**STAGE 1:** Manually standardize column names in datasets

**Naming convention:** name\_dataset\_standard\_1.csv

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Field available in dataset?** | | | | |
| **FIELD** | **EEBO** | **ESTC** | **OPEN-SYLLABUS** | **OTA** | **PROJECT GUTENBERG** |
| author\_name |  |  |  |  |  |
| count |  |  |  |  |  |
| fraction\_total |  |  |  |  |  |
| birthdate |  |  |  |  |  |
| deathdate |  |  |  |  |  |

**STAGE 2:** First round of splitting using Excel “text to columns” feature

**Naming convention:** name\_dataset\_standard\_2.csv

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Can data be transformed to fit field?** | | | | |
| **FIELD** | **EEBO** | **ESTC** | **OPEN-SYLLABUS** | **OTA** | **PROJECT GUTENBERG** |
| first\_name |  |  |  |  |  |
| last\_name |  |  |  |  |  |
| count |  |  |  |  |  |
| percent\_total |  |  |  |  |  |
| birthdate |  |  |  |  |  |
| deathdate |  |  |  |  |  |

|  |  |
| --- | --- |
| **DATASET** | **SPLIT POINT** |
| EEBO | Comma, dash |
| ESTC | Comma |
| Open Syllabus | Space |
| OTA | Comma |
| Project Gutenberg | Comma |

**STAGE 3:** Clean and consolidate columns

* Sort dataset based on the last column that has content (Z-A)
  + When rows are more fractured across more columns, they require more cleaning. Highest rate of error when splitting (e.g. “Mornay, Philippe de, seigneur du Plessis-Marly, 1549-1623.” is split across 6 columns when it should only be split across 4)
* Move values to appropriate columns
  + e.g. Mornay, Philippe de, seigneur du Plessis-Marly, 1549-1623.

Original:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **last\_name** | **first\_name** | **birthdate** | **deathdate** |  |
| Mornay | Philippe de | seigneur du Plessis-Marly | 1549 | 1623 |

Cleaned:

|  |  |  |  |
| --- | --- | --- | --- |
| **last\_name** | **first\_name** | **birthdate** | **deathdate** |
| de Mornay | Philippe | 1549 | 1623 |

* Cleaning standards for names:
  + For names without surnames, use title (“Saint of”, “Prince of”, “King of”, “of Garland”, etc.) in last\_name column
  + For surnames with nobiliary particles (“de”, “di”, “von”, etc.), put particle at the beginning of last\_name column
    - e.g. “Philippe de Mornay” = “de Mornay, Philippe”
  + For records with *floruit* (fl.) instead of birthdate, use numerical *floruit* value only
  + For records with approximate dates (“ca.”, “approximately”, “active”), use numerical date only
  + For records from B.C.E., keep “B.C.” in birthdate and deathdate columns
  + For records with varying birthdates or deathdates (e.g. “1553 or 4”, “1579 or 80”), keep both values

**Naming convention:** name\_dataset\_standard\_3.csv