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

**Naming convention:** name\_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\_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)

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