Smithsonian Libraries

Reconciling Smithsonian Library data with VIAF

Allyson Ota, Intern
8-9-2016

Contents

Software Requirements

OpenRefine

Java 1.7 or higher

Installing the VIAF Reconciliation Service

There are multiple open source versions of VIAF reconciliation services, but the most accurate in testing was written by Jeff Chiu, named **refine_viaf**. If needs are low, and a small amount of queries are being made, a public server can be used in place of installing this service here: http://refine.codefork.com/reconcile/viaf**.** I chose to install the service locally since I would potentially be making large requests.

These instructions will allow you to install the service locally. For additional information, see: https://github.com/codeforkjeff/refine_viaf

1. In order to run the service, you should have Java 1.7 or higher installed on your computer
    a. To check your java version, open a command prompt and type: **java –version**

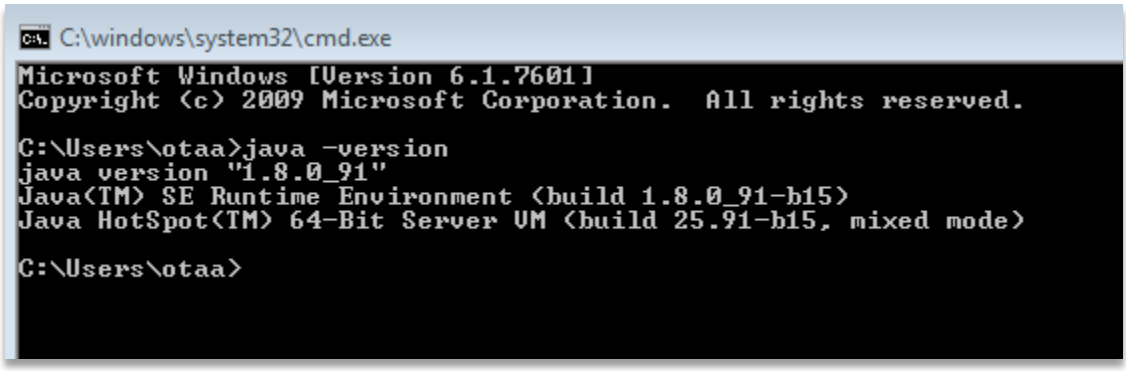

Figure 1. Check java version

2. Download the latest version of **refine_viaf** from: https://github.com/codeforkjeff/refine_viaf/releases and place it in a location you'd like to run it from
3. From the command prompt, CD to the directory where you saved the **refine_viaf.-1.5.1.jar** file
4. Type **java –jar refine_viaf-1.5.1.jar** to execute the service
5. The service should startup; Ctrl+C stops the service

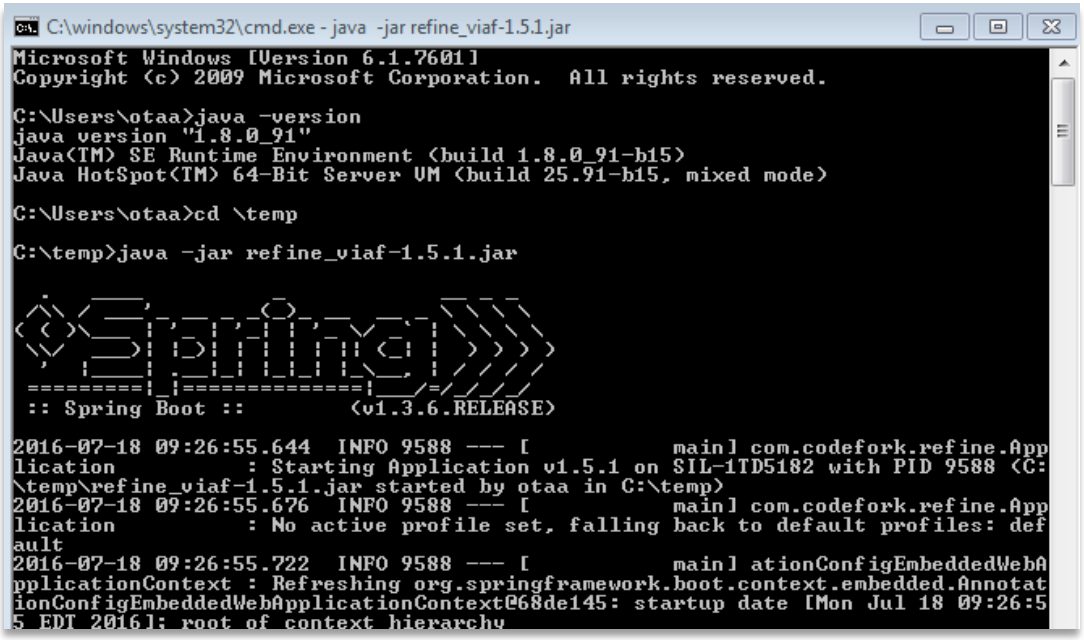

*Figure 2. The service **viaf_refine**, starting up*

Character Encoding in Excel

When opening the .csv files, check diacritical marks in author/organization names and ensure they are not displaying garbled. If this seems to be happening, you should:

1. Open a new workbook in Excel, and select **Data -> From Text** and select the .csv to be imported
2. In the Text Import Wizard, select **delimited**, **My data has headers**, and choose File origin: **65001 : Unicode(UTF-8)**

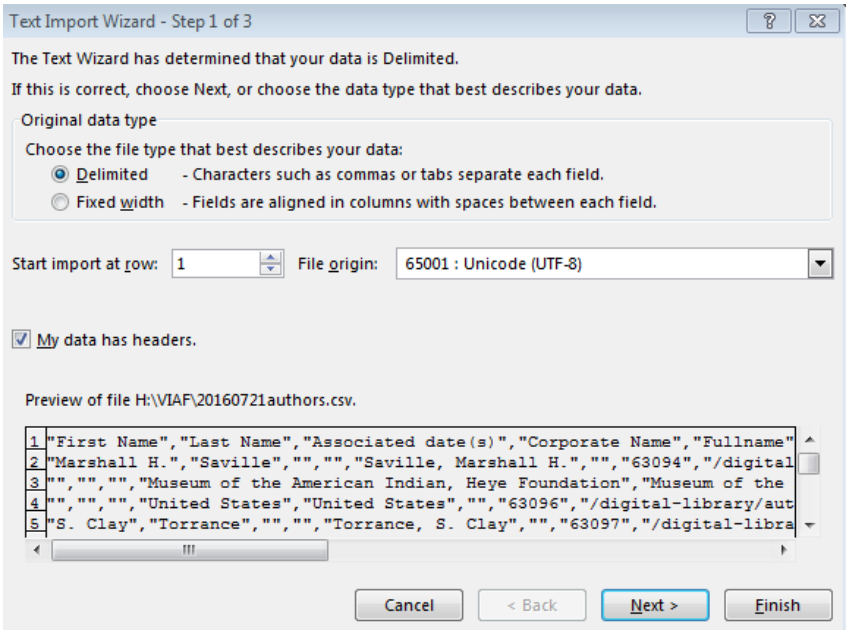

Figure 3. Importing .csv as text to retain character encoding in Excel

3. Click Next
4. Under **Delimiters**, select **Comma**
5. Click Next
6. For **Column Data Format**, select each column in the pane below and select **Text**
7. Click Finish
8. On the **Import Data** window, click OK
9. Spreadsheet should populate with UTF-8 encoding retained

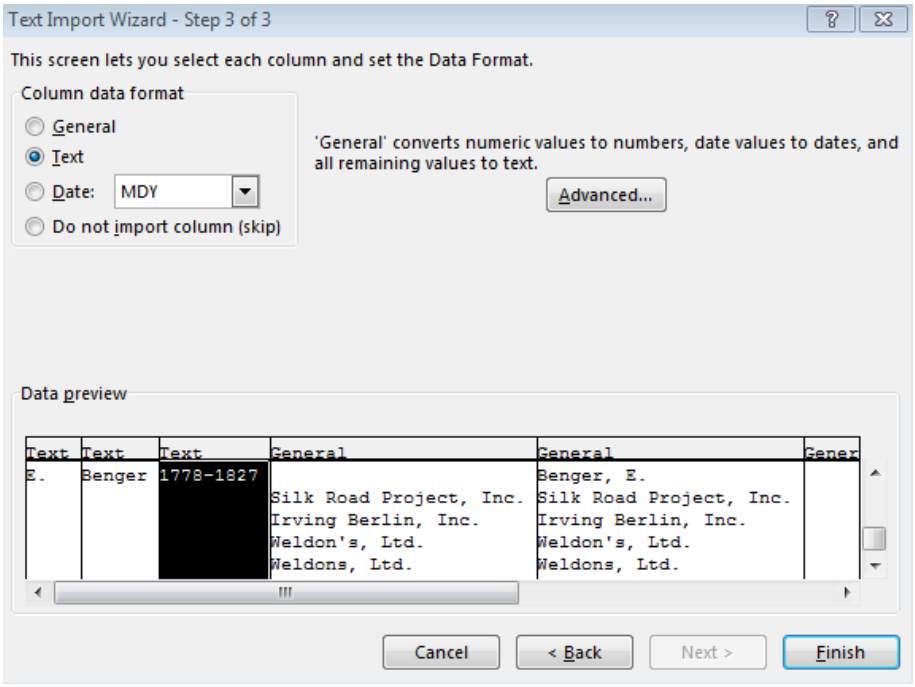

Figure 4. Setting format as text

Separating Authors by type: corporations vs. persons

We need to separately reconcile persons and organizations since the reconciliation service can only handle one entity type at a time. Organizations should already be separately listed in the export under the heading **Corporate Name**. Persons have their names split across the columns **First Name** and **Last** Name. For persons containing only a **Last** Name, they presented some issues in OpenRefine, so they should be handled separately from authors with both a first and last name. The **Fullname** column mixes both persons and organizations and should not be used to reconcile against VIAF. All persons must have at least a **Last Name**, and if an author only has a first name or pseudonym, that should be updated in the catalog before being included in the reconciliation process.

1. In excel, do a sort on the **Corporate Name** field, cut and paste them into a new spreadsheet, e.g. organizations.xslx
2. With the remaining cells which are all now persons, do a sort on the **First Name** field, and cut and paste the blank entries into a new spreadsheet where the authors only have a **Last Name**.
    a. Delete the **First Name** column, and ensure your first column is **Last Name**, since OpenRefine seems to have trouble with leading blank cells.
    b. Save

3. Save the spreadsheet that lists all persons with both a **First Name** and **Last Name** into its own worksheet, e.g. persons.xlsx
4. Each spreadsheet should follow the reconciliation processes below

Reconciling authors in OpenRefine

1. In OpenRefine, create a new project and select your file
2. If it is a .csv you can designate **Character encoding** by clicking the empty field towards the bottom-left, and from the list choose **UTF-8;** otherwise check to see if your spreadsheet is displaying diacritical marks properly and make adjustments before importing into OpenRefine

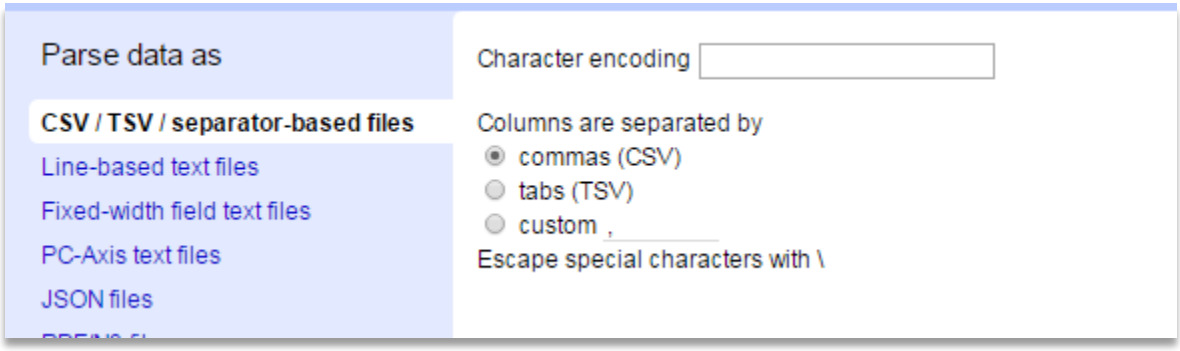

Figure 5. Choosing UTF-8 for .csv imports in OpenRefine

3. Create a new column named **Person Name** (or any name you'd like) and add it by choosing the arrow on the **Last Name** column heading -> **Edit column -> Add column based on this column**

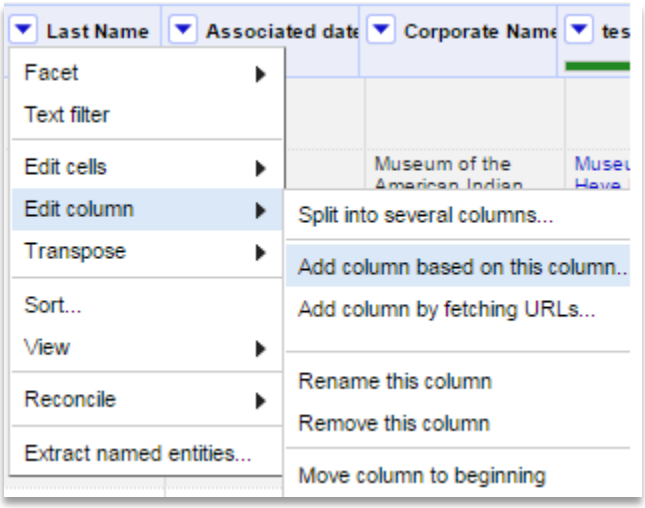

Figure 6. Add a columnbased on another column in OpenRefine

4. Input a new column name, e.g. Person Name, and in the Expression box, type:

```
value + ", " + cells["First Name"].value
```

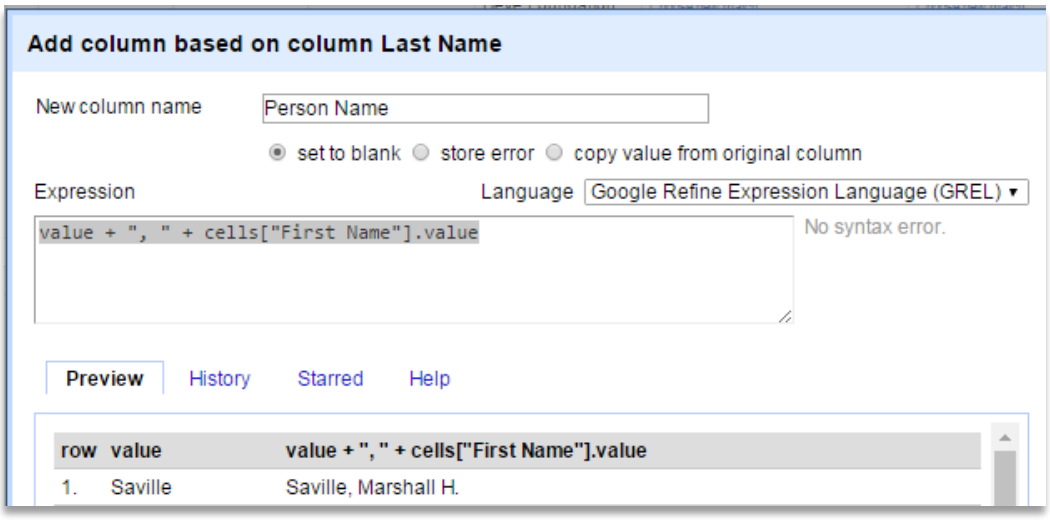

Figure 7. Using GREL to create content for the new column

5. You should get back a new column with author names populated in the format **LastName, Firstname**
6. Select the column you created, choose -> **Reconcile -> Start reconciling…**

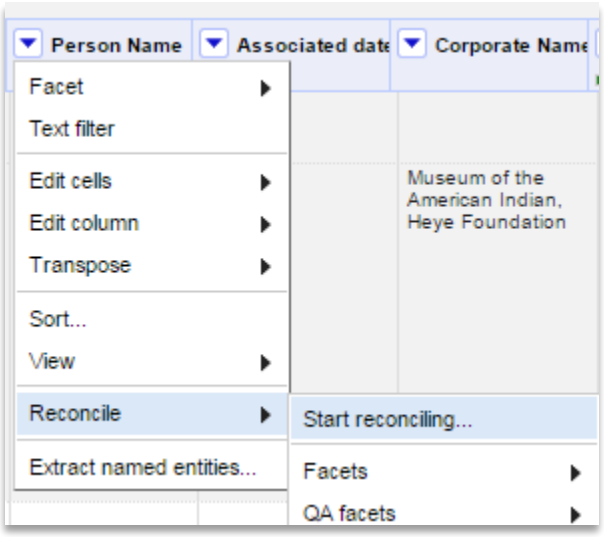

Figure 8. Start reconciliation

7. If the VIAF Reconciliation service doesn't appear as an option, add it by clicking **Add standard service**
8. Input the service URL: **http://localhost:8080/reconcile/viaf** and click **Add Service**
9. Select the service from the left-hand column; select the entity type; and for persons you can select the **Associated date(S)** (for persons) column under relevant details, then choose to **Start Reconciling**

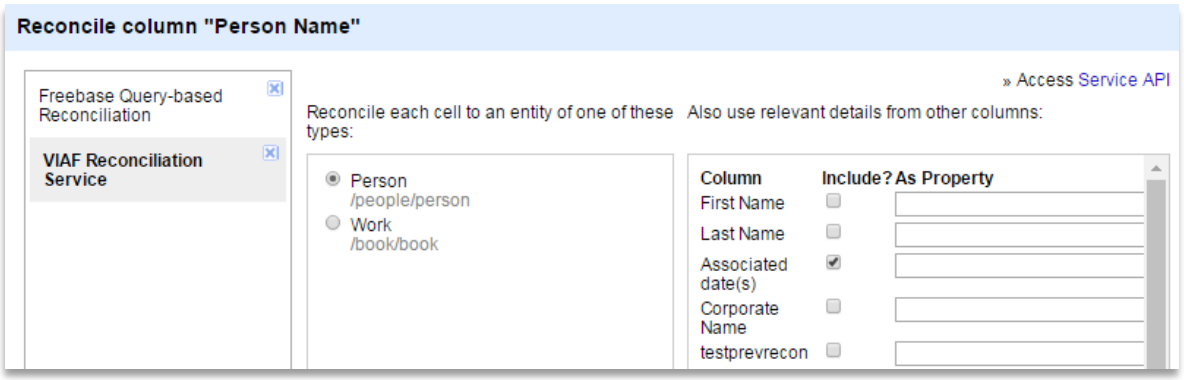

Figure 9. Choose the VIAF Reconciliation Service, entity type, and any additional columns to include

10. For persons with only surnames, follow the steps above, skipping steps to join the first and last names.

Reconciling organizations as authors in OpenRefine

12. For **Corporate Names** reference the steps above, but choose **Corporation** as the entity type when reconciling

Service Reconciliation should be manually verified

When refine_viaf returns a match

Authors that are matched will appear as a blue hyperlink in OpenRefine. Manually verify each authority (corporation/person) by following the link on matched entities to ensure this is the correct entity. If it is not, you should remove the link.

When multiple candidates are returned

A listing will appear, showing a name, and a listing of possible candidates, along with the probability of each candidate offered being a match. The best match should be selected by clicking the single-checkbox which should then verify a match against VIAF. The difference between the single-checkbox, and the double-checkmark box, are that the single-checkmark denotes only that particular cell. Double-checkmarks denote any other cells that would hold the same value (which in theory should not apply to this particular process).

When zero matches are returned

When no matches are made, zero candidates are offered. A search should still be manually done in VIAF to see if a listing can be found. Occasionally, a typo or other diacritical may obscure results.

Retrieve VIAF IDs

1. After you've matched as many persons/corporations as possible, select the column and choose **Edit column -> Add column based on this column**
2. In the expression box, name the new column, then type the following GREL expression: **cell.recon.match.id**
3. You should get back a listing of VIAF IDs for all matched rows
4. Do this for both the reconciled persons and corporations

5. Export by selecting **Export** at the top right

Lessons Learned

Service Reconciliation with refine_viaf

Corporations and Persons should be reconciled separately since queries with **refine_viaf** only allow for choosing one entity type at a time.

Reconciling Persons

In testing, more matches were made when the author's name was in the format: LastName, FirstName (vs. FirstName LastName). In testing both, I included the **Associated dates** field when reconciling. The results were that 23.0% matches obtained were reconciled with the format LastName, FirstName as opposed to 7.7% when the format was FirstName LastName. See Table 1 below for more details.

Table 1. Testing name formats with refine_viaf service reconciliation. Results against **3,458 total** names.

	LastName, FirstName	FirstName LastName
matched	613	207
none	2059	2465
(blank)	786	786

Once a column has been manually checked, and someone has gone through to select from candidates, and verify matches made by refine_viaf, you can reconcile. However, any time you choose to reconcile again, it will undo previous selections. It is recommended to create a new column for anything being adjusted that needs further reconciliation so manually performed work is not undone.

Sometimes I was able to find a person in VIAF that received no matches, or no correct candidates by going directly to the VIAF site and searching for individuals. It may be that the service had glitches during the reconciliation process?

Data cleanup issues

While each entity should be unique, there were instances of authors in the database export that appeared to be referencing the same person/corporation. Also, at times there were typos that caused two authors to be listed, but having a comma or part of a title at the end of a person's name caused two entries. These issues should be rectified in the DB.

Limitations with OpenRefine

I noticed an issue when importing the authors.csv file I was given, where the first column was "First Name" for persons. This field was blank for a few persons who only had surnames, as well as all organizations listed. In OpenRefine, persons with only a surname were being appended to the row directly preceding it. As a fix, I split the organizations out in Excel and did them separately from persons. When a person has only a surname, importing into OpenRefine combined these rows with the line the previous row, so procedures were updated to separate surname-only persons from all other persons, and organizations with these columns being the leading column since there should be no blank cells at the start of a new row in OpenRefine.

Software differences on Mac vs PC

I was able to do everything required on a Mac as well as a PC. With a Mac, however, the Java JDK has to be installed (not just the JRE).

Possibilities for retrieval of other IDs aside from VIAF

According to the developer's site, you can retrieve other IDs aside from VIAF. For example, if you just wanted Library of Congress IDs. This is directly from the website: http://refine.codefork.com/

How to Use It

- In OpenRefine, select "Reconcile" and "Start Reconciling..." in the pull-down menu beside a column whose values you want to reconcile.
- Click "Add Standard Service..."
- To reconcile against names from any VIAF source, type in:

  http://refine.codefork.com/reconcile/viaf
- OR, to reconcile against a specific VIAF source, add its code to the end of the path. For example, to search only names from the Bibliothèque nationale de France, type in:

  http://refine.codefork.com/reconcile/viaf/BNF
- OR, to retrieve the IDs used by source institutions, rather than VIAF IDs, use "proxy mode." For example, to search only names from the Library of Congress and retrieve their IDs, type in:

  http://refine.codefork.com/reconcile/viafproxy/LC