Fetching Volume OCR Content in HTRC Data Capsule (Secure Mode)

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;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.