Ellery Bruns

Data Librarianship and Management

Professor Vicky Rampin

December 12, 2024

How Much Do We Share? A Hathitrust V. Pratt Libraries Collection Comparison

Introduction

As a Reference Graduate Assistant at Pratt Institute Libraries (Pratt), I participate in collection evaluation projects. These projects involve a process of identifying resources in a collection for withdrawal or retention to maintain the collections' relevance and quality for patrons (Gregory, 2011; Kohn, 2015). Pratt uses a resource' 3rd party availability in WorldCat, the New York Public Library (NYPL), or Hathitrust Digital Library (Hathitrust) as a criterion in collection evaluation. If a resource has a digital copy, it's ideal that the resource is fully viewable and downloadable from the website providing access.

Initially, I attempted to find the accuracy of WorldCat's data on 3rd party institutions' holdings for Pratt's resources. However, credential restrictions made the necessary data inaccessible. Since collection evaluation procedures at Pratt frequently use Hathitrust—a large, free, digital library of content contributed by libraries around the world—to identify digital availability of print resources, I pivoted focus. Inspired by Pratt's collection evaluation procedures, my research aims to find resources within Pratt's print collection that are viewable and fully downloadable from Hathitrust and find, among the shared items, the institutions that provided the digital copies to Hathitrust.

Methodology

Data Collection

My research required the use of quantitative analysis and descriptive statistics, for which I acquired 4 datasets.

The first is Pratt's collection's data. Associate Director for Collections Management, Johanna Bauman, queried Pratt's library management system to create a dataset with bibliographic information for print resources with OCLC numbers, sending the resulting CSV file to me on October 8, 2024 (Bauman, 2024). Bauman refined the dataset to include only resources with OCLC numbers and to exclude rare, archival, or periodical resources. Because the initial project considered resources recorded in WorldCat shared between Pratt and the NYPL, these refinements aligned with the former project's scope. Pratt usually only lists resources on WorldCat that have an OCLC number, and Pratt infrequently deaccessions rare or archival materials (J. Bauman, personal communication, 2024). The dataset also excludes periodicals because utilizing a standard identifier would not be sufficient to identify shared issues or volume holdings between institutions. Upon the scope change, I utilize Bauman's (2024) dataset to compare the holdings between Hathitrust and Pratt. Since the current research is framed by deselection procedures, the dataset's refinements do not prohibit investigation of the current research, though they pose limitations to the data and drawn conclusions must fall within these limitations.

The other three datasets are from Hathitrust. Each month, Hathitrust uploads to its website a TSV file containing bibliographic, rights, and access information for every resource in its collection (Hathitrust Digital library [Hathitrust], 2024a). As it's licensed under the CCO Public Domain Dedication (Hathitrust, 2024a), I downloaded the most current file at the time: the November 1, 2024, "hathi_full_20211101.txt.gz" (Hathitrust, 2024d). Because the file has no column names (Hathitrust, 2024a), I downloaded on November 4, 2024, a TXT file containing the necessary headers: "hathi_field_list.txt" (Hathitrust, 2022). The last dataset I downloaded from Hathitrust was "ht_institutions.tsv"; it provides the name of each institution that submitted resources to Hathitrust matched to a unique content provider code.

Data Cleaning

I used OpenRefine to clean and make the data as "Tidy" (Wickam, 2014), uniform, consistent, valid, and complete as necessary (Rampin, 2024). Common procedures performed on the datasets include consistently renaming columns, splitting multivalued cells, and filling down row values to keep corresponding data together. I also remove unnecessary columns and whitespace around cell values, normalize titles, and check for null and duplicate data. Null values are replaced with "NiV" (Not Informed Value) because investigation of the data made clear that while some values are uninformed, the data could exist. Absence of enough information to correct missing values that verifiably corresponds to what Pratt or Hathitrust would utilize prevents me from correcting missing values.

Each dataset required specific considerations. For Pratt's data, I created a subset to include only the OCLC number, title, and a normalized title. Because in analysis I would use the OCLC number as a shared identifier, I checked this column for duplicates. One emerged, for which information to correct it existed. The file's codebook—2024-11-23_results_pratt_print_collection_segment_v0001_clean_codebook.txt—records the event and includes permalinks to records establishing the values used to correct the data.

Hathitrust's institutional codes dataset was largely clean and tidy (Wickam, 2014). No duplicate codes existed. Institution names were unique, and no institution corresponded to more than one code, though OpenRefine was used to consistently format institution names.

Due to its massive size, I refined Hathitrust's collection data before cleaning it in OpenRefine. A Python script assigned column names from "hathi_field_list.txt" (Hathitrust, 2022) to the data, selected specific columns, and exported the subset as a CSV file. The project's research questions guided the selection of appropriate columns to keep: the "htid" (Hathitrust identifier), "access" (view status), "oclc_num" (OCLC number), "title," "bib_fmt" (bibliographic format), "content_provider_code" (the code for the institution that submitted the digital copy to Hathitrust), and "access_profile_code" (indicates download restrictions) (Hathtrust, 2024b). The script also filtered the data to only include bibliographic monographs that were fully viewable on and downloadable from Hathitrust, aligning with the scope of Pratt's data and my research inquiries.

OpenRefine analysis revealed multiple duplicate OCLC numbers within the subset of Hathitrust's data. Examining its layout made clear, however, that Hathitrust uses FRBR to record its data. When a resource or its volumes are provided to Hathitrust, each is assigned a new record and unique "htid," while the OCLC number(s) remain intact. Therefore, I did not deduplicate the OCLC numbers.

Analysis

With my clean and tidy data, I used OpenRefine to conduct comparative analysis. Referencing Wilson's (2014) and Lampron's (2023) instructions on joining datasets, I utilized a GREL function on Pratt's OCLC column to create a new column with values recording the total instances each OCLC number appeared in Hathitrust's data. If the number was greater than 0, a match existed. The quantity reflects the number of copies available in Hathitrust. With a similar GREL expression on the OCLC number column, I used Hathitrust's collection subset to find the unique content provider codes for each matched resource.

Verifying the data required several steps in OpenRefine. No instances of multiple OCLC numbers representing a single work among matched resources existed, so I verified matches by title comparison. Another GREL function added a column with Hathitrust's resource's titles into Pratt's data. I manually checked them against Pratt's titles to verify matches. In a duplicated column made for data verification, false matches had their respective new quantity-of-matches column updated to 0. Then, I adjusted the numbers in this new column for multi-volume works by finding discrepancies between the number of matches and the number of content providers for each resource. Where there were discrepancies, I used Hathitrust's website to search by "htid"s to determine if multiple instances of a work referred to multiple volumes or multiple full copies of a single work. If the former was true, I corrected the quantity of matches by counting distinct content providers to derive the adjusted value. When the latter was true, I similarly adjusted the new quantity-of-matches column. I also made a duplicate content provider column with GREL when faceted to resources identified to have multiple copies of a work provided by the same content provider. Then, I faceted to null values within the new column and merged the old and new content provider code columns together, simultaneously recording instances where content providers issued multiple full copies and retaining the unique codes for matched resources without this characteristic.

For later data visualization in Python, I created a final subset of the comparison data to split the multivalued cells in the content provider code column and listed out these institutions' full names and associated OCLC numbers.

Results

OpenRefine faceting on verified matches reveals 1,906 out of 135,976 resources in Pratt's print collection are fully viewable on and downloadable without restriction from Hathitrust. Remembering Pratt's dataset's limitations, this result does not include Pratt's resources without OCLC numbers, periodicals, archival or rare materials. Python script analysis shows this is 1.40% of Pratt's refined dataset (Figure one) and reveals that while 61.12% of these resources have one digital copy available in Hathitrust, 38.88% of these shared resources have multiple copies available. Some Pratt resources have five digital copies available in Hathitrust (Figure two).

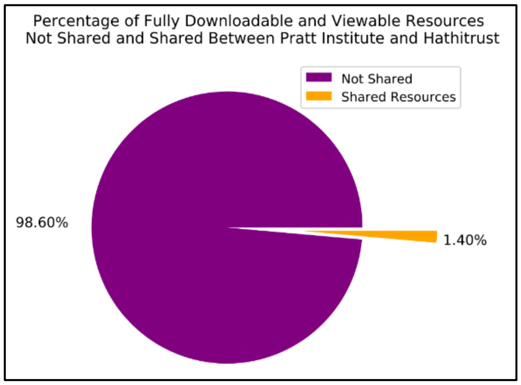

Figure 1

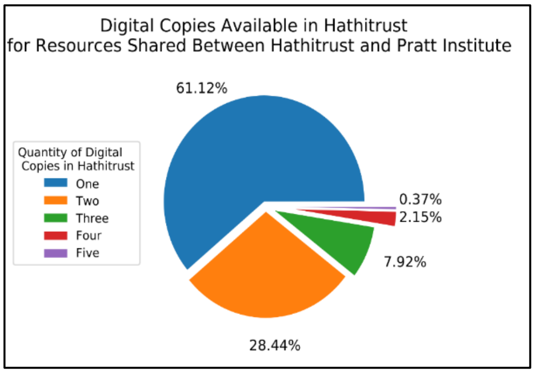

Figure 2

Another Python script found the total resources each content provider submitted to Hathitrust and visualized results in a horizontal bar graph and word cloud (Figures three and four). Figure four shows 23 libraries provided digital copies of shared resources between Hathitrust and Pratt. University of California, Cornell University, and Getty Research Institute are the top three institutions that provided the most digitized content shared among Pratt and Hathitrust.

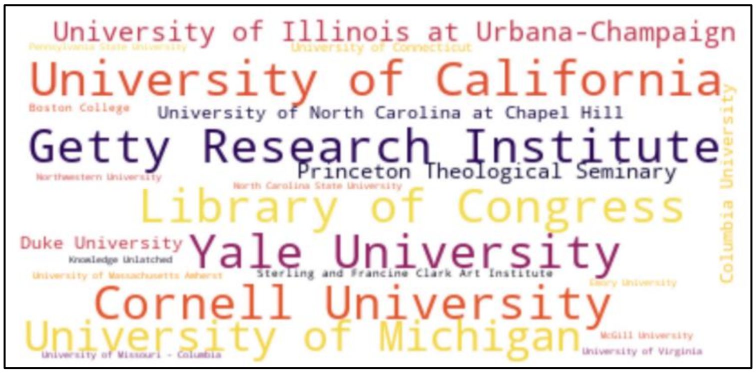

Figure 3

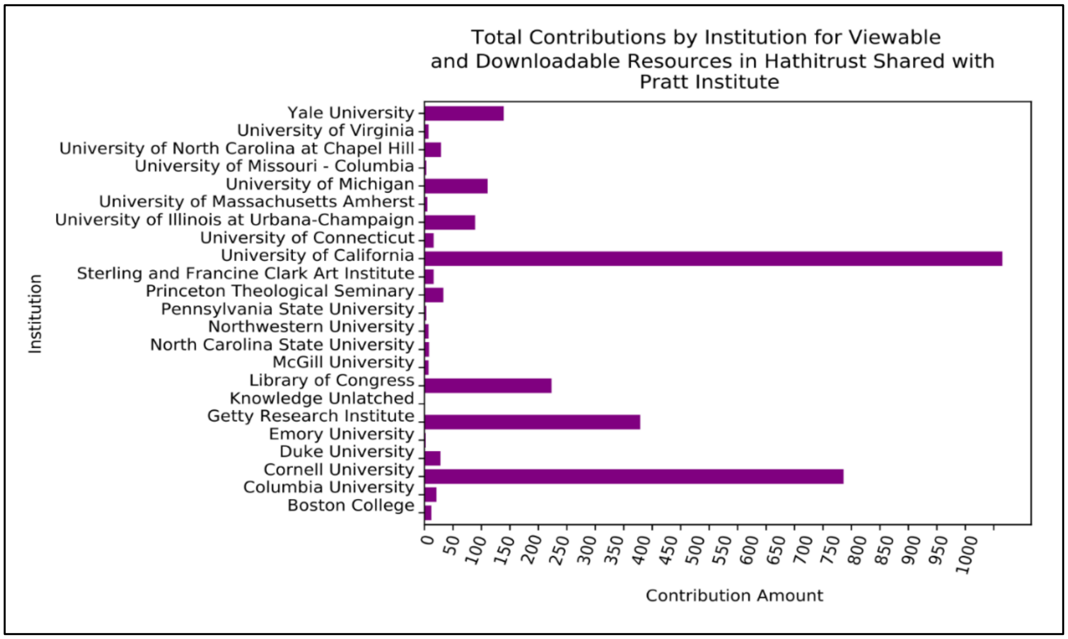

Figure 4

Conclusions, Limitations, and Avenues for Future Research

My research successfully found resources within Pratt's print collection that are viewable in Hathitrust without full-download restrictions and found the institutions providing these copies to Hathitrust. The results offer a contextual positioning of Pratt's non-archival, print collection in relation to Hathitrust which librarians can use to inform collection development decisions. Knowing at least 1,906 of Pratt's resources are accessible online in a high-quality, downloadable format, some of which have multiple copies, could impact decisions about retention or withdrawal of resources. It also indicates an additional mode of access for patrons in need of Pratt's print resources. Future research could repeat this research using various download or view statues for Hathitrust's resources to find additional matches, addressing a limitation of my research.

With Hathitrust as a common link, it's possible Pratt shares print materials with the 23 libraries that supplied digital versions of the shared resources to Hathitrust. Future research could see if these institutions retained print copies of their submitted resources. Because these shared resources have OCLC numbers, a study on 3rd party access for items cataloged in WorldCat could be conducted to find additional institutional access points, potentially impacting future retention of materials or indicating multiple modes of resource sharing for patrons.

Since Hathitrust regularly provides updates to its collections metadata (Hathitrust, 2024a), librarians could incorporate my methodology as an iterative workflow in policy. Significantly, this would enable efficient collection of evaluation data as both Hathitrust's and Pratt's libraries expand. Though it would require a systemization of my process through code, incorporating my methods into policy may also ease a significant burden of labor from librarians conducting collection evaluation projects. But, overall, both the results of this research and avenues for future research pose exciting possibilities for the future of Pratt's collection, patron's access to resources, and collection evaluation projects.

OSF Link: https://osf.io/qst5e/?view_only=ab8024e9026449daa53c4f0818eba639

References

Bauman, J. (2024). OCLCProjectExport_20241008 (Version no.1) [Dataset]. https://docs.google.com/spreadsheets/d/1jI7Z3G-VGKDypHgBnhmtAGvPymMaNXnxu9ijKrPj2jE/edit?usp=sharing. Accessed October 8, 2024.

Gregory, V. L. (2011). Collection development and management for 21st century library collections: An introduction. Neal-Schuman Publishers, Inc.

Hathitrust Digital Library (2022, April 1). hathi_field_list.txt.[Dataset]. Hathitrust. https://www.hathitrust.org/member-librariesresources-for-librarians/data-resources/hathifiles/. Accessed November, 4, 2024.

HathiTrust. Digital Library (2024a) *Hathifiles.* Retrieved October 21, 2024, from https://www.hathitrust.org/member- libraries/resources-for-librarians/data-resources/hathifiles/ .

Hathitrust Digital Library (2024b). *Hathifiles Description.* Retrieved November 4, 2024, from https://www.hathitrust.org/member-libraries/resources-for-librarians/data-resources/hathifiles/hathifiles-description/

Hathitrust Digital Library (2024c). ht_institutions.tsv. [Data set]. Hathitrust Digital Library. https://www.hathitrust.org/member-libraries/resources-for-librarians/institution-identifiers/. Accessed November 4, 2024.

Hathitrust Digital Library (2024d, November 1). hathi_full_20241101.txt.gz. [Data set]. Hathitrust Digital Library. https://www.hathitrust.org/member-libraries/resources-for-librarians/data-resources/hathifiles/. Accessed November 4, 2024.

Hathitrust Digital Library (2024e). *Institution Identifiers.* Retrieved November 4, 2024, from https://www.hathitrust.org/member-libraries/resources-for-librarians/institution-identifiers/ .

Kohn, K. C. (2015). Collection evaluation in academic libraries. Rowman & Littlefield.

Lampron. T. (2023). *OpenRefine: Joining Projects.* University of Illinois Urbana-Campaign.

https://guides.library.illinois.edu/openrefine/home

Rampin, V. (2024). "Week 5-Lecture Part 01." *Canvas.*

Wilson, K. (July 01, 2014). *Comparing Two Sets of Data in OpenRefine.* Open Library Foundation,https://openlibraryfoundation.atlassian.net/wiki/spaces/GOKB/pages/655657/Comparing+Two+Sets+of+Data+in+OpenRefine