The HathiTrust is one of the largest collections of digitized published work in the world. With nearly 14 million volumes (Currently Digitized: HathiTrust.org), it is a challenge to create mechanisms to understand and interpret a collection of this size.

The HathiTrust+Bookworm (HT+BW) project presents ways to behold that textual content through interactive visualization. Whereas HT+BW has previously been used in standalone contexts with pre-determined metadata, we present work in two new areas: allowing scholars to analyze custom personal collections from within the larger corpus, and 2) use of HT+BW as a supplement to other uses of the HathiTrust Research Center.

### Context
The sanguine of big text data was the realization that we had more books than a person can read in a lifetime (Cranes 2006). At the scale of current collections, such as that of the HathiTrust, it would take many years to read only the titles of all the digitized works. Such a scale presents opportunities for new forms of inference about historical, cultural, and linguistic trends. However, the collection may have biases and peculiarities which became apparent quickly when its contents are visualized. HT+BW seeks to address such problems.

HathiTrust+Bookworm (HT+BW) is a project seeking to adapt the generalized analytic tool Bookworm (https://bookworm.culturomics.org) to the large scales and unique needs in the HathiTrust Research Center. The HathiTrust is a consortium of institutions collecting millions of digitized works, and its research center seeks to support scholars in the large-scale insights that such a collection can provide. HT+BW is built on the HathiTrust Research Center’s Extracted Features Dataset (Capitanu et al. 2015), a publicly-released dataset of page-level counts for important features, such as text frequencies. The dataset contains nearly 8 terabytes of data for 4.8 million books, with plans to grow threefold in 2018.

The HathiTrust Bookworm: visualizing and comparing corpus-level trends by date. Screenshots from current public beta at https://bookworm.htrc.illinois.edu

Using book metadata, trends can be visualized in subsets of the corpus. This allows comparisons of language use in different countries or classes of books. Screenshots from current public beta at https://bookworm.htrc.illinois.edu

### References


### Worksets: Visualization over a personal research collection
Our current work extends faceting to custom groupings of the collection — personal subsets called worksets — in the HathiTrust Research Center (Jett 2015). Worksets can be visualized in Bookworm in a manner similar to visualizing pre-determined metadata-based facets. The affordances this allows include:

- generating descriptive statistics for a particular research domain, such as late 19th century best-sellers;
- creating two worksets, such as one of early-career and one of late-career authors, and comparing how thematic words occur in each case, or
- using HT+BW over a collection derived from other scholarship, such as Underwood’s work on identifying English-language literature by genre in the HT collection (2015).

Widget Access: HT+BW to support HathiTrust Research Center services
In addition to workset access, HT+BW is being used for ‘widget’ access to support other activities within the broader HathiTrust Research Center. One place where we are pursuing this is search. When searching for individual works, or building a workset, HT+BW can help contextualize the relevant results within the underlying contours of the dataset. For example, the widget-style visualization of a researcher’s query for ‘Stalingrad’ would reveal that the city name was only used for 26 years, showing that the given query only captures a narrow strand of the city’s history.

Advanced Faceting and Grouping
The API underlying HathiTrust+Bookworm keeps the advanced faceting seen in the main implementation, but does not restrict the numbers returned to date-based grouping. This opens the door to many more types of questions, such as ‘show me how many books match my query by geography’ or ‘what languages are represented most in my workset?’

The figures here illustrate the use of the API for understanding the corpus, i.e. the statistics are for the entirety of that part of the collection of the HathiTrust Digital Library.

The HT+BW includes books from all around the world in different languages. The materials held by HathiTrust are contributed to mainly from western institutions, meaning that English is the best-represented language in the collection, followed by other European languages. The best-represented Asian language is Japanese, with 73 thousand books, followed by Chinese with 12 thousand books. Bookworm supports extended Unicode characters, so Japanese is supported in the various uses of HT+BW. One limiting factor for scholars working with Japanese-language texts is that their metadata and coverage will not be as strong as for better-represented languages. For example, nearly no Japanese text in the current HT+BW have a subject class assigned.

### Figures

Adobe Illustrator and Adobe Photoshop

Publication location of volumes in a custom workset: in editions of Dickens’ work found in the HathiTrust. Visualized using workset created in HTRC Workset Builder and API call to forthcoming version of Bookworm.

**Figure 1.** Distribution of available pages for books matching a query. Some of these figures offer methods for data communication; more complex data graphics are more effective for exploratory uses, to aid hypothesis building by a scholar. Heatmap and map created using BookwormD3 library; other charts created from raw-count queries from the API.