HTRC, Help!

Need help? Browse anonymized and frequently-asked questions from the HTRC help email address below. Submit your own question to htrc-help@hathitrust.org. Or join us during the third Wednesday of each month at 3 ET for office hours at go.illinois.edu/htrchelp-live, where we have a Zoom conference line set-up for you to speak with a friendly expert from HTRC. You can also join office hours by phone using +1-669-900-6833 and meeting ID: 738781267 (International numbers available).

Upcoming office hours:

- August 21, 2019
- September 18, 2019
- October 16, 2019
- November 20, 2019
- December 18, 2019

Choosing a tool

Dear HTRC,

I'm pretty new to text analysis and heard HTRC might be a good resource for me because HathiTrust has materials I would like to study from the United States Environmental Protection Agency. How do I get started picking a tool for my research?

Sincerely,

Jane Researcher

***

Hello Jane,

There are a few factors to consider as you choose an HTRC tool or service for your research:

- How technically experienced are you?
- What are the rights statuses of the volumes you want to study?
- What research methods do you plan to use, and what form of access to the raw data does it require?

We can go through these 1-by-1:

- **Technical experience**

  HTRC's suite of tools and services are designed to meet the needs of a range of researchers. The easiest to use, requiring the least technical skills, are the web-based algorithms and visualization tools. These algorithms allow you to perform standard text analysis functions on a defined workset without having to program anything yourself. HathiTrust+Bookworm is another off-the-shelf tools for creating visualizations of word trends over time. You can pull any public collection from HathiTrust into HTRC and analyze public domain volumes with HTRC Algorithms. Or you can do metadata-based faceting in HT+Bookworm to narrow-in on volumes of interest.

  If, on the other hand, you feel like you have a decent command of a programming language, enough to manipulate data and carry out text analysis on your own, then you can turn to either the Data Capsule or HTRC Extracted Features. The Data Capsule environment will allow you to run the text analysis function of your choice on HTRC data in a secure environment. You will have control over how and which tools to use on the raw text, but you may only export derived data (i.e. results) from your capsule. The HTRC Extracted Features is a derived dataset that already exists for researcher use, and it contains enough data to perform bag-of-words analysis on your volumes, but not in a form that would allow semantic analysis or text reconstruction.

- **Rights considerations**

  For now, if your items are not in the public domain, the only provided mechanisms for analysis are the HTRC Extracted Features dataset and HathiTrust+Bookworm. Additionally, if you are interested in public domain texts but want to access them outside of HTRC, HathiTrust provides additional mechanisms for accessing datasets from the HathiTrust Digital Library: https://www.hathitrust.org/datasets.

- **Methods and data format**

  The methods you plan to use, and the data format they require, will also impact your choice in HTRC tool or service. For example, if you want a choice of predefined methods but do not need to do extensive data cleaning or parameter-setting, then choosing one of the HTRC Algorithms may be a good fit for you. (HT+Bookworm would be another option if want to track words over time in particular.)

  Additionally, if you keen to use Voyant Tools with HTRC data, then you can run Voyant from the Data Capsule. You will have to be comfortable using the command line in order to download your volumes of choice to your capsule, and you may want to have some knowledge of scripting if you will need to clean up your data before popping it into Voyant.

  If you are performing bag-of-words analysis or anything that requires primarily word and features counts, then the HTRC Extracted Features dataset will likely work for your needs. If you need complete control over your analytics process, then the HTRC Data Capsule secure compute environment will give you the flexibility you need.

This table summarizes the considerations you might want to have as you choose which tool or service is best for your needs:

HTRC Tool	Technical skills	Rights status	Methods	Data format
Web algorithms	Low	Public domain	Off-the-shelf	Can't see underlying data
HT+Bookworm tool	Low	All (13.7 million volumes)	Visualize trends	Can't see underlying data
Data Capsule environment	Medium to high	Public domain	Your choice, including Voyant	Raw OCR
Extracted Features dataset	Medium to high	All (15.7 million volumes)	Any requiring bag-of-words	Words and word counts in structured file

Finally, since you are studying the Environmental Protection Agency, you may also be interested to know that HathiTrust has created a collection (and corresponding HTRC workset) of volumes from the EPA. The collection can be found here: https://babel.hathitrust.org/cgi/ls?a=srchls;c=659388570, and the workset is called USEPA_Publications_2017Dec04 in HTRC Analytics.

Sincerely,

HTRC