About the Collection

Volumes

Items in HathiTrust are called volumes. A volume is a discrete object that was digitized and cataloged as one unit. In the case of HathiTrust and its collection, volumes are typically books (monographs), but they may also be one issue of a periodical, several issues of a periodical bound and described together, or even a musical score. Keep in mind that a volume may be an anthology containing multiple works! Currently, volumes in HathiTrust start as physical objects that are digitized and added to the HathiTrust Digital Library.

Volume Identification Numbers

HathiTrust volumes are identified via unique HathiTrust IDs. These alpha-numeric IDs track volumes across HathiTrust and HTRC systems. This volume of Jane Austen's letters has the volume ID hvd.32044021076179. When viewing a volume in the Digital Library, the volume ID can be found in the URL after "id=". The volume ID can be used to call metadata via the HathiTrust's Bibliographic API or to pull volume content via the HathiTrust's Data API. Additionally, the volume ID is often present in the file and/or directory name for content pertaining to a specific volume, and it also makes up the (pairtree) directory structure for volumes accessed via HathiTrust dataset requests or the HTRC Extracted Features Dataset.

HathiTrust volume IDs begin with a prefix code that identifies the library-of-origin (i.e. holding library) of the digitized item. For example, all volumes IDs that begin with **uiug** relate to objects held by the University of Illinois.

A list of volume IDs can be used to create an HTRC Workset, request a custom dataset from HathiTrust, or populate data in an HTRC Data Capsule, dependent, however, on the availability of those volumes in each system or service (see Copyright and Licenses below).

Catalog Record Identification Numbers

Catalog records, the library metadata for volumes, are also assigned unique identification numbers in HathiTrust. As a result of how libraries catalog material, some volumes share a catalog record with other volumes. Volumes may be cataloged under a single record if they are the same edition of the same work, in which case they may also share an identification number from OCLC, or if they are issues of a periodical or serial, such as issues of a magazine or annual publications from a government office. The record ID number can be helpful both for pulling all volumes associated with one intellectual entity (in the case of a serial) or to assist in weeding duplicative volumes (see Duplicates below). When viewing a catalog record in the Digital Library, record ID consists of the 9 digits following the final slash (e.g. 003910192).

Duplicates

There are many duplicative items in HathiTrust as some volumes were digitized multiple times, or they may be found in different forms in the Digital Library. Jane Austen's *Pride and Prejudice*, for example, is duplicated not only through repeat digitization of the same edition held by different partner libraries, but also through the digitization of different editions with unique paratextual content and of items such as Jane Austen anthologies. Researchers handle duplicates in different ways. For some, duplicates are considered either acceptable "noise" in their dataset or desirable representation of salience within the collection. Others prefer to remove duplicates, from their datasets prior to analysis. They do so either by hand-selecting volumes for the workset, or by programmatically winnowing items with shared record IDs, OCLC numbers, or author and title.

Data and OCR quality

Data quality in HathiTrust differs across volumes as well between HT and HTRC tools and services. OCR quality varies in part based on the original language of the volume, with modern English and European languages tending to be more accurate due to the software used for OCR. Read more: https://www.hathitrust.org/help_digital_library#OCRLikelihood. Additionally, data quality is impacted based on the parsers used in some HTRC tools and services. For example, the tokenization process used to create the current version of the HTRC Extracted Features–the data which underlies the HT+Bookworm too, as well–was not adept at CJK (Chinese, Japanese, and Korean) characters, and as such, the quality of the data in those languages in those services is quite low. Further, the HTRC off-the-shelf algorithms tend to handle modern English best due to the computational processes they rely on.

Many researchers would like to identify the "best" representative volume of a text they would like to include in their analysis. There is no right way to select volumes based on OCR quality without looking at the text, and HathiTrust and the HTRC currently do not have a quality score to facilitate selection based on OCR. A proxy for OCR quality that researchers have found helpful is to select volumes that were most recently digitized (as OCR processes have improved with time) by Google (as their OCR technique is considered to be high quality).

Metadata

While metadata for volumes in HathiTrust exists in a variety of formats and for a number of intended use cases, it generally begins as MARC metadata, the standard for library cataloging. It is often helpful to rely on the MARC specifications to navigate HathiTrust metadata for analysis, for example determining what certain codes mean or data structures imply. HathiTrust publishes specification for their metadata records that can be quite useful as there are HathiTrust-specific uses of some fields, particularly MARC field 975, that contain useful metadata about volumes: https://www.hathitrust.org/bib_specifications.

While HathiTrust does not facilitate bulk-download of full metadata records at this time, metadata is available in various formats and through several services that each can be useful depending on the use case:

- Hathifiles: tab-delimited files of reduced bibliographic metadata pulled from MARC records that are released daily for incremental additions to HathiTrust. On the first of each month, a file of every volume currently in HathiTrust is released.
- HathiTrust Bibliogrpahic API: for retrieving JSON-formatted MARC metadata via HathiTrust ID, HathiTrust record number, or OCLC number for up to 20 identifiers at a time.
- HTRC Extracted Features: volume-level JSON files include limited bibliographic metadata in addition to page-level metadata and features.

Additionally, this tables of MARC Coverage can help clarify the nature of the collection.

Copyright and Licenses

A volume's copyright and license status affects how researchers are permitted to interact with the data for that volume. Only public domain volumes are available via the HathiTrust Data API and custom data request process. Additonally, there are access restrictions based on the digitizing agent of volumes in HathiTrust that impact use if the HT Data API and custom data request procedures. Read more here: https://www.hathitrust.org/data. Only public domain volumes are available for research via the HTRC Analytics site and Data Capsulses compute environments. HTRC Extracted Features and the HT+Bookworm tool, however, do provide analytic access to derived data from the entire corpus.

This page features visualizations showcasing the nature of the post-1923 public domain volumes in HathiTrust, as including them in analysis will bias results due to the prevalence of content such as government documents.