**AUVIL — Exploration**

<1 figure>

**Exploration of Billions of Words of the HathiTrust Corpus with Bookworm: HathiTrust + Bookworm Project**

Auvil, L., Aiden, E. L., Downie, J. S., Schmidt, B., Bhattacharyya, S. and Organisciak, P.

Humanities scholars are traditionally concerned with close reading of a relatively small number of texts. Yet, as new textual resources such as the Google Books project and the HathiTrust Digital Library (HTDL) emerge, there is an increasing need for tools that analyze textual resources at scale. HTDL is the largest nonprofit digital book collection in the world, containing a total of 13,026,050 volumes in over 100 languages. The goal of this project is to integrate the HTDL corpus, processed at the HathiTrust Research Center (HTRC) (Downie et al., 2012), with the Bookworm platform for text analysis, developed at the Cultural Observatory.

Bookworm is an open-source platform that enables real-time analysis of repositories of digitized texts. Bookworm greatly extends the type of analysis that was popularized by the Google Ngrams Viewer (Michel et al., 2011), making it possible to slice and dice the data in an arbitrary corpus, in real time, using a greatly enhanced set of content-based and metadata-based features.

This poster will demonstrate initial results of this project (HT+BW)—in particular, a functional Bookworm interface displaying text data from HTRC. The HT+BW will greatly increase the value of the HTRC because it will assist humanities scholars and students in their effort to delve deeper into the HathiTrust corpus and to explore more complex, multifaceted research questions. At the same time, this project will continue to develop Bookworm as an open-source platform tool, not only for HTRC, but also as a potential portal for all libraries with extensive digital content. This collaboration includes the University of Illinois, Indiana University, Rice University, Baylor College of Medicine, and Northeastern University.

**Implementing Analytics at Scale**

One of our goals for this collaboration is to implement a greatly enhanced open-source version of the Bookworm text analysis and visualization tool designed to assist scholars to meet the challenges posed by the massive scale of the HT corpus. The enhanced Bookworm will enable many important new features—for instance, enabling scholars to better customize sets of text for their personal analyses (HTRC worksets), and to identify new HTRC texts to add to their corpora in real time. We will also be improving the APIs used by Bookworm to leverage a Solr index, an index used by many libraries and digital archives.

**Identify Valuable Metadata Formats for Humanities Scholars**

The effort to curate and deploy metadata is essential to any digital library effort, especially given the painstaking effort of cataloging by librarians. We have identified certain metadata fields that will be useful for examination of HTRC data. For instance, many raw MARC fields (year of publication, country, language) have been added. Some fields can be easily computed: page counts and word counts. Still other fields, such as author gender, can be recovered with high reliability by analysis of author names and comparison with external data repositories. We expect that the following metadata fields will be integrated into the HathiTrust Bookworm: Class, Subclass, Fiction, Genre, Language, Issuance, Author Gender, Page Count, Word Count, Publication Country, and Publication State. Thus, by combining the HathiTrust data with Bookworm analytics, scholars of English literature can study word frequencies in English novels, regional historians can limit their search to publishers from particular places, and historians of science can compare chemistry texts to those in biology. The use of facets can serve as an easy means for testing hypotheses that could previously have been probed only with extensive research. See Figure 1 showing the usage of the term ‘freedom’ with facets of ‘Genre’ in ‘government publication’ and ‘periodical’.

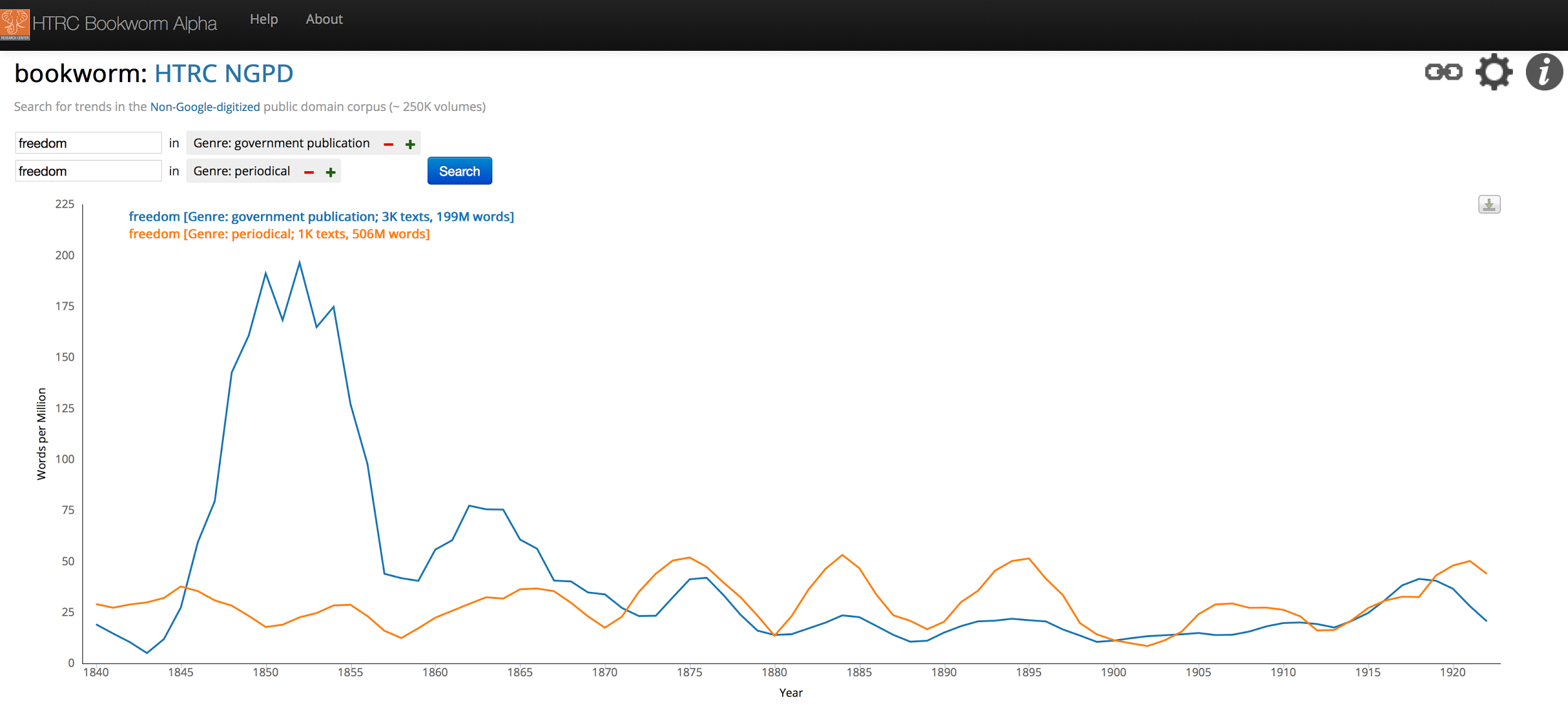


Figure 1. The HTRC Bookworm showing the usage of the term ‘freedom’ when faceting ‘Genre’ by ‘government publication’ and by ‘periodical’ for the Non-Google digitized Public Domain corpus from HathiTrust.

Textual data at a massive scale shifts the landscape of possibilities for the analysis of text corpora, allowing exploration to expand from the syntactic questions that linguistic corpora have excelled at answering, to capturing subtle cultural trends that underlie changes in the usage frequency of words or phrases. The goal of the HT+BW project is to create a tool that can help scholars realize this enormous potential.

**References**

**Downie, J. S., Plale, B., Kowalczyk, S., MacDonald, R. H., Poole, M. S. and Unsworth, J. M.** (2012). HathiTrust Research Center: Expanding the Frontiers of Large-Scale Text Analytics. *Proceeding of 2nd Annual Conference of the Japanese Association for Digital Humanities*, 15–17 September 2012, University of Tokyo.

**Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., . . . and Aiden, E. L.** (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, **331**(6014): 176–82.