

Introduction to HathiTrust and HTRC Tools for Text Data Mining

2023 University of Toronto TDM in Libraries Colloquium



HATHITRUST
research center

Presenter Information



Janet Swatscheno

Associate Director for Outreach and Education, HTRC

Digital Scholarship Librarian, HathiTrust

HathiTrust | <https://www.hathitrust.org>

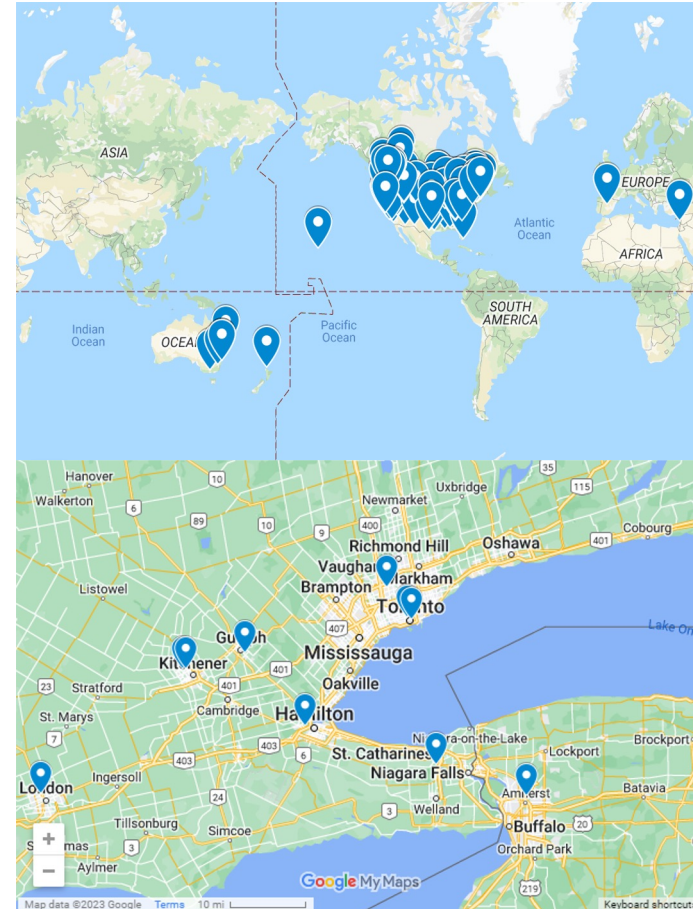
Email: jswatsch@hathitrust.org

Pronouns: she/her/hers

About HathiTrust

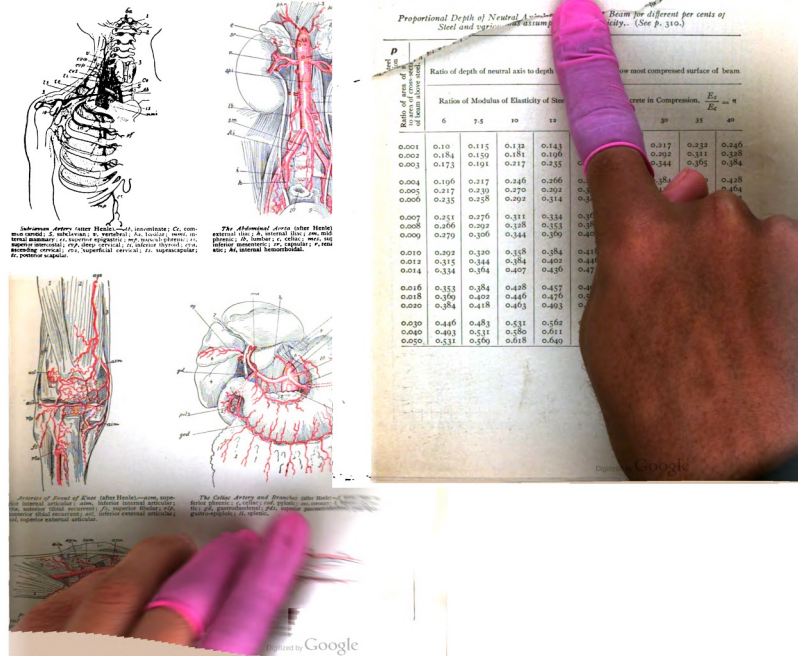
HathiTrust

- Non-profit academic partnership
- >200 member libraries
- Mission to support teaching, learning, scholarship
- Collaborative with members
- Balances *preservation* with *access*



Mass Digitization

- Minimizes curation
- Collection gathered in swathes of digitization
- 3+ billion page turns by thousands of scanning staff

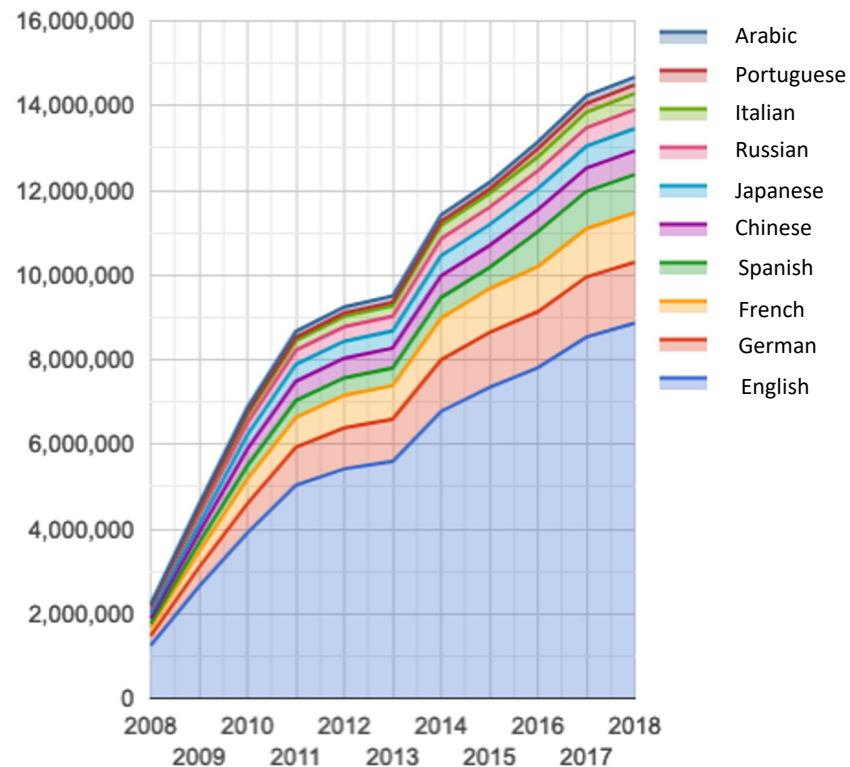


HathiTrust Digital Library

- 17+ million volumes
- Grows every day
- Composed of many sub-collections
 - Scripps Institute of Oceanography (> 100 thousand items)
 - University of California San Francisco
 - U.S. Federal Government documents



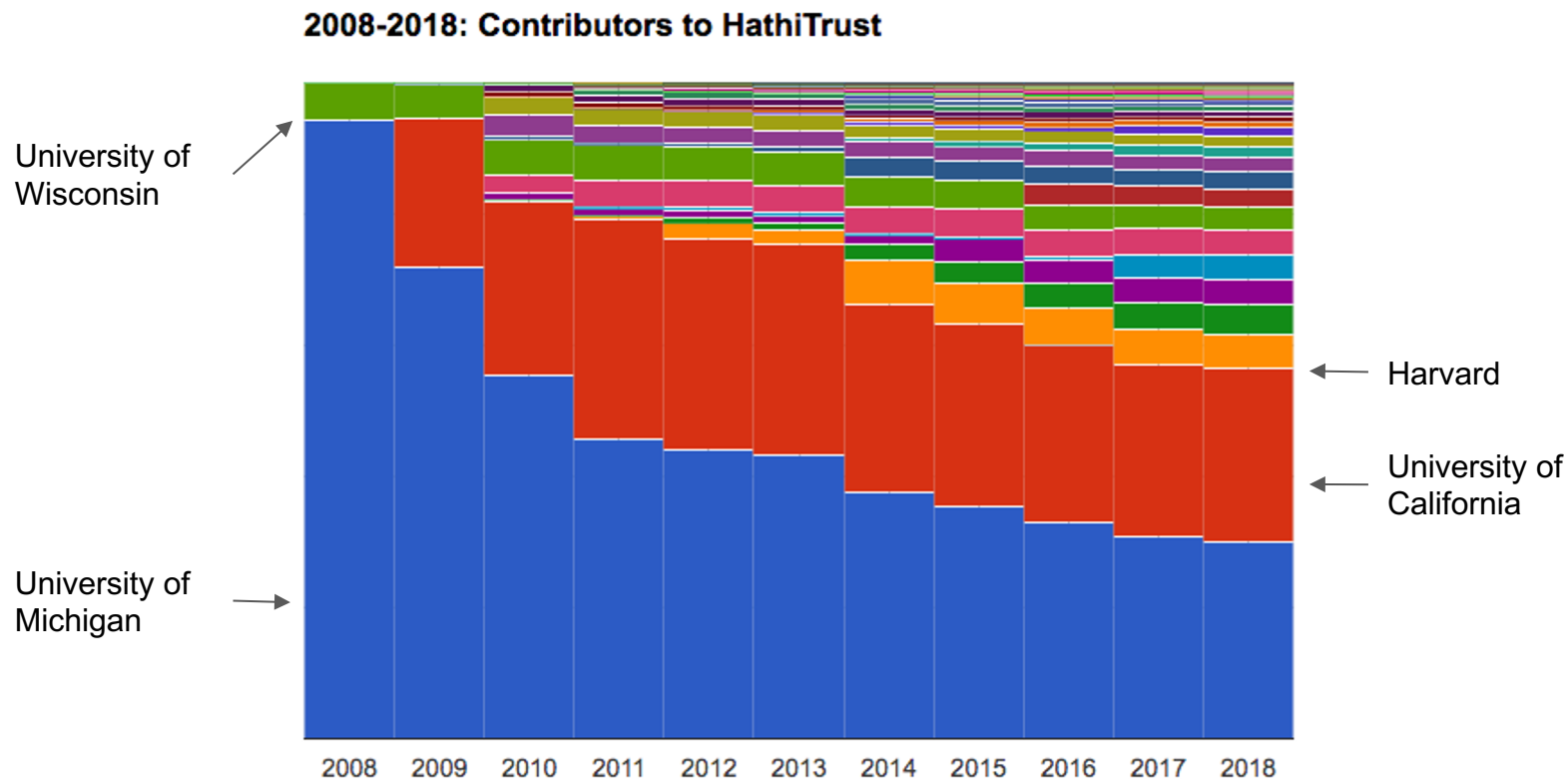
Publication dates for items in
HathiTrust



Languages for items in
HathiTrust



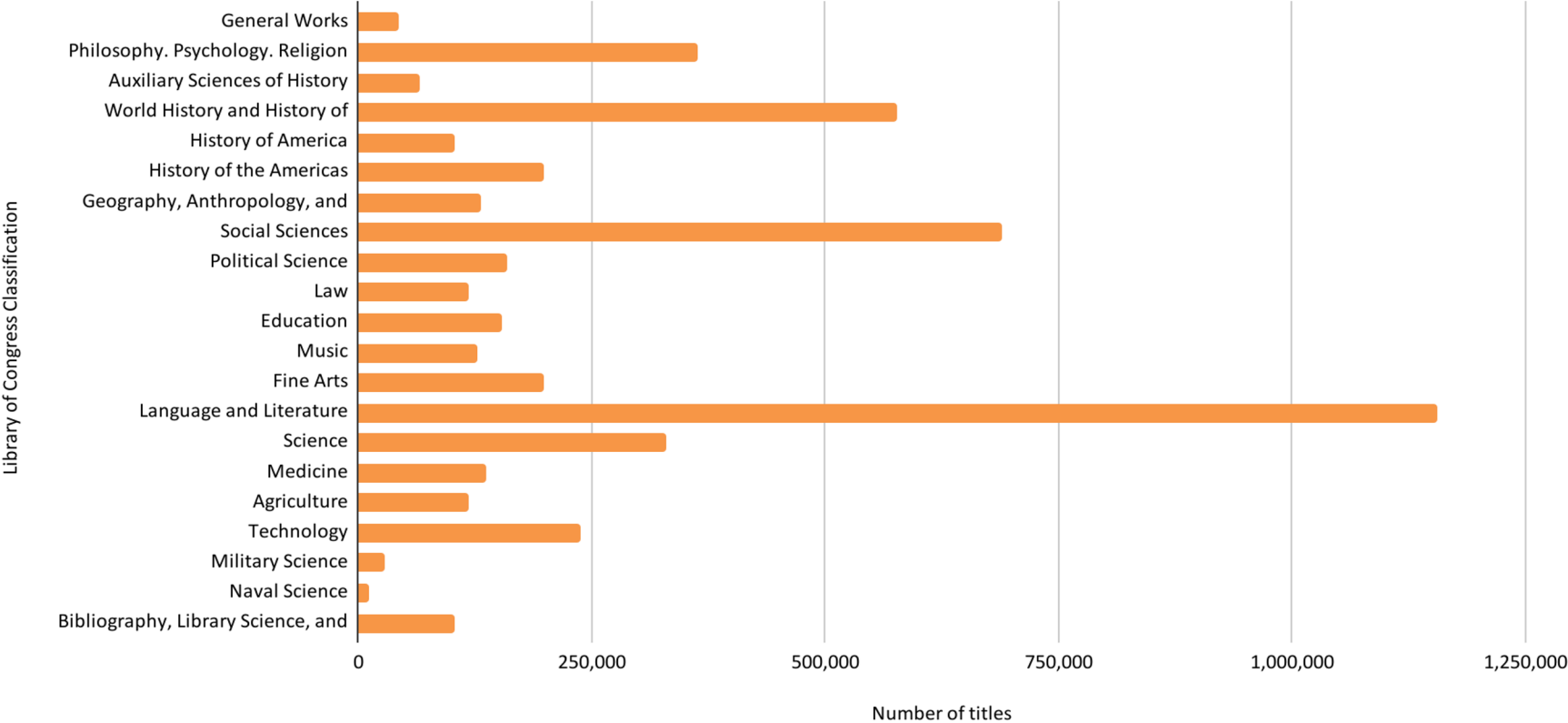
Member library contributions to the HathiTrust collection



From a snapshot analysis of the collection at: <https://is.gd/ht10collection>

Most recent contributor data at <https://www.hathitrust.org/visualizations> deposited volumes current

Subjects of HathiTrust books



HathiTrust Digital Library

- HathiTrust mirrors an academic library collection
- Library acquisition and publishing patterns impact content
 - Dearth of romance novels (Bode, 2019)
 - Gaps in speculative fiction, especially books by Black Female authors (Blecha, Kloster, and Gniady, 2020)
- Follows many library conventions

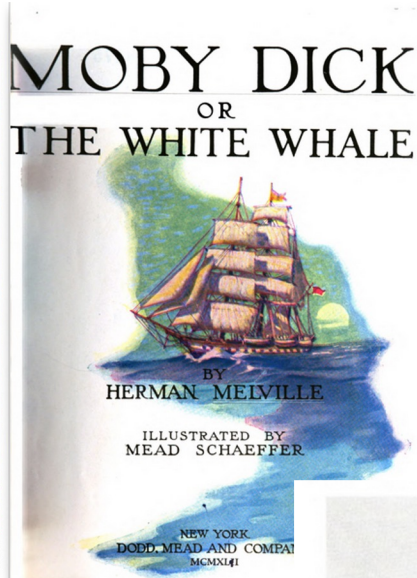
The cubby-bears in California, circa 1928
<https://hdl.handle.net/2027/uc1.b4594029>



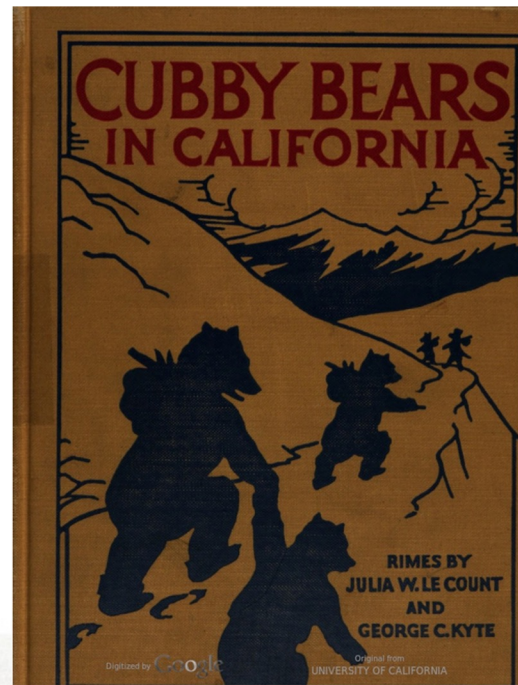
WE HANDED A GULL-LIKE CRY IN THE AIR. "THERE SHE BLOWS--THERE SHE BLOWS!"
 HUMP LIKE A SNOWHILL! IT IS MOBY DICK!" Page: 404

11

Moby Dick; or, The White Whale,
 illustrated by Mead Schaeffer, 1923
<https://hdl.handle.net/2027/mdp.49015002400035>



An Oklahoma family running for shelter during a storm in the 1930's. The Dust Bowl is an example of how America's agriculture practices have not always been harmonious with the environment.
 Arthur Rothstein/USDA CEN-170



*United States Department of
 Agriculture yearbook, 1991*
<https://hdl.handle.net/2027/mdp.39015022545605>



Diversified land use provides for conservation in many ways. Stripcropping, crop rotation, and pastures on steep slopes retard runoff and erosion on this farm in Carroll County, MD.
 Tim McCabe/USDA 0981X1234-32

About the HathiTrust Research Center (HTRC)


HathiTrust Research Center

- A virtual research center collaboration between **University of Illinois** and **Indiana University** with the mission to make the HathiTrust Digital Library as useful for research as possible while working within data and copyright limitations.
- Major initiatives center around:
 - Facilitating text and data mining (TDM) research with the HTDL
 - Instruction and user support for said TDM research
 - Research and development to design and create new tools, platforms, and datasets to support better and wider use of HTDL

HTRC Analytics Portal

HTRC Analytics Algorithms Data Capsules Worksets Datasets Explore

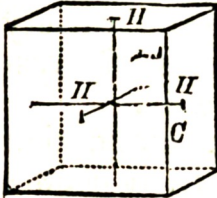
Help About Sign In Sign Up



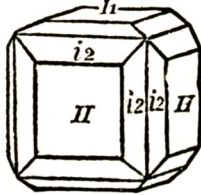
HathiTrust Research Center Analytics

Supports large-scale computational analysis of the works in the HathiTrust Digital Library to facilitate non-profit and educational research.


Featured Services



Extracted Features



Text Analysis Algorithms



Data Capsules

<https://analytics.hathitrust.org>



Non-consumptive research

Research in which computational analysis is performed on text, but not research in which a researcher reads or displays substantial portions of the text to understand the expressive content presented within it.

- Complies with copyright law
- Foundation of HTRC work
- Other terms: *non-expressive use*



Types of Non-Consumptive Research

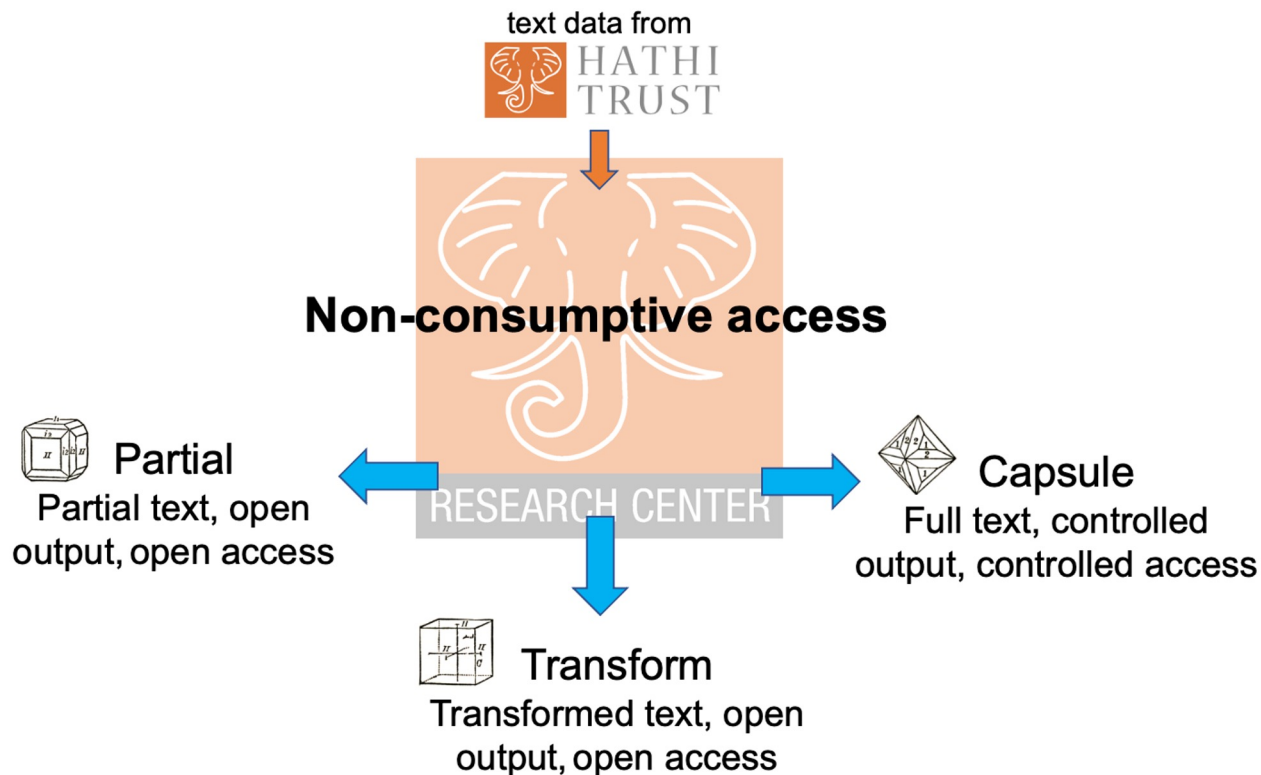
Includes such computational tasks as:

- text extraction
- textual analysis and information extraction
- linguistic analysis
- automated translation
- image analysis
- file manipulation
- OCR correction
- indexing and search

More here: https://www.hathitrust.org/htrc_ncup



Non-consumptive framework



Three approaches

Partial

- [Web-based tools](#): To analyze and visualize text data

Transform

- [Derived Datasets](#): Including Extracted Features dataset

Capsule

- [Secure Data Capsules](#): For flexible, self-directed research



Partial: HTRC Algorithms

- For analyzing HT data only
- Run against a workset
- Types of analysis:
 - Token count word cloud
 - Visualize themes (topic modeling)
 - List people, places, organizations, dates (named entity recognition)



Transform: Derived Datasets

- Preprocessed/transformed to be non-consumptive before access is provided
- Downloadable
- A means to analyzing in-copyright texts
- Output and access to the transformed data is open
- Open to all users, even non-HathiTrust members

Extracted Features File (for 1 volume from HathiTrust)

Volume metadata

Page features

Page 001

Word Frequencies in English-Language Literature, 1700-1922
Genre-specific wordcounts for 178,381 volumes from the HathiTrust Digital Library [v.0.1]

Description	Contents						
This dataset contains the word frequencies for all English-language volumes of fiction, drama, and poetry in the HathiTrust Digital Library from 1700 to 1922. Word counts are aggregated at the volume level, but include only pages tagged as belonging to the relevant literary genre. A full explanation of the dataset's features, motivation, and creation is available at the Genre dataset documentation page .	<table><tbody><tr><td>volumes of fiction</td><td>101,948</td></tr><tr><td>volumes of poetry</td><td>58,724</td></tr><tr><td>volumes of drama</td><td>17,709</td></tr></tbody></table>	volumes of fiction	101,948	volumes of poetry	58,724	volumes of drama	17,709
volumes of fiction	101,948						
volumes of poetry	58,724						
volumes of drama	17,709						
Download the data For each genre, we provide a metadata file, a corrections file, and a yearly summary, as well as tar.gz files that aggregate individual volume-level wordcount files, sorted by estimated date of publication. Here is a sample: Fiction, 1905-1909 . The full dataset can be downloaded from the documentation page .	Resources Spelling normalization and OCR correction Methods used for genre prediction Full documentation						
Attribution Ted Underwood, Boris Capitanu, Peter Organisciak, Sayan Bhattacharyya, Loretta Auvil, Colleen Fallaw, J. Stephen Downie (2015). <i>Word Frequencies in English-Language Literature, 1700-1922 (0.2) [Dataset]</i> . HathiTrust Research Center. http://dx.doi.org/10.13012/J8JW8BSJ .							

HTRC Extracted Features Dataset (EF)

The features are:

- Volume- and page-level
- Selected data and metadata
- Extracted from raw text

Position the researcher to begin analysis

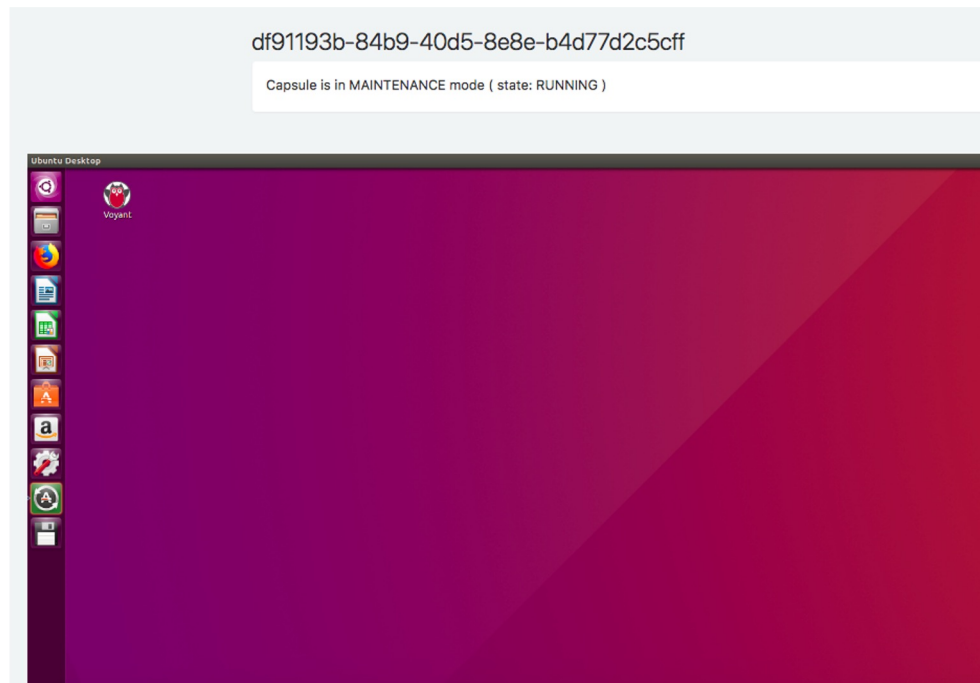
- Some of the preprocessing is already done

Form of non-consumptive access



Secure: Data Capsules

- A system of a secure computing environments for **performing researcher-driven text analysis** on the HathiTrust corpus
- All users may access public domain items
- In-copyright is available to **member-affiliated researchers**
- Out-of-the-box, Capsules are Ubuntu virtual machines with increased security settings
- Results are reviewed HTRC staff to prevent release of in-copyright data



Finding Textual Data

HathiTrust as Data

- Metadata
 - Primarily bibliographic (MARC) metadata
 - [Example](#)
 - Structural metadata (METS)
- Full text data
 - OCR text
 - Generated automatically during the digitization process
 - It's dirty (uncorrected)



HathiTrust data access options

Method	Data	Description	Rights status	Restrictions
HTRC Workset	Full text OCR	Analyzed using off-the-shelf algorithm	All	Data is not viewable
HTRC Data API	Full text OCR	Used to pull HathiTrust text into a Data Capsule	All for HT members by request	Use in Data Capsule only
HTRC Extracted Features	Abstracted text and metadata	JSON files for each of 17.1 million volumes in HathiTrust	All	Data is preprocessed
HT dataset request	Full text OCR	Download page images and plain text OCR	Public domain	Google-digitized by agreement only
HT Data API	Full text OCR, page images	Download page images and plain text OCR	Public domain	Non-Google digitized only



Dataset help

- Assistance crafting lists of volume IDs
- For HathiTrust custom dataset requests
 - support@hathitrust.org
- Or for HTRC worksets
 - htrc-help@hathitrust.org



Advanced Collaborative Support awards

- Competitively awarded “grants”
 - Time and resources awarded
- Available to HathiTrust Member Institutions
- New round of applications will open Fall 2023
- Read the descriptions and reports:

<https://wiki.htrc.illinois.edu/x/CADiAQ>



Documentation

- <https://wiki.htrc.illinois.edu/>
- Further information
- Technical documentation
- Step-by-step guides and video tutorials



Monthly Zoom Office Hours

- Every 3rd Wednesday from 3-4 p.m. ET
- Ask questions, connect with other researchers
- go.illinois.edu/htrchelp-join



Help email and listserv

- htrc-help@hathitrust.org
- For general inquiries, troubleshooting, and research consultations
- Join our listserv for general updates on HTRC tools and services:
 - htrc-announce-l@list.indiana.edu

