

# Embracing Bookness:

## Introducing Library Staff and Library Students to Text and Data Mining with HathiTrust Research Center

*Rachel N. Hogan and Patrick Williams*

### Introduction

In this chapter, we explore the affordances and benefits of teaching foundational text and data mining (TDM) skills to library staff and library school students using HathiTrust Digital Library (HTDL) and the associated HathiTrust Research Center (HTRC).<sup>1</sup> HTDL provides access to more than 18 million volumes sourced from academic and public libraries in order to advance the goals of scholars and researchers, independent from corporate interest.<sup>2</sup> HTDL represents an enormous source of texts familiar and unfamiliar to potential workshop audiences. HTRC makes it possible to explore basic concepts of TDM, like word frequency analysis, named-entity extraction, and topic modeling, without the need for specialized software or programming skills. Training in the HTRCs algorithms can provide a foundation for developing TDM literacies within libraries and serves as a strong and transferable introduction for library staff who may encounter the concepts across multiple platforms in their work.

We both currently serve in librarian roles that support digital scholarship and have identified instructional opportunities offered by HTRC as a chance to introduce TDM to library and information science (LIS) students and staff in contexts familiar to their fields. This chapter provides an overview of our experiences offering multiple iterations of a workshop introducing TDM and HathiTrust at Syracuse University. The workshops, which we planned and delivered, allowed participants to build on their experiences as LIS professionals to explore HathiTrust and compare HTRC to other proprietary TDM software.

HTDL and HTRC are rooted in fundamental LIS professional practices: cataloging, preservation, and usage rights—topics within which we work every day. Additionally, their familiar metadata approaches preserve items’ “bookness” by retaining pagination, paratextual apparatus, and autonomy as volumes. Both platforms engage our professional experience and help us to consider concepts like data structures, data cleaning, and data curation with a common vocabulary rooted within common practices. HTDL and HTRC offer library workers opportunities to curate collections, extract metadata, and use the HTRC tools with the intention to expand their insights as their work in TDM unfolds.

However, as a site for exploring TDM within a community of library staff and library students, HTRC offers much more. It’s our belief that engaging the LIS community’s prior knowledge of library organization, metadata, cataloging, and access presents a rich opportunity for confident exploration of TDM by new learners. In this chapter, we provide background on TDM and skills training in libraries. We also recount our experiences developing, delivering, iterating, and assessing a workshop approach for introducing TDM via HTRC that makes use of common vocabulary and shared professional practices.

As a digital library, HathiTrust is infused with key values in alignment with the LIS profession: trust; access; equity, diversity, and inclusion; resilience; constructive, meaningful engagement; and leadership. HathiTrust demonstrates these values by incorporating practices such as commitments to metadata transparency, emphasizing collection provenance, and maximizing shared ownership, access, and preservation. Its tools and collections can benefit library workers’ training, collection management, provision of service, and their own research agendas as they learn TDM techniques. Materials in HTDL are structured, arranged, described, and connected, and work within the tool is scaffolded by many phases of work that mirror print and electronic collections in traditional libraries. HathiTrust also positions library staff to make skilled use of this widely available resource within their own communities.

Furthermore, HTRC allows for the focused application of TDM techniques within single titles or the work of a single author of the user’s choosing and can help them quickly access authentic and well-formed results of TDM-based analysis. Our workshops have attempted to balance a fully formed introduction to TDM via HTRC within a manageable timeframe for a single session. To do this, we planned with a few themes in mind: 1) relevance and resemblance to traditional library work, and 2) a scaffolded instructional approach that makes use of instructive and easily reproducible examples and activities. HathiTrust advocates for the practice of non-consumptive research, meaning “any research that involves computational analysis on books wherein the researchers do not, themselves, have direct access to the text of those books such that they might read them or reproduce

large portions of text from them.”<sup>3</sup> This enables users to apply non-consumptive research methods at scale through easy-to-use tools on a centralized and accessible platform.

## Literature Review

Like many emerging technologies, TDM has been a topic of interest in academic libraries and requires training staff and patrons alike. Not only does TDM offer applications for researchers in a wide variety of disciplines, but TDM also approaches can be useful in many phases of library work. In the LIS field, we often deal with large datasets, including structured and unstructured textual data and even full collections of books, periodicals, manuscripts, and other works. As evidenced by the chapters in this section, TDM capabilities are increasingly present in the collections, products, and applications academic libraries support and provide access to among their patron communities. Such support and access, of course, demand staff knowledge and training.

Over the past decade, a growing literature has pointed to TDM, data literacy, and software training needs for library staff. Koehl and Dubineck write that “[s]hifting norms such as these [library and information professionals supporting digital scholarship research] require librarians and information professionals to redevelop their skills as they explore potentially new and novel models for library-based research.”<sup>4</sup> In the 2015 issue of *Library Technology Reports*, “Coding for Librarians: Learning by Example,” Yelton suggested that librarians interested in learning coding applications should “[l]ook for projects that would be helpful in the context of your work, and use those to guide what you need to learn.”<sup>5</sup> It follows that authentic, practice-based tasks and contexts could be key in developing adjacent skills like those represented in TDM methodologies.

In fact, in their piece documenting the development of Library Carpentry, the LIS-focused adaptation of the Data and Software Carpentry models, Baker, et al. highlight the importance of a professional audience “who self-identify with the benefits of developing their own software skills in relation to their organizational needs”<sup>6</sup> as a preferable situation to some of the more open-ended or comprehensive training approaches available. The literature highlighting training approaches and applications for TDM through HTRC in library-related contexts is as varied as the many ways that library staff and students can go about gaining these skills.<sup>7</sup> Further studies, such as Bainbridge et al.<sup>8</sup> and Parulian and Worthey,<sup>9</sup> look at the Extracted Features dataset function in detail and provide open resources for learning and using this specific tool of HTRC. All of these works together provide examples and frameworks for learning and teaching HTRC in library contexts.

In our experience, it can be difficult to build a shared, transferable foundation in TDM because the approaches can be so broadly applied. TDM involves large-scale corpora for which users may have little prior knowledge or context. The way those corpora are structured can be obscured behind “black box” interfaces, which so many digital scholars and library staff routinely work to demystify. As well, the open-ended nature and the varying possibilities for interpretation of the results of TDM methods can leave users further confused and frustrated with using these tools. While essential and widely used tools and resources like the Voyant suite,<sup>10</sup> SAGE Research Methods Data Visualizations,<sup>11</sup> and Carpentries curricula<sup>12</sup> provide incredible support for TDM activities, we believe that

introducing TDM in library contexts via HTRC can reduce the friction and confusion due to its familiar collections, flexibility, structure, and built-in tools.

Patrick, one of the chapter's co-authors, was a part of the IMLS-funded New York Data Carpentries Library Consortium and noticed that participants, while successful in workshop activities, often struggled to connect the more generic exercises and example datasets available in the lessons to concrete and practical applications of the same concepts in their familiar library work contexts—that is, the usefulness of such methods was clear, but the direct applicability to immediate professional tasks was lacking. The development of the Library Carpentries curriculum, as recounted by Cope and Baker,<sup>13</sup> prompts us to think about the opportunities HathiTrust offers an LIS audience: a learning experience informed both by their prior professional knowledge and their aspirations for what they may want to use TDM to accomplish.

## Affordances and Benefits of HathiTrust Digital Library and HathiTrust Research Center

HathiTrust is supported by over 200 member libraries that have access to both HTDL and HTRC.<sup>14</sup> The preexisting access to both of these products makes HathiTrust a strong choice for TDM training in libraries. Similar TDM tools—ITHAKA's Constellate,<sup>15</sup> ProQuest TDM Studio,<sup>16</sup> and Gale's Digital Scholar Lab<sup>17</sup>—all require an additional subscription in order to use their tailored TDM tools on their respective collections: JSTOR and Portico, ProQuest, and Gale. As LIS professionals already know, databases can be expensive, and a library newly investing in digital scholarship support and subsequent training for their staff may not have the resources to license any or all of these new subscriptions. In contrast, libraries with HathiTrust subscriptions do not need to do any additional purchasing to begin supporting and learning TDM. Other databases that libraries commonly subscribe to, like Readex, have recently added built-in TDM tools that do not require an additional subscription; however, they are not as robust as HathiTrust in their current forms.

A benefit of using HathiTrust for training in library contexts is that HTDL follows familiar library customs and data structures. Texts in HTDL are sourced from academic and public libraries and contain the bibliographic metadata, including the MARC records, from those original libraries—even pagination is preserved in HTDL texts and HTRC outputs. All of the metadata are available for viewing when searching and opening HathiTrust records. This makes it easy to compare a copy of a text stored within HTDL to another library's copy that a researcher has already accessed. For library staff and students, this means that they can get a low-barrier entry into TDM in the HathiTrust environment. People who have already worked with HTDL will have an easier time learning the additional steps required for their TDM purposes. For those already doing research in HTDL or on a volume that is also contained within HTDL, it makes the transition from traditional scholarship to newer digital approaches that much easier. HTDL provides cover-to-cover scans of distinct expressions of a work, and all of that information is captured when

the data is analyzed in HTRC. This makes HTDL unlike some other digital libraries with public domain texts, such as Project Gutenberg, where users may not be able to choose or identify the exact published version of the text that they seek. Furthermore, for version comparison and volume selection, which is easy to do with the many scanned volumes in HTDL, users can compare not only the main text but the copyright and publishing information as well. This is an important feature because HTDL retains paratextual elements of the scanned books, and that material is also subject to analysis done in HTRC.

Collections built within HTDL, both those pre-existing public collections already available on the platform and newly conceived ones created by individual users, can easily be converted into datasets for TDM exploration in HTRC with a few simple steps. This saves users the initial and cumbersome steps of downloading, preparing, and uploading files before they can begin exploring TDM topics. HTRC is great for TDM beginners because it does not require advanced programming skills or prior knowledge. The HathiTrust team also hosts training sessions, virtual open office hours, and a user support team to help library staff and researchers use its tools.

For these reasons, HathiTrust is a more than suitable environment for introductory TDM training in libraries. Because the digital surrogates in the collection maintain their “bookness,” workshop participants can quickly identify questions they would like to explore and are positioned to understand the information the surrogates contain. Prior knowledge of the structure and content of books and collections aids in their interpretation of the HTRC algorithm outputs, which may include information about locations where entities appear in a text as well as the impact of paratextual elements like title pages, tables of contents, indexes, running headers, and chapter titles. These facets of texts and their application in TDM can be obscured when the basic unit of analysis is split into articles, chapters, or other chunks of text as dictated by the structure and content of competing platforms.

## Workshop Context and Approach

An introductory training session using HathiTrust can be an ideal way to introduce library staff to TDM approaches because it resembles a typical library catalog and database.

As with many digital scholarship workshops and trainings, TDM with HathiTrust requires that participants have access to a computer so that they can follow along with the steps of creating their accounts, creating or picking a collection in HTDL, building and validating a workset in HTRC, and applying the various analytical tools. We have found that hybrid arrangements are not an ideal option for this type of training because they can split attention between modes and can undermine the conversation.

Between spring 2020 and spring 2023, Syracuse University Libraries offered four iterations of a workshop entitled Digital Humanities Workshop: Introduction to Text Mining with the HathiTrust Research Center. Each session was co-led by Patrick Williams, humanities librarian and digital scholarship lead, and a graduate student assistant. The first session was co-led by graduate student employee Zhiwei Wang, and the remaining three sessions were co-led by Rachel N. Hogan, then master’s student in library and information science and information literacy scholar. The first three sessions were open to anyone

interested in TDM and marketed to the whole campus. The first was the only one offered online because of the COVID-19 pandemic. We saw the most interest from library staff in these sessions, and the fourth session was specifically offered as an online workshop for Syracuse's LIS students through the Library and Information Science Student Association (LISSA). The session for LISSA was explicitly tailored to LIS students and highlighted the "bookness" of HathiTrust texts as described above. These different iterations gave us insight into what led to the best conditions for the workshop. While the first three sessions were open to a wider audience, the turnout varied and the attendees were mostly library staff, graduate students, and faculty that we had worked with previously. The LISSA session had a strong turnout and was also coordinated with the graduate-level course being offered that semester: Digital Humanities for Librarians, Archivists, and Cultural Heritage Workers, which had just completed a text analysis unit. We discuss the lesson plan for the LISSA session below.

While we have discussed the benefits of teaching TDM in library contexts with HathiTrust, there were also some limitations. We tended to have groups of no more than ten at each of the workshops and found that to be beneficial for two reasons: this ensured that all participants followed along without issue and reduced slow run times with HTRC. In our experience, the HathiTrust system can be slow to respond when multiple users run some of the computing-intensive operations simultaneously. In some cases, one person experimenting with the parameters of an algorithm caused a system freeze. In other cases, having multiple people executing the same algorithm simultaneously appeared to overwhelm the system. Once we experienced this in a live workshop, we made sure to have screenshots and examples of each step and the outputs of each algorithm that we showcased; this advanced preparation may also benefit online or asynchronous workshops. Once we identified and anticipated this problem, we had no further problems with server stability for the LISSA-focused session, though we still made use of the prepared examples to ensure that users saw a variety of outputs.

To keep everyone on task and pace the workshops, we focused on previously curated HTDL collections to build our worksets. Participants chose from the existing Early American Cookbooks, African American Fiction collections,<sup>18</sup> or compact, topical collections we built prior to the workshops. These included a small set of serialized Victorian novels, selections of presidential papers, and several volumes of local Syracuse University history.<sup>19</sup> This allowed us to quickly establish how collections are curated and to provide multiple examples for participants to explore without the need for each person to create their own well-formed dataset during the workshop (though some have done so!). We chose the collections we prepared and shared as part of our workshop materials because they were relatively small, intentionally curated for a particular aim, and would hopefully help us avoid long processing times in-session.

## Workshop Planning and Delivery

Workshops for library staff and students were designed with the following learning objectives in mind:

- Participants will be able to describe three common approaches to TDM (term frequency, named-entity recognition, and topic modeling).
- Participants will recognize the familiar data structures and “bookness” of texts provided by HTDL.
- Participants will be able to identify, collect, and create datasets from HTDL for analysis with HTRC algorithms.
- Participants will consider outputs and applications for the HTRC algorithms in their own work and contexts.

In all cases, we planned for workshops to last between ninety minutes and two hours, with a balance of time for both discussion and hands-on experimentation. We produced a brief online guide to serve as a shared home base, and that guide included information on HTDL/HTRC account creation, links to the collections we chose for the sessions, examples, screenshots, and outputs (in case of server issues), and materials and links for further exploration.<sup>20</sup> We created a LibGuide for this, but a shared document could work equally well and would allow for additional interactivity among participants. This guide served as both a resource used during the session and as a post-workshop resource and refresher.

In advance of the workshops, and at the beginning of each, we encouraged users to set up their HathiTrust accounts and to navigate to HTDL and HTRC in separate tabs. Once participants arrived and were settled in and logged on, we introduced HTDL and the broader concepts of TDM through examples and discussion. Once we established the range of roles TDM can play for researchers, we narrowed our focus to just token counting, named-entity extraction, and topic modeling, noting that these three approaches are well aligned with the basic HTRC algorithms.

Following the introduction, we engaged participants in an exploration of HTDL collections and how to locate and create them. In this segment, we emphasize “bookness” by urging our library worker participants to consider using advanced catalog searches rather than merely searching within the full text of the HTDL to optimize their corpus curation. We have found audiences quickly grasp the implications of these intentional datasets more quickly than on-the-fly corpora composed of search results across the wide range of materials on some other TDM platforms. Since HTDL treats individual volumes as the unit of analysis (unlike tools such as ProQuest’s TDM Studio, which presents items at the individual article level), collections and worksets can be understood in a more cohesive way. We invited participants to consider how they might query their own libraries for useful information in the stacks, for meaningful paratextual information, and to apply those questions to their exploration of HTDL. After some brief collection exploration time, we pointed our participants to our curated HTDL collections to begin creating worksets in HTRC.

We then walked through the workset creation and validation process, first with step-by-step guidance screenshots, and then with hands-on practice. As this was the most complex portion of the workshop, we found it important to ensure that all participants arrived at the validated workset step together, before each set off to explore their own selections via execution of the HTRC algorithms described below. We also emphasized how researchers can locate and process their previously created worksets, acknowledging that they can return to prior work for comparison’s sake.

We then spent about twenty minutes each on the Token Count and Tag Cloud Creator, Named Entity Recognizer, and InPhO Topic Model Explorer algorithms, asking participants to execute each up front so their outputs could process while we discussed what was happening. We concluded each of these segments with five minutes for users to explore their own results (or, in the case of a technical issue, the example outputs we shared on the guide). It was typically in this section of the workshop—handling the outputs—that participants’ observations and questions guided the conversation, and we considered the possibilities of curating collections and worksets with these outputs in mind. We feel that the intentional choices of worksets really set the stage for meaningful TDM observations and ideas because users were able to share and theorize their expectations alongside the results that were generated. We closed the session with a preview of some additional TDM workshops and resources to engage the foundational TDM knowledge they developed in the session.

## Lessons Learned

Across multiple sessions, we noticed participants’ awareness of and interest in the role that choices made while building collections and worksets play in how they were able to interpret HTRC algorithmic outputs. For example, processing a smaller workset of volumes on a specific person or familiar topic brought certain expectations or interpretations, while a larger less-curated workset can invite more exploratory and speculative interactions. Library staff were quite keen to think about the ways that applying TDM approaches to their library work might yield interesting results, often in combination with other collections, information resources, or working knowledge. For example, one librarian envisioned a project for assessing collections of published music using named-entity recognition on musical volumes in HTRC. In other TDM workshops, particularly with more general audiences, we have often found that participants struggle with considering how the session applies to their own research goals or the kinds of questions they can explore and what choices and decisions they can make about what is included in the analysis. Interfaces that foreground enormous corpora, keyword-based source selection, and novel visualizations can be difficult to conceptualize in an introductory session. Particularly in library contexts, HTRC and HTDL can offer familiar and less overwhelming options. As library workers ourselves, we have found the ability to talk through TDM concepts with a common vocabulary, embedded in professional practices, and at a library-sized scale and with the transparency and familiarity implicit in HTDL and HTRC allow us to connect emergent concepts with established ones.

## Conclusion

As TDM becomes more common in academic research, library staff will need a way to quickly get up to speed on TDM practices to assist patrons. We believe that the HathiTrust Digital Library and HathiTrust Research Center offer a strong set of affordances and low-barrier tools for providing such a foundation. As prior library technology training

approaches have shown, strategies that consider library workers' skills, needs, practical contexts, and prior knowledge are often the most successful. We believe that HTDL and HTRC, with their roots in library organization and their ease and familiarity of use, provide us with a preferred platform for foundational TDM training among library staff. While some products and platforms may have wider or more contemporary coverage for exploring deeper TDM questions and may output more dazzling and diverse visualizations, the ease-of-use and simple and reusable workflows of HTDL and HTRC offer library workers a chance for sufficient experience with the full range of TDM applications within a short time. In our experience teaching TDM workshops using HTDL and HTRC, we've found that emphasizing connections between library work and selecting, validating, processing, and interpreting their outputs has been a fruitful route to establishing a working knowledge of TDM.

Authors' note: During the preparation of this chapter, HathiTrust announced upcoming changes to the HathiTrust Research Center's funding and offerings. While there have been no specified changes to the features described in this chapter, users can expect a different HTRC experience after 2026, when the changes go into effect. The authors anticipate that, due to the theme of "bookness" present in how HathiTrust is designed, strategies outlined in the above chapter will continue to present the collection as a strong foundation for digital scholarship for Library Staff and Library Students in the future. For more information about the changes, including Frequently Asked Questions, please visit HathiTrust's Plans for the HathiTrust Research Center memo released in late 2024.<sup>21</sup>

## Notes

1. In this chapter, we refer to HathiTrust Digital Library and HathiTrust Research Center jointly as "HathiTrust."
2. "Our Mission & History," HathiTrust Digital Library, accessed January 16, 2024, <https://www.hathi-trust.org/about/mission-history/>.
3. Jacob Jett, Timothy W. Cole, Christopher Maden, and J. Stephen Downie, "The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections," *Journal of Open Humanities Data* 2 (0) (2016), <https://doi.org/10.5334/johd.3>.
4. Eleanor Dickson Koehl and Ryan Dubniecek, "Text Mining with HathiTrust," in *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2019, 451–52, <https://doi.org/10.1109/JCDL.2019.00115>.
5. Andromeda Yelton, "Chapter 1. Introduction," *Library Technology Reports* 51, no. 3 (April 6, 2015): 5–8.
6. James Baker et al., "Library Carpentry: Software Skills Training for Library Professionals," *LIBER Quarterly: The Journal of the Association of European Research Libraries* 26, no. 3 (November 28, 2016): 141–62, <https://doi.org/10.18352/lq.10176>.
7. Koehl and Dubniecek, "Text Mining with HathiTrust"; Ryan Dubniecek and Deren Kudeki, "Introduction to and Hands-On Use Cases with HathiTrust Research Center's Extracted Features 2.0 Dataset," in *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2021, 352–53, <https://doi.org/10.1109/JCDL52503.2021.00073>; Sarah Sutton and Kelly Swickard, "Text Mining 101," *The Serials Librarian* 78, no. 1–4 (June 1, 2020): 3–8, <https://doi.org/10.1080/0361526X.2020.1715775>.
8. David Bainbridge, David M. Nichols, Annika Hinze, and J. Stephen Downie, "Using the HTRC Data Capsule Model to Promote Reuse and Evolution of Experimental Analysis of Digital Library Data: A Case Study of Topic Modeling," in *Proceedings of the 18th Joint Conference on Digital Libraries, JCDL '19* (Champaign, IL: IEEE Press, 2020), 463–64, <https://doi.org/10.1109/JCDL.2019.00124>.
9. Nikolaus Nova Parulian and Glen Worthey, "Identifying Creative Content at the Page Level in the HathiTrust Digital Library Using Machine Learning Methods on Text and Image Features," in

- Diversity, Divergence, Dialogue*, ed. Katharina Toeppe, Hui Yan, and Samuel Kai Wah Chu, Lecture Notes in Computer Science (Cham: Springer International Publishing, 2021), 478–89, [https://doi.org/10.1007/978-3-030-71292-1\\_37](https://doi.org/10.1007/978-3-030-71292-1_37).
10. Stéfán Sinclair and Geoffrey Rockwell, “Voyant: See Through Your Text,” Voyant Tools, 2024, <https://voyant-tools.org/>.
  11. “Data Visualizations,” Sage Research Methods, 2024, <https://methods.sagepub.com/data-visualization>.
  12. “Our Lessons,” Library Carpentry, 2024, <https://librarycarpentry.org/lessons/>.
  13. Jez Cope and James Baker, “Library Carpentry: Software Skills Training for Library Professionals,” *International Journal of Digital Curation* 12, no. 2 (December 30, 2017): 266–73, <https://doi.org/10.2218/ijdc.v12i2.576>.
  14. “Member Libraries – HathiTrust Digital Library,” n.d., accessed January 16, 2024, <https://www.hathitrust.org/member-libraries/>.
  15. “Constellate: Explore Text Analysis for Research,” ITHAKA, 2024, <https://constellate.org/docs/exploring-text-analysis-for-research-video>.
  16. “TDM Studio Text and Data mining Solution,” ProQuest, 2024, <https://tdmstudio.proquest.com/home>.
  17. “Removing Barriers to Digital Scholarship,” Gale, 2024, <https://www.gale.com/primary-sources/digital-scholar-lab>.
  18. The Early American Cookbooks Collection is available at <https://babel.hathitrust.org/cgi/mb?a=listis;c=1934413200>; the African American Fiction Collection is available at <https://babel.hathitrust.org/cgi/mb?a=listis;c=51682948>.
  19. The small Victorian Serialized Novels Collection is available at <https://babel.hathitrust.org/cgi/mb?a=listis&c=195250793>; the Five Presidents Collection is available at <https://babel.hathitrust.org/cgi/mb?a=listis&c=1211455659>; and the SU History Collection is available at <https://babel.hathitrust.org/cgi/mb?a=listis&c=924748279>.
  20. Patrick Williams and Rachel Hogan, “DH Workshop: Introduction to Text Mining with HathiTrust Research Center: Home,” Syracuse University Libraries, accessed April 23, 2024, [https://researchguides.library.syr.edu/htrc\\_tm](https://researchguides.library.syr.edu/htrc_tm).
  21. Mike Furlough, “Plans for the HathiTrust Research Center,” HathiTrust, accessed May 15, 2024, <https://www.hathitrust.org/press-post/plans-for-hathitrust-research-center/>

## Bibliography

- Bainbridge, David, David M. Nichols, Annika Hinze, and J. Stephen Downie. “Using the HTRC Data Capsule Model to Promote Reuse and Evolution of Experimental Analysis of Digital Library Data: A Case Study of Topic Modeling.” In *Proceedings of the 18th Joint Conference on Digital Libraries*, 463–64. JCDL ’19. Champaign, IL: IEEE Press, 2020. <https://doi.org/10.1109/JCDL.2019.00124>.
- Baker, James, Caitlin Moore, Ernesto Priego, Raquel Alegre, Jez Cope, Ludi Price, Owen Stephens, Daniel van Strien, and Greg Wilson. “Library Carpentry: Software Skills Training for Library Professionals.” *LIBER Quarterly: The Journal of the Association of European Research Libraries* 26 (3) (2016): 141–62. <https://doi.org/10.18352/lq.10176>.
- Cope, Jez, and James Baker. “Library Carpentry: Software Skills Training for Library Professionals.” *International Journal of Digital Curation* 12 (2) (2017): 266–73. <https://doi.org/10.2218/ijdc.v12i2.576>.
- Dubnick, Ryan, and Deren Kudeki. “Introduction to and Hands-On Use Cases with HathiTrust Research Center’s Extracted Features 2.0 Dataset.” In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2021, 352–53. <https://doi.org/10.1109/JCDL52503.2021.00073>.
- Furlough, Mike. “Plans for the HathiTrust Research Center,” 2024. <https://www.hathitrust.org/press-post/plans-for-hathitrust-research-center>
- Gale. “Removing Barriers to Digital Scholarship.” 2024. <https://www.gale.com/primary-sources/digital-scholar-lab>.
- HathiTrust Digital Library. “Member Libraries.” n.d. Accessed January 16, 2024. <https://www.hathitrust.org/member-libraries/>.
- . “Our Mission & History.” 2021. <https://www.hathitrust.org/about/mission-history>.

- ITHAKA. “Constellate: Explore Text Analysis for Research.” 2024. <https://constellate.org/docs/exploring-text-analysis-for-research-video>.
- Jett, Jacob, Timothy W. Cole, Christopher Maden, and J. Stephen Downie. “The HathiTrust Research Center Workset Ontology: A Descriptive Framework for Non-Consumptive Research Collections.” *Journal of Open Humanities Data* 2 (0) (2016). <https://doi.org/10.5334/johd.3>.
- Koehl, Eleanor Dickson, and Ryan Dubnick. “Text Mining with HathiTrust.” In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2019, 451–52. <https://doi.org/10.1109/JCDL.2019.00115>.
- Library Carpentry. “Our Lessons.” 2024. <https://librarycarpentry.org/lessons/>.
- Parulian, Nikolaus Nova, and Glen Worthey. “Identifying Creative Content at the Page Level in the HathiTrust Digital Library Using Machine Learning Methods on Text and Image Features.” In *Diversity, Divergence, Dialogue*, edited by Katharina Toeppe, Hui Yan, and Samuel Kai Wah Chu, 478–89. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021. [https://doi.org/10.1007/978-3-030-71292-1\\_37](https://doi.org/10.1007/978-3-030-71292-1_37).
- ProQuest. “TDM Studio Text and Data mining Solution.” 2024. <https://tdmstudio.proquest.com/home>.
- SAGE. Sage Research Methods. “Data Visualizations.” 2024. <https://methods.sagepub.com/data-visualization>.
- Sinclair, Stéfan, and Geoffrey Rockwell. “Voyant: See Through Your Text.” Voyant. 2024. <https://voyant-tools.org/>.
- Sutton, Sarah, and Kelly Swickard. “Text Mining 101.” *The Serials Librarian* 78 (1–4) (2020): 3–8. <https://doi.org/10.1080/0361526X.2020.1715775>.
- Williams, Patrick, and Rachel Hogan. “DH Workshop: Introduction to Text Mining with HathiTrust Research Center: Home.” Accessed April 23, 2024. [https://researchguides.library.syr.edu/htrc\\_tm](https://researchguides.library.syr.edu/htrc_tm)
- Yelton, Andromeda. “Chapter 1. Introduction.” *Library Technology Reports* 51 (3) (2015): 5–8.

