HTRC data access

Learn about the different access points and formats to the data HTRC provides, as well as the various affordances and limitations of each method. Your research project will largely dictate which method is best suited for your needs.

HTRC provides access to data from the HathiTrust corpus in several forms across its suite of tools and services for computational text analysis. Data is periodically synced from HathiTrust, but not all HTRC tools and services are updated on the same schedule. Additionally, copyright, user agreements, and security concerns impact data availability and format.

- HTRC algorithms and HTRC Data Capsules are capable of analyzing the entire HathiTrust corpus, and additionally make use of each volume's MARC bibliographic and METS metadata. Both the HTRC algorithms and Capsule-environments draw from the HTRC Data API described below.
- The HTRC makes available also two datasets, the HTRC Extracted Features Dataset and a dataset of Word Frequencies in English Language Literature, 1700-1922. HTRC Extracted Features includes metadata and extracted page-level data (words and word counts) for 15.7 million volumes.
- HathiTrust+Bookworm visualizes data for 15.7 million volumes.

HathiTrust text data

- Stats about the HTRC collection
- Stats about the HathiTrust corpus

Access options

The following table disambiguates textual data access, including availability and format, within the HTRC ecosystem.

Tool or service	# volumes available (as of 8/19)	rights status	data access mechanism	file format	data format	permissions required
HTRC Analytics algorithms	17 million	All	Via HTRC Workset; researcher runs tool without accessing underlying data	(no file access)	Uncorrected OCR text data; only results are exposed to researcher (not the underlying data)	HTRC Analytics account
HT+Bookworm tool	15.7 million	All	Via web-interface; researcher visualizes data without accessing underlying data	(no file access)	Unigrams (single words), based on HTRC Extracted Features dataset; underlying data not exposed to researcher	(none)
HTRC Data Capsule	6.5 million for everyone; 17 million for affiliates of HT member institutions upon approval or request	Public domain for everyone; In-copyright limited to affiliates of HT member institutions upon approval of request	HTRC Data API	Zipped text files in PairTree directory structure	Uncorrected OCR text data	HTRC Analytics account
HTRC Extracted Features dataset	15.7 million	All	rsync	JSON files in PairTree directory structure	Volume- and page-level metadata and part-of-speech tagged page-level "bags of words"	(none)
HathiTrust dataset request	6.5 million, dependent on institutional agreements	Public domain; accessibility of Google-digitized volumes dependent on whether researcher's home institution has signed agreement	rsync	Zipped text files in PairTree directory structure	Uncorrected OCR text data	Custom dataset request application
HathiTrust Data API	~800 thousand available; practical limit for retrieval is 10 thousand volumes	Public domain volumes not digitized by Google	HathiTrust Data API	Zipped text files in PairTree directory structure	Uncorrected OCR text data and page images	Key required to use the API outside of the Web client: read the documentation

Building a workset/dataset of HathiTrust text data

- Search and select volumes to build a collection in HathiTrust. Import to HTRC as a workset to use with algorithms or call data into your Capsule using the HTRC Workset Toolkit.
- Query the HathiTrust Bibliographic API.
- Utilize the HathiFiles.
- Need more help? HTRC can help you build a list of volume IDs to analyze. Contact htrc-help@hathitrust.org.

HathiTrust and HTRC APIs

This table outlines the differences between the HTRC Data API and HathiTrust Data API. As a researcher-accessible service, HTRC Data API functions within the HTRC Data Capsules environment.

	HTRC Data API	HathiTrust Data API
purpose	to provide access to HathiTrust text data within an HTRC Data Capsule AND to serve high-performance large-scale algorithms and programs (not publicly-accessible)	to provide public users some volume retrieval capabilities
data available	entire HathiTrust corpus	public domain volumes not digitized by Google
use	In-capsule via the HTRC Workset Toolkit	Get more information from HathiTrust
throttling enforcement	no	yes
security	JWT	OAuth
bulk retrieval of volumes	yes	up to 10,000 volumes
metadata available	METS	METS, MARC