TECHNOLOGIES AND SKILLS

**Research Question:**

How do the authors cited in English-language dictionaries evolve over time, especially compared to the English literary canon?

My research interests lie in the development of the Western literary canon and the lasting impact of its historical exclusion of minority writers. I'm interested in using digital methods to explore who gets read, why, and how we can broaden that scope. My most recent project is positioned at the intersection between canonicity and lexicography, specifically English-language dictionaries.

Many contemporary language dictionaries use illustrative quotations from written sources to demonstrate the language's use in context. Each of these illustrative quotations is paired with an author citation, which includes the author's name and the title of the cited work. These citations form an explicit link between language authority and literary authority. Dictionaries are often regarded as supplemental texts in the literary sphere, but in reality they are fundamental to forming the very linguistic foundation that creates literature.

In the burgeoning print culture of 18th century British society, increased literacy brought along concerns of language degradation. Dictionaries acted as authorities in the language standardization process during this time, and Samuel Johnson effectively bound this process together with literature by incorporating literary quotations into his *Dictionary of the English Language* (1755). James Murray's *Oxford English Dictionary* (1884) continued Johnson's tradition of literary quotation evidence. By granting certain literary works a presence in his record of the English language, Johnson established the dictionary as a powerful evaluative force in both language standardization and literary canonization.

James Murray was tasked to replace and improve upon Johnson when he was given charge of the *Oxford English Dictionary* (1884), but he and the Oxford University Press made a fundamental mistake in claiming to sever themselves from the rich lexicographic history that Johnson established. The *OED* cannot claim neutrality any more than it can reject the dictionary's essential nature as a codifying text. The *OED*'s literary record is undeniably canonical and inescapably an influence of that canon. The *Oxford English Dictionary* ultimately sustained Johnson's canon through the 19th century and continues to uphold 18th century ideals of canonical literature as the reigning authority quotations dictionary today.

With this historical background in mind, I'm using digital humanities methods to explore how authoritative language texts like dictionaries are linked to the broader development of literary history. For the larger project, I will compare a corpus of authors cited in English-language dictionaries to a corpus of authors present in the literary canon. For the project in this class, I will aggregate different literary corpora datasets to identify trends in authorship attribution and build a representative corpus of author names with additional metadata on the authors and their works.

# DATASETS

**NAME:** Oxford Text Archive (OTA)

**AUTHORSHIP:** Oxford Text Archive

**ACCESS:** https://ota.ox.ac.uk/

**LICENSE:** All material is protected by copyright, with exception that "material may be duplicated by you for your research use or educational purposes in electronic or print form" with citation and attribution to the Oxford Text Archive. (https://ota.ox.ac.uk/documents/user_agreement.xml)

**VALUE:** Since the catalog is made for 'research and teaching' at Oxford, I think it works well to represent the academic side of literary representation. The type of research and teaching that would use this corpus are likely the same types of literary studies I'm hoping to analyze critically. These are the works deemed important enough to digitize and preserve for further study at Oxford, and therefore represent a canon within this particular (and prestigious) section of the academy.

**WEAKNESSES:** Author counts might be skewed a bit in this dataset because of how the catalog was constructed. Many titles are split into several parts, and each part counts as a record and an additional author attribution.

| | | |
|---|---|---|
| 3656 | Letters of Mr. Wycherley & Mr. Pope, from the year 1704 to 1710: [pt.2] | Pope, Alexander, 1688-1744. |
| 3655 | Letters of Mr. Wycherley & Mr. Pope, from the year 1704 to 1710: [pt.1] | Pope, Alexander, 1688-1744. |

Similarly, some work titles contain more than one record because the records are based on physical copies.

| | | |
|---|---|---|
| 4436 | Advice to all parties: By the author of The true-born English-man. | Defoe, Daniel, 1661?-1731. |
| 3332 | Advice to all parties: By the author of The true-born English-man. | Defoe, Daniel, 1661?-1731. |

Record 4436 is from the British Library, while Record 3332 is from the Harvard University Houghton Library. Furthermore, authorship attribution can get tricky with things like edited anthologies or translated works.

| | | | |
|---|---|---|---|
| 4379 | The Iliad: of Homer. Translated by Mr. Pope. [pt.6] | Homer. | 1715-20. |
| 4378 | The Iliad: of Homer. Translated by Mr. Pope. [pt.5] | Homer. | 1715-20. |
| 4377 | The Iliad: of Homer. Translated by Mr. Pope. [pt.4] | Homer. | 1715-20. |
| 4376 | The Iliad: of Homer. Translated by Mr. Pope. [pt.3] | Homer. | 1715-20. |

While the original work was written by Homer, Alexander Pope should be the attributed author in the literary history context I'm working within. I'll have to look more closely at how the count distributions turn out to see if I want to keep that element of context when I construct my final dataset or ignore it.

# DATASETS

**PROCESSING:** I've managed to extract all the author names and cleaned the entries a little bit to split up lines that had multiple names. The current CSV has about 2,760 rows and one column. Each author cell is formatted as “LastName, Firstname, birthyear-deathyear”.

| |
|---|
| Holcroft, Thomas, 1745-1809. |
| Burke, Edmund, 1729-1797. |
| Trusler, John, 1735-1820. |

The catalog is separated by work titles, so many authors are listed multiple times. I plan to write a program to consolidate the names into a master list with counts, then separate the last name, first name, and life dates into columns for better flexibility later on. The author cells aren’t all formatted perfectly – many lack some part of the metadata (no life/death dates, single name, etc.) so I’ll need to make sure my cleaning program can accommodate those differences. I’ll need to develop a schema to determine exactly how reformat the final cleaned dataset for future comparison against my other datasets.

---

**NAME:** Project Gutenberg

**AUTHORSHIP:** Project Gutenberg

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

**LICENSE:** Main site is intended for human users only. There are some additional resources for automated processing (mirror sites, index files, machine-readable database) and a roboting guidelines page with instructions. (https://www.gutenberg.org/wiki/Gutenberg:Terms_of_Use)

**VALUE:** Project Gutenberg welcomes volunteers to take charge of the submission process and expands control of the catalog content to the general public rather than just authorized academics or professionals. I think its data will offer a different perspectives than my other datasets, as it is a bit more representative of public reading habits and interests rather than just academic research/ education interests.

**WEAKNESSES:** I think it’s important to mention that the population of eBook submitters to Project Gutenberg is necessarily of a particular class: individuals with enough time, resources, technical knowledge, and personal interest to spend on the very long process of submitting an eBook. The Gutenberg wiki even warns that “creating an eBook is a lot of work, and Project Gutenberg's requirements are rather strict” (https://www.gutenberg.org/wiki/Gutenberg:Public_Domain_eBook_Submission_How-To) and directs volunteers to proofread rather than submit “if these steps seem daunting.” Thus, it’s disingenuous to imply that this is an entirely egalitarian platform that equally welcomes a truly diverse variety of submissions – structurally, it privileges

groups that may have particular (and similar) literary tastes. Nevertheless, its submission process is the most "open" of the datasets I have and the most "open" that I could find, given the fact that digital accessibility is a systemic issue that affects all digital repositories.

**PROCESSING:** I extracted all the author names from the site using XPATH on September 18th. The current CSV has 27 columns and each column has a varied number of rows, ranging from 31 to 3326, for a total of about 34,000 author cells. Each author cell is formatted as "LastName, Firstname, birthyear-deathyear".

| |
|---|
| Aaberg, J. C. (Jens Christian), 1877-1970 |
| Aakjær, Jeppe, 1866-1930 |
| Aalto, Ari, 1876-1938 |

This dataset has the same inconsistencies as the OTA dataset, which is actually encouraging – it seems that both datasets follow roughly the same metadata schema for recording author names, so I may be able to recycle the program I use for cleaning the OTA dataset and use it on this one. However, this dataset won't need a counting function because the catalog is separated works by author names and each name occurs as a single instance.

I'm not familiar enough with computing to know how much more time I'll need to process this dataset, so I may need a quick check of my program to make sure it's as efficient as possible before I run it.

---

**NAME:** English Short Title Catalogue (ESTC)

**AUTHORSHIP:** HathiTrust

**ACCESS:** https://babel.hathitrust.org/cgi/mb?a=listis&c=247770968

**LICENSE:** I only need the author metadata for the collection, not any of the collection items themselves. HathiTrust says that "Yes, you can download the metadata from any Collection Builder collection that you can view" (https://www.hathitrust.org/help_digital_library#CBDownload)

**VALUE:** This dataset is pretty similar to the EEBO dataset, but it's hosted by the British Library (a national library) rather than a commercial database (ProQuest). I can use this dataset to explore whether how databases with similar materials but different origins compare.

**WEAKNESSES:** Because of concessions I had to make during processing (detailed below), the coverage of my dataset is significantly smaller than the full scope of the catalog. I don't have much information on how exactly these particular volumes were

added to HathiTrust and what the selection parameters were. I've chosen to accept these shortcomings rather than abandon the dataset fully.

**PROCESSING:** My original inventory included the original ESTC. I found a blog post that details how to use Google Sheets to extract record information from the ESTC, but I don't think their method will be efficient enough for my needs. The collection supposedly has 480,000 total works, but even when I conducted a quick search for titles published between 1400-1750 (roughly the time period I'm working with for Johnson's dictionary) the collection still has 252,689 records. The original catalog seems resistant to scraping, so I think the HathiTrust collection will be a better investment of my time. It holds about 2% of the total catalog (10,474 records), which I think is sizable enough for what I need. I already have the JSON file downloaded and ready for extraction. The raw dataset is separated by work titles and includes multiple authors, so hopefully I can recycle the same program from the OTA dataset to clean this one once I extract all the author names and put them in a CSV. This dataset is about midway in size between the previous two, so I'm not sure how much time I'll need to process it.

---

**NAME:** EEBO

**AUTHORSHIP:** Early English Books Online

**ACCESS:** ProQuest

**LICENSE:** The terms of use state that "Customer and its Authorized Users are permitted to display and use reasonable portions of information contained in the Service for educational or research purposes, including illustration, explanation, example, comment, criticism, teaching, or analysis." Although the end of the terms of service state that users cannot "text mine, data mine or harvest metadata from the Service," I don't think my manual extraction of authors from this specific collection breaks that rule. My methods are all contained within the search feature, and I think my combination of those searches is within the definition of "research purposes."

**VALUE:** The EEBO is a pretty famous project for digitizing works from the period I'm working with and was recommended to me by Ted Underwood. Not utilizing this dataset in some way would be a pretty big oversight in the theoretical foundation of literary studies and canon studies that I'm building off of.

**WEAKNESSES:** I'm a little concerned by the concessions I had to make in the processing phase (detailed below). I don't think they will make a massive difference in my data, but I will have to sacrifice some of the nuance and holistic-ness I'm trying hard to maintain.

**PROCESSING:** I conducted a search in ProQuest for all records between 1470-1759, which produced 95,552 results. I then narrowed the search results to a chunk of years,

then clicked the “Authors” panel on the left-hand side. The table that pops up features the authors and their counts, which I can copy and paste into an Excel file. First I tried going through the year chunks chronologically, but the authors table maxes out at 100. As I tried to keep the author table results under 100, the year chunks got smaller and my collecting process slowed down to a point where I realized that collecting all the authors would be a bad investment of my time. I didn’t want to completely abandon this dataset, so I decided to only count authors who are cited at least 6 times. The process started to slow down again around the year 1600, so I upped my citation count minimum to 10 counts. I think this will still give me a sample that’s representative enough of major authors and will hold up against my other datasets. I’ve extracted authors from all records with a publication between 1473-1625 and I’m hoping to go up to at least 1750.

---

**NAME:** Wikipedia, writers born before 1775

**AUTHORSHIP:** WikiData

**ACCESS:** https://query.wikidata.org

**LICENSE:** Data is classified under Creative Commons Public Domain Dedication 1.0 and anyone can "copy, modify, distribute and perform the data, even for commercial purposes, without asking for permission" (https://www.wikidata.org/wiki/Wikidata:Introduction)

**VALUE:** If I can extract the data and find a way to link it to the other datasets, I think this dataset can add a lot of additional context to my other datasets. It serves a very different purpose than my other datasets – instead of functioning as another piece to aggregate, it will help me contextualize and organize the final aggregated dataset. Primarily, I’ll need this dataset to add in birth/death years to my Open Syllabus dataset, which is the only dataset that lacks this data point.

**WEAKNESSES:** Because this dataset is so different, it will also be the most difficult to fit with the other datasets. Once I get my dataset together, I’ll have to evaluate whether the context I might gain from fitting it together with my other datasets is worth the time I’ll have to invest into the process.

**PROCESSING:** Admittedly, I’m still confused by how this query service works. I’m planning on asking Elizabeth if she has experience (or knows someone with experience) on our 10/2 class to see if I can get over the learning curve more quickly. I’m hoping to spend the immediate days after on this dataset to determine conclusively whether it’s viable. Ideally, I’ll be able to extract a spreadsheet with the author names and their birth/death years and format the dataset to fit with my Open Syllabus data so I can pair the data points together and produce an aggregate dataset.

# DATASETS

---

**NAME:** Open Syllabus

**AUTHORSHIP:** The Open Syllabus Project

**ACCESS:** https://opensyllabus.org/

**LICENSE:** I won't be working directly with any of the syllabi, just different facets of the Explorer. OSP terms of use state that "much of the content of a syllabus is not protectable under copyright and is therefore in the public domain ... All of the metadata made available through the Explorer fall into this category." (http://blog.opensyllabus.org/terms-and-policies)

**VALUE:** This dataset is theoretically significant to the project because it represents taught literature in higher education. While my other datasets are based solely on active contribution, this dataset collects its data “primarily by crawling and scraping publicly-accessible university websites.”

**WEAKNESSES:** These authors are classified by appearances in syllabi and number of titles rather than record counts, so fitting the statistical significance of these authors will take a bit more time. Additionally, this list spans a much longer time span and includes many contemporary authors. The Explorer feature doesn’t have way to filter out authors by year, so I’ll have to find some way to weed out the authors who don’t fit the time period I want. After all that weeding, I’m a bit worried that my dataset will be too small to have any real impact on the final aggregated dataset.

**PROCESSING:** I used the Explorer feature and XPATH to pull the author names from two different searches: “Authors, English Lit” (156,489 authors) and “Authors, English Lit, U.S.A” (78,537 authors). Each sub-dataset holds 5,000 of the top authors from each search. I’m still deciding which dataset to use. The author names are formatted slightly differently than the other datasets (FirstName LastName), so I’ll need to write a different program to split them in the same way as the other datasets. This dataset also lacks birthdates and deathdates, so I’ll either have to add them in using my WikiData dataset or ignore that element in my other datasets.

## TECHNOLOGIES AND SKILLS

- Project scoping
- Web scraping
- Python (string methods, dictionaries, counting, data cleaning)
- XPATH queries
- SPARQL queries
- JSON (reading and manipulating)
- Dataset evaluation and comparison
    - Finding multiple sources for similar data (i.e. ESTL vs HathiTrust)
    - Choosing between two similar datasets
    - Evaluating exactly how a dataset benefits a project
- Developing a data standardization schema for fitting together disparate datasets
- Data cleaning and transformation
- Data management and organization
- Workflow planning and execution
- GitHub / version control
- Jupyter notebooks
- Pandas
- Statistical analysis
- Data visualization

## WHAT I LIKE

I'm glad I finally have a chance to look at the literary canon part of my project. It's making the whole project feel a bit more relevant and modern. The data is a bit difficult to gather but it doesn't require the level of cleaning that previous parts of my project did. I'm really enjoying learning different web scraping techniques and exploring different data sources that actually have structure and organization.

## CHALLENGES

- **Statistical analysis and data visualization** – I realized last month that I have very little experience with statistical analysis and data visualization, so I plan on visiting the Scholarly Commons and meeting with a librarian to collect resources and get a head start on the end of my project.
- **WikiData Query** – I'm still baffled by this tool. I need to talk to Elizabeth about it and see if she's familiar with it or knows someone who is.