1) **RESEARCH AREA:** English literary canon, large literary corpora, public reading vs academic reading

2) **FEATURES:**
    a. Which authors appear across multiple datasets
    b. How many times they are included (in datasets that count multiples)
    c. Similarities between authors along time period and gender

---

NAME: Oxford Text Archive author corpus

ATTRIBUTION: Oxford Text Archive

ACCESS: https://ota.ox.ac.uk/catalogue/index.html

WHY:

EFFORT: EASY/MODERATE

I viewed the page source, which has the entire catalog, and saved the HTML. Since the authors aren't explicitly tagged, I'll need to write a short program to work around the structure and extract the names. The structure is pretty consistent (see screenshot), so I don't anticipate having too much trouble writing a program that can pull what I need. There might also be an easier way to pull data that I'm not familiar with?

```
<thead><tr>
<th>ID</th>
<th>Title</th>
<th>Author</th>
<th>Date</th>
<th>Language</th>
<th>Availability</th>
<th>Genre</th>
</tr></thead>
<tr>
<td><a href="/id/3359">3359</a></td>
<td>The man of ten thousand: a comedy. As it is acted at the
Theatre-Royal, Drury-Lane. The second edition. By Thomas Holcroft.</td>
<td>Holcroft, Thomas, 1745-1809.</td>
<td>1796.</td>
<td>English</td>
<td>CC BY-SA</td>
<td></td>
</tr>
<tr>
<td><a href="/id/4017">4017</a></td>
<td>A letter from the Right Hon. Edmund Burke, M.P. in the kingdom of
Great Britain, to Sir Hercules Langrishe: Bart. M.P. on the subject of
Roman Catholics of Ireland, and the propriety of admitting them to the
elective franchise, consistently with the principles of the constitution
as established at the Revolution.</td>
<td>Burke, Edmund, 1729-1797.</td>
<td>1792.</td>
<td>English</td>
<td>CC BY-SA</td>
<td></td>
</tr>
```

NAME: Project Gutenberg author corpus

ATTRIBUTION: Project Gutenberg

ACCESS: https://www.gutenberg.org/

WHY: Since Project Gutenberg is volunteer-run, I think it will provide an interesting perspective on what gets preserved (and canonized) when readers are making the decision. Since PG's scope only covers out-of-copyright books, there will be fewer "modern" books that would serve as "noise" in the corpus I'm trying to create.

EFFORT: EASY/MODERATE

Project Gutenberg has an "Authors" page group that separates people alphabetically by surname, and all the names for each alphabet section are listed fully on each page. Based on some testing, I can use an XPath Helper to cleanly pull all the names out page-by-page with a query [`/html/body/div[@class='contents']/div[@class='body']/div[@class='pgdbbyauthor']/h2`], then add all the names to a spreadsheet.

NAME: English Short Title Catalogue (ESTC)

ATTRIBUTION: British Library / HathiTrust

ACCESS: https://babel.hathitrust.org/cgi/mb?a=listis&c=247770968 / http://estc.bl.uk

WHY: The ESTC has 480,000+ items of written works published in England between 1473 and 1800. This time frame places it squarely within the canon Johnson would have pulled from and offers the largest corpus of works from that period (that I've found so far).

EFFORT: MODERATE/HARD

I've found two versions of the ESTC online – the source published by the British Library, and a collection in HathiTrust. The ESTC search engine is similar to the Oxford Text Archive, but my "view page source" technique didn't work and I'm still fiddling with ways to pull out the data I need. If I'm unable to pull what I need efficiently from the original ESTC, the Hathitrust collection has part of the catalog (10,474 items). I've downloaded the metadata in a JSON file and can write a program to extract all the authors and the number of times they appear.

NAME: Early English Books Online

ATTRIBUTION:

ACCESS: http://eebo.chadwyck.com.proxy2.library.illinois.edu , https://search-proquest-com.proxy2.library.illinois.edu/eebo?accountid=14553

WHY: This is the largest corpus of titles I was able to find and probably one of the most famous corpuses for the literary time period I'm focusing on.

EFFORT: HARD

Pulling the data I want is proving fairly difficult due to the platform(s) it's house on and the sheer size of the dataset. I haven't found an efficient way to extract what I need, so this is the resource I'll most likely need to abandon. However, I'm still exploring ways to modify my search parameters and get a useful subset.

NAME: Wikipedia, writers born before 1775

ATTRIBUTION: WikiData Query Service

ACCESS: https://query.wikidata.org

WHY: This dataset will allow me to add context to my canon corpus and organize/group the author names / work titles by year. Since I'm hoping to cross-reference the canon corpus against my Johnson's Dictionary and OED corpuses later, I'll need to identify which authors could have feasibly been cited in which dictionary (i.e. any post-1750s canonical authors cannot not match up with Johnson's Dictionary and should not be factored into the analysis).

EFFORT: MODERATE

To extract my data, I'll need to learn how to use the WikiData Query Service to specify my parameters and form my dataset. WikiData already has classifications that will help me form my queries (occupation=writer, date of birth, etc.), I'll just need to learn how to write a query along these classifications. This may be my hardest dataset to develop because I've invested a bit of time trying to learn the Query Service and the syntax is still confusing to me.

NAME: Syllabus Authors

ATTRIBUTION: The Open Syllabus Project

ACCESS: https://opensyllabus.org/

WHY: This resource aligns perfectly with my “school reading lists” dataset idea. Their website states that they aggregate syllabus data from “6 million classes, 4700 schools, and 79 countries” and the most recent version of the Explorer was published in July 2019. This source will help me build a ‘taught canon’ dataset to compare against the other datasets I’m discovering/building.

EFFORT: MODERATE/HARD

I don’t think I’ll be able to pull the raw data from the site, so I’m planning on using the Explorer feature to identify relevant pages with data and use XPath to pull the author names. I’ve tested this on the https://opensyllabus.org/result/field?id=English+Literature page and figured out what XPath query would retrieve the author names and I’m hoping to repeat this for other pages. The “Authors” section of the Explorer also looks promising: the URL (https://opensyllabus.org/results-list/authors?size=) allows me to enter in a number to change the number of results, and I figured out an XPath query to cleanly slice out the names. The maximum number of results this method can pull is 5,000, which is 0.3% of the total author corpus (1.4 million authors) and 3.2% of the English Literature author corpus (156,489 authors).

NARRATIVE

For my larger research project, I want to explore the link between dictionaries and the literary canon through the cited illustrative quotations included in English-language dictionaries from the 1700s to the present. On the data side, this will involve comparing corpora of authors cited in English-language dictionaries to authors present in the literary canon. For this project, I'm hoping to cover the second part. I want to develop a literary canon corpus by comparing different literary corpora and finding commonalities between them. I also want to bring in another dataset of author information to contextualize the canon corpus and explore commonalities between authors.

I'm hoping to speak to Ted Underwood and get some help in identifying domain areas with relevant (and theoretically sound) datasets, but in the meantime I'm going by my own understanding of where 'canonical' literature is often held.

First, my search was guided by resources I've encountered in the past that address canonical literature. I started with an anthology I remember from high school, the *Norton Anthology of Literature*, but I wasn't able to find any indexes for any editions online (probably due to copyright). After some time, I took a break to explore digital libraries. I picked re3data.org from the Lecture 2 Sources slide and checked out a couple of databases from the Humanities and Literary Studies subject fields. Of the databases I looked at, the Oxford Text Archive (OTA) seemed the most straightforward.

After our discussion last class, I decided to use Wikipedia to build a dataset of author context (living/death dates, gender, etc.) that I can overlay on my canon aggregate. After fiddling with the WikiData Query Service for a bit, I took a break and looked around for digital repositories, libraries, and catalogs similar to the OTA. After some Googling, I found the English Short Title Catalogue. I remembered Project Gutenberg from IS452, and Early English Books Online came up in my 590DH class, so I decided to add those two to supplement the digital libraries side of my dataset group.

My dataset group was looking very heavy on the digital library side, so I decided to go back to finding a dataset to represent school data. After more Googling, I found an article on the Open Syllabus Project and visited their website. Their aggregated data seems perfect, and I'm optimistic that I can draw out at least something useful.

I'm pretty happy with the datasets I've collected so far, and I'm hoping to collect more after my meeting with Ted Underwood. I'm fairly confident I can extract the data I need. Although I graded most moderate or hard (because of the amount of manual and repetitive extraction or programming), I'm fairly familiar with the methods I want to use. I'm a bit concerned that my methods will take more time than I anticipate, but I think finding different datasets or researching new methods would take just as long, if not longer.