

*Fig 1. Top 20 “authors” of final dataset*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **POSITION** | **LAST\_NAME** | **FIRST\_NAME** | **BIRTHDATE** | **DEATHDATE** |
| 4 | Shakespeare | William | 1564 | 1616 |
| 3 | Defoe | Daniel | 1660 | 1731 |
| 5 | Swift | Jonathan | 1667 | 1745 |
| 15 | Addison | Joseph | 1672 | 1719 |
| 6 | Pope | Alexander | 1688 | 1744 |
| 11 | Johnson | Samuel | 1709 | 1784 |
| 16 | Griffiths | Ralph | 1720 | 1803 |
| 12 | Smollett | Tobias | 1721 | 1771 |
| 8 | Goldsmith | Oliver | 1728 | 1774 |
| 19 | Burke | Edmund | 1729 | 1797 |
| 13 | Trusler | John | 1735 | 1820 |
| 20 | Paine | Thomas | 1737 | 1809 |
| 7 | More | Hannah | 1745 | 1833 |
| 14 | Hayley | William | 1745 | 1820 |
| 18 | Didbin | Charles | 1745 | 1814 |
| 17 | Pratt | Samuel Jackson | 1749 | 1814 |
| 10 | Kipling | Rudyard | 1865 | 1936 |

*Fig 2. Top 17 named authors, sorted by birthdate*

Initial analysis of the dataset shows that almost all of the top 20 authors were active during the Restoration and the Augustan literary eras. William Shakespeare and Rudyard Kipling are slight outliers in terms of life dates but are otherwise fairly canonical authors. Most of the authors are literary (poets, novelists, essayists) but some were primarily prolific in other areas (i.e. Didbin, Griffiths, and Paine in music, publishing, and politics, respectively).

An additional phase of cleaning would probably be helpful to remove the non-named entities (“Anonymous”, “Unknown”, “Various”) for a clearer picture of an actual author corpus. Another next step for this corpus would be re-incorporating life dates into the dataframe and developing a visualization of authorship distribution across eras.

Upon consideration of the final dataset and reflection on the project process as a whole, I have three conclusions:

1. The value of the dataset is limited in part due to the scope of the project. The intended use of this dataset was as a representative corpus of canonical authors for future data-driven research on literary history, lexicography, and canonicity. Upon reflection, I conclude that five relatively small catalog datasets are not sufficient to build a dataset for those original purposes. This project would benefit significantly from incorporating more datasets of a greater variety, as well as more particular processing of each dataset before final aggregation.
2. Despite the limited value of the dataset itself, I believe this project is a promising pilot program for gathering and standardizing catalog data. The cleaning, standardizing, and processing methods exercised during this project have a strong potential for broader application in the future, especially for building on the work conducted during this project. These processes are iterable for other structured catalog datasets, which could then be incorporated with the intermediate datasets built during this project.
3. This project was an excellent exercise in familiarizing myself with humanities data, particularly catalog and name data, and applying new cleaning and processing methods. I now have a clearer understanding of how to execute these workflows and can extend them to other data processing projects.