11-1-2015

Identifying potential headings for Authority work using III Sierra, MS Excel and OpenRefine

Lynn K. Whittenberger
Marquette University, lynn.whittenberger@marquette.edu

Identifying potential headings for Authority work using III Sierra, MS Excel and OpenRefine

Lynn Whittenberger, Metadata Librarian
Raynor Memorial Library, Marquette University
Lynn.whittenberger@marquette.edu

In order to participate in a newly established Wisconsin NACO funnel project, Marquette University's Raynor Library needed to identify author/corporate names currently in the Catalog (http://libus.csd.mu.edu/) or Institutional Repository (http://epublications.marquette.edu/) that might be good candidates for creating new Authority records.
Two collections immediately came to mind that would benefit from additional authority control: Marquette Faculty authors (represented in the Institutional Repository and in our Faculty Publications Collection), and Milwaukee musicians (represented in the Cujé Collection)

III Sierra: pulling lists & exporting records
Lists of names represented in the catalog were easy to pull. Faculty Publications and Cujé Collection materials have unique location codes in our catalog, which made it easy to gather the relevant bibliographic records using the create lists function in III Sierra. From those lists, I exported only the author and added author fields (MARC fields 100/110 and 700/710). I stripped subfields e and t from the export, as that made it easier to identify and remove duplicates in Excel and OpenRefine. The author and added author fields were exported as a | (pipe) delimited file. Sierra offers the option of specifying delimiters for repeated fields, and I used the same delimiter for both fields and repeated fields, which made it easier to convert the resulting file into one long list in Excel.

MS Excel: consolidating the catalog data
Once the file exported, I used the Excel wizard to convert the delimited file into an Excel spreadsheet. The initial conversion gave me many columns of data. To consolidate the columns of data down to one, I sorted the last column, used the Excel "remove duplicates" function on that column, then copied the de-duplicated cells and pasted them onto the bottom of the column to the left. I repeated the process of sorting and deduplicating, then copying and pasting, moving left on the spreadsheet until I was left with one column of data. A time consuming process, but not terrible.

Digital Commons data
Our Institutional Repository (IR) is built on the Digital Commons platform and the list of faculty names is natively exported as an Excel spreadsheet. Here, the author name exports into three columns: last name, first name, middle name/initial. Using the concatenate function in Excel, I was able to 'glue' the pieces of the author name together into the same format as the data from the catalog (Last Name (comma space) First Name (space) middle name/initial). To convert the column of built names from a formula to text, I had to copy the column, and then paste as values into a new column.

MS Excel: more data cleanup

I took the list of Faculty names from the IR and merged it with the list of Faculty names from the catalog. I then cleaned up the list by running the Excel sort and "remove duplicates" functions on the merged list.

The end result was two spreadsheets: a Cujé sheet and a Marquette Faculty sheet. For the Cujé sheet, I left all the headings in the list. For MU Faculty, I did a little more cleanup to remove corporate names. In Excel, I used the "Conditional Formatting"-"Highlight Cells Rules"- "Text that contains" functionality. Using a list of common corporate name stopwords (e.g. Association, Inc., University, etc..) I was able to find corporate body names in the list and remove them. To identify obvious non-Marquette authors, I cleaned the list by using the same "Conditional Formatting …" functionality and searching for "Text that contains" birthdates in the 1600s, 1700s, and early 1800s.

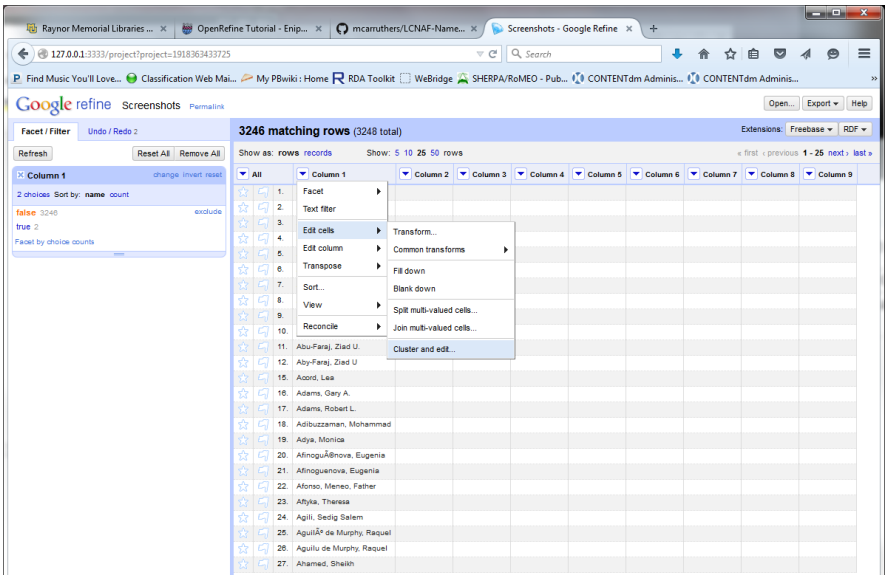
Open Refine: data cleanup

Once the data was in one column and the preliminary cleanup work done, I was ready to use Open Refine (OR) to clean up the data more thoroughly. From there, I could reconcile the names against VIAF (Virtual International Authority File: https://viaf.org/), and LCNAF (Library of Congress Name Authority File: http://authorities.loc.gov/).

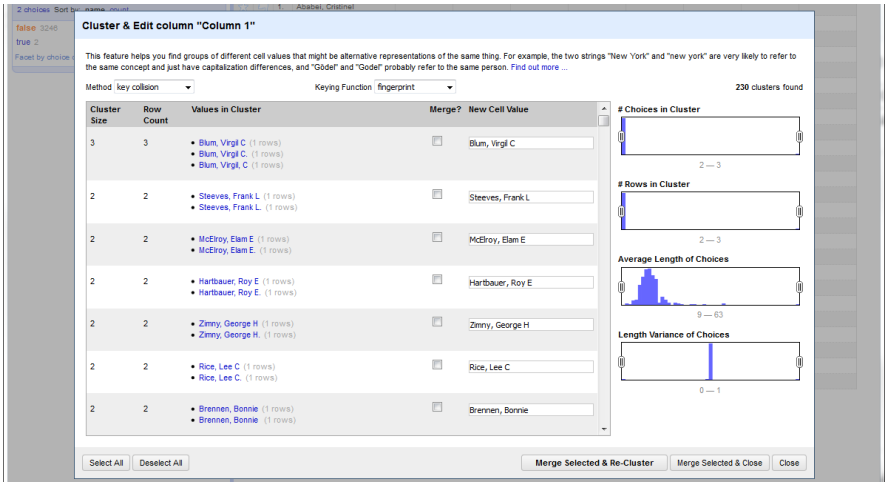
Each spreadsheet became a separate project in Open Refine.

I first used Open Refine to eliminate duplicates that Excel couldn't: ending punctuation differences, the presence or absence of middle initials, single letter typos in the name, etc.

To de-duplicate values in OR, I used the "Edit Cells" and "Cluster and Edit" functionality.

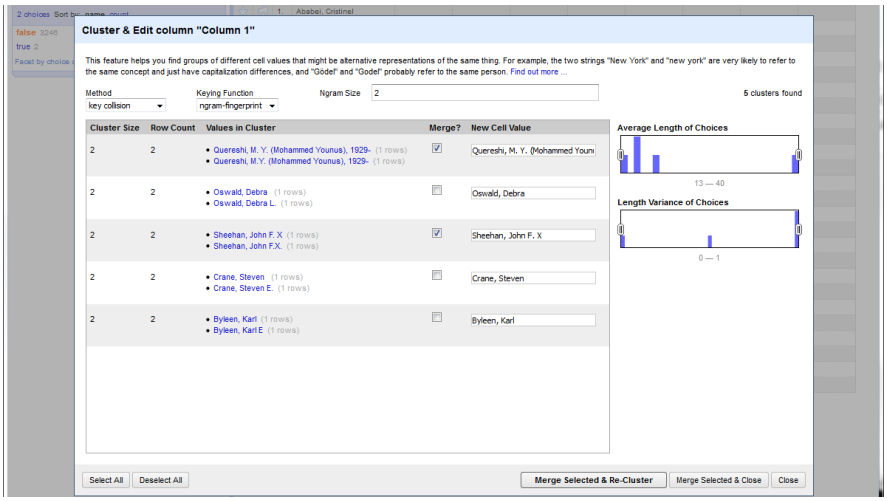


The default clustering algorithm is method = key collision and keying function = fingerprint. This algorithm caught many variants of the same name. In my case, they were typically with punctuation variations.

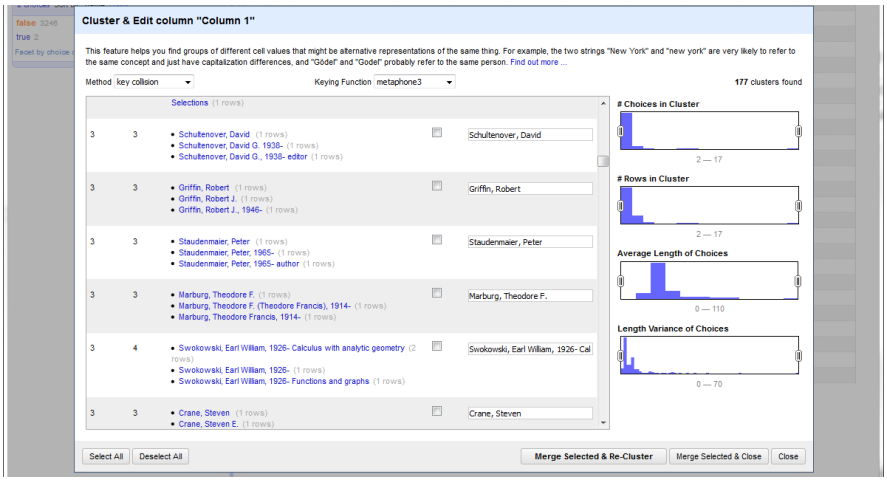


For the key collision/fingerprint clustering, I would simply click “Select All”, then “Merge Selected & Re-cluster” to merge the clusters and check for stragglers. Once OR stopped finding clusters with the fingerprint method, I changed the keying function to ngram-fingerprint, with ngram size=2. This seemed to catch headings where one had a middle initial and another didn’t. You might or might not want to merge these, as an entry with/without a middle initial could match in VIAF/LCNAF.

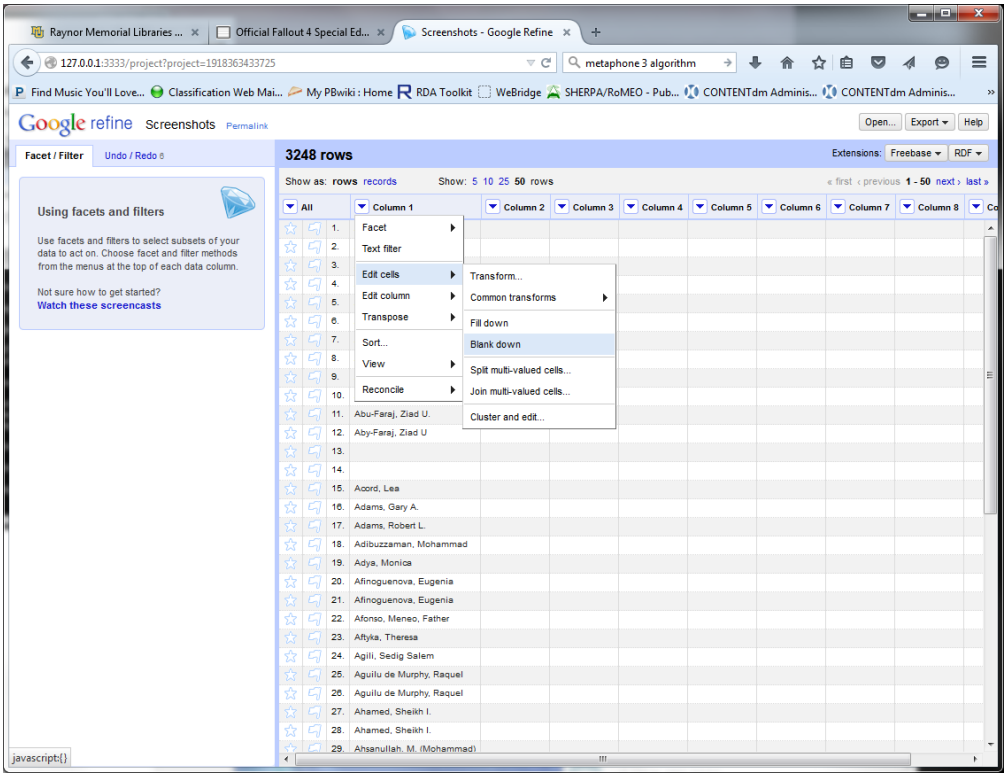
In this example, there were only two headings I wanted to merge.



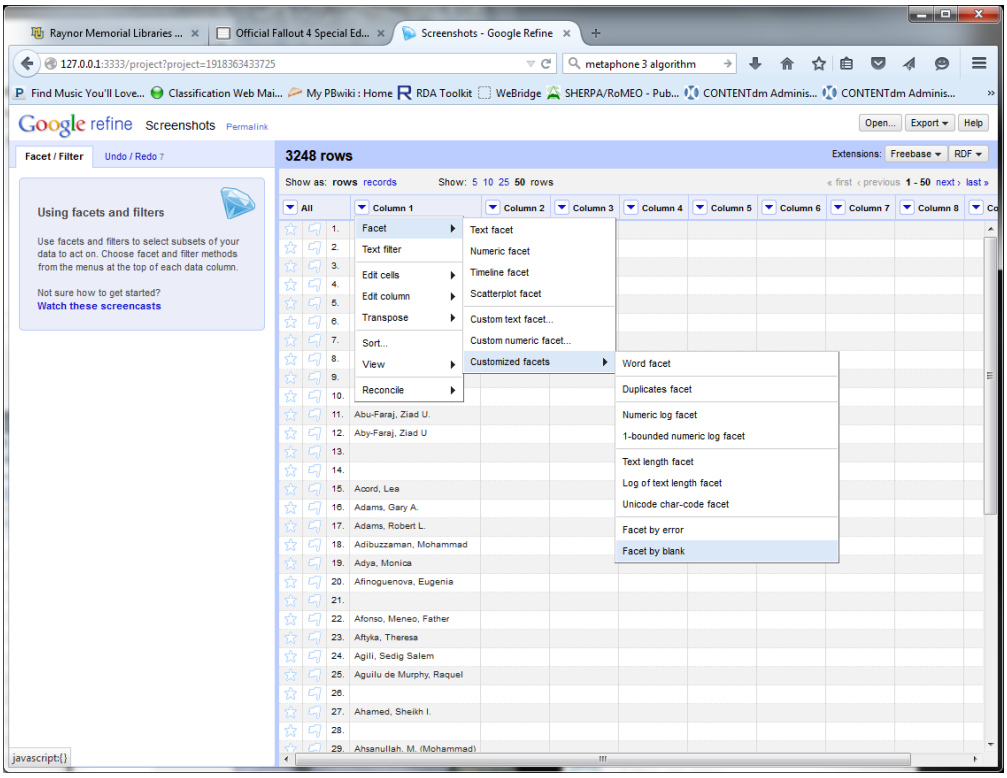
For my last de-duplication pass, I selected the “metaphone3” keying function. This caught bigger variations: instances where there was a date present, where middle names spelled out, or there were names with uniform titles.
This method casts a much broader net, so it is worth the time to scan through the list and select the headings you want to merge.



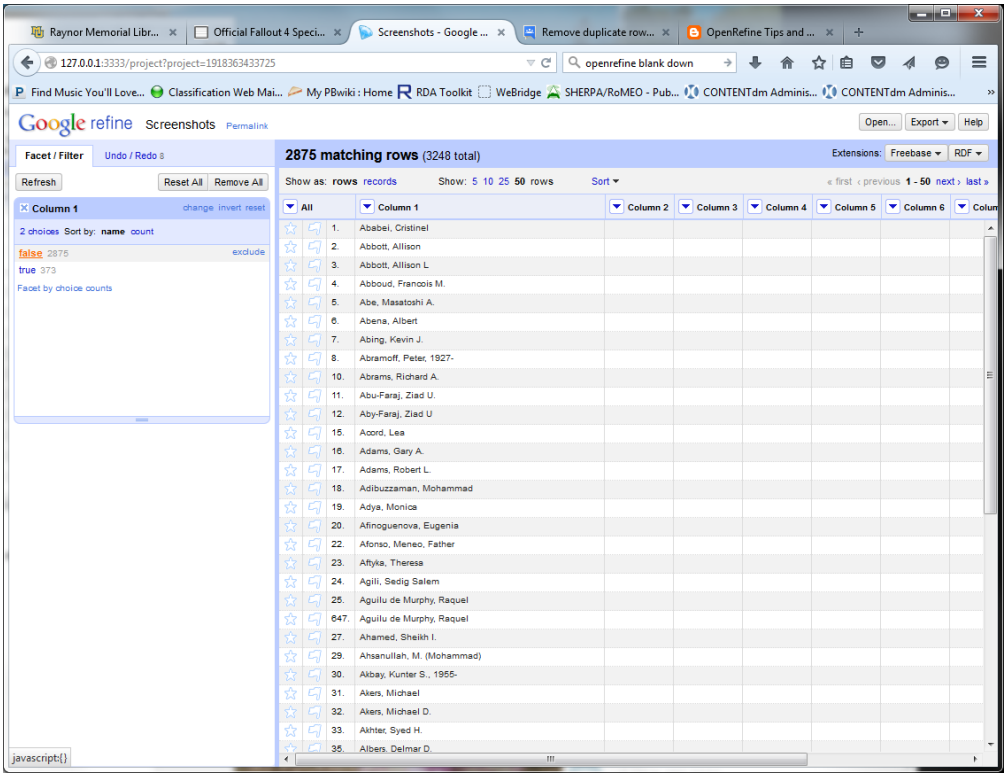
Once I had the cells clustered and edited to my satisfaction, I re-sorted the column first, just in case a name changed enough to move it in the list.
After sorting, it was time to remove duplicates. In OR, this is done by “Edit cells” – “Blank down.” The “Blank down” function works by blanking every row that matches the value of the preceding one based on the row ID. (http://googlerefine.blogspot.ca/2011/08/remove-duplicate.html)



Once duplicate values were blanked out, I hid the blank rows by applying “Facet” – “Customized facets” – “Facet by blank”



Selecting “false” in the facet box on the left will result in a display of rows that have data. In my case, this was 2875 rows out of 3248.



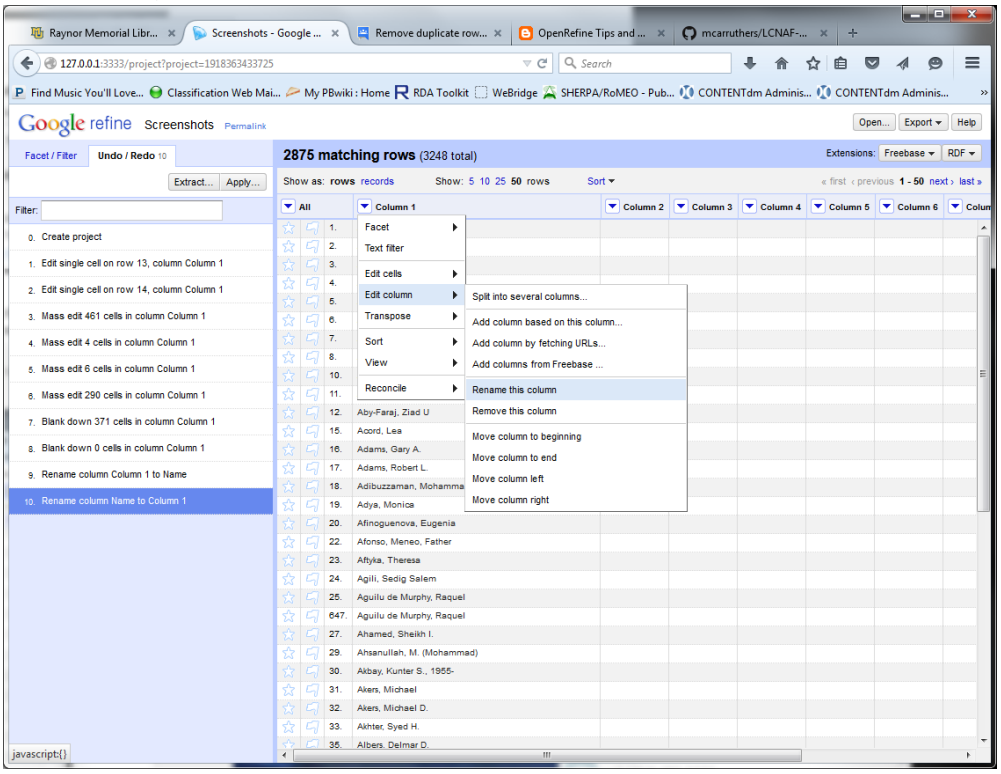
Open Refine: reconciliation

Once I had a fairly clean list of names, I wanted to check them against VIAF/LCNAF to see if possible authority records already existed.

Luckily for me, Matt Carruthers (Metadata Projects Librarian at University of Michigan) has done a huge amount of heavy lifting in creating JSON scripts to run in OpenRefine to accomplish this very task. The scripts can be found on GitHub at https://github.com/mcarruthers/LCNAF-Named-Entity-Reconciliation. There are three different flavors of script: Personal Names, Corporate Names, and Generic Names. I recommend reading through Matt's instructions on using the scripts before proceeding.

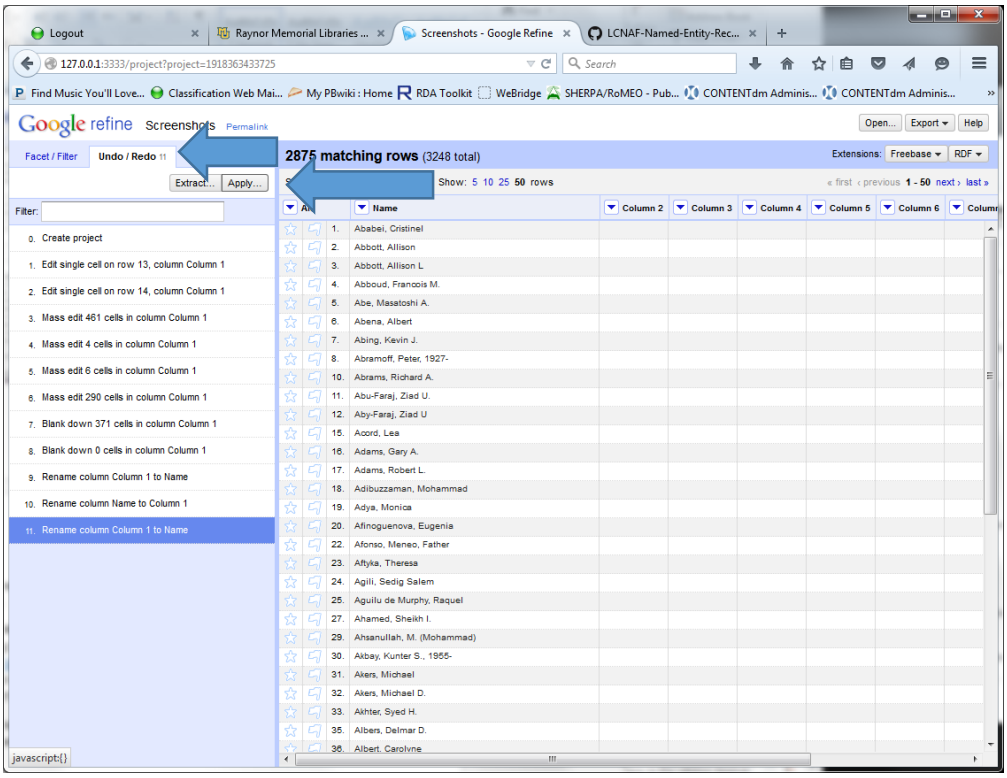
I used the Personal Names script for the Marquette Faculty list, and the Generic Names script for the Cujé list as that one was a mix of personal and corporate names.

The first thing I needed to do was to change the OR column name to "Name" ("Edit column" -- "Rename this column")

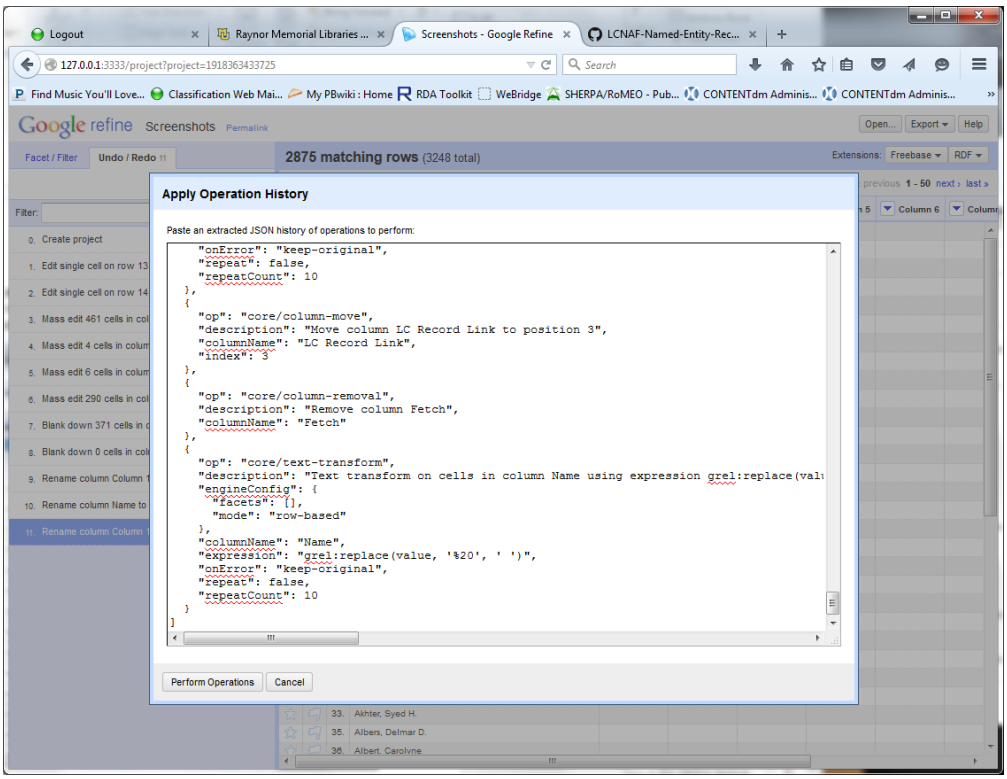


In this example, I was working with the Marquette Faculty list, so I copied the Personal Names script from the GitHub .txt file, and then pasted it into the "Apply Operation History" window.

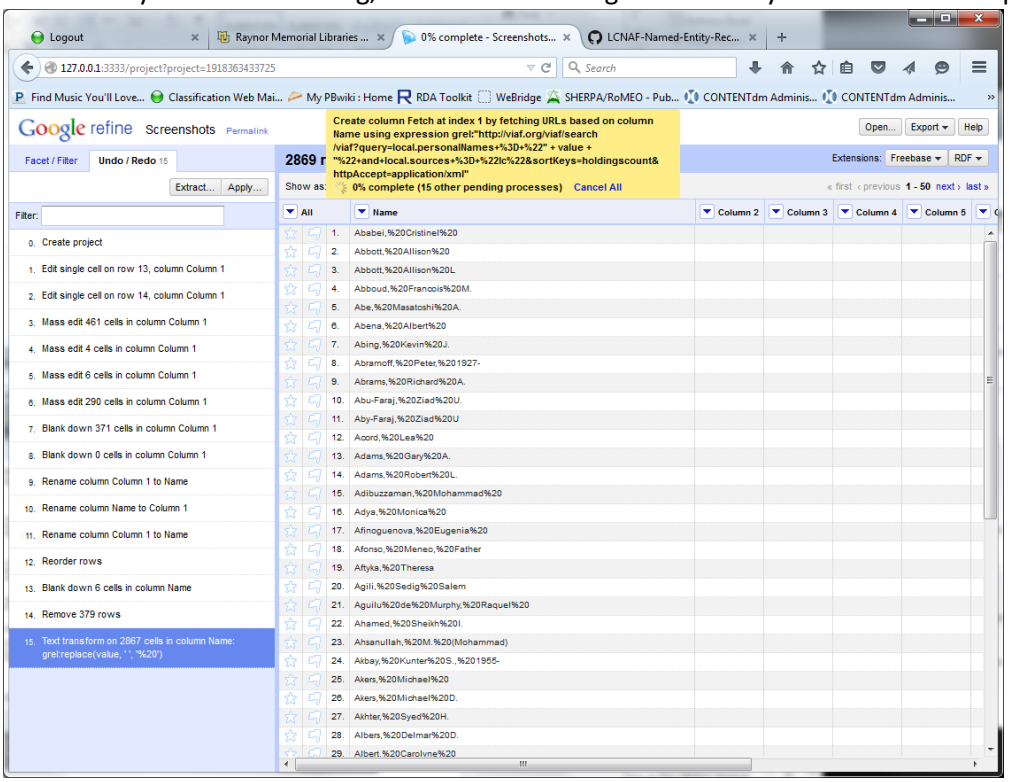
To get to the "Apply Operation History" window, click the "Undo/Redo" tab (next to the "Facet/Filter" tab), and then click the "Apply" button



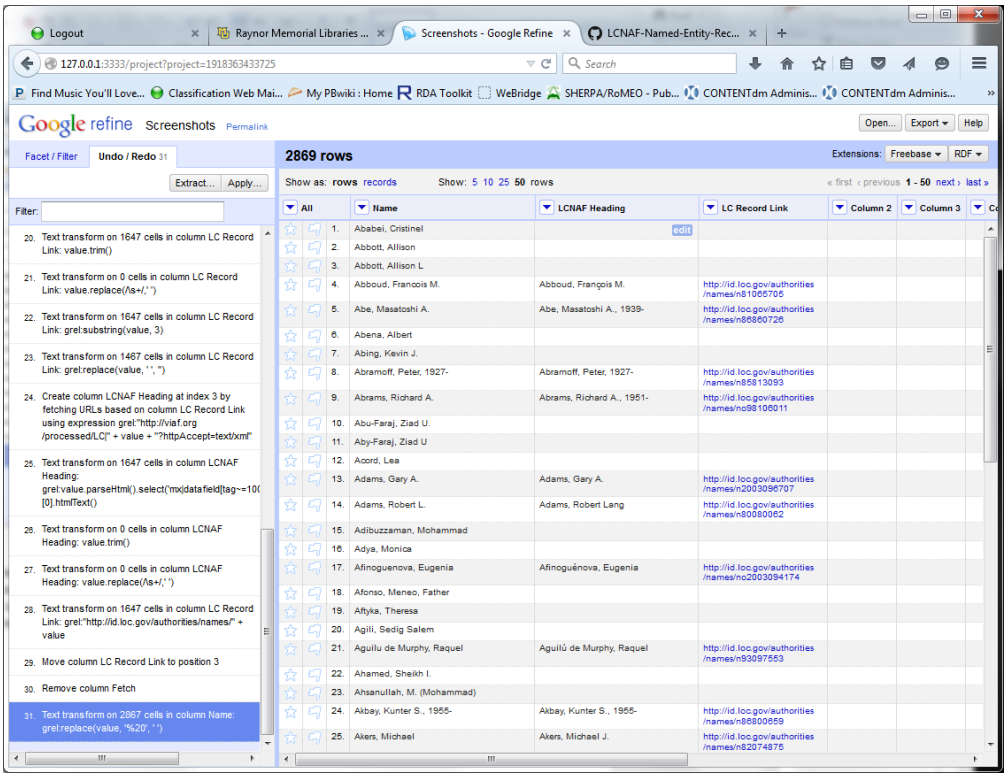
Once the script is pasted, click the “Perform Operations” button and OR will be off and searching!



OR will tell you what it is doing, and how far it has gotten in the yellow box at the top of the window.



Fetching the names by URL *does* take a little time. The list of almost 3000 names took about an hour to process, as the script has OR fetch twice: once from VIAF to get a list of matches, and then again from VIAF to build the LC link if the VIAF record(s) have LCNAF numbers. So if you have a large dataset, you may want to break it into smaller chunks before running it through OR.



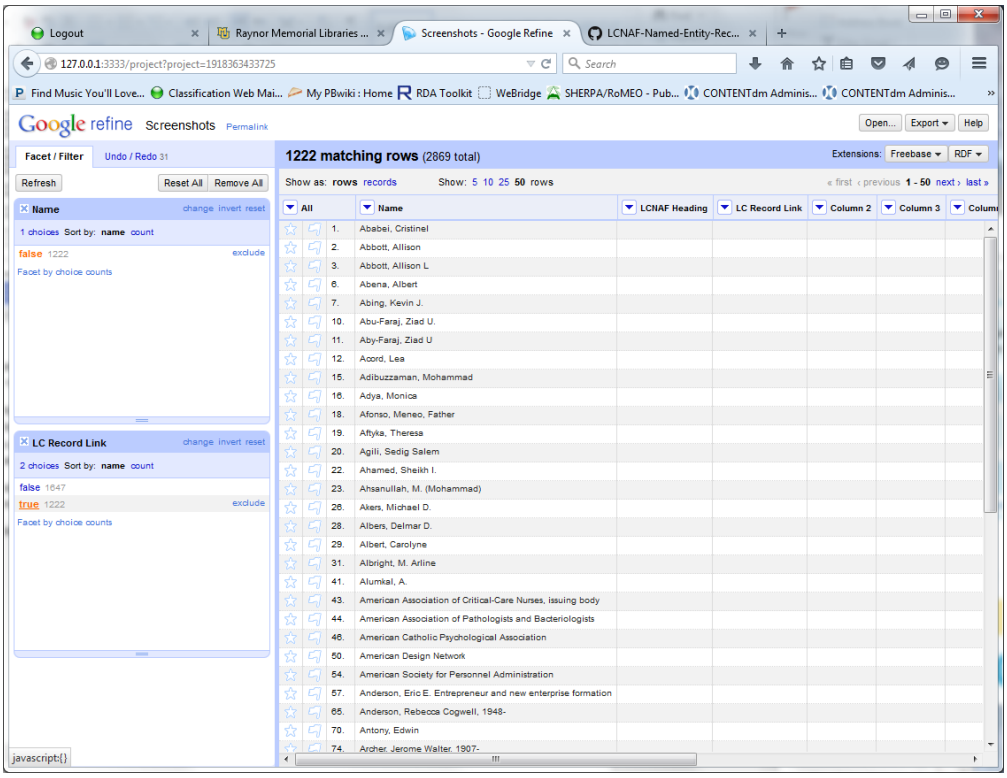
The end result looks like this: a spreadsheet with the name from your database, the LCNAF heading, and the LCNAF link – nice! Thanks Matt!

Open Refine: creating a list of potential Authority names

For the purposes of an Authority record creation project though, I was interested in the names where the process *didn't* match anything.

As an aside, let's not forget that just because OR/VIAF found a match, it doesn't mean it was the right match. These are only machines, after all. I've decided to focus first on things that aren't matching anything. After that project is complete, we can go back and check on names that matched to see if they were matched correctly.

To find names in my list that didn't match in VIAF/LCNAF, I used the column "Facet by blank" option, faceting either on the LCNAF Heading column, or the LC Record link Column.
This time, I selected "true" from the facet box



These remaining 1222 headings are where we will invest our initial work, first by verifying that no authority record actually exists for the heading, and then, if there is no authority record, by creating the new authority record.

Sources consulted

Open Refine (formerly Google Refine):

- Software : http://openrefine.org/
- Free Your Metadata video tutorial (if you're new to Open Refine, this is a good place to start): https://www.youtube.com/watch?v=NnCA1dnCT-c also at http://freeyourmetadata.org/
- Google Refine video tutorials : https://www.youtube.com/channel/UCqwSVsJ8CWD9pQUZDbJC1ew
- Google group: https://groups.google.com/forum/#!forum/openrefine
- GitHub wiki: https://github.com/OpenRefine/OpenRefine/wiki
- Stephens, O. on behalf of the British Library. (2014, July 1). Introduction to OpenRefine. Retrieved October 7, 2015, from http://www.meanboyfriend.com/overdue_ideas/wp-content/uploads/2014/11/Introduction-to-OpenRefine-handout-CC-BY.pdf

GitHub:

- VIAF/LCNAF reconciliation: https://github.com/mcarruthers/LCNAF-Named-Entity-Reconciliation