Giving Shape to Large Digital Libraries through Exploratory Data Analysis

Peter Organisciak[1, 2]
University of Denver
Denver, CO, USA
peter.organisciak@du.edu

Benjamin M. Schmidt[2]
New York University
New York City, NY, USA
bs145@nyu.edu

J. Stephen Downie
University of Illinois at Urbana-Champaign
Champaign, IL, USA
jdownie@illinois.edu

Abstract

The emergence of large multi-institutional digital libraries has opened the door to aggregate-level examinations of the published word. Such large-scale analysis offers a new way to pursue traditional problems in the humanities and social sciences, using digital methods to ask routine questions of large corpora. However, inquiry into multiple centuries of books is constrained by the burdens of scale, where statistical inference is technically complex and limited by hurdles to access and flexibility. This work examines the role that exploratory data analysis and visualization tools may play in understanding large bibliographic datasets. We present one such tool, HathiTrust+Bookworm, which allows multi-faceted exploration of the multi-million work HathiTrust Digital Library, and center it in the broader space of scholarly tools for exploratory data analysis.

Keywords: Digital libraries, exploratory data analysis, visualization

Introduction

The rapid recent development of scanned text digital libraries provides the material for new scales of historic and humanistic inquiry into the published word. However, while the size of collections such as the HathiTrust Digital Library, Google Books, and Internet Archive allows for more comprehensive, aggregate-level insights of culture and language across eras (e.g., Michel et al 2011, Aiden and Michel 2014, McConnaughey et al. 2017, Manovich 2018; Evans and Wilkens 2018), the burdens of scale also limit the flexibility and approachability of those insights. This paper argues for flexible and easy-to-use exploratory tools to understand massive text collections, in order to support new forms of corpus-based scholarship. The insights that large text collections may offer about history and culture should not be limited to a limited band of deep questions that domain experts with the time and compute power choose to ask, nor should ordinary users have the choice only of interacting with texts through search engines; rather, there should be options to 'read' corpus trends quickly and flexibly. Through the case of the HathiTrust+Bookworm project, this paper presents one such approach.

Exploratory data analysis was first described by Tukey (1977), as a practice of using summative statistics and data visualization in a hypothesis-building workflow. It has been leveraged in big data scientific contexts for building an understanding of large datasets outside of a traditional hypothesis-testing approach (Ahrens et al 2000, Ahrens et al 2005, Fisher et al 2012). As library collections become more fully digitized in both full-text and metadata, we argue that exploratory data analysis be applied in information science to lessen the challenges that arise with scale. Particularly, it can lend clarity to the underlying biases and textures of a corpus, and it can help avoid the technical and computational burdens of pursuing large-scale corpus analysis by providing a fast, iterable way to initially assess research questions. Further, in disciplines concerned with corpus analysis, exploratory data analysis continues an established history of ludic, exploratory tools for text scholarship (e.g. Ramsay 2011, Rockwell and Sinclair 2016).

Our recent work with the HathiTrust+Bookworm (HT+BW) is a demonstration of exploratory data analysis in large digital libraries. Bookworm is a tool that visualizes language use trends at large scales, while HT+BW implements that functionality over one of the largest digital book collections: the HathiTrust Digital Library. In a manner that is easier and quicker to use than a traditional analysis workflow, Bookworm allows for multi-faceted exploration of a dataset against a series of content-based and metadata-based features, giving scholars the ability to explore language in the full HathiTrust corpus across features as subject classes, place of publication, genre, and language. Further, our work also provides tools for improved implementations of Bookworm over other text collections. Unlike online library catalogs, it presents works based on their full text and in the context of statistical aggregates; unlike prior art in corpus-

[1] Corresponding author
[2] Authors contributed equally

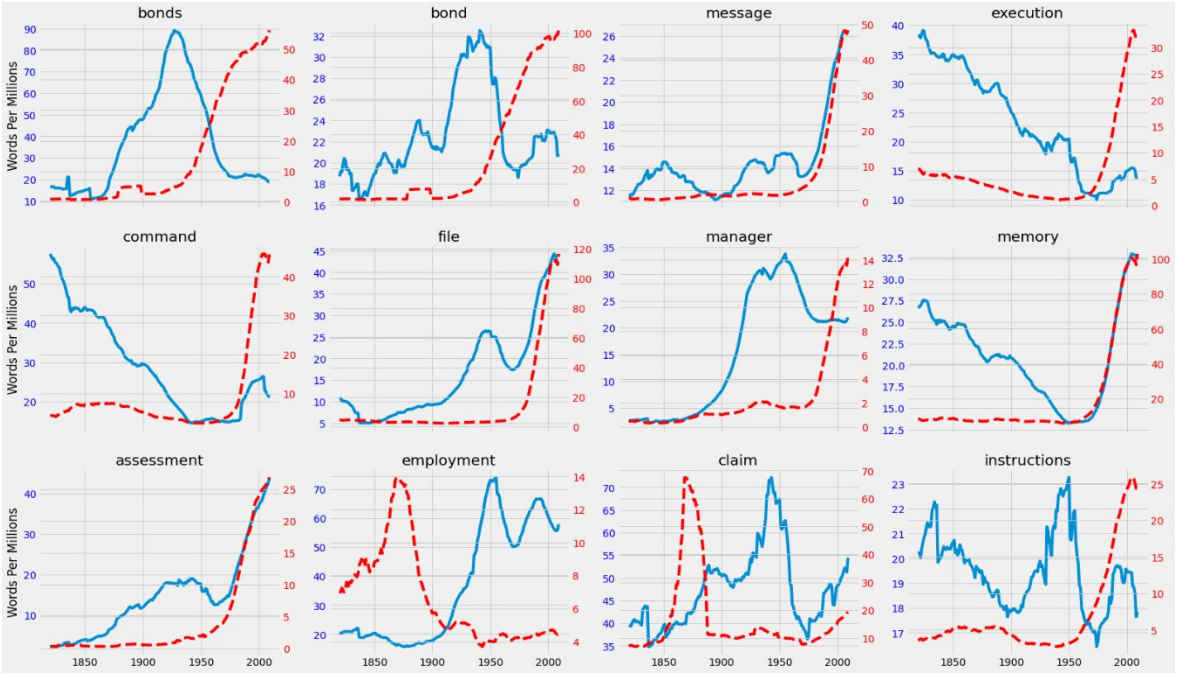


Figure 1: Demonstrative comparison of word trends between class facets, comparing words in general use (solid line) and the Library of Congress class for 'Science' (dashed line). Shown are words with the most divergent distributions, according to Jensen-Shannon Divergence.

based data visualization, it is fully able to work with gold-standard library metadata.

This work introduces Bookworm and the large digital library challenges that is addresses, reports on its application with the HathiTrust (HT+BW), and discusses the broader use of text analysis in exploratory and hypothesis-building settings. The strengths of HT+BW include a topically broad era- and domain-spanning corpus of books; flexible and faceted properties for more nuanced access, based on professionally-created library metadata; and quick and iterable turnaround on research queries. HT+BW is demonstrated in Figure 1 with an example comparison across facets and years, showing word trends in scientific texts compared to general use over the past two centuries.

Many large digital corpora consist of born-digital materials, limiting them to recent decades, or are in a specialized domain. In contrast, the HT+BW tool provides a stronger look into the published record of the past few centuries, particularly in English, spanning the breadth of the archive held in libraries across the world. The strengths of a statistical query tool and our particular implementation are also balanced by some limitations, which we will discuss, including difficulty to communicate corpus trends relative to underlying corpus biases, and challenges to valid inductive learning.

Background on Digital Libraries

The HathiTrust Digital Library is a digital repository of over 17 million volumes of digitized materials contributed by dozens of university and public libraries. The scans and optical character recognition (OCR) were created by the libraries themselves or by partners such as the Google Books project and Internet Archive. This combined corpus provides scholars with a wealth of opportunities for text analysis research, much of which is supported through the HathiTrust Research Center (HTRC). The Research Center supports a number of venues for scholarship over the HathiTrust Digital Library, including the public release of the Extracted Features dataset on which HT+BW is built (Jett et al., 2020). The HathiTrust corpus has seen especially strong scholarly adoption by those interested in cultural analytics and digital humanists (e.g., Underwood 2019, Manovich 2018; Evans and Wilkens 2018; McConnaughey et al. 2017), but the breadth of languages and subjects combined with the depth of the collection makes it very broadly useful.

There are unique obstacles of working with a corpus of such a scale. Making sense of billions of pages is difficult: both development and computation can be slow and resource-intensive. HT+BW provides solutions to speed up a scholar's early hypothesis-building exploration through exploratory data analysis (Tukey 1977), lowering the skill barrier to asking complex quantitative questions. It also now provides a method through which scholars can both visualize

their worksets and identify new items of interest for inclusion in scholarly worksets.

One promising aspect of exploratory data analysis over texts is abstracted-yet-useful access to otherwise inaccessible copyrighted works, as HT+BW offers with the two-thirds in-copyright HathiTrust corpus. In the US, most works created after 1925 are still under copyright and therefore much harder to acquire and analyze. Copyright law changes have extended terms significantly, meaning that new works entering the public domain was effectively stalled for decades until 2019, and the works that do enter are far removed from contemporary times. Beyond the US context, accessing public domain works is even more difficult, due to labyrinthian differences in copyright terms. At the HathiTrust, a single conservative worldwide public domain date (1881) is used for determining full-text availability outside of the US. Exploratory data analysis tools can avoid the issue by offering insightful abstracted or partial views which do not need full-text access.

The partner institutions contributing to the HathiTrust are predominantly libraries, and all works benefit from professionally-created metadata ingested in a standard format at the level of individual volumes. This richness of information about the collection is used for detailed faceting and filtering options in HT+BW. Whereas its predecessor, Google Ngram Viewer, can only narrow down searches by language and display results faceted by year, HT+BW can control for publication country and state, subject class, resource type, target audience, and others. This grants researchers the ability to ask more nuanced questions as well as controlling for and identifying biases within the dataset itself.

Related Work

Visualization and corpus analytics

Visual and statistical tools are increasingly used to support scholarship over large text collections, stimulated by the growth of text corpora and the growing role of digital libraries in text scholarship. Yet, the underlying approach of building tools to remix and explore texts has a long tradition in corpus-based scholarship in fields such as arts, languages, and history. These tools align with the approaches seen in both sub-areas of exploratory data analysis: graphical and non-graphical (i.e. statistical) tools for gaining an understanding of a corpus (Komorowski et al 2016).

An early form of exploratory text tool was a concordance program, which shows keywords of interest in context; for example, all instances of 'love' in a Bible concordance. Following from a long tradition of hand-developed concordances for studying texts, the mid-20th century saw early applications of computing to creating them, such as religious studies projects led by Ellison and Busa (Sprokel 1978, Hockey 2004). Subsequently, general purpose tools were developed to allow other scholars to inspect text corpora through concordances. *COCOA* was released in 1967 for Fortran (Corcoran 1974), and the machine-independent *Oxford Concordance Program* (OCP) was first developed in 1978 (Hockey 1987). Both tools provided frequencies of top words in addition to concordances, and by 1987, OCP had been licensed by 240 institutions (Hockey).

More recently, tools such as Voyant (Rockwell and Sinclair 2016) combine multiple text analysis modules with a suite of visualization interfaces to encourage exploratory use more directly. Those modules include statistical features, such as term frequencies, cooccurrence patterns, and information on vocabulary density and unique words. They also include modeling algorithms, such as correspondence analysis and topic modeling. Finally, they include visualization models, such as frequency trendlines, principal component analysis plots, and word clouds.

Voyant follows in the multi-tool tradition of TACT (Text Analysis Computing Tools, Lancashire et al 1996) and TAPoR (Text Analysis Portal for Research, Rockwell 2003; Rockwell et al 2010), which Rockwell describes as "a hermeneutical tradition that incorporates play in method" (2003). This paradigm has grown and the TAPoR directory of text analysis tools lists 1,587 tools. Examples include MALLET, a machine learning tool with a robust technique for learning conceptual topics (topic modeling) from a corpus of texts (McCallum 2002), and Rich Prospect Browsing, a visual language for browsing and faceting cultural heritage (Ruecker, Radzikowski, Sinclair 2016).

The tradition of exploratory tools in fields that rely on corpus analysis has remained strong. However, it is worth noting that most such tools do not scale to millions of texts, limiting the methodological access that they provide to smaller corpora. This reality has led to newer tools that rely on the statistical paradigm of exploratory data analysis, including HT+BW.

Exploratory Data Analysis for Large Bibliographic Collections

Though often overlooked for this use, by far the most important tool for examining large corpora is the search engine. As scholars in humanities disciplines have recently acknowledged, large-scale textual search has a transformative effect on research in non-quantitative fields. But this is a double-edged sword. Putnam (2016) describes the ways that search has facilitated a particular type of historical research that pinpoints individuals in time while failing to take into account the larger social context; Underwood (2014) criticizes the use of search engines by

literary scholars for enabling confirmation bias and for a tendency to "filter out all the alternative theses" that a scholar brings. The easy-to-use aggregate access that text search offers should be acknowledged, but it is clear that more sophisticated tools are needed in that space.

In their 2011 article, *Quantitative Analysis of Culture Using Millions of Digitized Books*, Michel et al. used large scale textual analysis techniques to investigate cultural trends. From a corpus of five million books, they focused on "…linguistic and cultural phenomena that were reflected in the English Language between 1800 and 2000". This style of emergent, corpus-based study was called *culturomics*, and was accompanied by the release of the Google Ngram Viewer. Both projects attempt to help answer questions in the social sciences and the humanities by tracking word usage across time, location, and language. Bookworm also has a direct lineage following from these projects. The time-series chart has remained a fixture for exploring word trends, such as the tool *How the Internet Talks*, which tracks phrases and words used on Reddit over time (King and Olson 2015), and the search habits tool *Google Trends* (Google).

Outside of bibliographic domains, concern over the effect of scale in stifling flexibility and exploration have been present in scientific visualization for years. Fisher et al. argue that "we have reverted to a batch-job era... a step backward from the interactive querying that we expect in exploratory data analysis" (2012). One solution explored has been parallelizing visualization processing, as is commonly done in data processing (Ahrens et al. 2000). Others have shown the effectiveness of subsampling strategies in representing the whole, such as Fisher et al. (2012) with gradually expanding slices. Balancing fast and accurate visualization, Ahrens, Geveci, and Law have proposed tools with 'exploratory' and batch processing modes, allowing for different fits within a scientific workflow (2005). The HathiTrust Digital Library exists at a smaller scale than larger datasets in the sciences, which may grow to petabyte scale. Still, the challenge of scale persists, particularly due to the nature of text, which is semi-structured and deals with a wide feature set, and the public user community that HT+BW serves. HT+BW is able to meet these needs without parallelization, but rather through preprocessing and optimizing the data prior to visualization by reducing to discrete problems.

Another approach for enabling easier access to large corpora avoids full-on statistical or visual tools and instead offers derivative datasets of pre-processed statistical information about texts. Two prominent derivative bibliographic datasets are the *Extracted Features Dataset*, on which HT+BW is built (Organisciak et al 2017, Jett et al 2020), and *JSTOR Data for Research*, a service for downloading build metadata and word counts for publications in the JSTOR repository (JSTOR, n.d).

Corpus-scale word visualizations

Beyond the time-series precedent of the Google Ngram Viewer, the current suite of functionality in HT+BW follows from a set of traditions in visualization and exploratory data analysis designed to support a carefully limited subset of the ambitions of some of the corpus analytics tools described above. Its design follows the Visual Information-Seeking Mantra: "overview first, zoom and filter, then details on demand" (Shneiderman 1996). Its facetable API and primary visualization interface allows both an overview of the full collection to be seen, as well as filtered subcollection views, and the ability to retrieve underlying volumes for a statistic is intended to satisfy *detail on demand*. It also satisfies Shneiderman's supplements to the mantra, particularly in its ability to compare and contrast different facets of the data.

Querying data in HT+BW, as well as visualizing data in its 'Advanced' interface is performed through a declarative grammar, a style of computer interaction where a user describes what they want rather than how it should be computed. Many declarative languages are common, such as HTML, but the most fitting precedent is declarative grammars for visualization. Popularized by Wilkinson's *The Grammar of Graphics* (1999), they provide a data-oriented rather than system-oriented way of interaction, binding visual elements directly to data (Wickham 2010, Heer and Bostock 2010). Low-level grammars such as *D3* and *Vega* allow a great deal of flexibility over the output, while higher-level grammars (e.g. *ggplot2, Vega-lite, Plot.ly*) use sensible defaults to reduce the amount of input needed (Satyanarayan et al. 2014, Satyanarayan et al. 2016). In contrast, non-grammatical approaches include templating tools, such as the visualization tools in Excel. These differing approaches balance expressiveness, efficiency, and accessibility differently (Heer and Bostock 2010). Bookworm approaches these trade-offs through multiple tools: graphical interfaces with pre-selected options for guided useful visualization, with a declarative visualization grammar and programmatic tools and as options for increasing expressiveness at a cost to accessibility.

Overview of Bookworm

Bookworm grew from the Google Ngram Viewer (Michel et al 2011). The first version of Bookworm expanded on this tool, which focused on time-series trend lines, to allow other corpora to be used (Aiden and Michel 2014). Since then, Bookworm has been generalized to a more flexible quantitative querying tool (Schmidt 2011, Gorges 2011), and scaled up and iterated with HT+BW. This work reports on the modern tool, Bookworm, and an exemplar implementation, HT+BW, in the context of the broader paradigm of fast corpus exploration methods.

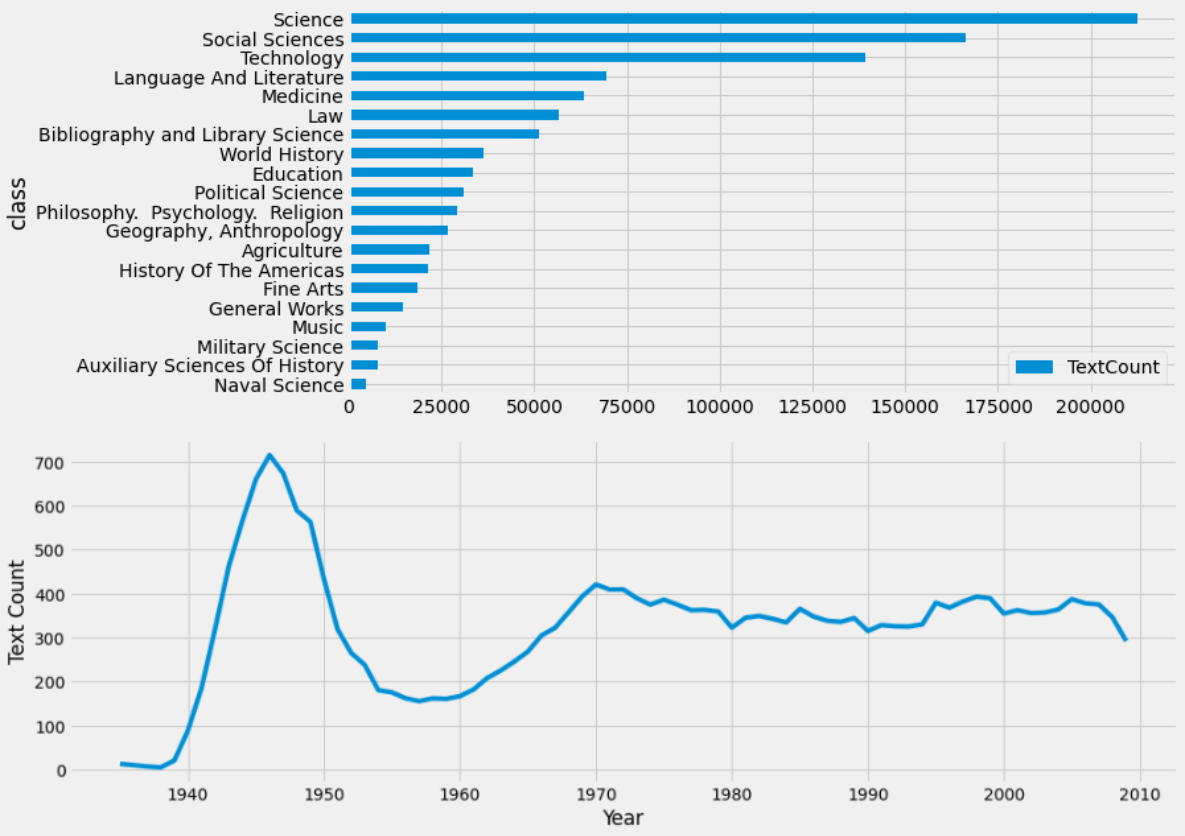


Figure 2: Two queries with HT+BW. Top: Class distribution of books mentioning 'computer', rendered as bar graph. Bottom: Metadata query showing World War II books in HathiTrust by year, rendered as line chart. This chart tracks total books in the facet, not a particular keyword.

Data Model

There are two components to how a document is represented: the metadata — information about the volumes in the collection — and the data — information about the actual words in the documents. Fundamentally, a build of Bookworm is a powerful data query engine, one that allows you to ask quantitative questions about the books in the collection conditioned across various metadata facets. One can write a data-based query, such as: "What is the class distribution of books mentioning computers", or a metadata query, such as: "How many books belong to the Library of Congress subclass 'World War II', by year?" (Figure 2). The underlying engine is accessed through an API (application programming interface) accessible online through the standard HTTP protocols; with it, anyone can structure a query in their browser or a web-connected tool.

A number of web-based tools are available to more easily use the API for statistics or visualization. The most popular view is a time-series line chart, which can show the frequency with which a word appears in texts over time. Still, other available interfaces include ways to bind queries to visualizations such as maps, bar charts, and heatmaps, and more advanced tools streamline programmatic querying through Python, R, or JavaScript. All of these tools connect to the same API and require no privileged access by the HT+BW project. This has two effects on general reuse and access:

1) Scholars may craft their own queries or build their own tools against the entire HT+BW instance. The existing interface tools necessarily make decisions for the user, but scholars with their own unique questions can nevertheless ask them.

2) The tools implemented for HT+BW can be reused for custom, non-HathiTrust implementations of Bookworm. Beyond its application with the HathiTrust collection in HT+BW, the Bookworm toolset is openly available to apply over other large collections. Scholars with even modest technical ability may build their own Bookworms.

A quantitative API allows a great degree of flexibility in rapidly exploring a corpus, while supporting a spectrum of user expertise. Curious casual users or scholars in early exploratory stages can use some of the out of the box visualization interfaces, intermediate users can use a declarative grammar for retrieving custom visualization views, and the most advanced users can pull raw statistics for a large list of words directly through the query language. At the same time, the novelty of fast corpus exploration introduces challenges to use, such as contextualizing HT+Bookworm corpus trends within its biases, considering

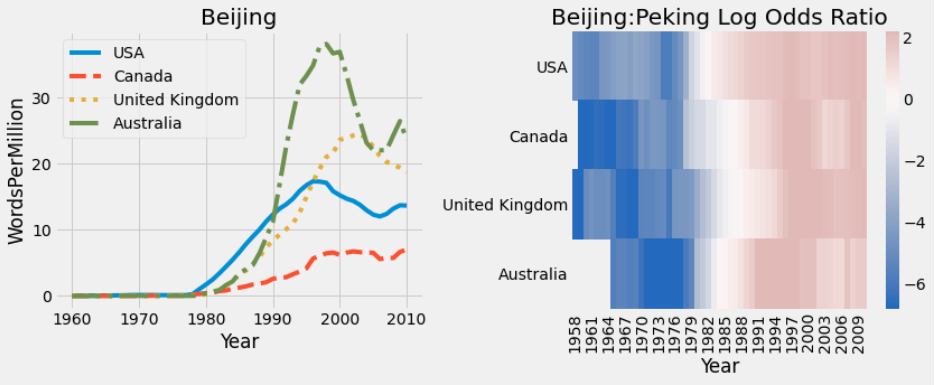


Figure 3: Use of the word 'Beijing' in English-language countries (left) and a log-odds ratio comparing its use to the pre-1949 official transliteration, 'Peking'.

its place in a large corpus analytic pipeline, and onboarding novice users and leading them toward more advanced uses.

Access and Use

To ground discussion of exploratory data analysis over text corpora, we now consider the case study of how HT+BW implements it. HT+BW supports a number of approaches for developing queries and interacting with the data, including the Bookworm GUI, Bookworm Playground, and the Advanced interface, as well as a library for calling the data programmatically. Since different scholarly uses will differently balance expressiveness and concision, these provide different ways to interact with the underlying quantitative API.

Bookworm GUI

The Bookworm GUI allows plotting multiple word trend lines by year. The trends can look across texts, or be specified as subfacets of the collection. Across the entire collection, the only sensible search comparisons are between different words: e.g. how 'telephone' and 'typewriter' have ebbed and flowed in our published works. Through subfacets, however, it is possible to consider the *same* words in different *contexts*: for example, to compare how quickly 'Beijing' was adopted in US books versus British or Canadian books after the official Pinyin transliteration was adopted in China in 1958 (Figure 3, left). By extension, it is also possible to map word distributions in highly particular subfacets, when you have a question about a specific domain or subset of the collection (e.g. the rise and fall of 'documentation' in library science, Figure 4).[1]

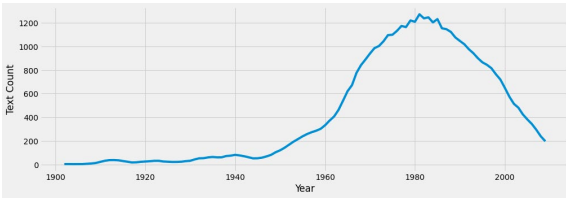


Figure 4: Bookworm GUI query, looking at the rise and fall of 'Documentation' specifically in 184k books classified as Bibliography, Library Science, and Information Resources.

The Bookworm GUI always plots information across time. The default metric is a words per million ratio, though it can be changed to the percentage of all texts, count of all matching texts, or total count of occurrences. The relative measures (words per million and text percentage) provide a more valid comparison between differently sized subsets of the collection.

The facets that can be used for selecting subsets of the data include language, publication country, state, Library of Congress class/subclass/most narrow class, resource type, author name, place, and publisher, among others. While a portion of these library-specific facets will be opaque for some users, they present an opportunity for practiced users to carefully create a corpus of their own.

Bookworm Playground

The Bookworm Playground is intended to fill a need between the GUI and the advanced and programmatic interfaces. There is a value to the "quick to learn, quick to use" interface of the time series visualization that makes it more popular than programmatic or declarative access. However, the latter is much more powerful as a tool for cultural and critical inquiry. The Playground offers user

[1] Figures are redrawn for print using the same queries as in the web interfaces.

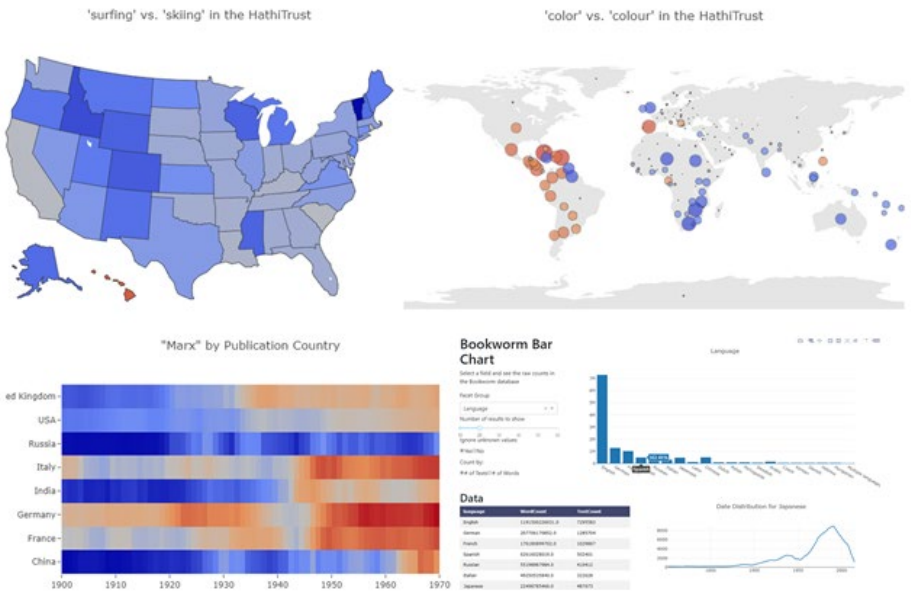


Figure 5: Views in Bookworm Playground. Map (top), Heatmap (bottom left), Metadata Explorer (bottom right).

interfaces oriented towards more types of visualizations than line charts. While an out-of-the-box interface requires a certain amount of decision-making on behalf of the user, this series of smaller tools offers a more extensive sampler of what can be accomplished with Bookworm. The 'playground' branding also communicates it as a space for rapid deployment, allowing the HT+BW team to publish potentially useful, easy-to-use but less polished tool, while signaling that trade-off to the user. Currently, the Playground includes examples of maps, heat maps, and bar chart interfaces.

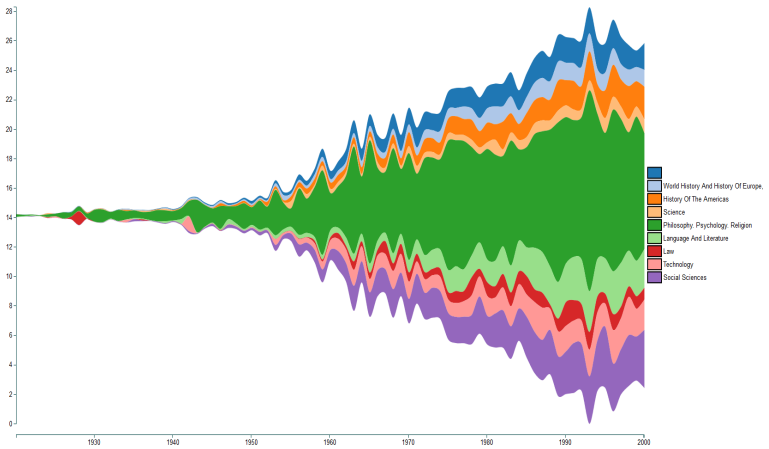
The map visualization allows plotting of word trends by publication country or state, and optionally allows two words to be compared. Selecting a location returns a list of books that contribute to the statistic.

The heatmap visualization plots a search across three dimensions: year (y-axis), words per million (color), and user-selected values from any of the Bookworm facets (x-axis). It is a more elegant alternative to the time-series line charts for instances where there would be too many lines to compare.

Finally, the bar chart page of the playground offers a dashboard of metadata information. Rather than searching for a word, it provides a glimpse into the distributions of books by a particular facet. In the example shown in Figure 5, we see the number of texts per language, the corresponding data in table form, and the date distribution for a selected language.

Advanced Bookworm Interface

Rounding out visualization tools, the HT+BW project hosts an advanced interface, which uses a declarative visualization grammar to draw various types of data graphics (e.g. (Satyanarayan et al. 2014)). Consider the streamgraph in Figure 6, which uses stacked area to demonstrate usage of a word ('creativity') over time, and uses color to show which



```
{
    "database": "Bookworm",
     "search_limits": {
        "word": ["creativity"],
         "date_year": {
             "$gte": 1920,
             "$lte": 2000
         }
      },
     "plotType": "streamgraph",
      "aesthetic": {
         "x": "date_year",
         "fill": "class",
         "y": "WordsPerMillion"
      }
}
```

Figure 6: Bookworm Advanced declarative query, with screenshot of resulting figure. Note the interpretability of the parameters in the query: graph type, faceting groups, word and date query, and data to map to visual elements for x, y, and color fill.

Library of Congress classes that usage falls in. On the right is the query that generates this view.

The declarative query described both the data criteria and visual layout. For data, we see it ask for words per million statistics grouped by year and class, counting the word 'creativity' in texts between 1920 and 2000, where '$gte' and '$lte' describe "greater/less than or equal to". Visually, this data is plotted to a streamgraph with the x-axis as year, y-axis as count, and fill-color faceted by class.

Libraries for Programmatic Access

To aid more advanced use of Bookworm, the HT+BW project developed Python, Javascript, and R libraries for programmatic access to any Bookworm index. They provide scaffolding around the API to assist in using it, validating for errors, and handling the output. This allows Bookworm queries to connect to scholarly statistical workflows. For example, the Python implementation allows export to a Pandas DataFrame, a common data representation format that is used in the SciPy ecosystem of data science tools ('Scientific computing tools for Python' 2021).

Figure 3 demonstrates a case where scholars may prefer to work with the raw data in a statistical language or library. The BookwormGUI shows the growth of the use of the word *Beijing*, and provides evidence that the spelling did not pick up traction after the 1958 country-wide adopting of the Pinyin system for transliteration until 1979, when the system - associated with communist rule - started seeing international acceptance (Wiedenhof 2005). For a fuller look at the change, though, a comparison between *Beijing* and the previous transliteration of the word, *Peking*. Bookworm Advanced does allow for comparative ratios between two searches, but scholars seeking other strategies for quantifying the difference between terms would need to calculate it from the raw data. Figure 3 (right) shows log-odds-ratio (Monroe, Colaresi, and Quinn 2008), which takes the odds of each word occurring in a year (i.e. $O_w = \frac{P(w)}{1-P(w)}$), and takes the log of the ratio between the odds ($LOR_{w1,w2} = log\ \frac{O_{w1}}{O_{w2}}$). We see that the majority use of *Beijing* was led by the USA, with the United Kingdom lagging, and that the relative use of the word is still lower than the inverse was before 1979, as *Peking* continues to persist in certain contexts, such as the name of *Peking University*.

This approach to exploratory data analysis makes the underlying data openly available, providing a backup option when the visual tools are insufficient for a particular use. Indeed, even for this paper, the programmatic libraries were used to redraw Bookworm figures in a style more appropriate for print, where we used the same API queries as the web interfaces, but presented the output in a differently-formatted way.

Links to the described views of HT+BW are provided in the resources section.

Working with a Quantitative API

The core of HT+BW, and Bookworm in general, is an engine that allows queries about word and document counts to be crafted for a given corpus of text, faceted or grouped by common features of the documents. Given the typically large quantities of data usually involved, visualizing those counts is the most sensible action to take with that response, but certainly not the only one.

For HT+BW, the API for the engine is publicly accessible, so other scholars can build applications or tools with their own questions.

A quantitative query is comprised of a few parts:

- Facets: Facets are used to disaggregate results by a metadata property. For example, while it is possible to get a single numeric count of all books mentioning a word in the HTDL, it may be more valuable to retrieve that information per year, subject class, subclass, publication country or state, type of resource, or even author. Facets can stack, so you can ask for counts of every permutation of a set of facet groups, as demonstrated in Figure 7.
- Metadata filters: Whereas its predecessors only allowed filtering to a specified span of years, HT+BW enabled more flexibility. Rather than seeing trends across the full collection, one may choose to focus on a slice of the corpus (e.g. only books from Canada).
- Word filters: Optionally, a word can be searched for, from the *content* of books. Without a word filter, as shown with a visualization of average page count in Figure 7, queries tend to be about bibliographic structure or corpus properties.
- Count statistics: Bookworm offers multiple ways of returning its results, such as 'text count' (how many books match each facet of the filtered search), 'words per million' (how often does the filtered word occur out of every million), and 'text percent' (what percent of books include the word).

Table 1 shows how research questions convert to Bookworm queries, and how that data may be displayed.

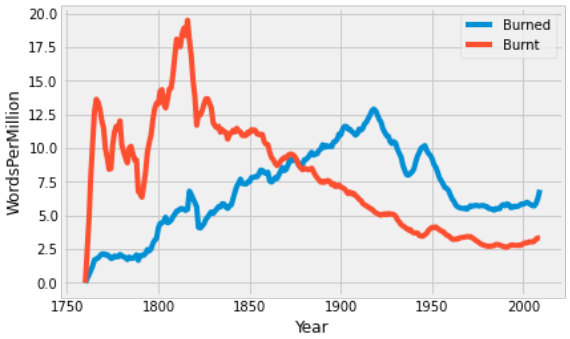
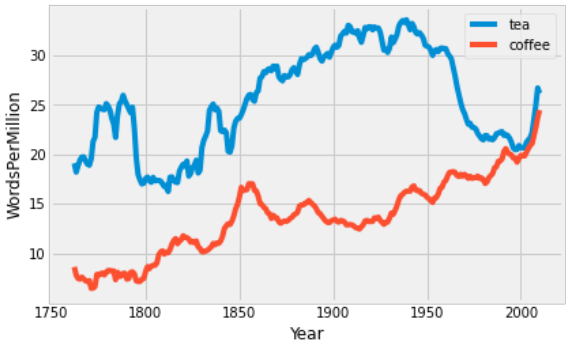
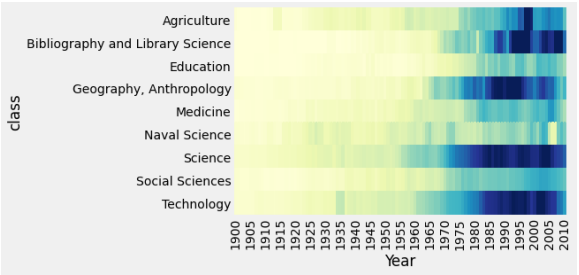
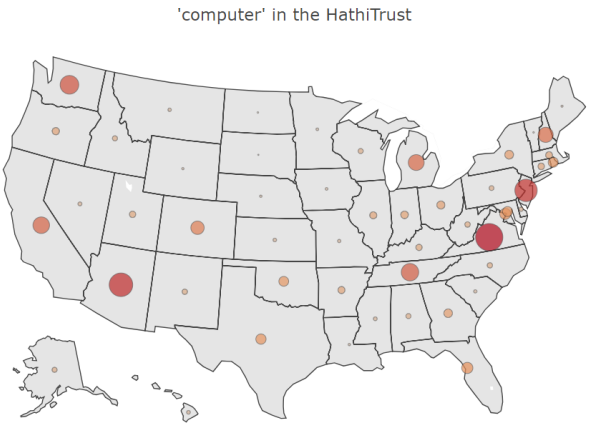
Question	Query	Visualization Example
How do English spelling conventions change over time? *Comparing words across time*	Words: burned, burnt Facet on: Date Limit: Date(1760-2010)	
What are cultural preferences within societies? *Comparing words across time for a specific country*	Words: tea, coffee Group by: date Limit by: publication country (UK), date (1760-2010)	
How do vocabularies differ over academic fields?	Words: data Group by: class, year Limit by: date (1900-2010) Bookworm Advanced parameters: y representing: class x representing: date year color representing: word uses per million	
Where are books on computing more likely to be published? *Tracking words by geography*	Words: computer Group by: publication state Limit by: publication country (USA) Bookworm Playground parameters: Map Scope: USA Map Type: Scatter	

Table 1: Examples of Questions to Queries to Visual Output

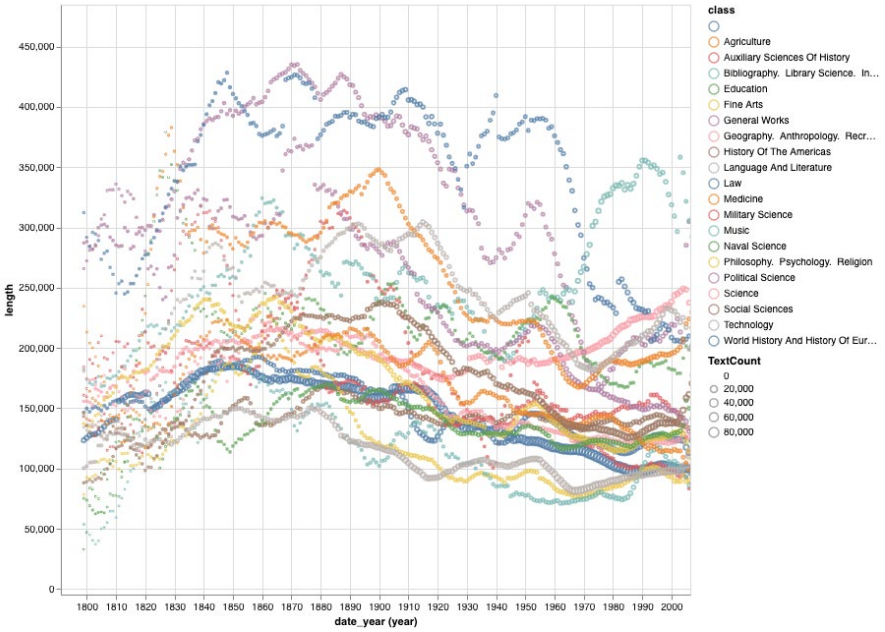


Figure 7: HT+BW used to simultaneously present multiple interacting facets. Average page length of books over time by LC class, with marker sized by the number of volumes represented by each point and smoothed over a 10 year window. Books without a class included.

Discussion

We explore a body of solutions to a challenge to scholarship - the complexity and inflexibility of working with large and extensive text datasets. Our case study, HT+BW, continues the paradigmatic tradition of exploratory data analysis (Tukey 1977), encouraging flexible and fast exploration of large datasets. Its outputs are provided in the form of common graphical and non-graphical exploratory data analysis (Komorowski et al. 2016), such as heatmaps, line plots, and summary statistics.

Despite its classic roots, the use of HT+BW can be novel or even unfamiliar. Learning to use HT+BW and similar tools within a scholarly workflow presents its own set of interpretive challenges. For many of these issues, the burden lies with the user, but given the goals of HT+BW in being easy-to-use and flexible, it bears considering how the tool and visualizations may guide scholars in their proper use. In other words, the facade of simplicity in such tools may cause a user to overlook the considerations of bias and inference necessary for serious application of the tool.

Challenges of Inference on Large Libraries

The paradigm of exploratory data analysis challenges traditional inference. It has been argued that exploration may indirectly encourage 'fishing' for results that contradicts multi-hypothesis expectations (Gelman et al 2013). For example, if a user of an exploratory data analysis tool views a large number of trends before settling on one that seems interesting, the result should be viewed in the context of the results that came before it. However, it has been argued that when used and reported properly, exploratory data analysis is distinct from improper fishing or *p*-hacking activities (Jebb et al 2016). The difference is in the distinction between hypothesis-forming and hypothesis-confirming – whereas p-hacking seeks to find data to support a hypothesis, exploratory tools instigate learning about the dataset and discovering trends or patterns worthy of further study. It is certainly not possible to enforce such proper use, but in working with scholars at HT+BW trainings and workshops, we note that users generally embrace the exploratory mode.

As Milo and Somech describe, "the purpose of [exploratory data analysis] is to better understand the nature of data and to find clues about its properties and quality" (2021). In adhering to those principles, this project approaches HT+BW as a formative research tool, an initial step in a longer theory building processing. That approach is likely not the only one; exploratory data analysis tools can be leveraged for inferential protocols that have been developed for visualization (e.g. Wickham et al 2010). The core of HT+BW as a quantitative API and its extensive faceting can be used to properly contextualize and compare results in a fully contained analysis. As tools like HT+BW develop, it will be imperative to think how those skills can be trained and communicated without greatly undermining their primary goals of enabling easy and flexible access to large text corpora.

Lincoln Mullen poses a challenge when presenting trends learned in digital history. In response to a common argument

that linguistic trends simply communicate what is already known, they ask their audience whether they can predict a trend line *a priori* (Mullen 2018). Such predictions are difficult to get correct, yet while the exercise may protect against charges of self-evident results, the fact that they *seem* self-evident raises questions about honestly communicating to a user that is constantly seeking meaning-making and narrative.

If a trend is unexpected, we may seek to make it expected. These types of questions can be productive, if they open up a new thread of study, or they may misconstrue noise for signal. Hartwig and Dearling argue that the key principles of exploratory data analysis should be *skepticism* and *openness:* skepticism of what may be concealed by summative measures and openness to unexpected results (1979). Keeping to these principles is difficult because what is shown is much more present than what may be missing. Yet, Ramsay argues that the tradition of close reading in the humanities is a similarly mediated activity, noting that "the critic who endeavors to put forth a 'reading' puts forth not the text, but a new text in which the data has been paraphrased, elaborated, selected, truncated, and transduced" (2011).

Selection Biases and Omissions

As digital library collections and other primary source digital humanities corpora grow, the most pressing challenge for exploratory data analysis tools like Bookworm is in communicating the underlying structure of the data.

HT+BW benefits from a digitization project with fairly low selection bias. Where early scanning projects were selective in choosing which books to scan, the digitization projects underlying much of the HathiTrust's data was minimally discriminatory, making the texts more representative of the libraries where they originate. Still, those libraries are not unbiased themselves, as their collections are a result of language, audience, and mission. Where a reader may be critical of a conclusion about history draw from a handful of books, the sheer scale and data-driven approach with HT+BW lend a sheen of objective verifiability that may undercut that critical eye. Omissions and over-representations are still necessary considerations.

The libraries contributing to the HathiTrust are predominantly in English-speaking countries. Subsequently, while nearly 200 languages are represented in the HathiTrust, about half of the texts are English-language. Contributors are also predominantly academic institutions, which prioritize some types of texts over others. In literature, for example, prestige works are strongly represented, yet that may be to the detriment of culturally notable popular works. Further, in the face of a gradual split between popularity and prestige in the 20th century (Underwood and Sellers 2016), while countless copies of popular classical literature may be found in the HathiTrust, it may be challenging to find copies of contemporary popular works. It has been noted that large amounts of repeating texts can mislead text mining models (Schofield, Thompson, et al 2017). A related concern is books being republished, which may represent their linguistic trends in a newer year than when the work was first written.

Recent work has sought to better identify duplication in the HathiTrust (Organisciak et al 2019). Additionally, it has been found that date of first publication can often be inferred from the earliest known duplicate in the HathiTrust (Bamman et al. 2017). This work will inform future iterations of HT+BW, reducing duplication bias and better aligning texts and dates.

The most challenging biases in the collection are not ones of overrepresentation, but underrepresentation. Singh refers to the disparity of what we preserve in digital archives as the *archive gap* (2015, 2019), the tendency of preservation attention toward western white authors. This challenge extends past digital archives toward many institutional collection development policies (Sadler and Bourg 2015, Quinn 2012). The published record is not fully representative of historical and cultural trends, because it is biased on who is provided the opportunity to publish. Among works that are published, there are inequities in what cultural institutions collect. This selection bias is a broader challenge in library and information science. It needs to both be contextualized and, as Sadler and Bourg argue, actively challenged (2015).

The HathiTrust Research Center is currently leading a major multi-institutional effort, funded by the Mellon Foundation, to address some of these biases. This scholar-led project focuses specifically on historically under-resourced and marginalized textual communities (Dickson-Koehl et al. 2021). The HathiTrust, in the depth afforded by its comprehensive scale and multi-institutional provenance, is a reasonable site to address systematic collection marginalization.

In some cases, tools like HT+BW may also assist in contextualizing the bias of their underlying collection, showing gaps and omissions which cultures, people, and topics are discussed in the collection. Michel et al. (2011) demonstrate this with censorship during wartime Germany, showing a distinct, multi-year erasure in mentions of censored authored in German texts.

Wickham et al. (2010) explore the tendency toward graphical narrative construction in the context of inferential learning: can visualization be used to infer a truth, or does a viewer presume a truth? One strategy proposed reverses the Mullen predictive approach, calibrating our meaning-making tendencies by asking viewers about the story told by graphics of random noise data; another asks viewers to choose a real graphic from a lineup alongside narrative imposters. Bookworm does not aim for statistical inference,

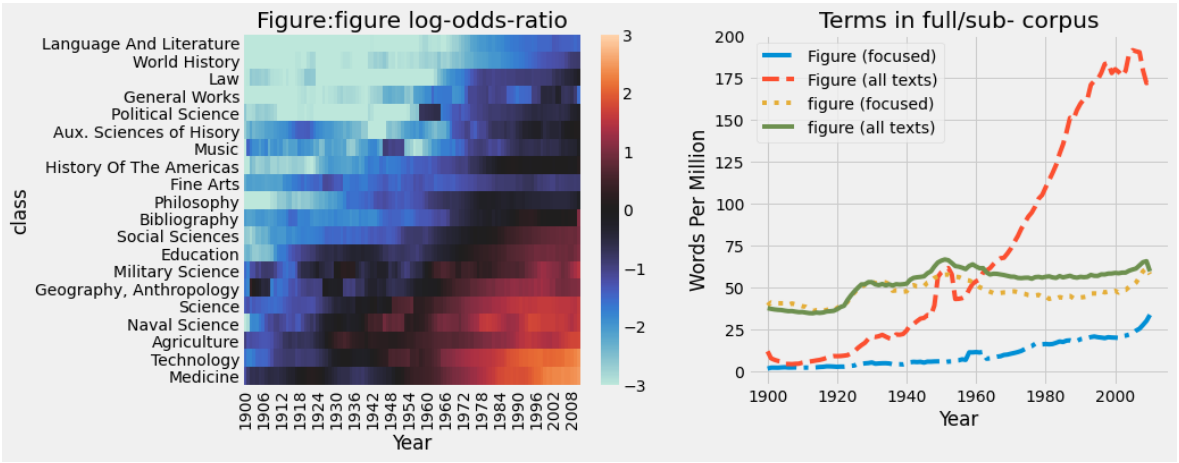


Figure 8: Log odds ratio of 'Figure' versus 'figure' in classes, showing the academic text bias predominantly in specific classes (left). Trends of the words shown using a Bookworm query that excludes the most 'Figure'-biased classes, those listed from 'Social Sciences' to 'Medicine' in the vertical axis (right).

but a similar multi-view approach may be adopted to better contextualize the underlying contours and strengths of the collection. For example, one elegant example criticizing the underlying bias of the Google NGrams Viewer (and in turn, HT+BW) focuses on the capitalized word 'Figure' - common in academic publishing and more prominent in the collection than 'figure' (Pechenick, Danforth, and Dodds 2015). Demonstrated in Figure 8, that bias toward academic publishing is partially temporal, growing through the twentieth century, but much more predominantly localized to specific classes, such as Medicine, Science, and Technology. With HT+BW, a scholar seeking to minimize that bias may filter specifically to less academic-biased classes, also shown in Figure 8. Another type of collection bias that may be uncovered from within Bookworm is temporal errors due to OCR issues. A search for common words such as *the*, *and*, or *as*, shows that they have been relatively stable since about 1800, which coincides with an improvement in printing press technology as well as standardization in typefaces. This suggests that the per-year distributions of scanning errors are only notable prior to the 19th century, after which their effect on the corpus is fairly unchanging.

Conclusion

Digital humanists and other corpus scholars can benefit greatly from computational access to massive, era-spanning digital libraries. However, the scale of such collections presents an obstacle to their widespread use, counteracting inferential affordances with technical hurdles. In this paper, we argue that the paradigm of exploratory data analysis used in other data-intensive domains may be valuably applied to digital libraries. We apply it to HathiTrust+Bookworm, a declarative data retrieval system paired with an array of higher-level visualization tools. On an immediate level, HT+BW presents scholars, students, and the public with enhanced analytic access to the HathiTrust Digital Library collection, an unprecedented aggregation of digitized print materials in hundreds of different languages. More broadly, information science and the digital humanities are increasingly grappling with issues of scale, and we continue to work with ever-growing collections, Bookworm presents both a model and a set of tools for using them effectively.

Resources

- The Bookworm GUI is available at http://bookworm.htrc.illinois.edu.
- The Bookworm Playground is available at https://bookworm.htrc.illinois.edu/app.
- The advanced interface is an instance of the Bookworm D3 library (Schmidt 2015). It is available at https://bookworm.htrc.illinois.edu/advanced and its declarative grammar is documented by Schmidt (2015).
- BookwormPython is available at https://github.com/organisciak/BookwormPython.
- Bookworm R Library is available at https://github.com/bmschmidt/edinburgh.

Acknowledgements

The project was made possible by the National Endowment for the Humanities, award number HK-50176-14 (PI: J. Stephen Downie). Any views, findings, conclusions, or recommendations expressed in this article do not necessarily represent those of the National Endowment for the Humanities. Bookworm was initially developed at the Harvard Cultural Observatory under the direction of Erez Lieberman Aiden and Jean-Baptiste Michel. Notable contributions to Bookworm have been made by Benjamin Schmidt, Martin Camacho, Billy Janitsch, Neva Cherniavsky, Erez Aiden, Matt Nicklay, JB Michel, Peter Organisciak, and

Colleen Fallaw. Further funding and institutional support has been provided by the Harvard Cultural Observatory, the Digital Public Library of America, the HathiTrust Research Center, University of Illinois, Northeastern University, and Rice University. Additional thanks to Loretta Auvil for contributions to the project, Andy Lawder, Adrienne VandenBosch and Danielle Francisco Vasquez Albuquerque for assistance in preparing this manuscript, and Danielle Albers Szafir and Glen Worthey for notes and advice.

References

Ahrens, J., Geveci, B., & Law, C. (2005). Paraview: An end-user tool for large data visualization. *The Visualization Handbook*, *717*.

Ahrens, J., Law, C., Schroeder, W., Martin, K., & Papka, M. (2000). A parallel approach for efficiently visualizing extremely large, time-varying datasets. *Los Alamos National Laboratory, Tech. Rep.# LAUR-00-1620*.

Aiden, E., & Michel, J.-B. (2014). Uncharted: Big Data as a Lens on Human Culture. Penguin.

Bamman, D., Carney, M., Gillick, J., Hennesy, C., & Sridhar, V. (2017). Estimating the Date of First Publication in a Large-Scale Digital Library. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 1–10. https://doi.org/10.1109/JCDL.2017.7991569

Corcoran, P. E. (1974).COCOA: A FORTRAN program for concordance and word-count processing of natural language texts. *Behavior Research Methods & Instrumentation*, 6(6), 566–566. https://doi.org/10/b2wnv8

Dickson-Koehl, E., Downie, J.S., Dubnicek, R., Graham, M., Harrison, J., Walsh, J., & Worthey, G. (2021). Recovering Spectral Presences in the "Universal" Digital Library. 2021 Global Digital Humanities Symposium, Michigan State University.

Evans, E., & Wilkens, M. (2018). Nation, Ethnicity, and the Geography of British Fiction, 1880-1940.

Fisher, D., Popov, I., Drucker, S., & Schraefel, M. (2012). *Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster*. https://www.microsoft.com/en-us/research/publication/trust-me-im-partially-right-incremental-visualization-lets-analysts-explore-large-datasets-faster/

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, *348*.

Google Trends. (n.d.). Google Trends. Retrieved January 2, 2020, from https://trends.google.com/trends/?geo=US

Gorges, B. B. (2011). Review of Bookworm, produced by Benjamin Schmidt, Martin Camacho, et al. *Journal of Digital Humanities*, *1*(1). http://journalofdigitalhumanities.org/1-1/bookworm/

Hartwig, F., & Dearling, B. (1979). *Exploratory Data Analysis*. SAGE Publications Inc. https://doi.org/10.4135/9781412984232

Hearst, M. (2009) *Search User Interfaces*. Cambridge University Press. Retrieved May 28, 2021 from https://searchuserinterfaces.com/book.

Heer, J., & Bostock, M. (2010). Declarative Language Design for Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 1149–1156. https://doi.org/10.1109/TVCG.2010.144

Hockey, S., & Martin, J. (1987). The Oxford Concordance Program Version 2. Literary and Linguistic Computing, 2(2), 125–131. https://doi.org/10/db3xdh

Hockey, S. (2004). The History of Humanities Computing. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *Companion to Digital Humanities*. Blackwell Publishing Professional. http://www.digitalhumanities.org/companion/

Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, *27*(2), 265–276. https://doi.org/10/f93pfc

Jett, J., Capitanu, B., Kudeki, D., Cole, T. W., Hu, Y., Organisciak, P., Underwood, T., Dickson Koehl, E., Dubnicek, R., & Downie, J. S. (2020). The HathiTrust Research Center Extracted Features Dataset (2.0). https://doi.org/10.13012/R2TE-C227

JSTOR. (n.d.). *JSTOR Data For Research*. JSTOR. Retrieved February 5, 2020, from https://www.jstor.org/dfr/

King, R., & Olson, R. (2015, November 18). *How The Internet* Talks*. FiveThirtyEight. https://projects.fivethirtyeight.com/reddit-ngram/

Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory Data Analysis. In MIT Critical Data (Ed.), *Secondary Analysis of Electronic Health Records* (pp. 185–203). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_15

Manovich, L. (2018). *The science of culture? Social computing, digital humanities and cultural analytics*. https://doi.org/10/ghs8xg

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. Http://Mallet. Cs. Umass. Edu.

McConnaughey, L., Dai, J., & Bamman, D. (2017). The labeled segmentation of printed books. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 737–747.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Milo, T., & Somech, A. (2018). Deep Reinforcement-Learning Framework for Exploratory Data Analysis. *Proceedings of the First International Workshop on*

Exploiting Artificial Intelligence Techniques for Data Management, 1–4. https://doi.org/10/gkfm6j

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis*, *16*(4), 372–403. https://doi.org/10/cb486t

Mullen. (n.d.). *Isn't it obvious?* Retrieved February 5, 2020, from https://lincolnmullen.com/blog/isnt-it-obvious/

Putnam, L. (2016). The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast, *The American Historical Review*, Volume 121, Issue 2, April 2016, Pages 377–402, https://doi.org/10.1093/ahr/121.2.377

Organisciak, P., Shetenhelm, S., Vasques, D. F. A., & Matusiak, K. (2019). Characterizing Same Work Relationships in Large-Scale Digital Libraries. *International Conference on Information*, 419–425.

Quinn, B. (2012). Collection Development and the Psychology of Bias. *The Library Quarterly*, *82*(3), 277–304. https://doi.org/10/f33t8h

Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press.

Rockwell, G. (2003). What is text analysis, really? *Literary and Linguistic Computing*, 18(2), 209–219. https://doi.org/10/cchv8h

Rockwell, G., & Sinclair, S. (2016). Hermeneutica: Computer-assisted interpretation in the humanities. MIT Press.

Rockwell, G., Sinclair, S., Ruecker, S., & Organisciak, P. (2010). Ubiquitous Text Analysis. *Poetess Archive Journal*, *Dec. 2010*.

Ruecker, S., Radzikowska, M., & Sinclair, S. (2016*). Visual Interface Design for Digital Cultural Heritage: A Guide to Rich-Prospect Browsing*. Routledge. https://doi.org/10.4324/9781315547961

Sadler, B., & Bourg, C. (2015). Feminism and the Future of Library Discovery. *The Code4Lib Journal*, *28*. https://journal.code4lib.org/articles/10425

Satyanarayan, A., Moritz, D., Wongsuphasawat, K., & Heer, J. (2016). Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 341–350.

Satyanarayan, A., Wongsuphasawat, K., & Heer, J. (2014). Declarative interaction design for data visualization. *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, 669–678.

Schmidt, B. M. and Camacho, Martin (2011, October 21). *Bookworm* [Beta Sprint Competition Selection Presentations]. Digital Public Library of America Plenary Meeting, Washington, DC.

Schmidt, B. M. (2015). *Bookworm D3 layouts*. http://bookworm.benschmidt.org/posts/2015-10-20-D3-bookworm-plottypes.html

Schofield, A., Thompson, L., & Mimno, D. (2017). Quantifying the Effects of Text Duplication on Semantic Models. *Conference on Empirical Methods on Natural Language Processing*. EMNLP 2017, Copenhagen, Denmark.

Scientific computing tools for Python. (2021). SciPy.org.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, 336–343. https://doi.org/10.1109/VL.1996.545307

Singh, A. (2015). *The Archive Gap: Race, the Canon, and the Digital Humanities*. http://www.electrostani.com/2015/09/the-archive-gap-race-canon-and-digital.html

Singh, A. (2019). Beyond the Archive Gap: The Kiplings and the Famines of British Colonial India. *South Asian Review*, *40*(3), 237–251. https://doi.org/10/gkfpzz

Sprokel, N. (1978). The "Index Thomisticus." *Gregorianum*, 59(4), 739–750.U

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, Mass.

Underwood, T. (2014). Theorizing research practices we forgot to theorize twenty years ago. *Representations 127(1)*: 64–72. https://doi.org/10.1525/rep.2014.127.1.64

Underwood, T., & Sellers, J. (2016). The Longue Durée of Literary Prestige. *Modern Language Quarterly*, *77*(3), 321–344. https://doi.org/10.1215/00267929-3570634

Wickham, H. (2010). A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, *19*(1), 3–28.

Wickham, H., Cook, D., Hofmann, H., & Buja, A. (2010). Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 973–979. https://doi.org/10.1109/TVCG.2010.161

Wiedenhof, J. (2005). Purpose and effect in the transcription of Mandarin. *Proceedings of the International Conference on Chinese Studies*, 387–402.

Wilkinson, L. (1999). *The grammar of graphics*. Springer.