MODULE 2.2 Gathering Textual Data: Bulk Retrieval

KEY TOOLS & PLATFORMS

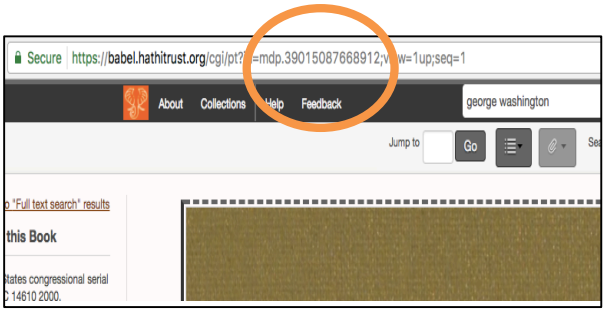
PythonAnywhere
A browser-based programming environment that's also a code editor and file hosting service. It comes with a built-in Bash shell and does not interact with your local file system.

wget
A command line tool for retrieving files from a server. It can scrape the contents of a website, with options that can be modified to tailor the parameters of the request.

ACTIVITY: Get experience with API calls

☞ ***Slide M2.2 – 10***

Can you use the HathiTrust Bibliographic API to retrieve the metadata for a volume of your choice?
More information: https://www.hathitrust.org/bib_api



1. Search HTDL for a volume: hathitrust.org
2. Click "Full view" or "Limited (search-only)". Which link is available depends on the volume's rights status.
3. Look at the URL of the page and find the volume ID. It should consist of all the characters between "id=" and ";".
4. The first characters of the identifier (i.e. pst or miua) are the code for the digitizing institution. An example of a volume ID is in **bold:**

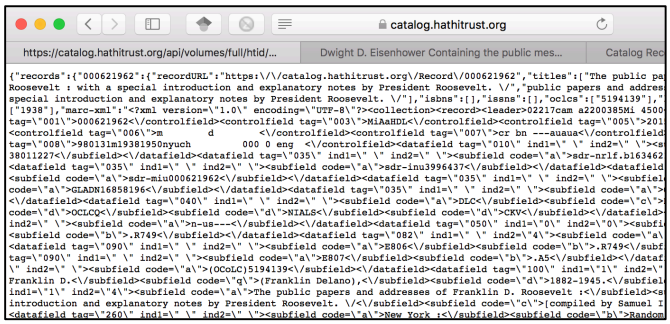
   https://babel.hathitrust.org/cgi/pt?id=**pst.000023137875**;view=1up;seq=9

   - An API call generally includes a structured URL. Here's the format for the HT Bibliographic API:

     http://catalog.hathitrust.org/api/volumes
     _____/brief **OR** /full
     _____/<ID type>
     _____/<ID.json>

     Example: https://catalog.hathitrust.org/api/volumes/full/htid/mdp.39015005337046.json

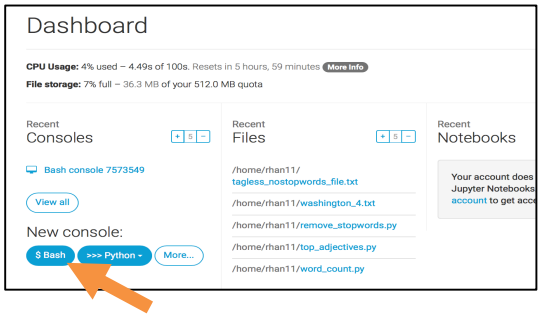
5. Type your structured URL into a new tab in your web browser and hit enter to call the metadata.



ACTIVITY: Introduction to the command line

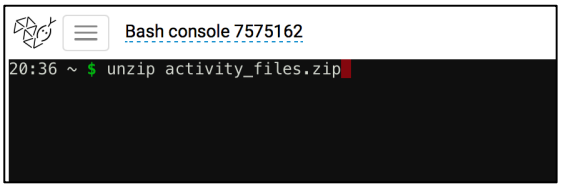
We will practice with PythonAnywhere's built-in Bash shell. It's useful for teaching and learning!

1. Start a new Bash console in PythonAnywhere by going to the Dashboard and clicking on the "$ Bash" button under "New console".



2. Try basic commands to:
   - See your working directory: Type and enter `pwd`
   - See what is in your working directory: Type and enter `ls`

3. Unzip and move the activity files we uploaded to PythonAnywhere at the start of the workshop.



   - To unzip activity_files.zip, type and enter:

     `unzip activity_files.zip`

   - To move the files to your user directory, type and enter:

     `mv activity_files/* /home/`*`USERNAME`*`/`

     Substitute your PythonAnywhere user name for `USERNAME`*!*

4. Use `ls` to check if the files were successfully unzipped and moved.

5. Practice more basic commands:
   - Make a directory called 'test': `mkdir test`
   - Change into and then back out of that directory:

     `cd test`

     `cd ..`

Tips for working in a shell

- **Directory = folder**
- **Case, spaces, and punctuation matter**
- **Tab to autocomplete a line**
- **Hit up/down arrow to see last commands entered**
- **Use "q" to quit viewing a file**

NOTE: *Make sure you are back in your main directory after finishing the activity.*

Command	Function	Example
pwd	see which directory you're in	`pwd`
mkdir	make a directory	`mkdir test`
cd	change directory	`cd /home/`*`your username`*
ls	list files and directories	
mv	rename a file	mv README.txt README1.txt
less (*Press "q" to quit)*	view contents of a file,	less FILENAME
touch	create a new, empty file	touch file1
nano	edit a file	nano file1.txt

ACTIVITY: Web scraping with wget

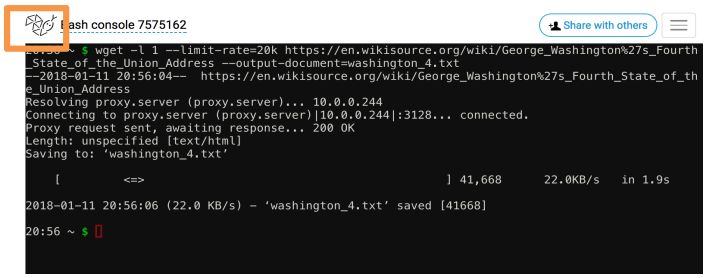
Practice web scraping to access a single text that may be of interest to our researcher.

1. Webpage to be scraped:
   https://en.wikisource.org/wiki/George_Washington%27s_Fourth_State_of_the_Union_Address
2. Make sure you are back in your main directory.
3. In your Bash shell, type the command:

```
wget -l 1 --limit-rate=20k https://en.wikisource.org/wiki/George_Washington%27s
_Fourth_State_of_the_Union_Address --output-document=washington_4.txt
```

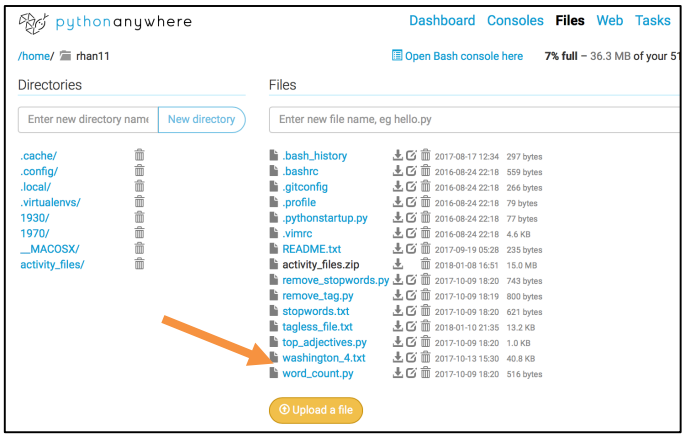
NOTE: *Ensure the apostrophe has been "escaped" using the code %27s as seen above.*

4. Click the "PythonAnywhere" logo in the top left corner to go back to the Dashboard.



5. Click on the "Browse files" button in the Files column, or click on the "Files" option in the upper right corner of the page.

6. Note your new washington_4.txt file in your list of files. This was the one just generated.



7. Click on the file name to view the file contents.

- Notice that all the html tags are still included.
- You could also view the file using the command `less washington_4.txt` in your console. (Press "q" to quit viewing when you finish.)

ACTIVITY: More web scraping with wget

☞ ***Slide M2.2 - 31***

- *Now you try! Can you modify the command to scrape George Washington's second State of the Union Speech?*
- *Can you view your scraped files using the* `less` *command?*

READ AND REFLECT: Collections as Data

☞ ***Slide M2.2 - 36***

Santa Barbara Statement on Collections as Data (Collections as Data National Forum, 2017)
https://collectionsasdata.github.io/statement/

Select points from the Statement:

- With a few exceptions, cultural heritage institutions have rarely built digital collections or designed access with the aim to support computational use. Thinking about collections as data signals an intention to change that.
- While the specifics of how to develop and provide access to collections as data will vary, any digital material can potentially be made available as data that are amenable to computational use. Use and reuse is encouraged by openly licensed data in non-proprietary formats made accessible via a range of access mechanisms that are designed to meet specific community needs.
- Ethical concerns are integral to collections as data.
- **Principle 2 for collections as data**: "Collections as data development aims to encourage computational use of digitized and born digital collections."

Discussion questions:

- *Does your library provide access to digital collections as data?*
- *How so? Why not? How could it?*