What News is New?: Ads, Extras, and Viral Texts on the Nineteenth-Century Newspaper Page

Ryan Cordell
rccordell@gmail.com
Northeastern University, United States of America

David Smith
dasmiq@gmail.com
Northeastern University, United States of America

Newspapers became a truly mass medium in the nineteenth century due to the steam press, the post, the telegraph, and, especially in the U.S., mass political parties. Many scholars working on newspapers and other nineteenth-century media have noted the high level of "exchange" between different publications and the importance of understanding what was reprinted for understanding what readers and editors of the period valued (see, for instance, McGill, 2003; or Garvey, 2012).

In the first stages of the Viral Texts Project, we developed efficient clustering methods based on statistical language modeling and alignment to identify reprinted texts in digital archives of newspapers and magazines, beginning with the Chronicling America corpus and expanding to include papers from the UK, Australia, and Europe. Approaching the newspaper corpus through the lens of text reuse, Viral Texts has led to substantial insights into low-level problems of text reuse analysis in errorful OCR archives, higher-level network analysis of cultural circulation (Smith et al, 2015), the informational mode of nineteenth-century newspaper reprinting, the network effects of authorship in the nineteenth-century newspaper medium (Cordell, 2015), the circulation of popular "fugitive poetry" during the period (Cordell and Mullen, forthcoming 2017), the bibliographic implications of errorful OCR (Cordell, forthcoming 2017), among other things (project publications and press are available on the project site, and the project data sets are available on Github).

The computational methods of the Viral Texts Project produce a database view of the textual field. We read nineteenth-century newspaper snippets as "clusters" of reprints in a spreadsheet or database. We do not read them in the contexts of their original publications, but as disambiguated segments of text excerpted, aggregated, and listed. Such a database view is not *substantially* distinct from much older bibliographic views. The "clusters" of reprints generated by this algorithmic approach are still, essentially, enumerative bibliographies of textual snippets that circulated in nineteenth-century newspapers. Organized in a database, their appearance and presentation echo the conventions of printed bibliographies that list, for example, all known witnesses of a particular author's works.

Considering a text's bibliography allows us to speak of it in terms of circulation, audience, and influence, but the meanings of those corpus-scale phenomena are not evenly distributed among the witnesses that comprise its bibliography, when considered individually at the codex scale. Our current work seeks to map what we know of reprinting at a systemic level back onto the newspaper page, taking up the challenge in Matthew Philpott's recently articulated "understanding of the conceptual dimension of a periodical as a statistically self-similar, fractal form across all levels of scaling, from the periodical as a whole in its full publication run, to the year or annual volume, and down to the individual issue or number." (Philpotts, 2015) Likewise, what we are learning about disambiguated reprinted texts can help us understand the generic conventions and material operations of particular newspapers from our corpus over time, and to compare such features among papers.

Since developing our methods to detect reprints, project Ph.D. student Jonathan Fitzgerald has developed classification methods for sorting the millions of resultant clusters into meaningful categories: separating poetry from prose, for instance, or fiction from advertising testimonials and hard news (for discussion of some of this early classification work, see Fitzgerald, 2016). Building on Fitzgerald's classification work, we can now ask what generic trends we can spot within particular newspapers, or across the corpus. Some genres prove quite amenable to automatic classification, with success rates well over 90% for classifying advertisements and news and 84% for classifying poetry. Nineteenth century scholars and book historians, for instance, would expect to see poems in particular corners of the newspaper: the top left corner of page one in some papers, or the top left corner of page 4 in others. Drawing on tens of thousands of automatically identified poems, we can test those ideas more broadly, asking just how consistently the *Lewisburg Chronicle, and the West Branch Farmer* reproduced poems in the page one poetry corner, or just what percentage of papers chose one or the other of these two conventional placements. Perhaps more interestingly, we look for outliers in the data: newspapers that printed poetry in ways that defied generic norms, which may prove to be particularly interesting for periodicals and book historians.

Another structural aspect of periodicals addressed by this approach could be summed up with the question, "What news is new?" In short, while we know newspapers relied on reprinting and other kinds of textual recycling to fill column inches with limited staff, we can now compare habits of reuse within papers and among papers, asking for instance which newspapers relied more or less on reprinting—which were primarily producers or consumers of content—and how reprinted content populates the pages of particular newspapers. Again, we ask what patterns of new vs. reprinted content we can map across issues of particular papers, or how those patterns compare among different papers. Finally, we investigate how reprinted material migrates within particular newspapers over time; as stories age, can we trace them moving from page one through the latter pages of subsequent issues?

At an even finer level, we can ask how much total content was kept set up in type from one newspaper issue to the next. In the early and middle nineteenth century, many newspaper advertisements ran continuously for weeks or months at a time. Some advertisements even had short codes printed at the end for start date and duration, so that compositors could check whether an ad should be kept. In addition, news stories might be only slightly updated from day to day. By tracing this reuse among consecutive issues of the same newspaper, we can better estimate how much work by editors and compositors went into putting together the paper. We can also see how these practices changed over the course of the century as wire services and national advertising campaigns increased the freshness of newspaper content (figure 1). By illuminating precisely what text was kept in standing type, both in particular newspapers and comparatively across the corpus, we generate new insight into the workings of the nineteenth-century printing office, as well as the priorities and work habits of editors and other newspaper employees. Given the new internationalism of our corpus, we can also compare such practices across national and linguistic boundaries.

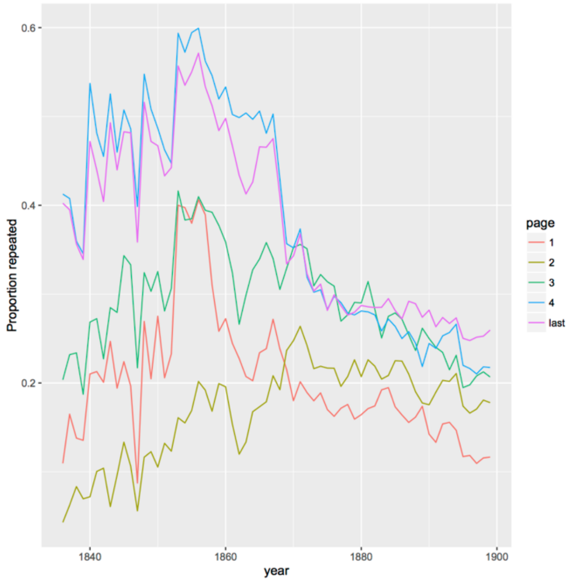

Figure 1: Average proportion of text on a page repeated from the previous issue of the same newspaper in the Chronicling America dataset, 1835–1900. Separate lines show reprint proportions on page 1, 2, 3, and 4. Through the early part of this historical period, most newspaper issues had four pages. A separate line shows reprint proportion on the last page, whatever it is. Repeated content is measured by globally aligning each page with all pages of previous issues in the last seven days.

For this paper, we build on the alignment and clustering methods developed in the Viral Texts project in two ways. First, by applying text categorization at the cluster level, we are able to make more robust inferences about genre than we would by scanning an individual newspaper page. Since each cluster of reprints preserves the location on the original page of each witness, we can then easily map these inferences back onto each issue. Second, we implement new, efficient global alignment algorithms to trace repeated ads, boilerplate, and news updates across successive issues of all newspapers in the corpus. We can prune the search space of this alignment algorithm more aggressively since texts kept in set type will retain the same line breaks, spelling errors, etc., as their previous printings. For robustness in the face of OCR errors, we align each issue to one or two weeks worth of previous issues. While global, linear (monotonic finite-state) alignment gives us promising initial results (see figure 1), we are also experimenting with greedy top-down inference that allows block moves of passages. Finally, we note that this procedure for detecting repeated passages across consecutive issues gives us yet more evidence about the boundaries of stories, which might not be typographically marked by, e.g., headline fonts.

These are just a few examples of how we are modeling the structure of newspaper layout and production. Such modeling allows us to test ideas about newspaper materiality advanced by scholars working with particular publications and ask new questions about both trends and outliers in the newspaper system of the nineteenth century.

Bibliography

Cordell, R. (2015) "Reprinting, Circulation, and the Network Author in Antebellum Newspapers," *American Literary History* 27.3 (August 2015), pre-print available at http://ryancordell.org/research/reprinting-circulation-and-the-network-author-in-antebellum-newspapers/.

Cordell, R. (forthcoming, 2017) "'Q i-jtb the Raven': Taking Dirty OCR Seriously,", forthcoming in *Book History*.

Cordell, R., and Mullen, A. (forthcoming, 2017) "'Fugitive Verses': The Circulation of Poems in Nineteenth-Century American Newspapers," forthcoming in *American Periodicals* 27.1 (Spring 2017), preprint available at http://viraltexts.org/2016/04/08/fugitive-verses/.

Fitzgerald, J. D. (2016). "Computationally Classifying the Vignette Between Fiction and News". *Jonathan D. Fitzgerald.* Blog post. 10 October 2016. Available at http://jonathandfitzgerald.com/blog/2016/10/10/the-viral-vignette.html

Garvey, E. G. (2012) *Writing with Scissors: American Scrapbooks from the Civil War to the Harlem Renaissance* (Oxford: Oxford University Press).

Philpotts, M. (2015) "Dimension: Fractal Forms and Periodical Texture," *Victorian Periodicals Review* 48.3 (Fall 2015): 413.

McGill, M. (2003) *American Literature and the Culture of Reprinting, 1834-1853* (Philadelphia: University of Pennsylvania Press).

Smith, D., Mullen, A., and Cordell, R. (2015) "Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers," *American Literary History* 27.3 (August 2015), pre-print available at http://viraltexts.org/2015/05/22/computational-methods-for-uncovering-reprinted-texts-in-antebellum-newspapers/.