Articles

HathiTrust as a Data Source for Researching Early Nineteenth-Century Library Collections: Identification, Coverage, and Methods

Julia Bauder

ABSTRACT

An intriguing new opportunity for research into the nineteenth-century history of print culture, libraries, and local communities is performing full-text analyses on the corpus of books held by a specific library or group of libraries. Creating corpora using books that are known to have been owned by a given library at a given point in time is potentially feasible because digitized records of the books in several hundred nineteenth-century library collections are available in the form of scanned book catalogs: a book or pamphlet listing all of the books available in a particular library. However, there are two potential problems with using those book catalogs to create corpora. First, it is not clear whether most or all of the books that were in these collections have been digitized. Second, the prospect of identifying the digital representations of the books listed in the catalogs is daunting, given the diversity of cataloging practices at the time. This article will report on progress towards developing an automated method to match entries in early nineteenth-century book catalogs with digitized versions of those books, and will also provide estimates of the fractions of the library holdings that have been digitized and made available in the Google Books/HathiTrust corpus.

INTRODUCTION

Digital libraries such as Google Books and HathiTrust have created tantalizing opportunities for research into the history of American culture: automated analyses of the entire corpus of books published at a given point in time. The attraction of this prospect is most clearly demonstrated by the avalanche of papers written using the Google Books Ngram data, which provides counts over time of the words and phrases used in the works that make up the Google Books corpus. As soon as this data became available in 2009, it was used to make arguments about social, linguistic, and other changes over time as reflected in changes in the words used in print.[1] However, for nearly as long, other researchers have been cautioning that the Google Books corpus is not a representative sample of publishing output, let alone of what the public at large was actually reading in a given year, and that its unrepresentativeness makes it dangerous to draw sweeping conclusions from this data.[2]

One potentially feasible solution to the problem of unrepresentativeness in the Google Books corpus would be to use corpora based on the holdings of a specific library or a group of libraries. Using library holdings to form corpora helps to remedy some known issues with using the Google Books corpus as an indicator of social change, such as the fact that many books did not become

Julia Bauder (bauderj@grinnell.edu) is Associate Professor and Social Studies and Data Services Librarian, Grinnell College.

popular and/or widely available until well after their official publication date, and that some prolific authors who contributed hundreds of thousands of words to the Google Books corpus were never as widely purchased and read as authors who wrote a single, short, best-selling work.[3] Although using books held by a set of libraries at a given time as the corpus has its own problems of unrepresentativeness—particularly, for long-established libraries, the fact that the books on the shelf at a given time represent not only works of interest to current users but also those of interest to users from decades past—triangulating this data with that provided by the Google Books Ngram data would at least give some sense of whether and where these different corpora disagree.[4]

Creating corpora using books that are known to have been owned by a given library at a given point in time is potentially feasible because digitized records of the books in several hundred nineteenth-century library collections are available in the form of scanned book catalogs: a book or pamphlet listing all of the books available in a particular library. However, there are two potential problems with using those book catalogs to create corpora. First, it is not clear whether most or all of the books that were in these collections have been digitized, incorporated into Google Books and HathiTrust, and hence made available for Ngram analyses. Second, the prospect of identifying the digital representations of the books listed in the catalogs is daunting, as both widely agreed-upon cataloging standards and universal identifiers were not adopted until late in the nineteenth century. This article will report on progress towards developing a fully-automated method to match entries in early nineteenth-century book catalogs with digitized versions of those books, and will also provide estimates of the fractions of the library holdings that have been digitized and made available in the Google Books/HathiTrust corpus.

METHODS

Practical considerations dictated using data from HathiTrust rather than from Google Books for this research. The HathiTrust corpus, although not perfectly coextensive with the Google Books corpus, has very substantial overlap with it. The HathiTrust digital archive was founded in 2008, when a group of large academic libraries formed a collaboration to archive and disseminate their digitized books. The vast majority of those digitized books—around 95 percent, as of mid-2017—had originally been scanned as part of the Google Books project; the agreements that Google Books entered into with the libraries typically stipulated that Google had to provide the library with a digital copy of each book scanned from that library.[5] It was necessary to use HathiTrust rather than Google Books as the comparison corpus because the metadata for the titles in HathiTrust is readily available in ways that the Google Books metadata is not, including as bulk MARC-data downloads.

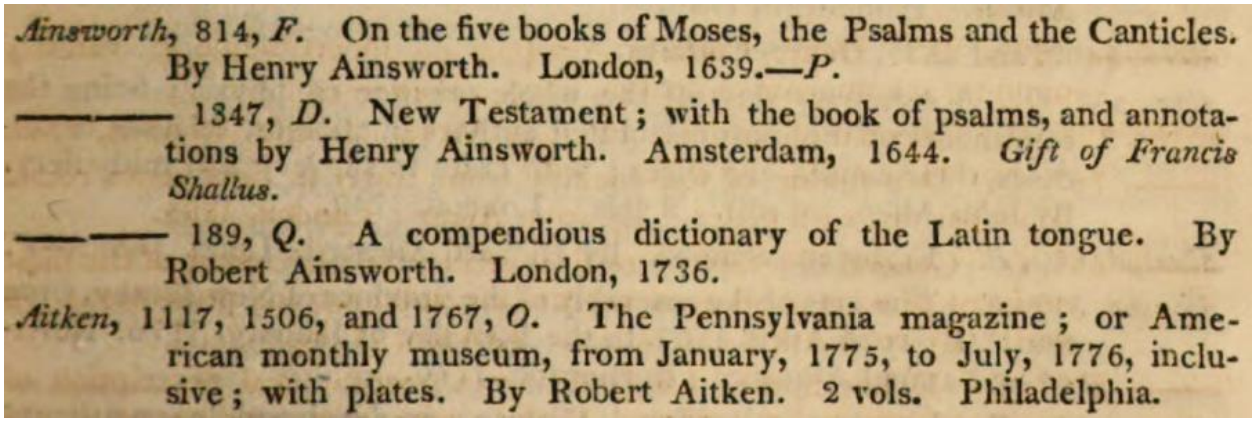
The libraries included in this analysis are social libraries, which were a type of quasi-public library that predated the now-standard, tax-supported public library in the United States. These libraries were privately owned and operated, but were open to some large portion of the population of a particular area who were willing and able to pay a fee or buy a share to belong to the library. Although the presence or absence of a book in social library collections is not a perfect indicator of the book's popularity—most social libraries pointedly refused to purchase the "trashy" but widely read sensational fiction of the day—it is a defensible proxy (although with some caveats, as noted above) for the popularity of the "serious" literature and nonfiction works that made up the bulk of these libraries' collections.

https://doi.org/10.6017/ital.v38i4.11251

Roughly one hundred social library book catalogs published between 1800 and 1860 can be found in HathiTrust.[6] For the purposes of the present study, attention was focused on the thirteen library catalogs from ten different American libraries that were published between 1776 and 1825. (A list of these catalogs can be found in appendix A.) These catalogs were chosen because they are likely to present the worst-case scenario in terms of both of the challenges mentioned above: the highest percentage of rare and extremely old books, which Google's partner libraries would have been least likely to permit to be scanned by Google, and, presumably, the most primitive and eclectic cataloging practices.

To the extent that it was possible to do so, this analysis focused on book-length monographs. When serials or pamphlets were listed in a separate section of the catalog, those catalog pages were excluded from the process by which entries were extracted from the catalogs and parsed into CSV files. Serials present particularly intractable matching problems: not only are the original catalogs often unclear about which specific volumes were held, but also HathiTrust's MARC data does not always clearly indicate which volumes are available in HathiTrust either. Pamphlets have limited coverage in HathiTrust.

The selected catalogs were downloaded from HathiTrust as PDFs, and the pdftotext software was used to extract the OCR data from the relevant pages of the scans as hOCR (a file format for OCR that includes information about where each word is located on the page in addition to the words themselves). [7] Then cleaning scripts were created that parsed the hOCR data into CSV files for analysis, with one catalog entry per line of the CSV file.[8] Given the widely varied cataloging practices of the early nineteenth century, several different cleaning scripts were written, each tailored to a particular catalog format. For example, many of the catalogs had entries that spanned multiple lines (see figures 1 and 2), so the scripts for those catalogs had to be able to identify when each new entry started. Many catalogs had extraneous information, such as the name of the donor of the book or the size of the book, that had to be filtered out (see figure 1; F, Q, O, and D refer to the size of the book: folio, quarto, octavo, or duodecimo). In addition, various forms of dittoes were frequently used in these catalogs (see figures 1, 2, and 3), so one of the tasks for the cleaning scripts was to identify the dittoes and replace them with the correct words from the previous entry.


Ainsworth, 814, *F.* On the five books of Moses, the Psalms and the Canticles. By Henry Ainsworth. London, 1639.—*P.*

——— 1347, *D.* New Testament; with the book of psalms, and annotations by Henry Ainsworth. Amsterdam, 1644. *Gift of Francis Shallus.*

——— 189, *Q.* A compendious dictionary of the Latin tongue. By Robert Ainsworth. London, 1736.

Aitken, 1117, 1506, and 1767, *O.* The Pennsylvania magazine; or American monthly museum, from January, 1775, to July, 1776, inclusive; with plates. By Robert Aitken. 2 vols. Philadelphia.

Figure 1. Library Company of Philadelphia, *A Catalogue of the Books Belonging to the Library Company of Philadelphia: To Which Is Prefixed, A Short Account of the Institution, with the Charter, Laws, and Regulations* (Philadelphia, PA: Printed by Bartram & Reynolds, 1807), 5.

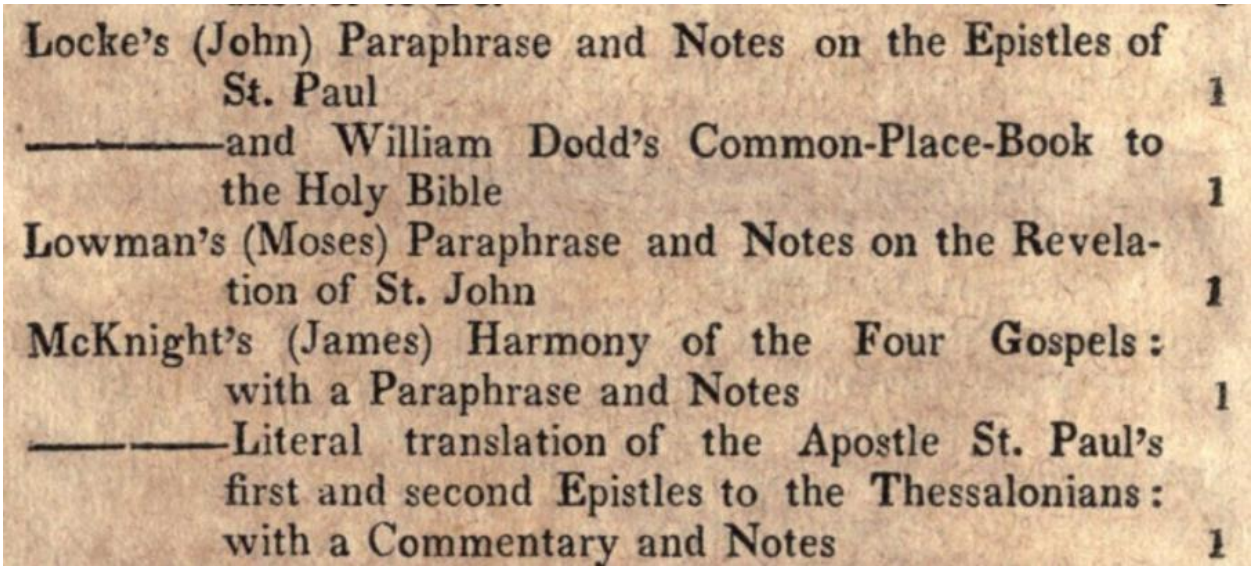

Locke's (John) Paraphrase and Notes on the Epistles of St. Paul 1
——————and William Dodd's Common-Place-Book to the Holy Bible 1
Lowman's (Moses) Paraphrase and Notes on the Revelation of St. John 1
McKnight's (James) Harmony of the Four Gospels: with a Paraphrase and Notes 1
——————Literal translation of the Apostle St. Paul's first and second Epistles to the Thessalonians: with a Commentary and Notes 1

Figure 2. Library Company of Baltimore, *A Catalogue of the Books, &c. Belonging to the Library Company of Baltimore: To Which Are Prefixed the Act for the Incorporation of the Company, Their Constitution, Their By-Laws, and an Alphabetical List of the Members* (Baltimore, MD: printed by Eades and Leakin, 1809), 46.

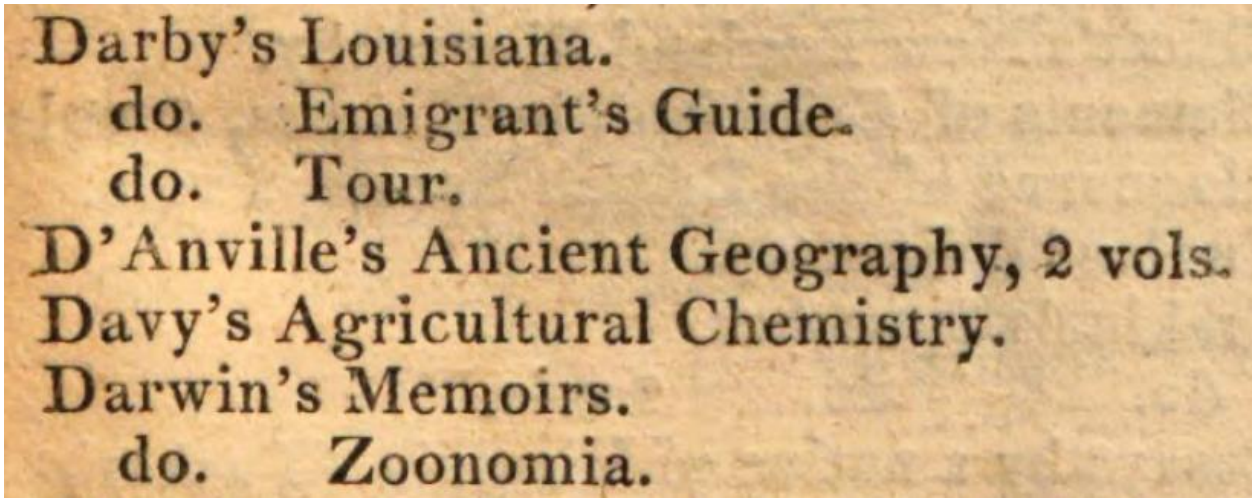

Darby's Louisiana.
do. Emigrant's Guide.
do. Tour.
D'Anville's Ancient Geography, 2 vols.
Davy's Agricultural Chemistry.
Darwin's Memoirs.
do. Zoonomia.

Figure 3. Washington Library Company, *Catalogue of Books in the Washington Library* (Washington, DC: printed by Anderson and Meehan, 1822), 17.

Unfortunately, the horizontal-line dittoes seen in figures 1 and 2—a type of ditto which is used in seven of the thirteen catalogs—are represented inconsistently or not at all in the hOCR, so they cannot reliably be used to identify places where words need to be carried down from the previous entry. For the catalog of the Library of Company of Philadelphia, from which figure 1 was taken, the numbers after the horizontal-line dittoes (which identify the books' locations on the shelves) can be used to distinguish between a line that is indented because it is a continuation of the entry above and a line that is indented but is the start of a new entry. In theory, a cleaning script for the catalog of the Library Company of Baltimore (figure 2) could use a similar process to identify the last line of an entry by watching for the right-justified count of volumes at the end of each entry.

https://doi.org/10.6017/ital.v38i4.11251

When a right-justified digit was encountered, the script could then carry down the first word from that entry if the first word in the next entry was indented. However, these isolated digits were also not handled well by the OCRing process: many do not appear in the hOCR file at all, and those that do are as likely to be OCRed as a colon, an exclamation point, a capital I, etc., as they are to be a digit. Hence, the three catalogs of the Library Company of Baltimore, which use this format and have this OCR issue, were not analyzed for this project.

Table 1. Results of verification

Library	Date founded if known, or inc. if not known[9]	Date catalog printed	Number of spreadsheet entries	Number of entries hand-verified	Hand-verified entries that cannot be positively identified	Hand-verified, positively identifiable entries that are not in HathiTrust	Positively identifiable entries successfully matched when work was in HathiTrust
Library Company of Philadelphia	1731	1807	7619	128	0%	16.9%	79.8%
Horsham Library Company	1808	1810	143	143	28.4%	5.1%	79.8%
Salem (MA) Athenaeum	Inc. 1810	1811	1585	130	0.8%	11.3%	72.3%
New York Society Library	1754	1813	4522	135	5.7%	17.9%	76.1%
Providence Library Company	1753	1818	688	688	17.1%	9.4%	87.2%
Apprentices' Library (New York, NY)[10]	1820	1820	1811	124	34.4%	15.0%	69.7%
Washington (DC) Library Company	Inc. 1814	1822	900	124	12.9%	3.2%	83.7%
Boston Library	Inc. 1794	1824	2273	138	4.1%	11.1%	82.5%
Mercantile Library (New York, NY)	1820	1825	1386	138	0%	11.3%	86.0%

The catalogs of the other nine libraries could all be parsed with an acceptable success rate and, with one exception, were included. The exception was the Salem Athenaeum's 1818 catalog, which was identical in format and nearly identical in content to the Athenaeum's 1811 catalog. Given the overwhelming similarity it was decided to include only one of the catalogs; given that the goal of this analysis was to try to use the worst-case-scenario catalogs, the older of the two catalogs was chosen for inclusion.

Once the catalogs were parsed into CSV files, they were run through another script that attempted to match each entry in the catalog against metadata from HathiTrust. In February 2019, MARC records containing metadata for 2,824,875 public-domain titles in HathiTrust were downloaded from HathiTrust via their OAI feed and ingested into a local Apache Solr index for searching and matching, using code from the SolrMarc and VuFind projects.[11] Because of OCR errors in the catalog files and mistakes in the original catalogs, many of the words in the entries have one or more character-level errors. Therefore, Solr's fuzzy searching option was used, which allows words to match as long as the Levenshtein distance between them is two or less. (The Levenshtein distance is the number of edits, such as changing one letter to another or adding or deleting a letter, it would take to turn one word into the other.) No attempt was made to match specific editions; as can be seen from the excerpts in figures 2 and 3, many of the catalogs do not contain sufficient detail to do so, even if it was desirable. The goal was merely to determine whether the text of that work, from *any* edition, was contained in the HathiTrust corpus.

Once the catalogs had been checked against HathiTrust, a sample of the entries was hand-verified. For the two smallest catalogs, the Horsham Library Company and the Library Company of Providence, all entries were hand-verified. For the other catalogs, a random sample of approximately 130 items (+/- 10) was selected. Microsoft Excel's random-number generator was used to assign each line in the CSV file a number between 0 and 1, and then the lowest 1.5 percent to 12.5 percent (depending on the number of items in the catalog) were examined.

RESULTS

Percentage of Works Included in HathiTrust

A minimum of four of the books in every catalog examined was missing from HathiTrust. As can be seen in table 1, the fraction of books from the hand-verified sample that was missing from HathiTrust ranged from 3.2 percent for the Washington Library Company to just shy of 18 percent for the New York Society Library. The Library Company of Philadelphia, at 16.9 percent missing, had the second-highest missing number. It is not surprising that these two libraries, as two of the oldest and most venerable libraries in the United States at the time, owned the most books that are not represented in HathiTrust, as both have a high percentage of very old and rare works. However, not all of the books from these collections that are not represented in HathiTrust fall into that category. Only six of the twenty missing works from the Library Company of Philadelphia sample, and no more than eight of twenty-two from the New York Society Library, were published before 1700, for example.[12]

Percentage of Works That Cannot Be Positively Identified

As can be seen in figures 1 through 3, some catalogs provided relatively full titles (figures 1 and 2), while others described the works in only two or three words each (figure 3). As might be expected, it is much easier to positively identify the works when fuller titles are provided, although two or three words proved to be enough to identify the work unambiguously the

https://doi.org/10.6017/ital.v38i4.11251

majority of the time. (All of the titles shown in figure 3 can be positively identified, for example.) In the samples taken from the nine catalogs, the percentage of titles that were unidentifiably ambiguous ranged from 0 percent (Library Company of Philadelphia, Mercantile Library of New York) to more than one in four (Apprentices' Library of New York, 34.4 percent; Horsham Library Company, 27.9 percent). The Apprentices' Library of New York and the Horsham Library Company were particularly problematic because they frequently omitted the name of the author, in addition to greatly compressing the title; without an author name, titles such as *Modern Geography* (Apprentices' Library) and *History of Rome* (Horsham Library Company) present far too many potential matches. However, even including the author's name does not make all greatly compressed entries identifiable. One particularly egregious example comes from the Library Company of Providence's 1818 catalog, which contains an entry reading "*Bell's Inquiry*." The list of candidates for this work includes *A Practical Inquiry into the Authority, Nature, and Design of the Lord's Supper*, by William Bell; *An Inquiry into the Causes Which Produce, and the Means of Preventing Diseases Among British Officers, Soldiers, and Others in the West Indies*, by John Bell; and *Inquiry into the Policy and Justice of the Prohibition of the Use of Grain in Distilleries*, by Archibald Bell.

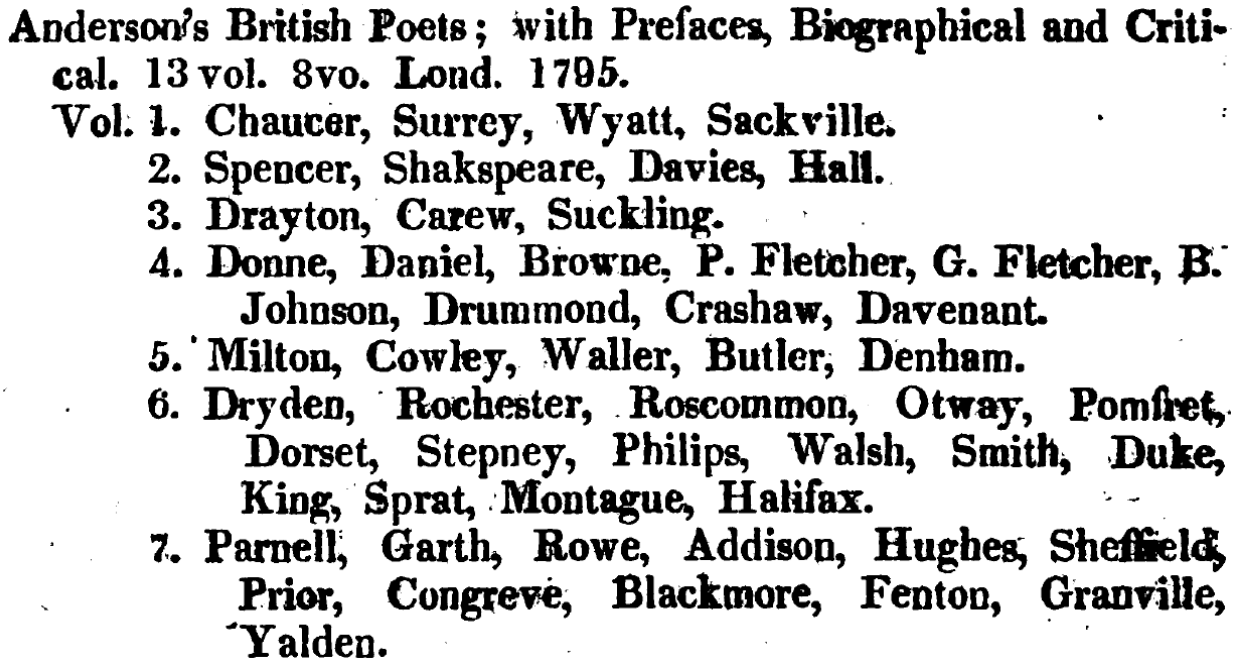

Anderson's British Poets; with Prefaces, Biographical and Critical. 13 vol. 8vo. Lond. 1795.
Vol. 1. Chaucer, Surrey, Wyatt, Sackville.
2. Spencer, Shakspeare, Davies, Hall.
3. Drayton, Carew, Suckling.
4. Donne, Daniel, Browne, P. Fletcher, G. Fletcher, B. Johnson, Drummond, Crashaw, Davenant.
5. Milton, Cowley, Waller, Butler, Denham.
6. Dryden, Rochester, Roscommon, Otway, Pomfret, Dorset, Stepney, Philips, Walsh, Smith, Duke, King, Sprat, Montague, Halifax.
7. Parnell, Garth, Rowe, Addison, Hughes, Sheffield, Prior, Congreve, Blackmore, Fenton, Granville, Yalden.

Figure 4. New York Society Library, *A Catalogue of the Books Belonging to the New-York Society Library* (New York: printed by C. S. Van Winkle, 1813), 139.

Success Rates for the Parsing and Matching Scripts

When there was a single, identifiable work that matched the catalog entry, and that work was in HathiTrust, the matching scripts identified it at least 70 percent of the time for every individual catalog. Unsurprisingly, catalogs such as those of the Horsham Library Company and the Apprentices' Library of New York that had entries that were difficult to positively identify were also more difficult for the script to properly match, although the matching script still succeeded between roughly 70 and 80 percent of the time.

For two other libraries with below-average matching results (the Library Company of Philadelphia and the New York Society Library), many of the matching problems were caused by issues with the scanned catalogs that the data-cleaning scripts did not handle well. The New York Society Library catalog listed out the contents of multivolume sets in a way that was difficult for the cleaning script to identify and remove (see figure 4); instead, it was common for each volume of the set to end up with its own entry in the dataset. Since the HathiTrust records generally do not list out the contents of each volume, it was very rare for the cleaning script to correctly match a set based on an entry for one volume in the set. Twenty-seven percent (six out of 22) missed matches from that sample failed because of this table-of-contents issue.

For the Library Company of Philadelphia, the problem lies with a quirk in the hOCR where the character heights for many of the horizontal-line dittoes are extremely high—around twenty pixels, when the text around those dittoes is typically around ten pixels high. It appears as if the OCR program may have treated each horizontal-line ditto as an em dash and assigned it a height that would be proportional for an em dash of that length. These extra-tall line heights for the first "word" on the line cause issues with the algorithm that processes the text line-by-line, causing some entries to be inappropriately divided across two entries in the data spreadsheets. Unsurprisingly, the matching script had difficulty correctly identifying the correct work in HathiTrust when it was trying to match based on only half of the book's title.

CONCLUSIONS

Although not a complete success, the results of this study provide hope that it might be possible to create full-text corpora based on the works in individual libraries with minimal manual labor, with a few caveats. The first caveat is that the digitized catalogs of those libraries must meet certain specifications:

1) The catalog is formatted, and has been OCRed, in such a way that it is consistently possible to parse the catalog line-by-line and to identify algorithmically where each entry starts and ends.
2) The catalog provides at least the authors' last names, if not their full names, plus a more-or-less complete and accurate transcription of the title proper.
3) Either the catalog contains minimal extraneous information (such as tables of contents or donors' names), or the extraneous information is consistently formatted in a way that it can be algorithmically identified and removed.

The second caveat is that even if all of these conditions are met, the full-text corpora that can be created will probably still be missing some small percentage of the books available in that library. One potential direction for future research could be more closely examining the books that are absent from HathiTrust to see if there are any commonalities among them that might bias research done using these corpora, or if the missing works can safely be treated as random omissions. On the other hand, as was noted above, the catalogs used in this study represent a likely worst-case scenario for being able to positively identify the works listed in the catalogs and for those works being present in HathiTrust. Another promising avenue for future research would be to repeat this analysis on catalogs from the mid-to-late nineteenth century to see if the works in those catalogs are in fact more likely to exist in the HathiTrust corpus.

https://doi.org/10.6017/ital.v38i4.11251

APPENDIX A: AMERICAN LIBRARY CATALOGS FROM 1776 TO 1825 INCLUDED IN HATHITRUST

Boston Library, *Catalogue of Books in the Boston Library, June, 1824,* Boston: Munroe and Francis, 1824, http://hdl.handle.net/2027/hvd.32044080249337.

General Society of Mechanics and Tradesman of the City of New York, *Catalogue of the Apprentices' Library, Instituted by the Society of Mechanics and Tradesman of the City of New-York, on the 25th November, 1820: With the Names of the Donors: To Which Is Added, an Address Delivered on the Opening of the Institution by Thomas R. Mercein, a Member of the Society*. New York: printed by William A. Mercein, no. 93 Gold-Street, 1820, http://hdl.handle.net/2027/nnc2.ark:/13960/t8md1cv2t.

Horsham Library Company, *The Constitution, Bye-Laws, and Catalogue of Books, of the Horsham Library Company.* Philadelphia, PA: J. Rakestraw, 1810, http://hdl.handle.net/2027/nnc1.cu55910696.

Library Company of Baltimore, *A Catalogue of the Books, &c. Belonging to the Library Company of Baltimore: To Which Are Prefixed the Act for the Incorporation of the Company, Their Constitution, Their By-Laws, and an Alphabetical List of the Members.* Baltimore, MD: printed by Eades and Leakin, 1809, http://hdl.handle.net/2027/nyp.33433069263907.

Library Company of Baltimore, *A Supplement to the Catalogue of Books, &c. Belonging to the Library Company of Baltimore.* Baltimore, MD: printed by J. Robinson, 1816, http://hdl.handle.net/2027/nyp.33433069263899.

Library Company of Baltimore, *A Supplement to the Catalogue of Books, &c. Belonging to the Library Company of Baltimore.* Baltimore, MD: printed by J. Robinson, 1823, http://hdl.handle.net/2027/nyp.33433069263899.

Library Company of Philadelphia, *A Catalogue of the Books Belonging to the Library Company of Philadelphia: To Which Is Prefixed, A Short Account of the Institution, with the Charter, Laws, and Regulations*. Philadelphia, PA: Printed by Bartram & Reynolds, 1807, http://hdl.handle.net/2027/nyp.33433075914816.

Mercantile Library Association of the City of New York, *Catalogue of the Books Belonging to the Mercantile Library Association of the City of New-York: To Which Are Prefixed, the Constitution and the Rules and Regulations of the Same*. New York: printed by Hopkins & Morris, 1825, http://hdl.handle.net/2027/nyp.33433057517090.

New York Society Library, *A Catalogue of the Books Belonging to the New-York Society Library.* New York: printed by C. S. Van Winkle, 1813, http://hdl.handle.net/2027/mdp.39015023478822.

Providence Library Company, *Charter and By Laws of the Providence Library Company, and a Catalogue of the Books of the Library.* Providence, RI: printed by Miller and Hutchens, 1818, http://hdl.handle.net/2027/nyp.33433059555346.

Salem Athenaeum, *Catalogue of the Books Belonging to the Salem Athenæum, with the By-Laws and Regulations.* Salem, MA: Printed by Thomas C. Cushing, 1811, http://hdl.handle.net/2027/hvd.32044080252174.

Salem Athenaeum, *Catalogue of the Books Belonging to the Salem Athenæum, with the By-Laws and Regulations.* Salem, MA: Printed by W. Palfray, 1818, http://hdl.handle.net/2027/hvd.32044080252174.

Washington Library Company, *Catalogue of Books in the Washington Library, July 20, 1822.* Washington, DC: printed by Anderson and Meehan, 1822, http://hdl.handle.net/2027/chi.098498263.

REFERENCES

[1] See, e.g., Jean-Baptiste Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science*, 311, no. 6014 (January 11, 2011): 176-82, https://doi.org/10.1126/science.1199644; Jean M. Twenge, W. Keith Campbell, and Brittany Gentile, "Male and Female Pronoun Use in U.S. Books Reflects Women's Status, 1900-2008," *Sex Roles* 67, nos. 9-10 (November 2012), 488-93, https://doi.org/10.1007/BF00287963; Patricia M. Greenfield, "The Changing Psychology of Culture from 1800 through 2000," *Psychological Science* 24, no. 9, 1722-31, https://doi.org/10.1177/0956797613479387.

[2] Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds, "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-cultural and Linguistic Evolution," *PLOS One* 10, no. 10 (October 7, 2015): e0137041. https://doi.org/10.1371/journal.pone.0137041; Alexander Koplenig, "The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets—Reconstructing the Composition of the German Corpus in Times of WWII," *Digital Scholarship in the Humanities* 32, no. 1 (April 2017): 169-88, https://doi.org/10.1093/llc/fqv037.

[3] Pechenick et al., 2015; Lindsay DiCuirci, *Colonial Revivals: The Nineteenth-Century Lives of Early American Books* (Philadelphia: University of Pennsylvania Press, 2019).

[4] Robert A. Gross, "Reconstructing Early American Libraries: Concord, Massachusetts, 1795-1850," *Proceedings of the American Antiquarian Society,* 97, no. 1 (January 1, 1987): p. 331-451.

[5] Jennifer Howard, "What Ever Happened to Google's Effort to Scan Millions of University Library Books?," *EdSurge*, August 20, 2017, https://www.edsurge.com/news/2017-08-10-what-happened-to-google-s-effort-to-scan-millions-of-university-library-books.

[6] Book catalogs fell out of favor in the latter half of the nineteenth century as library collections became larger and more dynamic, making book catalogs much more difficult and expensive to compile and to keep up to date. By the end of the nineteenth century, book catalogs had largely been replaced by the card catalog system that remained in use through most of the twentieth century. Although card catalogs were far superior for their primary purposes—maintaining an inventory of books presently owned by the library and allowing library users to locate the books that they wanted—they leave no permanent record of the books listed in the catalog at any particular point in time.

[7] Information about pdftotext can be found at https://manpages.debian.org/testing/poppler-utils/pdftotext.1.en.html.

[8] The cleaning scripts, as well as data and other code used in this project, are available in https://github.com/julia-bauder/library-catalog-analysis-public.

[9] The founding and incorporation dates were taken from the prefatory texts in the book catalogs used in this analysis, as listed in appendix A.

[10] The scan of this catalog that is available from HathiTrust is missing pages 3-6.

[11] Apache Solr is a widely used, open-source search platform. SolrMarc is a utility that can be used to index MARC records into Solr. VuFind is an open-source library discovery layer built in part on Solr and SolrMarc. For more information, see http://lucene.apache.org/solr/, https://github.com/solrmarc/solrmarc, and https://vufind.org/vufind/, respectively. The HathiTrust OAI feed is available at https://www.hathitrust.org/oai.

[12] Five of the missing works from the New York Society Library sample were undated in the catalog.