FILE FORMATS:

FILE TYPE	USE
.csv	Datasets (raw and cleaned)
.py	Code for processing data
.txt	README files, miscellaneous notes
.docx	Documentation, visualizations, presentation script
.xlsx	Datasets (raw and cleaned), documentation, visualizations
.json	English Short Title Catalogue original data
.html	Oxford Text Archive original data
.tar	Gutenberg original data archive
.rdf	Gutenberg original data files
.pdf	Documentation, class assignments
.ai / .jpg / .png	Visualizations? (data analysis, workflow)
.ipynb	Reproducible notebook
.pptx	Final presentation slides
.mp4	Final presentation recording

DATA STORAGE

- **LAPTOP:** This is my primary system of data storage and processing. It contains the most up-to-date version of my data.
- **GITHUB:** This is my secondary backup. I push changes at least every fifteen minutes or so when I'm writing code, less often if I'm writing documentation or notes. If my laptop fails, I should be able to download my working files back onto a new system.
- **BOX:** This is my tertiary backup. It contains all of my working files at slightly more spread out stages, as well as the larger source files that can't be housed on GitHub. Whenever I make significant progress, I'll drop the updated files into Box. If needed, I can download my versions from Box and pick up where I left off.
- **EXTERNAL HARD DRIVE:** This is my last resort backup. I cleaned and backed up all my laptop documents onto an external hard drive on October 18, including all my files for this project. I won't be updating this backup as often as the others, but I intend to update it at least at the end of each week. If needed, I can plug the drive into a new system and pick up where I left off.

DATA PRESERVATION/VERSIONING

I'm using Github as my primary means of managing and tracking changes. My current gitignore settings are as follows:

```
8 lines (6 sloc)  231 Bytes

1
2  dataset-inventory/project-gutenberg/2019-10-3/project-gutenberg_catalog.rdf
3  dataset-inventory/project-gutenberg/2019-10-5/project-gutenberg_catalog.rdf
4  dataset-inventory/project-gutenberg/2019-10-8/bookfiles.zip
5  *.rdf
6  *.rdf
7  *.rdf
```

This encompasses the Project Gutenberg data source files that I extracted from the website because they're too large to commit. Because I won't be transforming these files directly and I just need them as input to create my Gutenberg author counts dataset, keeping the originals in Box is sufficient to avoid any significant loss of progress over time. As long as I have my extraction code and these files backed up, I should be able to recreate the actual dataset if I lose it.

FINAL DATA FILE

My final data files will probably be a CSV and a PDF, and possibly an additional Excel/Word document. The CSV will contain the aggregated dataset of my top authors, listed from most significant to least. The PDF will include my analysis of the data and explanations of whatever calculations I do to produce the aggregate dataset. Depending on how far I get, I'll also include visualizations in the PDF and the Word/Excel document I used to make the visualizations.

DISSEMINATION

I don't have any concerns making the project public because all of my sources are publicly accessible (except maybe the data I pulled from ProQuest). I'm fine letting the data sit in GitHub until I need it for my larger project, but I'm open to dropping it in another repository if I want to present this project at a conference or showcase at some point in the future. If I do want to publicize my GitHub repo, I'll probably have to clean it up to make it more accessible for public viewing.

CURATION AND CLEANING

For most of my datasets, I'll have the raw data I extracted from my source, a cleaned version where I've consolidated author names or reorganized the data in some way, and a standardized version where I've transformed the data to fit a schema for later mashup. Every time I alter the data, I create a new file to hold the changed data, keep the original data file intact, and track the changes in Github.

- EEBO
    - **Cleaning:** Data was extracted manually in chunks and dropped into an Excel spreadsheet. I reorganized the data into one column, then consolidated repeated author names using OpenRefine.
    - **Documentation:** I've kept the original raw form of the chunked data and created a new file for the reorganized data. I also produced another file for the consolidated data. I plan on creating another file when I reformat the dataset according to my standardization schema.
- ESTC
    - **Cleaning:** Similar process as above – original extraction, consolidation, standardization.
    - **Documentation:** I've kept the original JSON file that I downloaded from HathiTrust. When I extracted the data using Python, I created a new dataset CSV. I then consolidated the author names using OpenRefine and created a new file. I plan on creating another file when I reformat the dataset according to my standardization schema.
- Open Syllabus
    - **Cleaning:** I extracted the data using XPATH and pasted it in an Excel spreadsheet. I won't be changing the extracted data until I reformat it according to my schema.
    - **Documentation:** When I reformat to fit the schema, I'll create a new file and keep the original extracted dataset file untouched.
- OTA
    - **Cleaning:** Similar process as above – original extraction, consolidation, standardization.
    - **Documentation:** As I extracted my data from the original HTML file using Python, I had to manually reformat some lines in the source file. I saved the changes as a new file and kept the original file intact. Again, I created a new file when I consolidated using OpenRefine and will create another file when I standardize.
- Project Gutenberg
    - **Cleaning:** Similar process as above – original extraction, consolidation, standardization.
    - **Documentation:** After I extract using Python, I'll keep the original extracted data and create new consolidated/standardized dataset files like above.

STAGES

Mapped out in data_processing_steps.xlsx

The spreadsheet is organized into two tabs: Stages and Weeks. The Stages tab is broken down by what I need to do at each stage for each dataset. The Weeks tab incorporates class assignments and spreads out the stage steps chronologically based on when I've finished them / when I intend to finish them. Depending on how smoothly Stages 5 & 6 go, I may need to push things back and subsequently abandon some of the final stages. I'm hoping to get at least some analysis done, but I'm fine if I don't manage to produce any visualizations.