**­­Exercise:**

**TEI Enrichment**

**Summary**

OpenRefine is a powerful tool for working with data. It is well suited to cleaning and correcting medium to large sets of data and accepts plugins for enrichment and export. These lessons cover some applied basics of working with OpenRefine: creating projects, cleaning & transforming data, and exporting for use in other applications. For our TEI exercises, we’ll be working with letters held by the University of Cincinnati’s Archives & Rare Books Library. The letters were written by Bierce to his friend Myles Walsh and Walsh’s sister Lily, a young deaf-mute woman.

**Exercise**

In this exercise, we’ll complete the following tasks:

* Examine TEI with OpenRefine
* Extract names and pronouns
* Associate the pronouns with names
* Reconcile names with name entities
* Export
* Use a Python script to add references to name entities

Example name tag from the letters:

<name type="person">I</name>

1. Start OpenRefine and create a new project
2. Select the 10 sample Bierce TEI files from [GitHub](https://github.com/crowesn/beyond-tei).
3. On the *Select Files to Import* screen make sure that all ten files are checked and click *Configure Parsing Options*
4. In *Click on the first XML Element* dialog, select the `p` node and leave the rest of the options unchanged.
5. Click *Create Project*

**Prepare the Data**

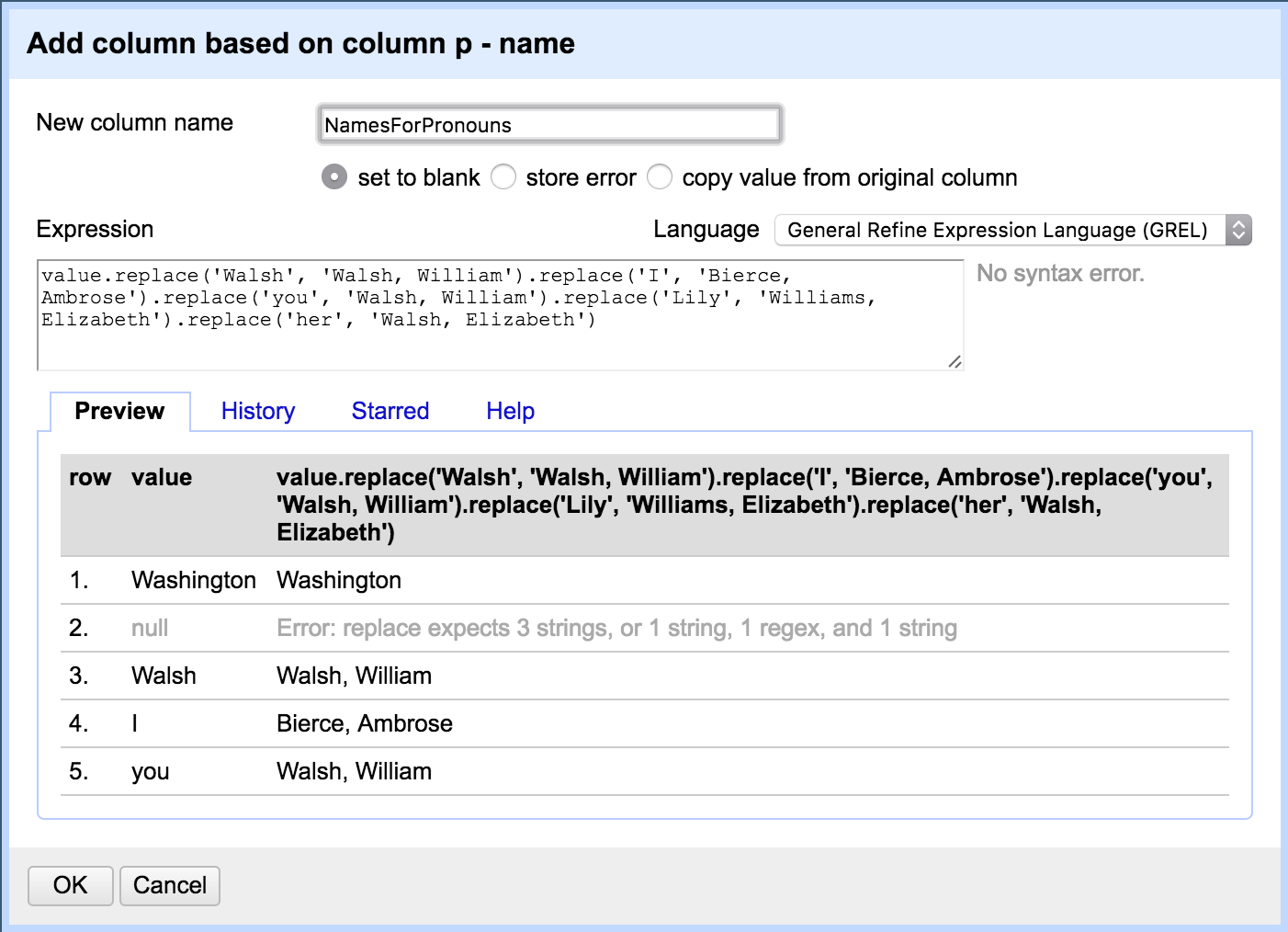
[GREL - https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions#replacestring-s-string-f-string-r](https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions#replacestring-s-string-f-string-r)

We’ll use the Google Refine Expression Language (GREL) to prep our pronoun data for name reconciliation. GREL includes several functions for complex manipulations of strings, much like Excel functions. The **replace()** function, coupled with Edit Column > Create column based on column allows us to generate a column of names to be associated with each pronoun.

replace(string s, string f, string r)

1. Select *Edit Column > Add Column based on this column* from the top of the column with names and pronouns
2. In the *Add column based on ...* dialog, input a new column name
3. Add a chained expression using the replace function to encompass all of your replacements for the column

e.g. value.replace('Walsh', 'Walsh, William').replace('I', 'Bierce, Ambrose') etc.



**Reconcile some names**

Now that we have a column of full names associate with the prounouns from the document, we can use a name reconciliation service to identify named entities for each name (if they exist).

Add some reconciliation services:

<https://tools.wmflabs.org/openrefine-wikidata/en/api>

<http://refine.codefork.com/reconcile/viaf>

1. Select *Reconcile > Start Reconciling* from dropdown at `p – names`
2. Click *Add Standard Service* on the bottom left and input one of the URLs from above
3. Under *Services* in the *Reconcile Column ...* dialog, select the new service you just added (likely bottom most)
4. After a short wait, you may see some type options such as `human, country, etc`; select the radio button for the most appropriate type and click *Start Reconciling*
5. Click *Start Reconciling*
6. Based on the default options, the reconciliation will match automatically based on high-confidence. If lower confidence, the service may return multiple options from which you can manually select.

**Extract URIs For Export**

OpenRefine currently will not expose the full URI for reconciled cells. Depending on the service used, you typically can only get the [identifier](https://groups.google.com/d/msg/openrefine/l1DJ76uKcXs/uU_vQcpzBAAJ). Users can however reconstruct the full URI with some wrangling.

1. Select *Edit Cells > Transform* at the top of the column with the reconciled terms.
2. Add the following to the *expression* field “if(cell.recon.match.id, “<uri>” + value”

**if(cell.recon.match.id, '', 'https://www.wikidata.org/wiki/' + value)**

1. This should result in a column with full URI

**Export**

We want to export a csv file with the following structure:

|  |  |
| --- | --- |
| Term (pronoun/name | Ref |
| I | https://www.wikidata.org/wiki/Q191050 |

1. In Open Refine, remove all columns except for **p­ – name** and **URI** (your column with URIs from reconciliation.
2. Click *Export > Comma Separated Value*  and save resulting csv.

**Add Ref Attribute to Name Tags with Script.**

<https://github.com/crowesn/beyond-tei-xml-demo>