Using Big Data to Ask Big Questions

Leah Weinryb Grohsgal
lgrohsgal@neh.gov
National Endowment for the Humanities
United States of America

How can you use open data to explore history?

Historic American newspapers illuminate the rich history and texture of life. They contain stories about politics, sports, music, shopping, food, health, science, movies, and everything in between. From the affairs of everyday life to major international events, newspapers enable humanists to delve into the past, providing vantage points in big cities and small towns, east and west, north and south, from various political, religious, and cultural standpoints, and multiple language and ethnic communities.

But what happens when cultural institutions make a large-scale set of historic newspaper data available? How are digital humanists able to use aggregated data from thousands of American news publications to tell stories and provide analysis, from the quotidian to the momentous? To find out, the U.S. National Endowment for the Humanities (NEH) asked the public "how can you use open data to explore history?" in the Chronicling America Historic American Newspapers Data Challenge. This paper describes the challenge and its results, which serve as powerful reminders of the possibilities for matching open data collections with the tools and questions of the digital humanities.

The nationwide competition, led by the NEH Division of Preservation & Access, interrogated the possibilities for using the data in *Chronicling America*, a digital repository of historic United States newspapers. *Chronicling America* is an open access, searchable database of historic U.S. newspapers, with a newly expanded date range from 1690 to 1963 (the date range at the time of the challenge was 1836 to 1922). The database and web site are produced by the National Digital Newspaper Program, a long-term partnership between the NEH and the Library of Congress (LC). *Chronicling America* includes millions of pages of digitized newspapers—nearly 12 million at the time of this writing, with more being added all the time—and descriptive information contributed by states and territories across the country. The LC supports the long-term management of the collection, including providing open access to the data through a well-documented API to enable exploration of the collection in a variety of ways beyond the site's web interface.

To spur use of the API and collection, the Chronicling America Historic American Newspapers Data Challenge encouraged researchers to create digital humanities projects, big or small, using the newspaper data. The results demonstrate exciting possibilities for connecting the creators of digital collections and the humanities research communities using them. The contest urged entrants to use the data to show trends, insights, themes, or stories from history using newspaper data. The challenge's parameters were broad, encouraging entrants to be creative in thinking about the humanities questions they find most compelling, and how they might approach them using open humanities data from newspapers.

The NEH awarded six prizes for the Chronicling America Historic American Newspapers Data Challenge in 2016. Far from being theoretical or speculative, the contest and the winning projects highlight the practical ways in which this data can be used to explore a variety of humanities themes. In brief, the winning projects were:

America's Public Bible: Bible Quotations in U.S. Newspapers, by Lincoln Mullen. The site tracks Biblical quotations in American newspapers to show how the Bible was used for cultural, social, religious, and political purposes, and how it was a contested yet common text.

American Lynching: Uncovering a Cultural Narrative, by Andrew Bales. This site explores America's long, dark history of lynching and the role of newspapers as both catalysts for killings and platforms for reform.

Historical Agricultural News, by Amy Giroux, Marcy Galbreath, and Nathan Giroux. The site is a tool for exploring information on farming organizations, technologies, and practices, as a window into social, economic, and political history.

Chronicling Hoosier, by Kristi Palmer, Caitlin Pollock, and Ted Polley. This site tracks the origins of the word "Hoosier," its geographic distribution, and its positive and negative connotations over time.

USNewsMap.com, by Claudio Saunt and Trevor Goodyear. This site allows users to discover patterns, explore regions, and investigate how words, terms, and news spread.

Digital APUSH, by Ray Palin and the A.P. U.S. History Students at Sunapee High School. This class used word frequency analysis to discover patterns in news coverage of several major issues such as secession, the KKK, and Plessy v. Ferguson.

This paper illustrates the potential impact of humanities collections such as *Chronicling America* when their data are made freely and easily available. It describes the data available in this huge data repository, the mechanisms researchers and students can use to access it, and some of the challenges inherent in its large span. The paper addresses some of the technical specifications and program guidelines for the National Digital Newspaper Program, in which metadata standards for both access and preservation were primary concerns for the NEH and the Library of Congress in creating and maintaining the program. It also explains how the program's decade-old community has cultivated shared practices and specifications that contribute both to the longevity of this dataset and to other newspaper digitization efforts across the country.

Then, the paper gives information about the winning entries in the first Chronicling America Historic American Newspapers Data Challenge in 2016, which touched on a variety of important humanities themes like religion, race, literature, violence, agriculture, law, and geography. It explains how winners used cutting-edge technology to produce maps, visualizations, tools, and data mashups. Winners reported their initial questions, the importance of this data in answering them, their methods, and their future plans for the projects they built. This paper highlights their processes and the variety of results they produced.

Finally, the paper discusses broader lessons for humanities users of "big data" in *Chronicling America*. It shows the value of a contest in publicizing data in the humanities and its uses. It showcases different modes of collaboration among humanities researchers, libraries, information technology professionals, and other partners in the research endeavor. The paper also explores some of the challenges for humanities scholars working with large datasets, including representation, bias, categorization, and documentation. The intellectual work of the digital humanities projects described here involves the same problem identifying and research that humanists have always pursued, but the methods and investigation present new questions and opportunities. The paper suggests ways of maximizing the benefits of large datasets such as *Chronicling America* to the humanities.