Exercise:

Create Controlled Vocabularies

**Summary**

In this exercise, you’ll take unstructured data from the House of Refuge dataset and create controlled vocabularies that can be used for indexing. These will include the “authorized” or “preferred” terms and “variant” terms. Standardizing historical data is particularly difficult, because language may be outdated, difficult to understand outside of the historical context in which it was created, or offensive. You may also be dealing with geographic entities that no longer exist.

For context, this dataset was created by multiple creators (likely clerks at the Cincinnati House of Refuge) between 1869-1902.

**Dataset 1**

*Parentage (Race/ethnicity)*

|  |  |
| --- | --- |
| **Raw Data Values** |  |
| 3/4 Colored | Irish |
| American | Israelite |
| American and Irish | Italian |
| Black | Jewish |
| Canadian | Mulatto |
| Colored | Negro |
| Colored and Irish | Negro Mulatto |
| Danish | Polish |
| Don’t Know | Prussian |
| Dutch | Russian |
| English | Scotch Irish |
| English Jewish | Scottish |
| French | Spanish |
| French and Scottish | Swedish |
| Georgia Negro | Swedish American |
| German | Swiss |
| German and French | U.S. Color |
| German and Irish | Unknown |
| German Jew | Welsh |
| Greek | White |
| Hebrew | Icelandic |

**Example**

* Note that this example makes MANY curatorial assumptions. For example, it takes granular information and makes it more general.

**Preferred Term:** Scandinavian

**Use for:** Danish

**Use for:** Icelandic

**Use for:** Swedish

**Use for:** Swedish American

**Task**

* Create a controlled vocabulary, using the dataset on page 1. First, organize your terms using the paper slips, then transcribe to the computer.
* What are your “preferred” terms (you can use contemporary language, not just the values in the dataset).
* What are the variant terms that you are using the preferred terms to represent?

**Discussion**

* What kinds of difficulties did you encounter?
* What kinds of curatorial decisions did you have to make?
* Is it possible to standardize historical data in ways that represent the authenticity of the historical data without being offensive?