HTRC Worksets

HTRC worksets are user-created collections of HathiTrust volumes to be treated as data and analyzed using HTRC tools and services. Worksets are curated by researchers, and they can be shared and cited to improve reproducibility.

Create and browse worksets Follow a Tutorial

Making a list/collection for your workset

One of the easiest way to generate a list of volume IDs for analysis is to first create a collection in the HathiTrust Digital Library. You can then bring your collection to HTRC as a workset using a one-click import

Here are the basic steps:

1. Select an existing or create a public collection in the HathiTrust Digital Library.
2. Import your public collection to HTRC by supplying the collection's unique URL (e.g. https://babel.hathitrust.org/cgi/mb?a=listis;c=490522989)
3. Add or edit the workset's metadata, and choose to make it private or leave it public.

Note: HTRC Analytics may display the order of your volumes differently than they appeared on the input file.

Other options for creating a workset include:

- Search for and select volumes in the (beta) Workset Builder 2.0 tool.
- Upload a list of volume IDs if you prefer to curate a list using other, more programmatic means. For example, you could work from the Word Frequencies in English Language Literature list, query the HathiTrust Bibliographic API, or make use of HathiFiles.

Workset Builder 2.0 Beta

The HTRC Workset Builder 2.0 Beta is a beta tool and interface built over the HTRC Extracted Features Dataset to enable both volume-level metadata search and volume- and page-level unigram (single word) text search in order to build worksets for use with HTRC tools and services. Read documentation for Workset Builder, or follow a tutorial for importing a workset from Workset Builder into HTRC Analytics.

The data

HTRC Algorithms can analyze volumes in a workset so long as they have been synched with HTRC from HathiTrust. While syncing happens regularly, there may be occasional discrepancies.

Worksets can contain in-copyright ("limited view") as well as public domain ("full view") volumes from HathiTrust. HathiTrust data is not exposed or viewable within HTRC Algorithms or worksets. A researcher applies an algorithm to their workset (collection) and the data is called and crunched behind the scenes.

Need more help?

HTRC can help you create a list of volume IDs if the options described above do not meet your needs. Contact htrc-help@hathitrust.org for assistance.

Workset format

Worksets start as lists of HathiTrust volume identification numbers (for example, hvd.hn5f64). If uploading a volume list file to create a workset, the file should be in CSV (comma-separated-value) or TXT format, and while it may contain other columns, it is only required to have your volume IDs in the first column. The file should contain a header row containing the text "volume" or "id".

volume_id
hvd.hn5f63
hvd.hn5f64
hvd.hn5f65
hvd.hn5f66

Workset toolkit

HTRC Workset Toolkit is a command line interface for use in the HTRC Data Capsule environment. It streamlines access to the HTRC Data API and includes utilities to pull text data and volume metadata into a capsule. Additionally, it allows a researcher to point OCR text data to analysis tools that are also available in the capsule.

It can also be used to manage volume IDs from HathiTrust collections, HathiTrust bibliographic record numbers, and more.

Additional documentation is also available here: https://htrc.github.io/HTRC-WorksetToolkit/cli.html