HTRC Data Capsule Tutorial

Overview of Generic Research Workflow in a Capsule

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

Step-by-Step Instructions

Register for an HTRC account if you do not already have one. Follow the steps in the "Sign Up and Sign In" link below for directions.

Sign Up

- Go to HTRC Analytics (https://analytics.hathitrust.org/)
- On the top right of the webpage, click on the "Sign Up" button.

- On the Sign Up page, enter requested information together with username you intend to use and password. The password must meet these requirements:
  - Password must be more than 15 characters long.
  - Password must contain characters from three of the following five categories:
    - Uppercase characters of European languages (A through Z, with diacritic marks, Greek and Cyrillic characters)
    - Lowercase characters of European languages (a through z, sharp-s, with diacritic marks, Greek and Cyrillic characters)
    - Base 10 digits (0 through 9)
    - Nonalphanumeric characters
    - Any Unicode character that is categorized as an alphabetic character but is not uppercase or lowercase. This includes Unicode characters from Asian languages.
  - Password must not contain any white spaces.
  - Password must not contain your user ID.
Trouble shooting: if you can't successfully sign up

You need to have an academic email to sign up. We maintain a growing list of allowed email domains, e.g. emails ending with .edu or edu.ac or edu.tw. If you find you can't successfully sign up with your academic email then it's probably your email domain is not on the list. In this situation, please submit an account request by clicking on the request an account button within the error message on the page.

Sign In

- You will receive an email in the email that you registered. Go to your email box, and follow the activation link in the email to activate your account.
- Now you have created an HTRC Analytics account, you can sign in. On the top right of the page, sign in with your username and password.
From the Analytics homepage, create a capsule by clicking on Data Capsules on the top menu. You will be asked to provide information about the capsule you would like to create. This step also explains how to create or convert an existing Capsule to one with access to the full HathiTrust corpus, for HathiTrust members only.

Create a Capsule

Your browser does not support the HTML5 video element

Check Capsule Status

Your browser does not support the HTML5 video element

Convert a Research Capsule

Your browser does not support the HTML5 video element

Navigate to the Capsule Creation Page

Navigate to "Data Capsules" on the top menu of HTRC Analytics.
Create a Capsule

Click *Create A Capsule* toward the top right.

You will be prompted to choose to create either a Demo Capsule or a Research Capsule.

**Create Demo Capsule**

*Note:*

- Demo Capsules are not configurable and can access public domain content only.
- You cannot request to export derived data from a Demo Capsule.

Hit the *Create Capsule* button. The capsule creation procedure usually takes about 1 minute to complete. Refresh your screen to see if it has finished.

You will be prompted to agree to the HTRC Data Capsules Terms of Use. Please review this document as it outlines policy for acceptable in-Capsule behavior.
Create Research Capsule

Note:

- Demo Capsules are configureable and by default can access public domain content only.
- You request to export derived data from a Research Capsule.
- Additional information is required to create your Capsule.
- During creation or after it's created, researchers from HathiTrust member institutions can request for their Research Capsule to be converted to one with computational access to the full HathiTrust corpus, including in-copyright content.

Fill out the form with the title of your research project, and choose the specs for your Capsule.

There are 2 images you can choose from, one that includes sample public domain volumes from HathiTrust for you to test with, and one that does not.

Capsule size can range from 2-4 VCPUs and from 4-16 GB of memory. The VCPUs and memory allocation you choose will affect the processing speed of your Capsule.

Add the description for your research project. These answers will be used to aid in reviewing requests to export results from your Capsule. The more information you can provide, the more easily we can assess your results for adherence to the HTRC’s Non-consumptive Use Research Policy.

Affiliates of HathiTrust member institutions can check the box to request a Capsule with access to the full HathiTrust corpus.
Checking that box will prompt you to fill out additional information about your project. Note: Creation requests from users who check this box will be routed for human review. Your request will be reviewed to verify that you are affiliated with a HathiTrust member institution and that your request demonstrates serious research intentions in compliance with the HTRC's Non-consumptive Use Research Policy and HTRC Data Capsules Terms of Use.

Include more information about your anticipated results to further assist in the human review of your data export requests.

If you like, you can choose to allow HTRC to communicate anonymized information about your research project. You must also agree that you will not share your log-in information for HTRC Analytics with anyone.

You will be prompted to agree to the HTRC Data Capsules Terms of Use. Please review this document as it outlines policy for acceptable in-Capsule behavior. You will be reminded of these terms regularly while using your Capsule.
Check Capsule Status

After creating a capsule, you will be taken back to the Capsules page. By default the capsule you just created is not running.

Convert a Research Capsule

HathiTrust member-affiliated individuals can request to convert existing Research Capsules into one with access to the full HathiTrust corpus.

From your Capsules page, click on the ID of the Capsule you would like to convert. Then, click the button to Request access to Full HathiTrust Corpus.
You will be taken to the Capsule creation form. If you submit answers when creating your Capsule, they will appear for you to review and, if desired, edit. You will also be asked to fill in additional information about your research use case. Your request will be reviewed to verify that you are affiliated with a HathiTrust member institutions and that your request demonstrates serious research intentions in compliance with the HTRC's Non-consumptive Use Research Policy and HTRC Data Capsules Terms of Use.

Start the capsule you created by clicking the Start Capsule button on the Capsules page.

On the Capsules page (found under Data Capsules on the top menu), click on the Start Capsule button.

Interact with the capsule either via Remote Desktop viewer or Terminal viewer.

To get the details of your capsule so that you can log in to it, click on Data Capsule ID link on the Capsules page.
You will see the details for your capsule. From this page, you can start, stop, or delete your capsule. You can also click to connect via Terminal (command line interface) or Remote Desktop (to see your capsule's Ubuntu desktop).

If you choose connect via Terminal, you will be take to a page showing a command line interface to interact with your capsule. (Note: This option available in Maintenance Mode only.)
If you choose to connect via Remote Desktop, you will be taken to a page from which you can interact with your capsule's desktop. (Note: This option is available in either Maintenance or Secure Mode.)

Watch the following video for rules for copying and pasting text/commands in the data capsule environment:

Your browser does not support the HTML5 video element

Overview of rules

- Regardless of your local operating system, use ctrl-c and ctrl-v to copy and paste IN the capsule.
- AND if you are using the terminal IN your capsule, regardless of your local operating system, use ctrl-shift-c and ctrl-shift-v
- You cannot copy and paste from your local desktop into the capsule window AND you cannot copy and paste from your capsule to your local desktop.
If you wanted to interact with your capsule via SSH from your personal machine, you can follow the directions to set up that access.

Alternatively, you can SSH into your capsule when it is in maintenance mode only.

Your browser does not support the HTML5 video element

First, you will need a public key. Click "Advanced Features" in the blue box to establish your public key at the bottom of your Capsules page.

You will be prompted for a key. If you do not yet have a public key set up, then entering one will establish your key. If you already have a key, resubmitting a response in this box will change your key.

You’ll find the command to SSH into your capsule in the blue “Advanced Features” box on each capsule’s status page.

Switch between maintenance and secure mode.

HTRC Data Capsules have two modes: Maintenance Mode and Secure Mode. In Maintenance Mode, the capsule can access the network (i.e. the internet) so that you can set up your capsule as you like, such as installing software or importing additional, non-HathiTrust data. In Secure Mode, the capsule can access HathiTrust Data.

HathiTrust data you import and/or work with in Secure Mode must be stored on the capsule’s Secure Volume, a storage location available in Secure Mode only, in order to persist in a capsule when it’s modes are switched or when it is turned off and back on. Data transferred or generated in the capsule in Secure Mode that is not saved to the Secure Volume will be deleted when the capsule switches modes or is turned off and on, for security reasons.

Your browser does not support the HTML5 video element

When viewing your capsule, you will see a blue button to either “Switch to Secure Mode” or “Switch to Maintenance Mode.”
Click the button to switch. You'll see the capsule's state change.

```
Capsule is in MAINTENANCE mode (state: SWITCH_TO_SECURE_PENDING)
```

Once it has switched, you'll see that you can click the blue button again to switch modes back.

```
Capsule is in SECURE mode (state: RUNNING)
```

Share your Research Data Capsule with up to 5 other researchers.

Your browser does not support the HTML5 video element

From your capsule listing page, click on the Data Capsule ID for the capsule you would like to share.

```
Data Capsules

<table>
<thead>
<tr>
<th>Data Capsule ID</th>
<th>Status</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>7ed263b-f39b-49ba-8a22-1f6d5e843e93</td>
<td>Mode: NOT_DEFINED</td>
<td>Start Capsule</td>
</tr>
<tr>
<td></td>
<td>Type: RESEARCH</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Role: OWNER_CONTROLLER</td>
<td></td>
</tr>
<tr>
<td>5ed288e1-57d7-47aa-9223-734de9e4162f</td>
<td>Mode: SECURE</td>
<td>Stop Capsule</td>
</tr>
<tr>
<td></td>
<td>Type: RESEARCH</td>
<td>Switch To Maintenance Mode</td>
</tr>
<tr>
<td></td>
<td>Role: OWNER_CONTROLLER</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Delete Capsule</td>
</tr>
</tbody>
</table>

In-capsule research is facilitated by the HTRC Workset Toolkit Python library, which comes pre-installed in all capsules created after March 18, 2018. For more information please refer the documentation.

Then, click the button that says Manage Collaborators.

```
7ed263b-f39b-49ba-8a22-1f6d5e843e93

Data Capsule ID: 7ed263b-f39b-49ba-8a22-1f6d5e843e93
Image Name: ubuntu-16-04
Capsule Type: RESEARCH
Memory[GB]: 8
VCPU: 4
Secure Volume Size: 10
State: SHUTDOWN
Capsule Mode: NOT_DEFINED
Role: OWNER_CONTROLLER
```
You will be taken to a new page, where you can input the email address for the user you would like to add.

The email address must be the one associated with their HTCC Analytics account or you will get an error.

When you successfully add a collaborator, that user's information will appear in the table of collaborators. By default, they will have the role of Contributor. Contributors can access the capsule and interact with it in its current state. You will have the role of Owner-Controller.
Before the new Contributor can access the capsule, they will need to agree to the Data Capsules Terms of Use. You will also be unable to delegate control of the capsule to them until they have agreed.

Once they have agreed to the Terms of Use, you can choose to make them a Controller of the capsule by clicking the "delegate control" button.

Once complete, you'll find that their role has changed to Controller. Only the Controller can start, stop, and switch the modes of the capsule. (The Owner-Controller likewise can do these tasks.)

Your role has changed to Owner. The owner can delete the capsule and revoke control from the Controller. Click on the "revoke control" button to resume Owner-Controller status.
Now the collaborator again has the role of Contributor and you are Owner-Controller.

If you no longer want to share your capsule with a user, click the red ‘X’ button.
When they are removed, you’ll see the collaborators table has returned to displaying only you as associated with this capsule.

Bring text data into your capsule.

Preferred method

HTRC has developed a Python library for loading volumes into the Data Capsule environment: the HTRC Workset Toolkit. The Toolkit is standard in all capsules created after March 18, 2018. If you have an earlier-created capsule then you will need to install or update the Toolkit.

Make sure you are in secure mode in order to prepare to fetch content into your Data Capsule; it won’t work in maintenance mode for security reasons.

You can use the Workset Toolkit’s "htrc download" command to transfer the volumes you would like to include in your dataset.

For example, the following command will import the volumes in the HathiTrust collection ‘Adventure Novels: G.A. Henty’.

```
htrc download 'https://babel.hathitrust.org/cgi/mb?a=listis;ö=c=464226859'
```

You can also curate a list of volumes whose data you would like to import by creating a file containing a HathiTrust volume ID list that you’re interested in, with one ID per line. Run the above command replacing the collection URL with your file name.

For example, if you had a file called myvolumes.txt, you would run the following command.

```
htrc download myvolumes.txt
```

In the above examples, the data will be transferred to “/media/secure_volume/workset/”. If you want to specify an alternative location, provide an output by including -o and the file path in your command.

Other functions of the Workset Toolkit

You can also use a volume ID, collection URL, or catalog record ID to import volumes. Additionally, you have the option to concatenate files, remove folders, and retrieve metadata using the functions of the Workset Toolkit.

For more examples, see the detailed guide.

For the technical documentation, see: https://htrc.github.io/HTRC-WorksetToolkit/cli.html

Optional method

Researchers can use the HTRC Data API to bring text data into their capsule, and can refer to the HTRC Data API guide for more details.
Perform your analysis. See the following use cases for examples of how to perform text analysis in the capsule.

- Running Voyant in an HTRC Data Capsule
- Perform Text Analytics Using Topic Explorer

If you will need more than one session to complete your research, save your interim data to the Secure Volume.

**Save data to the Secure Volume**

Make sure your capsule is in secure mode (see directions above if needed).

Open a terminal window in the capsule and navigate to the secure volume by typing:

```
cd /media/secure_volume
```

Between sessions, stop the capsule via the HTRC using the web browser on your personal desktop. The next time you log in, you can restart the same capsule and continue your work.

On the **Capsules** page (found under **Data Capsules** on the top menu), click on the Stop Capsule button.

When you are finished with your research, request to export your non-consumptive results. When you no longer need the capsule, delete it via the HTRC.

(This is the same as **Release results**)

If you'd like to export results out of the capsule, you must release them from your virtual machine.

First, switch the VM to secure mode in the Portal interface.

Second, open a terminal in the capsule, navigate to the secure volume by typing:

```
cd /media/secure_volume
```

Suppose the file you'd like to release is at `/home/demouser/demo/r/Rplots.pdf`

You can prepare the result data for release by first adding it to the release list:

```
releaseresults add /home/demouser/demo/r/Rplots.pdf
```

Repeat using this command if you have other files to add.
Finally, to complete the release of your data, type:

```
releaseresults done
```

The files will be delivered via email. You will receive them in the email that you registered for the HTRC Analytics account. The email link will be live for 12 hours.

**Shut Down the VM**

Go back to [HTRC Analytics](https://analytics.htrc.org), go to "Data Capsules" and click on the "Stop Capsule" button to shut down the VM you'd like to shut.

**Delete the VM**

If you do not need the capsule any more, you can delete it. On HTRC Analytics, navigate to "Data Capsules", click on the "Delete Capsule" button next to the virtual machine you want to delete.

**What to do if you need help troubleshooting problems with your data capsule:**

If you have questions about or are experiencing problems in your data capsule, you may submit a request for help directly from your data capsule. To locate the request help form, click on the dropdown menu located to the right of the screen labeled 'More Data Capsule Functions':
Click on the label ‘Request Data Capsule Help’.

You will be redirected to a form that you must fill out and submit.

Once you submit your form, a Jira help ticket will be created on your behalf and HTRC staff will be in touch with you via your institutional email on the status of your issue. We recommend putting in only one help request at a time before submitting a new form (i.e., wait for your initial problem to be addressed before putting in another request for help if another problem occurs in the same data capsule).
Questions?

- For questions not listed there, please contact htrc-help@hathitrust.org