**DATASET NAME:** eebo

**FULL NAME:** English Early Books Online

**AUTHORSHIP:** ProQuest

**SOURCE:** <https://search.proquest.com/eebo/index>

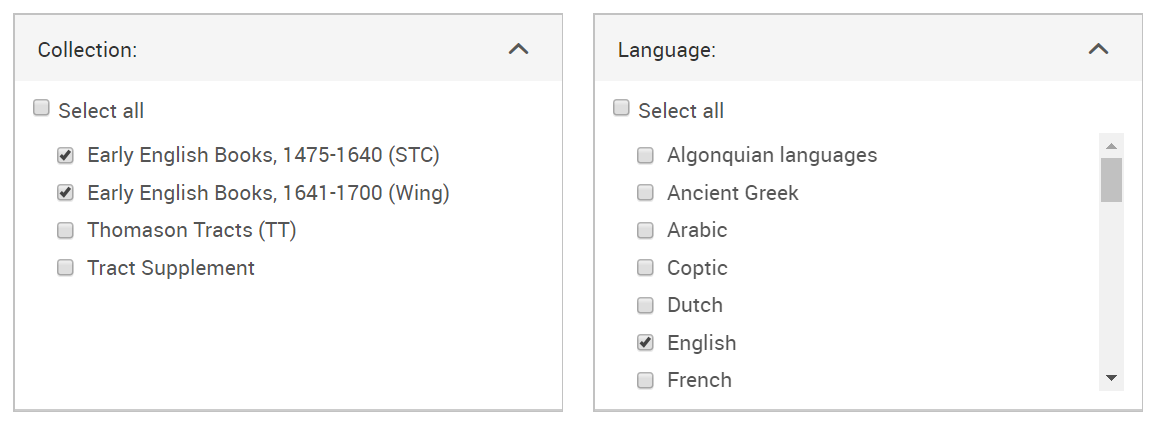
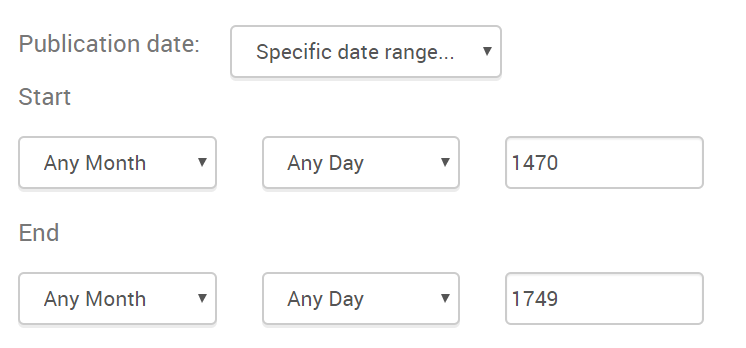
**DATE DOWNLOADED:** Oct. 3, 2019

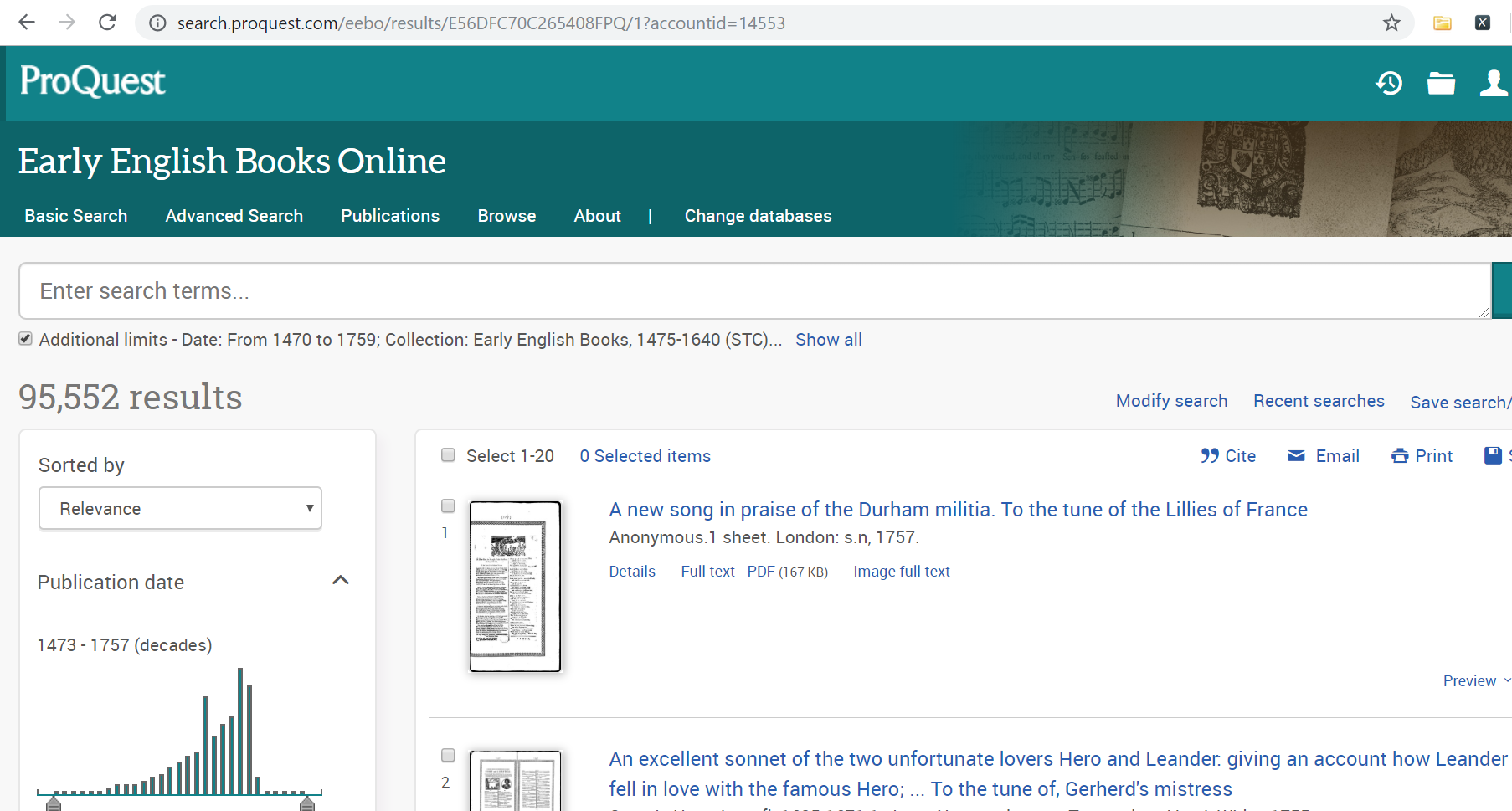
**QUERY WORKFLOW:**

1. Narrowing dataset:

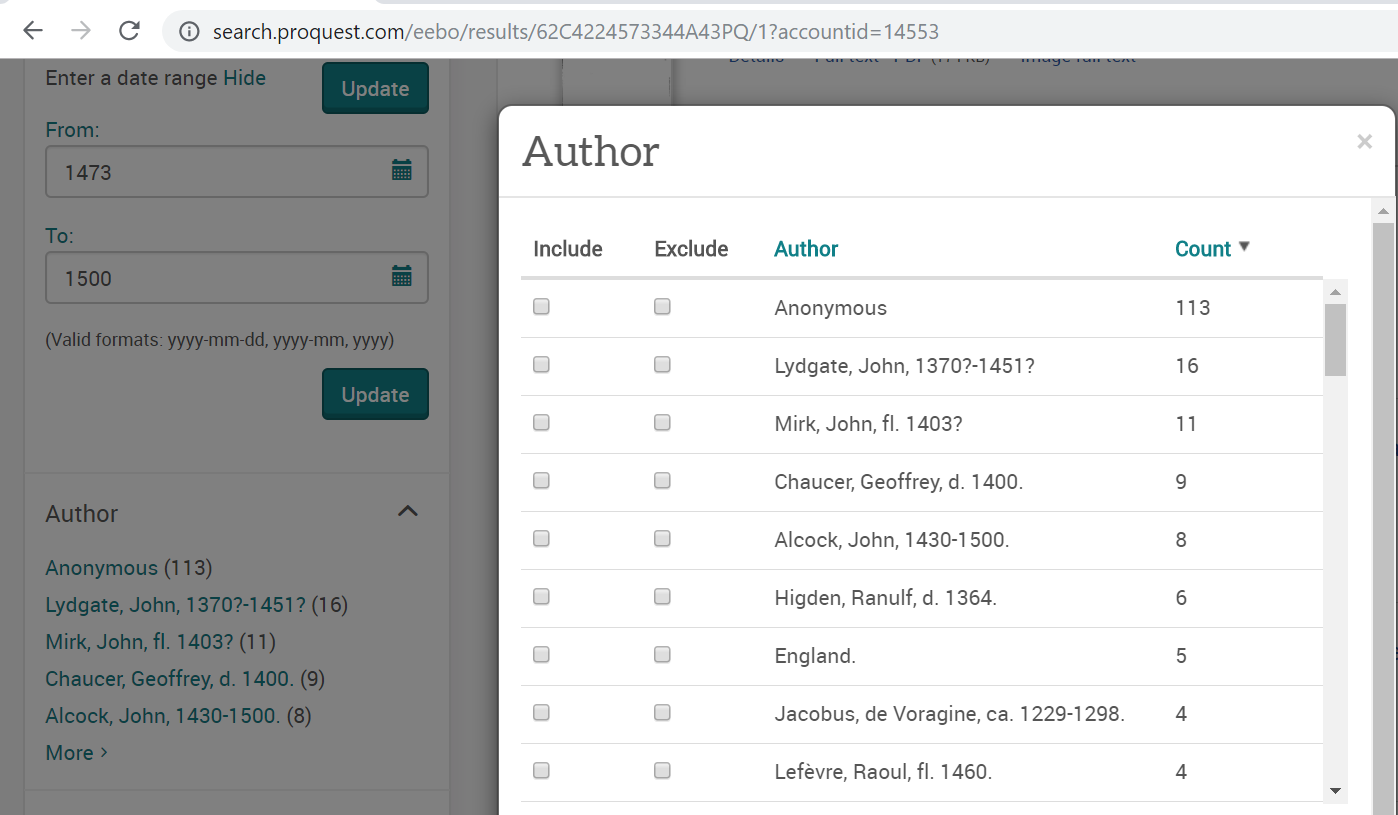
* Navigate to dataset source link (<https://search.proquest.com/eebo/index>)
* “Advanced Search” parameters:
  + Collection: Early English Books, 1475-1640 (STC), Early English Books, 1641-1700 (Wing)
  + Language: English
  + Publication Date (specific date range): 1470-1759

<https://search.proquest.com/search/1636116?accountid=14553>





2. Chunking author titles by year:



3. Copy “Author” and “Count” columns, paste into Excel

* For full documentation of raw data chunks and links to searches, see: eebo\_dataset\_raw\_DO-NOT-EDIT.xlsx

4. Manually reorganize data into two columns

**Product:** eebo\_dataset\_combined.csv

5. Use OpenRefine to consolidate author multiples and get total author counts

**Product:** eebo\_dataset\_cleaned.xlsx

**DATASET NAME:** estc

**FULL NAME:** English Short Title Catalogue

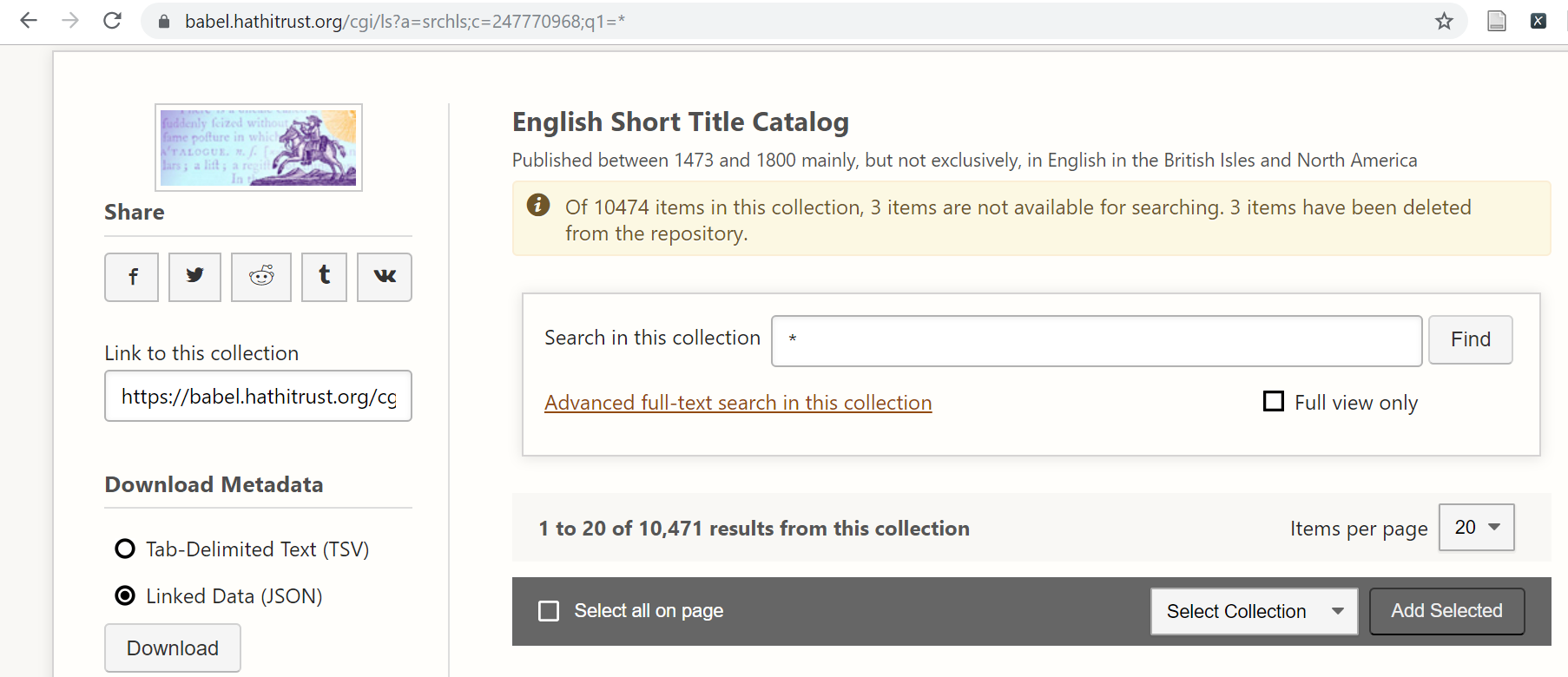
**AUTHORSHIP:** HathiTrust

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

**DATE DOWNLOADED:** September 15, 2019

**QUERY WORKFLOW:**

1. Navigate to HathiTrust collection record, downloaded JSON file



2. Use Python to extract text in “author” field

**DATASET NAME:** open-syllabus

**FULL NAME:** Open Syllabus Project

**AUTHORSHIP:** The Open Syllabus Project

**SOURCE:** <https://opensyllabus.org>

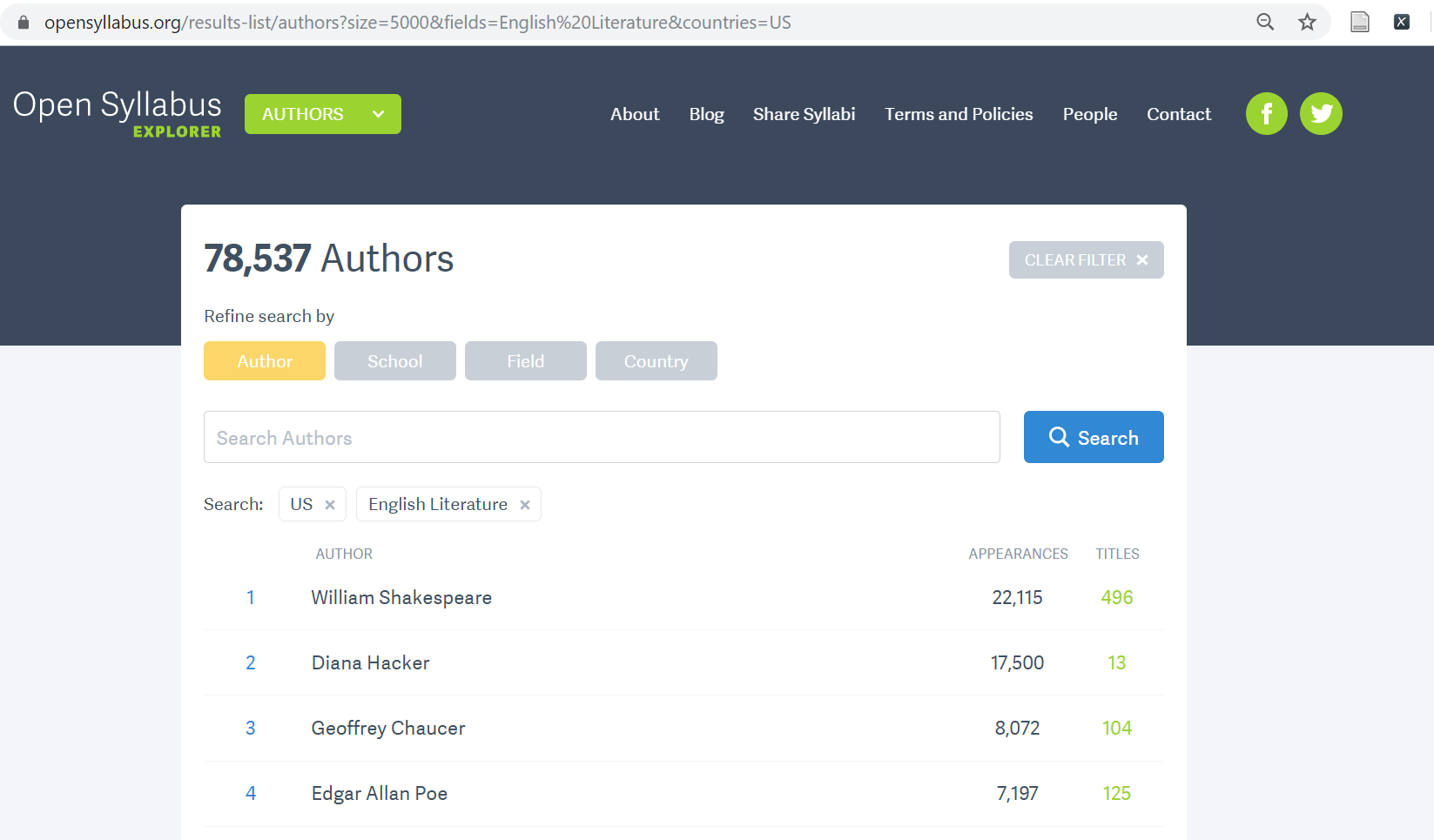
**DATE DOWNLOADED:** September 18, 2019

**QUERY WORKFLOW:**

<https://opensyllabus.org/results-list/authors?size=50&fields=English%20Literature&countries=US>

**Parameters:**

* Country = U.S.
* Field = English Literature



1. Use XPATH Helper to pull:

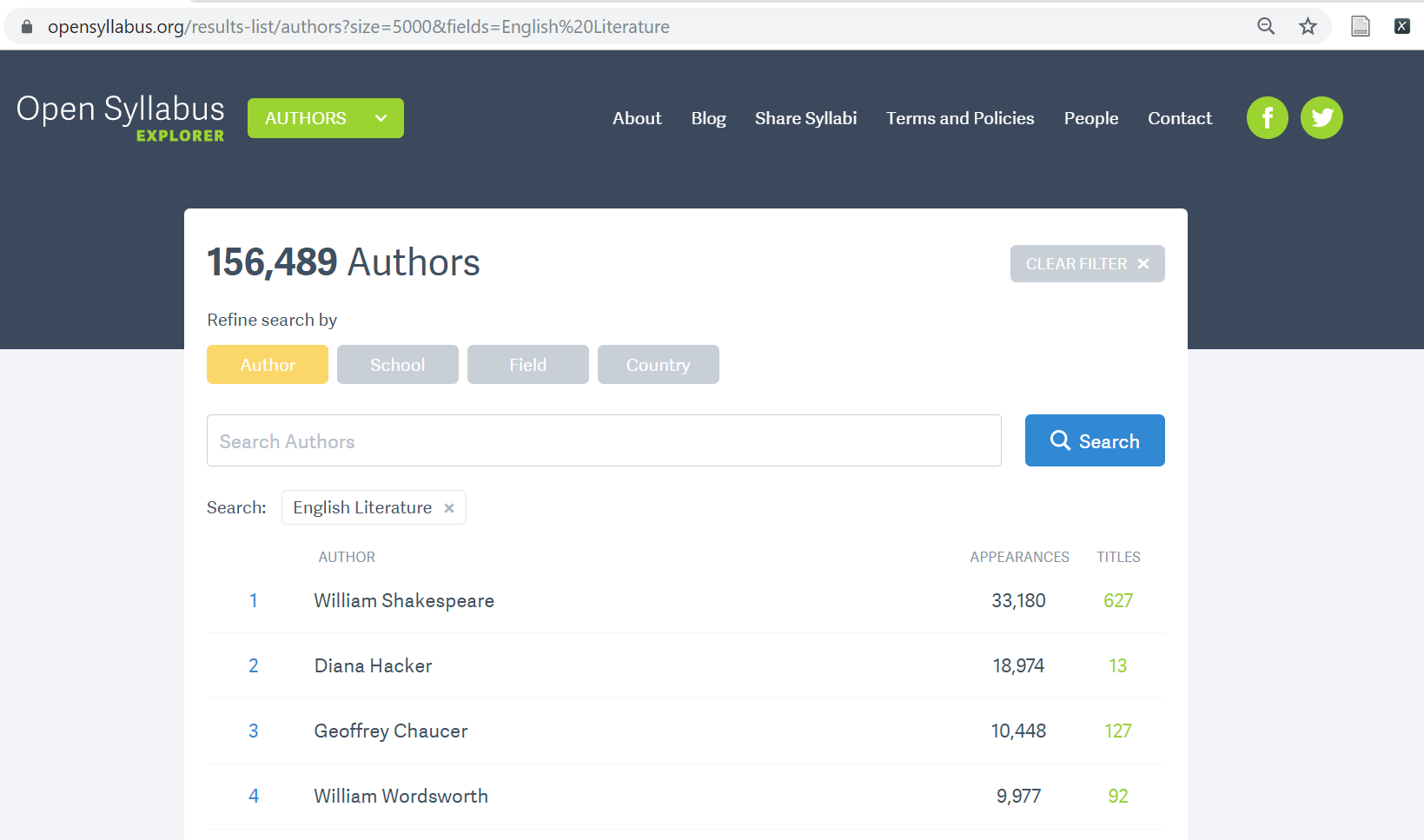
* **Author names:** /html/body/div[@id='root']/div/div[@class='main-layout']/div[@class='huge-container']/section/div[@class='result-list-container']/div/div[@class='result-list-view']/div[@class='center-content']/div[@class='container-list']/div[@class='author-list']/ul/li[\*]/div[@class='author-item']/div[@class='author']/a/p
* **Appearances:** /html/body/div[@id='root']/div/div[@class='main-layout']/div[@class='huge-container']/section/div[@class='result-list-container']/div/div[@class='result-list-view']/div[@class='center-content']/div[@class='container-list']/div[@class='author-list']/ul/li[\*]/div[@class='author-item']/div[@class='appearances']
* **Titles:** /html/body/div[@id='root']/div/div[@class='main-layout']/div[@class='huge-container']/section/div[@class='result-list-container']/div/div[@class='result-list-view']/div[@class='center-content']/div[@class='container-list']/div[@class='author-list']/ul/li[\*]/div[@class='author-item']/div[@class='titles']

2. Copy and paste results into Excel file

<https://opensyllabus.org/results-list/authors?size=5000&fields=English%20Literature>

**Parameters:**

* Field = English Literature



1. Use XPATH Helper to pull:

* **Author names:** /html/body/div[@id='root']/div/div[@class='main-layout']/div[@class='huge-container']/section/div[@class='result-list-container']/div/div[@class='result-list-view']/div[@class='center-content']/div[@class='container-list']/div[@class='author-list']/ul/li[\*]/div[@class='author-item']/div[@class='author']/a/p
* **Appearances:** /html/body/div[@id='root']/div/div[@class='main-layout']/div[@class='huge-container']/section/div[@class='result-list-container']/div/div[@class='result-list-view']/div[@class='center-content']/div[@class='container-list']/div[@class='author-list']/ul/li[\*]/div[@class='author-item']/div[@class='appearances']
* **Titles:** /html/body/div[@id='root']/div/div[@class='main-layout']/div[@class='huge-container']/section/div[@class='result-list-container']/div/div[@class='result-list-view']/div[@class='center-content']/div[@class='container-list']/div[@class='author-list']/ul/li[\*]/div[@class='author-item']/div[@class='titles']

2. Copy and paste results into Excel file

**Product:** open-syllabus\_dataset.xlsx

**DATASET NAME:** ota

**FULL NAME:** Oxford Text Archive

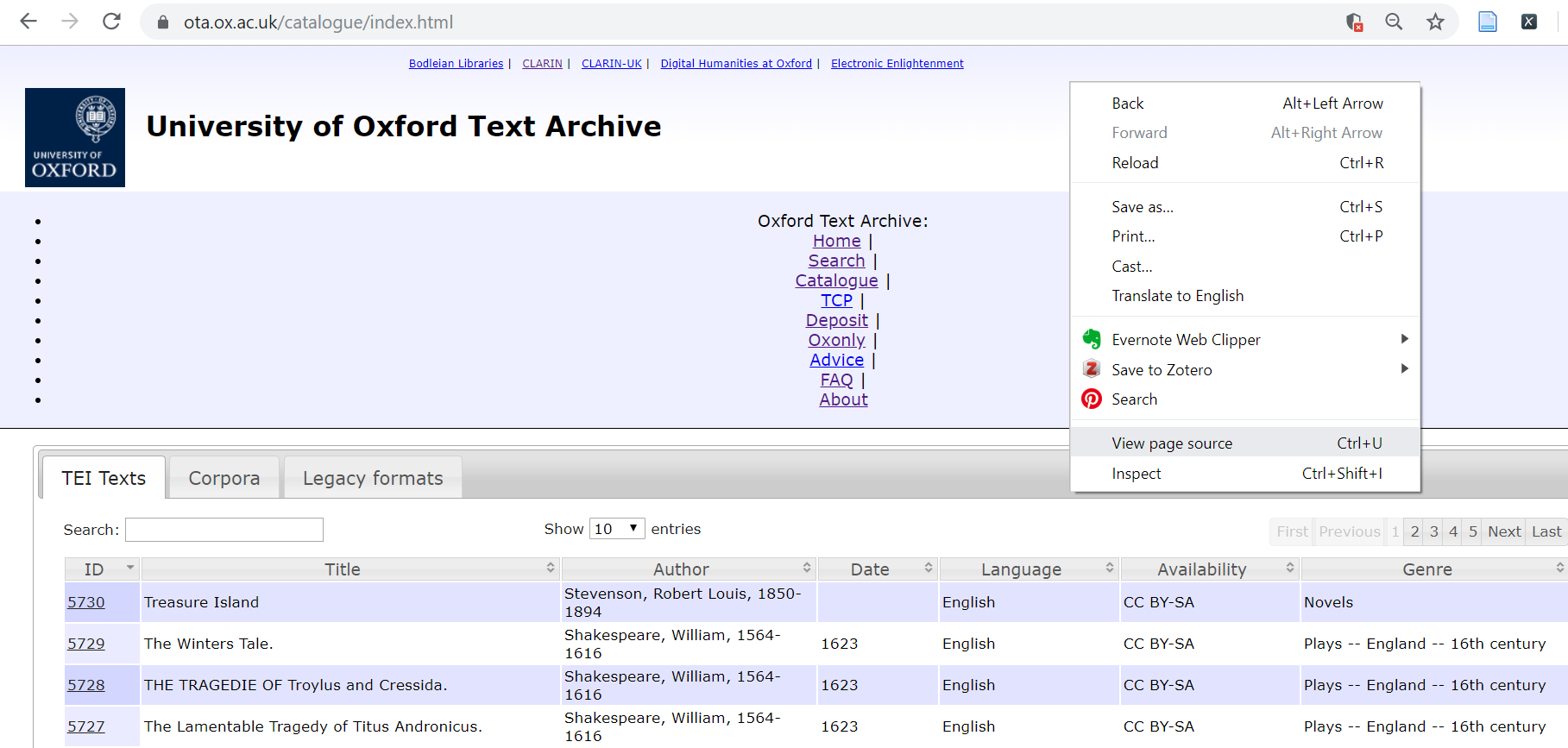
**AUTHORSHIP:** Oxford Text Archive

**SOURCE:** <https://ota.ox.ac.uk/catalogue/index.html>

**DATE DOWNLOADED:** September 15, 2019

**QUERY WORKFLOW:**

1. Download HTML of catalog



**Product:** view-source\_ota.ox.ac.uk\_catalogue\_index.html

2. Extract authors from HTML file using Python (ota\_extraction.py)

* Manually reformat HTML file as program gets stuck on inconsistencies (i.e. author names span more than one line)
* **Products:** ota\_source\_edited.html ; ota\_authors\_dataset.csv

3. Use OpenRefine to consolidate author multiples and get total author counts

**Product:** ota\_dataset\_cleaned.xlsx

**DATASET NAME:** project-gutenberg

**FULL NAME:** Project Gutenberg

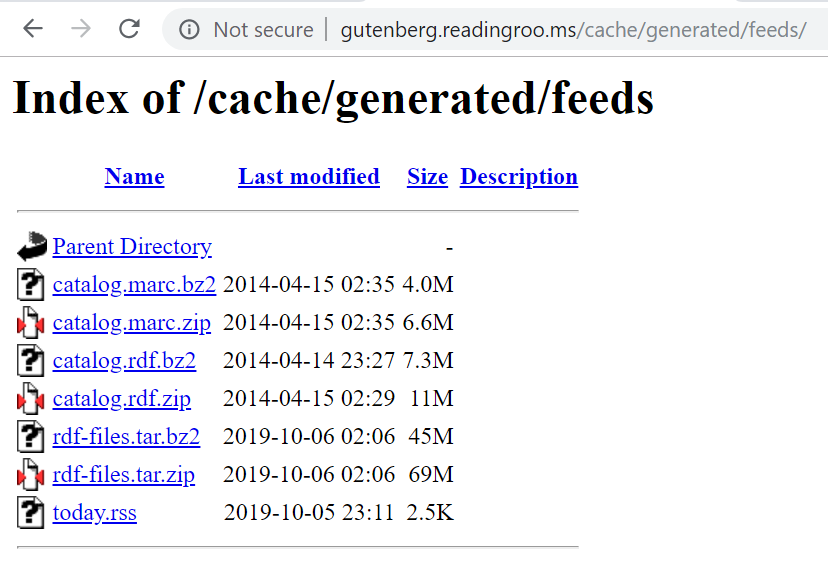
**AUTHORSHIP:** Project Gutenberg

**SOURCE:** <http://gutenberg.readingroo.ms/cache/generated/feeds/>

**DATE DOWNLOADED:** October 3, 2019

**QUERY WORKFLOW:**

1. Download RDF file of full catalog (catalog.rdf.zip)



2. Use Python to extract author names and put in CSV file

**Products:** project-gutenberg\_extraction.py ; project-gutenberg\_dataset.csv

3. Use OpenRefine to consolidate author multiples and get total author counts

**Product:** project-gutenberg\_dataset\_cleaned.csv

**DATASET NAME:** wikidata

**FULL NAME:** WikiData Query Service

**AUTHORSHIP:** WikiData

**SOURCE:** [https://query.wikidata.org](https://query.wikidata.org/)

DATE DOWNLOADED:

QUERY WORKFLOW: