

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

|  |  |  |
| --- | --- | --- |
| **AUTHOR NAME** | **BIRTHDATE** | **DEATHDATE** |
| Daniel Defoe | 1660 | 1731 |
| William Shakespeare | 1564 | 1616 |
| Jonathan Swift | 1667 | 1745 |
| Alexander Pope | 1688 | 1744 |
| Hannah More | 1745 | 1833 |
| Oliver Goldsmith | 1728 | 1774 |
| Rudyard Kipling | 1865 | 1936 |
| Samuel Johnson | 1709 | 1784 |
| Tobias Smollett | 1721 | 1771 |
| John Trusler | 1735 | 1820 |
| William Hayley | 1745 | 1820 |
| Ralph Griffiths | 1720 | 1803 |
| Samuel Jackson Pratt | 1749 | 1814 |
| Charles Didbin | 1745 | 1814 |
| Edmund Burke | 1729 | 1797 |
| Thomas Paine | 1737 | 1809 |

*Fig 2. Life dates of top 17 named authors*

Initial analysis of the dataset shows that almost all of the top 20 authors

Rudyard Kipling is an outlier in terms of life dates, but is a fairly canonical author otherwise.

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.