HTRC Worksets

What are 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. They are a foundational piece of all the work you will do in HTRC Analytics.

Create and browse worksets Follow a tutorial to create or validate a workset

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.

What can I do with a workset in HTRC Analytics?

Run algorithms. Worksets can be analyzed directly from HTRC Analytics using HTRC Algorithms.

Download and analyze extracted features. If you would like to download the Extracted Features of volumes in a collection for analyzing with your own code, you can do so easily using the Extracted Features Download Helper Algorithm, which will generate a shell script to download the Extracted Features for the volumes in a workset using rsync. (See more on downloading Extracted Features here.)

Access additional features in a data capsule. Volumes in a workset can be analyzed in the HTRC Data Capsule Environment using the command line interface HTRC Workset Toolkit. 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.

Workset Toolkit can also be used to manage volume IDs from HathiTrust collections, HathiTrust bibliographic record numbers, and more. Additional information is available in the Workset Toolkit documentation.

How do I build a workset?

See the Workset Tutorials page for instructions on the three methods of building a workset.

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.