Mining the Indianapolis Recorder: An Exploratory Study of a Digital Humanities Dataset

Ted Polley, Heather Coates, Jenny Johnson, Jere Odell, Kristi Palmer
IUPUI University Library
Indiana University-Purdue University Indianapolis

Established in the late 19th century, the *Indianapolis Recorder* is one of the nation's most enduring and important African American newspapers. In 2011, IUPUI University Library partnered with the owners of the Recorder to create a full text digital collection of the newspaper. Funding was provided by the Indianapolis Foundation Library fund. Today, researchers may search, browse and read over 5,000 issues covering 106 years of the newspaper's history in the library's open access digital collections.

The full-text contents of the collection can be accessed as a tab-delimited or XML metadata text file. As one approach for tapping into the rich potential of the text, we conducted an initial, exploratory study using basic approaches to text mining and visualization. For this descriptive exploration, we focus on characterizing the semantic content as it represents significant topics and concepts in Indianapolis, African American history and journalism. Visualizations were produced to make these concepts accessible and to demonstrate the value of this collection as a potential data set for humanities and social science research.

Methods: We exported the full text of more than 96,000 pages of the recorder in both a tab-delimited and XML text file. As large files, over 1.3 GB, these proved problematic for common, desktop computing software, such as Microsoft Excel. However, we imported the tab-delimited text file into R, a software program for statistical analysis, and divided the data set into more manageable portions. We created a text file from the transcripts of all the issues published in 1899 (3.42MB) and separate text file from the transcripts of all the issues published in 2005 (15.2MB). Using VOSviewer, a program for analyzing co-occurrences between words in a text corpus, we created a term map from each text file. We set the term occurrence threshold to 100 occurrences and removed terms, such as “page” or “Recorder”, that refer specifically to the paper itself.

Results: The visualizations generated from this data set show the topical character of the *Indianapolis Recorder*. Frequently used words are displayed in association with proximal vocabulary. Thus, as would be expected, “church” is frequently associated with “pastor.” We observed that church-related terms figure prominently in both the 1899 and 2005 term maps, providing a strong indicator as to their importance to the particular community this newspaper serves. We also observed unexpected associations; “teeth,” for example, was associated with “fact” in the term map created from the 1899 issues. A closer look at the text revealed columns on the subject of dentures (their manufacture and use) as well as advertisements for dentists serving the newspaper’s readership.

Discussion: This digital humanities data set represents a valuable asset for the exploration of the history of Indianapolis’ African American community, the evolution of language over time, and changes in the structure of a newspaper to reflect community interests and needs, among other things. The research use of this data, however, must overcome a number of limitations. These include the potential lack of computing resources for the average user that easily enable analysis of the entire data set, the need for improved metadata to facilitate finding and dividing the data, and lower OCR quality, especially for earlier issues. Additional studies using this data set will need to explore ways to address some or all of these limitations—perhaps best begun for a selection of the data. Nonetheless, the full text of the *Indianapolis Recorder* is currently available for innovative research and digital humanities projects.