HTRC Data Capsule Specifications and Usage Guide

HTRC Data Capsules are secure computing environments developed to facilitate non-consumptive text analysis research. Each Capsule is a virtual machine (VM) that provides researchers a desktop they can use to perform their investigation of volumes in the HathiTrust Digital Library.

Use a Capsule Follow a tutorial

HTRC Data Capsule Configurations

Kinds of Capsules

During creation, choose between a Demo Capsule, for testing and experimenting with the interface, or a Research Capsule, for conducting research.

Quick links

Non-consumptive Use Research Policy

HTRC Data Capsules Terms of Use

HTRC Data Capsule Step-by-Step Guides

HTRC Workset Toolkit

HTRC Algorithms documentation

Demo

- Capsule comes pre-loaded with sample volumes from the HathiTrust
- No options for Capsule size or specs
- Access to public domain corpus, only
- Results cannot be submitted for review to release
- No additional information required to create
- Expires after 30 days

Research

- Option for Capsule to come pre-loaded with sample volumes from the HathiTrust
- User can set the Capsule size (see 'Technical Specifications' below)
- By default, access to public domain corpus only
- All Research Capsules require additional information to create in order to aid in results export requests. Only the requests to create or convert to a capsule with full corpus access are subject to additional screening (as described above).
- **Members-only Benefit:** full corpus access for the Data Capsule service. Existing Data Capsule users from HathiTrust member institutions or new Data Capsule requesters from member institutions have the exclusive option to select “Full Corpus Access,” which includes copyrighted items.
    - Requests will be evaluated for demonstration of a legitimate research use, as well as understanding of the Non-consumptive Use Research Policy and HTRC Data Capsules Terms of Use.
- Expires 18 months from your last log-in date

Capsule Technical Specifications

Configuration options for Research Capsules:

- Data Capsule Image: there are two images (versions) of the standard Capsule desktop, one that comes pre-loaded with sample volumes from the HathiTrust and one that does not
- Virtual Machine CPUs (VCPUs): the number of virtual machine processors from 2-4 VCPUs for the Capsule
- Memory: between 4GB and 16GB

Shared Capsules

Data Capsules can be shared between up to 5 collaborators. The person who creates the Capsule has the most control over it, and they can add and remove other collaborators, assign permissions, and delete the Capsule.

There are 3 roles for users of a shared Capsule:

- **Owner (and Owner-Controller)**: By default each Capsule creator will get this role. It comes with the highest level of control. The Owner-Controller is able to perform all Capsule functions available in HTRC Analytics, including accessing, starting, stopping, switching modes, deleting, and managing the collaborators on the Capsule. By default the person who creates the Capsule will be the Owner-Controller until

they delegate control of the Capsule to a collaborator, at which point their role becomes Owner and the ability to start, stop, and switch the modes of the Capsule moves to the Controller (see below). The Owner can resume Owner-Controller status whenever they choose.
- **Contributor:** The Owner can share their Capsule with other HTRC Analytics users. New collaborators have the role of Contributor when they are added. This role has the lowest permission level. Contributors can connect to and conduct research in the Capsule, but cannot perform any of the Capsule management functions.
- **Controller:** The Owner can choose to give a Contributor the status of Controller in order to delegate some management tasks of the Capsule to that user, including starting, stopping, and switching modes. There can only be one Controller at a time, and the Owner can revoke control of the Capsule at any time.

Once a collaborator is added to a Capsule, the Capsule will appear for them on their Capsules listing page in HTRC Analytics. Before the new collaborator can access the Capsule, they will need to agree to the Data Capsules Terms of Use.

For Capsules with full-corpus access, HTRC will review the request to add a collaborator and either approve or deny it. The Capsule details will only appear on their Capsules listing page if the request is approved.

Pre-installed Software, Libraries, and Data

Each Capsule comes pre-loaded with the following software, libraries, and data. For more information, consult the ReadMe file on the desktop of your Capsule for more details about installed packages.

Software

Name	Version	URL	Note
Akka	2.4.14	http://akka.io/	
Anaconda 3	4.2.0	https://www.continuum.io/anaconda-overview	Supports both Python 2.X and 3.X. See list below for the Python libraries pre-installed (some via Anaconda)
Ant	1.9.7	http://ant.apache.org/	
Hadoop	2.7.3	http://hadoop.apache.org/	
InPho Topic Explorer		https://inpho.github.io/topic-explorer/	Project website: https://www.hypershelf.org/
Mallet	2.0.8	http://mallet.cs.umass.edu/	
R	3.3.9	https://www.r-project.org/	
Sbt	0.13.13	http://www.scala-sbt.org/	
Scala	2.12.1	https://www.scala-lang.org/	
Spark	2.0.2	http://spark.apache.org/	
Voyant Tools		https://voyant-tools.org/	

Python Libraries

HTRC-developed

- htrc-feature-reader
  Learn more
- htrc workset toolkit
  Learn more

General

- csvkit
- dask
- GenSim (currently running with warning)
- nltk
- numpy
- pandas
- pytables
- regex
- scipy
- theano
- toolz
- ujson

Sample data

- 3 sample HTRC worksets of 1000 volumes each: U.S. Government Documents, German language volumes, 19th Century English Literature.

User Quotas

There is an overall disk quota, a memory quota, and a CPU quota for each user in the Data Capsule environment. One user can consume up to 100 GB of disk space, ~20 GB of memory, and 10 CPUs. If you attempt to create a second or third Capsule that exceeds your quota in one of the areas above, then you will encounter an error.

Capsule Recall Practices

The HTRC Data Capsule service's maximum capacity flexes depending on the size of the Capsules it hosts. In the event that the Data Capsule service cannot satisfy all simultaneous demands for Capsules:

- A Capsule may be recalled (i.e. deleted) and the work environment will no longer exist.
    - Capsules will be identified for recall based on criteria such as date of last use and an individual's resource usage, with the goal being to extend the number of individuals afforded the opportunity to conduct research using a Capsule
    - A researcher whose Capsule is identified for recall will be notified via email regarding the pending recall, and they will have 5 days to respond to the recall notification.
- Priority in satisfying a new request for a Capsule will be given to researchers whose affiliated organization is a HathiTrust member.
    - At times when the Data Capsule service has reached capacity, incoming requests for Capsules will be screened based on institutional affiliation.
- Instructors who intend to use the HTRC Data Capsules in a course should contact htrc-help@hathitrust.org so that proper arrangements can be made.
- Users who do not abide by the Terms of Use will have their Capsule recalled immediately.

Using an HTRC Data Capsule

Follow a Tutorial

Administering a Capsule

Use the HTRC site to handle administrative tasks for your Capsule:

- Create - a Capsule is created, but it is not yet running
- Start - turn the Capsule on in maintenance mode
- Stop - shutdown a Capsule
- Delete - the Capsule is deleted (including its data and settings)
- Switch modes - change the Capsule from maintenance to secure mode, or vice-versa (see below)
- See status - view your Capsules and their statuses
- Interact - use your Capsule either through a desktop view or a terminal (command line) view

Maintenance vs. Secure Mode

The Capsules are configured with special security settings that allow you to interact with them in two modes: ***maintenance*** mode and ***secure*** mode

- In *maintenance* mode, you are allowed to access the network freely and install whatever software you want.
- In *secure* mode, general network access is restricted, but you can access the HTRC corpus repository, which is otherwise blocked. Any changes you make to the Capsule in secure mode will not persist. To save data from your analysis, you'll need to save your results in the *secure volume* storage on your Capsule. This storage option is not visible in maintenance mode.

Interacting with a Capsule

Access your Capsule in-browser from HTRC Analytics either by viewing the Remote Desktop (both modes available) or the Terminal command line interface (Maintenance Mode only). Earlier versions of the Capsule environment required a VNC viewer and passwords for both the VNC and the Capsule's operating system; those requirements are removed in the web-based version that was implemented in August, 2018.

You can also SSH into your Capsule in Maintenance Mode only if you've followed the directions under "Advanced Features" to set-up a public key.

To operate your Capsule, click on the Capsule ID from the Capsule list page. Then choose to either view the remote desktop or the terminal. The terminal will work in Maintenance Mode only.

If you've established a key for SSH access, you can also SSH into your Capsule when it's in Maintenance Mode by using the command viewable under "Advanced Features" on an individual Capsule's status page.

Importing data to a Capsule

Use the HTRC Workset Toolkit to import HathiTrust text data into your Capsule. Any outside data you plan to analyze in conjunction with HathiTrust data can be added to your Capsule from a web-accessible location when your machine is in Maintenance Mode.

Learn more

Passwords

Earlier versions of the Capsule environment required passwords for both the VNC and the Capsule's operating system; those requirements are removed in the web-based version that was implemented in August, 2018. If you use the included HTRC Workset Toolkit when in your Capsule to import data to your Capsule, you will be prompted for your HTRC Analytics username and password.

Generic Research Workflow

1. Create and start a Capsule in the HTRC
2. View your Capsule using the Remote Desktop view or Terminal view.
3. Configure the software environment of the Capsule as needed. Download the scripts or programs you plan to use in your analysis
4. Switch Capsule to secure mode through HTRC
5. Run your against the secure HTRC corpus repository
6. Move your results to the *secure volume* storage on the Capsule
7. Switch Capsule back to maintenance mode to regain normal network access

Non-consumptive Exports from an HTRC Data Capsule

Data and tools can easily enter a user's Capsule, but anything leaving a Capsule must undergo review prior to release to the user. The guidelines used during review of the outputs of a Capsule are as follows:

- Files containing any OCR text or images of pages or volumes will be prohibited from leaving a Capsule.
- Binary files are prohibited from leaving a Capsule
- Encrypted files are prohibited from leaving a Capsule
- PDF files or other format of file that contains images of OCR text or text images are prohibited from leaving a Capsule
- For any Capsule results directory that exceeds 1 MB in total, the collection will not be released pending discussion with the Capsule owner

Read the policy

The general rule-of-thumb is whether the export would create a substitute for human-reading the original text. (The full Non-Consumptive Use Research Policy is also available for your reference.) If you would like someone to pre-review a sample file that would represent the kinds of data you would like to export from a capsule before you begin your work, please contact htrc-help@hathitrust.org.

A release request **must** be under 67 MB, and any submitted requests over this size will fail due to technical limitations.

Release requests should include a README text file describing the files included in the request and their data structure.

If you have a directory of results files that you would like to export to be released, you can zip the directory and export the compressed file.

You will receive an email notifying you if your results export has been approved. The link for downloading results that have been approved for export will appear on the landing page for your capsule. Each approved request will be available for 2 weeks from the approval date. All collaborators on a capsule will get notification that approved results are ready, and will find the released results available to them on their capsule landing page.