**Using data dumps:**

Big data is becoming ever prevalent on the internet, as most data has been generated in the past few years. On the OpenLibrary, the data here is large and ever growing. Here is how to work with these large files:

**Notes:**

* Use highest performance mode on computer before starting
* Highlighted means fill in
* Using software is slow and can slow one’s computer down, but is effective
* Be careful of how much is copy/pasted, as not everything might get copied past a certain size. This includes from the Preview Pane of Windows Explorer, Microsoft Office, and Git Bash.
* Use CTRL+C to ‘kill a command’, as [CTRL+INS is copy and shift+INS is paste](https://superuser.com/questions/421463/why-does-ctrl-v-not-paste-in-bash-linux-shell) in the terminal.
* Use the [Linux manual](https://linux.die.net/man/) to find a complete list of capabilities.
* **To view, search, and do minor edits:**
	+ Quick view of first lines: Windows Explorer Preview Pane
	+ Everything: HxD Hex Editor: [https://mh-nexus.de/en/hxd](https://mh-nexus.de/en/hxd/)
* **To split up into columns:**
	+ Use EmEditor: [https://www.emeditor.com](https://www.emeditor.com/)
	+ choose delimiter (the punctuation (periods, spaces, etc.)) as the column separator
		- On EmEditor: 3rd row where it says ‘CSV/Sort’
* **Major edits:**
	+ **To remove unwanted text (warning, will take a while, so use CTRL+C to ‘kill the command’):**
		- ***For Users or Non-Developers:***
			* On HxD: “Search” -> “Replace” and type in words to replace and what to replace it with (leave blank if there is none).
			* Save after each one + close and open the program
		- ***For Developers:***
			* On Git Bash: sed -i -e "s:oldstring:newstring:g" filename.txt
				+ Oldstring - fill in string that exists in the file
				+ Newstring - the string that would replace the previous one

Leave this blank if there’s nothing

* + - * + Filename.txt - the file that the word replacement would happen
				+ Note

could use / instead of :, but:

requires using \ in front of punctuation, like \”

can’t use if there’s spaces

to replace multiple strings, use ;, like "s:oldstring1:newstring1:g; s:oldstring2:newstring2:g"

* + **To extract lines only with desired content (like lines that only have a word in it) and put the findings in a new file (developers would call this "redirecting"):**
		- ***For Users or Non-Developers - step 1 - rows:***
			* Step 1 - Download desired file
				+ On the data dumps page: https://openlibrary.org/developers/dumps, right click to save desired file to desired location
				+ Note: less reliable than Git Bash, as it prioritizes the file download over other processes
			* Step 2 - Extract file
				+ Download 7-zip https://www.7-zip.org/ and click 'extract' to extract the text file inside the .gz file
				+ optional: move (or preferably copy) file out from the folder that 7-zip generated
			* Step 3 - Condense file to desired content
				+ use command line on MAC or Linux/Unix. If in Windows, download Git for Windows: https://git-scm.com/download/win to get Git Bash
				+ Type in cd "\_\_\_\_\_\_\_\_\_\_" - copy/paste path from Windows Explorer of folder location of downloaded file to use it
				+ copy/paste: grep "\_string\_" editions.txt > \_file name\_.txt

additional features:

multiple strings: grep -E "\_string1\_|\_string2\_" editions.txt > \_file name\_.txt

unknown characters: use “.”, like "\_\_ab.d\_\_"

Note: you fill in:

\_string\_: word or phrase sought

\_file name\_: new file with only the lines that include the \_string\_ in it

Note: copy/paste is easiest through highlighting and right clicking, but the shortcut for copy is ctrl+insert and paste is shift+insert

Note: this step may take a while, so the best options are to:

Use the highest performance setting on computer to speed up the process

wait (to see updates, close the window then press 'cancel'. It'll be done when the prompt line appears)

press ctrl+c to cancel the request

* + - * Step 4 - view completed product
				+ Cat filename.txt
				+ Optional: use HxD or even EmEditor to view new file named filename.txt
		- ***For Developers - step 1 - rows:***
			* Step 1 - Download software
				+ use command line on MAC or Linux/Unix. If in Windows, download Git for Windows: https://git-scm.com/download/win to get Git Bash
			* Step 2 - download file from internet
				+ curl -L https://openlibrary.org/data/ol\_dump\_editions\_latest.txt.gz --output editions.txt.gz
			* Step 3 - extract file
				+ Type in cd "\_\_\_\_\_\_\_\_\_\_" - copy/paste path from Windows Explorer of folder location of downloaded file to use it
				+ gzip -d editions.txt.gz
				+ Note: much easier to do in 7-zip
			* Step 4 - Condense file to desired content
				+ grep "\_string\_" editions.txt > \_file name\_.txt

Note: you fill in:

\_string\_: string sought

\_file name\_: new file with only the lines that include the \_string\_ in it

* + - ‘grep’ is for searches
			* + Note: copy/paste is easiest through highlighting and right clicking, but the shortcut for copy is ctrl+insert and paste is shift+insert
				+ Note: this step may take a while, so the best options are to:

Use the highest performance setting on computer to speed up the process

wait (to see updates, close the window then press 'cancel'. It'll be done when the prompt line appears)

press ctrl+c to cancel the request

* + - * Step 5 - view completed product
				+ Cat filename.txt
				+ Optional: use EmEditor to view by row