by Peter McCracken

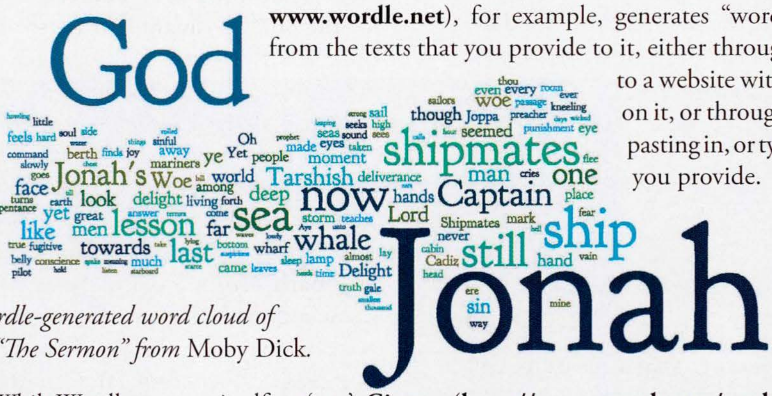
Text Mining and Digital Humanities Tools

The use of digital tools in the humanities has been growing by leaps and bounds in recent years. Some of the early tools could create a concordance for text, but these were just the beginnings of ways to analyze and explore texts. Today, new methods of using technology to explore textual works continue to appear, providing a host of tools to researchers in the humanities.

I'm a huge fan of serendipitous research—of just wandering into something unexpected, but relevant. But the wandering shouldn't quite be aimless; there should be something guiding it to maximize the chance of finding something both interesting *and* useful. In the library, this is often accomplished by roaming the stacks near other works. Online, this can be a bit tougher. **Serendip-o-matic** (at **http://serendipomatic.org**), however, does a great job of improving serendipitous research. It analyzes text that you provide—usually a paragraph or three of text that's of interest to you—and spits out a wide range of results, drawn primarily from the Digital Public Library of America, Europeana, Flickr Commons, and Trove. Each of these sites compiles a range of digital resources, usually from different geographical regions: photos, monographs, paintings, journal articles, maps, artifact images, government documents, and much more.

Serendip-o-matic shows what a few talented and dedicated digital humanists can do when locked up in a room together. The program was built over the course of a week in July 2013, during a digital humanities software development institute. The program is at version 0.1.1, and it looks like it won't be developed further. Content, however, is constantly added to the aggregated collections it mines, so results will continue to expand and improve as the program has more content to dive through.

The **Text Analysis Portal for Research** (**TAPoR**, at **http://www.tapor.ca**) gathers a collection of more than 300 tools for advanced text analysis and retrieval. **Wordle** (**http://www.wordle.net**), for example, generates "word clouds" from the texts that you provide to it, either through a URL to a website with full text on it, or through directly pasting in, or typing, text you provide.



Wordle-generated word cloud of "The Sermon" from Moby Dick.

While Wordle presents itself as a 'toy,' **Cirrus** (**http://voyeurtools.org/tool/Cirrus/**) presents itself as a serious tool for academic textual analysis. In addition to creating a word cloud, Cirrus, which is one part of a collection at **http://voyeurtools.org**, can analyze groups of works, links from the word cloud to all instances of the word in the text, and offers a range of other tools.

Google's Ngram Viewer (**http://books.google.com/ngrams**) makes it possible to analyze the distribution of terms across Google's massive collection of digitized volumes. Type in selected terms and see how they compare to each other. For example, a search comparing the terms "shanty" and "chanty" shows an increase in usage of "chanty" around 1810, and then a dropping-off from that point forward, while the use of the term "shanty" began in about 1825 and increased dramatically to 1890. (A search of "shanty" versus "shanty, excluding shantytown" showed very little use of the term "shantytown" before the 1970s, but more work could be done to ensure that "shanty" here refers only to music, and not dwellings.) One can also follow the growth and comparative usage of the terms "barquentine" and "barkentine" over time, and much more. Google provides extensive advanced tools, but one must also recognize the challenges inherent in ensuring that searches are doing what one intended.

Suggestions for other sites worth mentioning are welcome at **peter@shipindex.org**. See **http://shipindex.org** for a free compilation of over 140,000 ship names from indexes to dozens of books and journals.