints like OpenSearch and supports stable intelligible URLs.

To accommodate data harvesting activities, the Chronicling America Web site infrastructure and workflow includes several features specifically designed to support such work.

1. During data ingest, additional text-only data sets are created and stored separately ready for bulk download[vii].
2. To create transparency and ease of access to the bulk downloadable data, feeds for the downloadable files, in both ATOM and JSON format were added. Researchers can subscribe to the feed to ensure they get any new data that is added.
3. For the interactive API (JSON & RDF) caching was added to provide fast responses for pages that need to be created "on the fly" by the server (as opposed to the bulk processed data that exists in flat files).

Chronicling America is a valuable humanities dataset that has wide potential for use in scholarship, research, and education projects. Since building the collection and releasing NDNP data in open, standardized, and easily understood formats, researchers in the digital humanities have used the collection in a variety of ways to explore historic newspapers from different angles than used in close-reading approaches. One of the earliest projects, supported by Stanford University, harvested the newspaper bibliographic records and provided an interactive map and timeline interface to track the distribution of newspaper publishing across the US as the country expanded from 1690 to the present.[viii] Researchers at Virginia Tech harvested and analyzed text for a specific set of newspapers in a confined time range associated with the Influenza Epidemic of 1918 to evaluate the relationship between public news and the spread of the epidemic.[ix] Scholars at Northeastern University mapped the republication of articles across newspapers to identify how information traveled in the nineteenth century and when and what was reused.[x] And a University of Nebraska group is currently developing technology to extract poems from historic newspapers based on their sizes and sentence structures.[xi] Other examples of projects using Chronicling America's digitized historic newspapers may be found on the Library of Congress NDNP Extras page.[xii]

The diverse subjects and varied methodologies of these projects show the potential for using the data in Chronicling America to address important humanities questions. The National Endowment for the Humanities wished to inspire more projects using this open data to create innovative digital projects. The NEH Chronicling America Historic American Newspapers Data Challenge, administered by NEH's Division of Preservation and Access, sought to address this desire for further open data use.[xiii] The agency used challenge.gov, a platform for agencies across the federal government to list competitions inviting the public to use innovative techniques to solve a wide range of problems.[xiv] A response to the White House Office of Science and Technology Policy's 2009 call for agencies to promote innovation, the site has been available since 2010. Since then, hundreds of U.S. government agencies, including the Department of Health and Human Services, the House of Representatives, NASA, the National Science Foundation, and the Bureau of Labor Statistics, have used challenge.gov to run contests.

NEH posted the contest on October 15, 2015, with a long entry period leading to a submission deadline of June 15, 2016. This was by design: the agency sought to give enough time to allow entrants to conceive projects, and also to allow teachers to incorporate the challenge into their curricula. And unlike the calls for grant applications typically issued by NEH, this contest required that entrants submit projects, rather than applications.

The challenge asked members of the public, "How can you use open data to explore history?" Entrants were to create tools or visualizations using the data in Chronicling America. The challenge's parameters were purposely left broad in order to encourage entrants to be creative in thinking about what humanities themes interested them and how to approach those questions. Projects could involve maps, visualizations, or tools for using the data, and entrants were permitted (even encouraged) to mash up the data with other datasets. The challenge offered First, Second, and Third prizes as well as a special K-12 prize. All prizes also included a trip to Washington, D.C. to present the project at the National Digital Newspaper Program annual meeting.

NEH received many entries, and awarded six prizes, including ties for second and third places. A panel of three judges with expertise in digital humanities, historic newspapers, and educational resources assisted the agency in making decisions.

The First Prize winner was Lincoln Mullen, an assistant professor in the Department of History and Art History at George Mason University, for his site *America's Public Bible: Bible Quotations in U.S. Newspapers*.[xv] The site tracks biblical quotations in American newspapers to see how the Bible was used for cultural, social, religious, or political purposes. It shows how the Bible was a contested yet common text, with quotations in the expected sermons and Sunday School lessons frequently printed in newspapers, as well as arguments on all sides of social and political issues like slavery, women's suffrage, and capitalism.

A Second Prize winner was by Andrew Bales, a PhD student in Creative Writing at the University of Cincinnati, titled *American Lynching: Uncovering a Cultural Narrative*.[xvi] The site explores America's long and dark history with lynching, in which newspapers acted as both a catalyst for public killings and a platform for advocating for reform. Bales integrated the Chronicling America data with data sets on lynching from Project Hal, a national lynching database based on the NAACP Lynching Records at Tuskegee University and the Beck-Tolnay Confirmed inventory of that data. The site is intended to tell the story of lynching in America not only through maps and other visualizations, but also by connecting the data to the stories of the victims.

Another Second Prize went to the team of Amy Giroux, a computer research specialist at the University of Central Florida Center for Humanities and Digital Research, Marcy Galbreath, a lecturer in University of Central Florida's Department of Writing and Rhetoric, and Nathan Giroux, a programmer who works as a software engineer in the military simulation industry, for *Historical Agricultural News*.[xvii] This is a search tool for exploring information on the farming organizations, technologies, and practices of America's past. The site asserts that farming is a window into the social, economic, political, and cultural history of the United States. Additionally, the group sought to make the vast amount of data available in Chronicling America more manageable by focusing on one broad subject.

One Third Prize went to a team from Indiana University-Purdue University Indianapolis: Kristi Palmer, associate dean of digital scholarship, Caitlin Pollock, digital humanities librarian, and Ted Polley, social sciences and digital publishing librarian, for *Chronicling Hoosier*.[xviii] The site tracks the origins of the word Hoosier, its geographic distribution, and its positive and negative connotations over time. The project uses visualizations that are all connected back to stories in Chronicling America to tell the unexpected story of the word. Additionally, the group has documented and made all of their code open, so that anyone inclined to do so may apply the methodology to another word.

Another Third Prize went to Claudio Saunt, professor of American history and chair of the History Department at the University of Georgia and Trevor Goodyear, a research scientist in the Innovative Computing Division of the Georgia Tech Research Institute, for *USNewsMap.com*.[xix] This site allows the end user to enter a word or phrase, and trace its use in American newspapers over time and in different places geographically. Saunt and Goodyear assert that, because of their quick publication schedule, newspapers capture public discourse better than other textual evidence such as books. The site allows users to discover patterns, explore regions, and investigate how terms spread through the newspaper data in Chronicling America.

Finally, NEH offered a K-12 Educational Prize, which went to teacher Ray Palin and the A.P. U.S. History Students at Sunapee High School in Sunapee, New Hampshire for *Digital APUSH: Revealing History with Chronicling America*.[xx] The group of fifteen students used word frequency analysis, a kind of distant reading, to discover patterns in news coverage. Working in teams on subjects interesting to them, the students produced graphs and visualizations showing coverage of the U.S. Supreme Court case *Plessy v. Ferguson*, secession from the Union at the start of the Civil War, the novel *Uncle Tom's Cabin*, the Ku Klux Klan, and labor unions. Using data allowed students to do the traditional historian's work of identifying important historical questions and thinking about how best to research them, according to Teacher Palin, but also allowed them to learn data analysis skills in the process.

This collection of projects highlights the incredible diversity in subjects and humanities themes and questions that can be explored using this historic newspaper dataset. In September 2016, winners convened in Washington, D.C., at the invitation of NEH, to present their projects at the National Digital Newspaper Program annual meeting. The presentations were very well received, engaging the audience and soliciting many questions. NEH's Office of Congressional Affairs also arranged for the winners to visit their representatives on Capitol Hill. Winners were also honored at a reception hosted by the NEH and the LC.

The contest points to the importance of encouraging an active community of users to engage with this open digital resource. The National Digital Newspaper Program has been active since 2004, and builds upon an even older, decades-long partnership between NEH and LC. The challenge capitalized on an audience that was already familiar with the resource, and inspired them either to create new projects or to complete projects that were already conceived. Winners have also described plans for their sites involving expansion, continued scholarly use, and collaboration with museums and libraries. The contest also attracted both to solo scholars and researchers working outside of academic departments and across disciplinary lines. The entrants were scholars, librarians, digital humanities practitioners, programmers, and students. This indicates the wide audience of users for this humanities data set.

The contest garnered publicity and interest in our agencies and in Chronicling America. For a modest investment of NEH funds, the data challenge enabled both NEH and LC to showcase innovative projects using the collection and establish contact with members of our user communities. NEH's Division of Preservation and Access worked with the NEH Office of Congressional Affairs and Office of Communications to publicize the results. Several projects were also featured in university publicity and in major news publications, such as *The Washington Post*.[xxi] The Library of Congress and NEH jointly issued press releases, and publicized the contest's results on their web sites.[xxii] NEH and LC consider the challenge to be a successful experiment in encouraging significant and substantial use of an important

humanities dataset created by the NEH and LC. The contest also served to highlight the agencies' long partnership to save America's historic newspapers.

Over time, humanities researchers will continue to grow in understanding and interest in exploring computational techniques that enable them to ask new and different questions. NDNP, by utilizing standardized data specifications and open access infrastructure, can extend the research potential of this valuable cultural heritage content to support these new approaches to understanding American history.

Acknowledgments

The authors would like to thank David Brunton, supervisory information technology specialist at the Library of Congress, for his expertise and other contributions in supporting this research.

References

[i] National Digital Newspaper Program. http://www.loc.gov/ndnp/.

[ii] *Chronicling America: Historic American Newspapers*, http://chroniclingamerica.loc.gov/.

[iii] Bibliographic and library holdings data harvested from OCLC WorldCat regularly through open API and agreement.

[iv] Optical Character Recognition

[v] JPEG2000 (high-resolution, compressed image file format with embedded metadata); PDF (compressed image file format for printing with embedded metadata); METS-ALTO (XML-based schema for Analyzed Layout Text Object); MARCXML (XML-based schema for bibliographic MARC records)

[vi] MARCXML (XML-based schema for bibliographic MARC records)

[vii] See https://github.com/LibraryOfCongress/chronam/blob/master/core/management/commands/dump_ocr.py

[viii] Data Visualization: Journalism's Journey West, Rural West Initiative, Bill Lane Center for the American West, Stanford University. http://web.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us_newspapers. Captured 1 April 2017.

[ix] An Epidemiology of Information, Virginia Polytechnic Institute and State University. http://www.flu1918.lib.vt.edu/

[x] Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines, NULab for Texts, Maps, and Networks, Northeastern University. http://viraltexts.org/

[xi] Finding Poetry Amid Historic News Pages, Center for Digital Research in the Humanities, University of Nebraska-Lincoln. http://research.unl.edu/annualreport/2015/finding-poetry-amid-historic-news-pages/

[xii] https://www.loc.gov/ndnp/extras/

[xiii] NEH Chronicling America Historic American Newspapers Data Challenge. https://www.challenge.gov/challenge/chronicling-america-historic-american-newspapers-data-challenge/

[xiv] Challenge.gov. https://www.challenge.gov/list/

[xv] Lincoln Mullen, America's Public Bible: Biblical Quotations in U.S. Newspapers. http://americaspublicbible.org/

[xvi] Andrew Bales, American Lynching: Uncovering a Cultural Narrative. http://www.americanlynchingdata.com/

[xvii] Amy Giroux, Marcy Galbreath, and Nathan Giroux, Historical Agricultural News. http://ag-news.net/

[xviii] Kristi Palmer, Caitlin Pollock, and Ted Polley, Chronicling Hoosier. http://centerfordigschol.github.io/chroniclinghoosier/

[xix] Claudio Saunt and Trevor Goodyear, USNewsMap.com. http://usnewsmap.com/

[xx] Ray Palin and AP U.S. History Students at Sunapee High School, Digital APUSH. https://apush.omeka.net/

[xxi] Julie Zauzmer, "Newspapers Were Once Full of Bible Quotes – and a Local Professor's Tool Lets Us Learn From them." *Washington Post*, 3 August 2016. https://www.washingtonpost.com/news/acts-of-faith/wp/2016/08/03/newspapers-were-once-full-of-bible-quotes-and-a-local-professors-tool-lets-us-learn-from-them/?utm_term=.9b84abe27baa

[xxii] For example: NEH Announces the Winners of the Chronicling America Data Challenge, https://www.neh.gov/news/press-release/2016-07-25; The NEH "Chronicling America" Challenge: Using Big Data to Ask Big Questions, https://blogs.loc.gov/thesignal/2016/08/the-neh-chronicling-america-challenge-using-big-data-to-ask-big-questions/.