Marquette University e-Publications@Marquette

Library Faculty Research and Publications

Library (Raynor Memorial Libraries)

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

Unpublished paper. © 2015 The Author. Used with permission.

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 (<u>http://libus.csd.mu.edu/</u>) or Institutional Repository (<u>http://epublications.marquette.edu/</u>) 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 deduplicated 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: <u>https://viaf.org/</u>), 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.

						_ _ ×
😼 Raynor Memorial Libraries 🗙 👹 OpenRef	ine Tutorial - E	nip × 📿 mcarruth	ners/LCNAF+Name 🗙 🕟 S	creenshots - Google Refine 🗙	+	
127.0.0.1:3333/project?project=1918363433725			⊽ C'	Q, Search	🕹 🏠 🕹	ê ♥ ⋪ ♥ Ξ
🗜 Find Music You'll Love 😝 Classification Web Mai	🔑 My PBwi	ki : Home 尺 RDA Tooll	kit 🗍 WeBridge ᄶ SHERPA/R	oMEO - Pub 🚺 CONTENTdm	Adminis 🚺 CONTENT	dm Adminis >
Google refine Screenshots Permalink						Open Export - Help
Facet / Filter Undo / Redo 2	3246 mat	ching rows (3248 to	otal)		1	Extensions: Freebase 💌 RDF 💌
Refresh Reset All Remove All	Show as: row	vs records Show	r: 5 10 25 50 rows		e	first oprevious 1 - 25 next> last>
X Column 1 change invert reset	💌 All	Column 1	Column 2 Column 3	💌 Column 4 💌 Column 5 💌	Column 6 💌 Column 7	Column 8 Column 9
2 choices Sort by: name count	숤 티 1.	Facet +				
false 3248 exclude	\$ 92	Text filter				
true 2	☆ 의 3.	Edit calls b	Transform	1		
Facet by choice counts	☆ 되 4	Edit collar P	Transform			
	☆ 의 *	Edit column	Common transforms			
	☆ 다 8.	Transpose F	Fill down			
	☆ 딕 7.	Sort	Blank down			
	없 대 8.	View 🕨	Split multi-valued cells			
	었니 9. ☆ 니 10.	Reconcile 🕨	Join multi-valued cells			
	☆ 의 11.	Abu-Faraj, Ziad U.	Cluster and edit			
	公 5 12.	Aby-Faraj, Ziad U]		
	☆ 🗐 15.	Acord, Lea				
	☆ 🖓 18.	Adams, Gary A.				
	☆ 의 17.	Adams, Robert L.				
	☆ 🖓 18.	Adibuzzaman, Mohamma	d			
	☆ 🗐 19.	Adya, Monica				
	☆ 🧐 20.	Afinoguînova, Eugenia				
	☆ 딕 21.	Afinoguenova, Eugenia				
	없 더 22.	Afonso, Meneo, Father				
	\$2 - √ 23.	Aftyka, Theresa				
	23 4 24.	Agiii, Sedig Salem				
	· · · · · · · · · · · · · · · · · · ·	Aguilu de Murphy, Raque	*1			
	ン· G 27	Abamed Sheith				
	w ~ ~ ~.	Ponames, Shieldh				

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.

the same con	cept and just	a groups of different cell values that might be a have capitalization differences, and "Gödel" a	alternative representations of the s and "Godel" probably refer to the s	ame thing. For example, the two strings ame person. Find out more	"New York" and "new york" are very likely to refer to
Method key collision Cluster Row Values in Cluster Size Count		 Keying 	Function fingerprint -		230 clusters four
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	# Choices in Cluster
3	3	 Blum, Virgil C (1 rows) Blum, Virgil C. (1 rows) Blum, Virgil, C (1 rows) 		Blum, Virgil C	2-3
2	2	Steeves, Frank L (1 rows) Steeves, Frank L. (1 rows)		Steeves, Frank L	# Rows in Cluster
2	2	McEiroy, Elam E (1 rows) McEiroy, Elam E. (1 rows)	F	McElroy, Elam E	2 — 3 Average Length of Choices
2	2	Hartbauer, Roy E (1 rows) Hartbauer, Roy E. (1 rows)		Hartbauer, Roy E	
2	2	Zimny, George H (1 rows) Zimny, George H. (1 rows)		Zimny, George H	9 – 63 Length Variance of Choices
2	2	Rice, Lee C (1 rows) Rice, Lee C. (1 rows)		Rice, Lee C	0-1
2	2	Brennen, Bonnie (1 rows) Brennen, Bonnie (1 rows)		Brennen, Bonnie	

For the key collision/fingerprint clustering, I would simply click "Select All", then "Merge Selected & Recluster" 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.

Sort by: nar	me count		Ababei, Cristinel			
CIL	uster & Ed	it column '	'Column 1"			
the	is feature help e same concep	s you find grou and just have	ps of different cell values that might be alternative represen capitalization differences, and "Gödel" and "Godel" probable	tations of the si y refer to the si	ame thing. For example, the two string ame person. Find out more	s "New York" and "new york" are very likely to refer to
Me	ethod ey collision	•	Keying Function Ngram Size 2			5 clusters found
С	luster Size	Row Count	Values in Cluster	Merge?	New Cell Value	Average Length of Choices
2		2	Quereshi, M. Y. (Mohammed Younus), 1929- (1 rows) Quereshi, M.Y. (Mohammed Younus), 1929- (1 rows)	V	Quereshi, M. Y. (Mohammed Youni	
2		2	Oswald, Debra (1 rows) Oswald, Debra L. (1 rows)		Oswald, Debra	13 — 40 Length Variance of Choices
2		2	Sheehan, John F.X. (1 rows) Sheehan, John F.X. (1 rows)	V	Sheehan, John F. X	0-1
2		2	Crane, Steven (1 rows) Crane, Steven E. (1 rows)		Crane, Steven	
2		2	Byleen, Karl (1 rows) Byleen, Karl E (1 rows)		Byleen, Karl	
S	elect All D	eselect All			Merge Selected	& Re-Cluster Merge Selected & Close Close

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.

This feature the same co	helps you ncept and j	find groups of different cell values that m just have capitalization differences, and "	ght be alternative representati Södel" and "Godel" probably re	ons of the fer to the	same thing. For example, the two strings same person. Find out more	"New York" and "new york" are very likely to refer to
Method key	collision	•	Keying Function metaphone3	•		177 clusters foun
		Selections (1 rows)			^	# Choices in Cluster
3	3	Schultenover, David (1 rows) Schultenover, David G. 1938- (1 rows Schultenover, David G., 1938- editor () Irows)		Schultenover, David	
						# Rows in Cluster
3	3	Griffin, Robert (1 rows) Griffin, Robert J. (1 rows) Griffin, Robert J., 1946- (1 rows)			Griffin, Robert	
3	3	Staudenmaier. Peter (1 rows)			Staudenmaier, Peter	2 — 17
		Staudenmaier, Peter, 1965- (1 rows) Staudenmaier, Peter, 1965- author (1 r	ows)			Average Length of Choices
3	3	Marburg, Theodore F. (1 rows)			Marburg, Theodore F.	
		Marburg, Theodore F. (Theodore France Marburg, Theodore Francis, 1914- (1)	is), 1914- (1 rows) ows)			0 — 110
3	4	Swokowski Earl William 1926- Calculu	is with analytic geometry (2		Sunknucki Earl William 1926- Cal	Length Variance of Choices
		rows) • Swokowski, Earl William, 1926- (1 row • Swokowski, Earl William, 1928- Function	s) Ins and graphs (1 rows)		Shoonad, carringin, 1920 Car	0-70
3	3	Crane, Steven (1 rows) Crane, Steven E (1 rows)			Crane, Steven	

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. (<u>http://googlerefine.blogspot.ca/2011/08/remove-duplicate.html</u>)

🖲 Raynor Memorial Libraries 🗙 🔲 Official	Fallout 4 Special	Ed × 🕟 Screenst	iots - Google Refine	× +						×
127.0.0.1:3333/project?project=191836343372!	5		⊽ C	Q metap	hone 3 algorith	m → 1	▶ ⋒ ☆	e 🛡	1 O	=
P Find Music You'll Love 😝 Classification Web Ma	i 🔎 My PBwi	ki : Home 尺 RDA Tool	kit 🔅 WeBridge 🏅	SHERPA/R	MEO - Pub		m Adminis	CONTENT	lm Adminis	»
Google refine Screenshots Permalink								Open	Export 🕶	Help
Facet / Filter Undo / Redo 8	3248 row	5						Extensions:	reebase 👻 🛛 R	DF 🔻
	Show as: rov	vs records Show	r: 5 10 25 50 rows	•				« first « previou	is 1 - 50 next >	last »
Using facets and filters	💌 All	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Co
Use facets and filters to select subsets of your data to act on Choose facet and filter methods	2725	Â								
from the menus at the top of each data column.	☆ 더 3.	Edit cells	Transform							
Not sure how to get started? Watch these screencasts	☆ 년 5.	Edit column	Common transform	s 🕨						
	☆ 더 •.	Transpose >	Fill down							
	☆ 딕 8.	Sort	Blank down							
	☆ 덕 9.	Reconcile	Split multi-valued c	ells						E
	었 니 10. ☆ 니 11.	Abu-Faraj, Ziad U.	Cluster and edit							
	☆ 대 12.	Aby-Faraj, Ziad U	Cluster and colt							
	☆ 덕 13.									
	었 더 14.	Acced Las								
	었 더 10. 숫 더 18.	Adams, Gary A.								
	☆ 대 17.	Adams, Robert L.								
	숤 디 18.	Adibuzzaman, Mohamma	d							
	☆ 🧐 19.	Adya, Monica								
	☆ 🗐 20.	Afinoguenova, Eugenia								
	☆ 듸 21.	Afinoguenova, Eugenia								
	값 되 22.	Afonso, Meneo, Father								
	값 니 23.	Aftyka, Theresa								
	12 4 24.	Agili, Sedig Salem								
	∴ ∽ ^{20.}	Aguilu de Murphy, Raque								
	······································	Ahamed, Sheikh I.	-							
	☆ 대 28.	Ahamed, Sheikh I.								
	1 29.	Ahsanullah, M. (Mohamm	ad)							-
javascript:{}	•									- F

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

📲 Raynor Memorial Libraries 🗙 🔲 Officia	l Fallout 4 Speci	al Ed 🗙 🍗 Screensh	ots - Google Refine 🗙 🕂	-	-				x
127.0.0.1:3333/project?project=19183634337	25		⊽ C ^e Q metar	phone 3 algorithm) <i>></i> 1		i	4 0	=
P. Find Murie You'll Love	ai 🧢 May DRu	viki : Home 🗖 RDA Tooli	kit 🗍 WaBridge 🧟 SHEPDA/P	RoMEO - Pub		m Adminic		Im Adminic	
	al 🛏 Iviy Pov	NATIONE R NOA 100	ar webnuge 🚗 shear An	COMEO - Public V	CONTENTS	in Automatic	CONTENT		
Google refine Screenshots Permalin	C						Open	. Export 🕶	Help
Facet / Filter Undo / Redo 7	3248 rov	/s					Extensions:	Freebase 👻 🛛 Ri	DF 🕶
	Show as: ro	ws records Show	: 5 10 25 50 rows				« first < previou	is 1 - 50 next>	last »
Using facets and filters	💌 All	Column 1	Column 2 Column 3	3 💌 Column 4	Column 5	Column 6	Column 7	Column 8	Co
	숤 디 1.	Facet >	Text facet						<u>^</u>
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods	☆ 듸 2	Text filter	Numeric facet						
from the menus at the top of each data column.	었 더 3.	Edit cells	Timeline facet						
Not sure how to get started?	* 5	Edit column	Scatterplot facet						
watch these screencasts	☆ 대 6.	Transpose 🕨	Custom text facet						
	숤 디 7.	Sort	Custom numeric facet						
	☆ 57 8.	View 🕨	Customized facets	Word facet					
	었 더 9.	Reconcile +		Duplicates facet					
	ン	Abu-Farai, Ziad U.]	Numeric log facet	+				
	☆ 더 12	Aby-Faraj, Ziad U		1-bounded numer	ric log facet				
	☆ 🗐 13	ц. —		Taxt langth facat			_		
	☆ 되 14			Log of text length	h facet				
	값 더 18	Acord, Lea		Unicode char-cor	de facet				
	ン ジ ロ 12	Adams, Gary A.					_		
	☆ 역 18	Adibuzzaman, Mohamma	d	Facet by error					
	☆ 디 15	. Adya, Monica		r acer by blank					
	☆ 듸 20	Afinoguenova, Eugenia							
	☆ - 21								
	22 4 22	Afonso, Meneo, Father							
	12 5J 24	. Agili, Sedig Salem							
	☆ 대 25	- Aguilu de Murphy, Raquel							
	☆ 57 26	L							
	☆ 🤤 27	Ahamed, Sheikh I.							
	었 더 28	. Abaraullah M. (Maharara	- 40						-
javascript:{}	1 25	. Ansanullan, M. (Mohamm	801						•

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.

Raynor Memorial Libr × Official Fallout 4 Speci × Screenshots - Google × Remove duplicate row × OpenRefine Tips and × + I 27.00.13333/projecthproject=191836483725 <								
127.0.0.1:3333/project?project=191836343372	5	▼ C' (copenrefine blank d	own →	↓ ☆	☆ 自 ₪	, ∢ ⊝	=
				-				
P Find Music You'll Love 😝 Classification Web Ma	i 🔎 My PBw	iki : Home 尺 RDA Toolkit 🔅 WeBridge 🏔 Sl	HERPA/RoMEO - Pub.	🚺 CONTEN	ITdm Adminis	🚺 CONTEN	ITdm Adminis.	»
Google refine Screenshots Permalink	Alther, × Official Fallout 4 Specia, × Screenshots - Google × C Remove duplicate row × SopenRefine Tips and × + Vprojects 1918/3633323 × C openrefine blank down + A A B E vec. Classification Web Mai My Bowlis Home R RDA Toolkit. Webridge A SHERPA/Robel > Public & CONTENTid m Adminis * Contentid m Adminis<							
Facet / Filter Undo / Redo 8	emorial Libr X Official Fallout 4 Specie X Screenshots - Google X Remove duplicate row X OpenRefine Tips and X + 1:333/project/projects/1918/66/337275	2875 matching rows (3248 total)						
Refresh Reset All Remove All	Show as: ro	ws records Show: 5 10 25 50 rows	Sort -			« first < prev	rious 1 - 50 ne x	d⇒ last »
X Column 1 change invert reset		Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Colum
2 choices Sort by: name count	숤 듸 1.	Ababei, Cristinel						*
falso 2075 exclude	\$ 57 2	Abbott, Allison						
true 373	☆ 디 3.	Abbott, Allison L						
Facet by choice counts	☆ 되 4.	Abboud, Francois M.						
	☆ 대 5.	Abe, Masatoshi A.						
	☆ 다 .	Abena, Albert						
	숬 듸 7.	Abing, Kevin J.						
	☆ 듸 8.	Abramoff, Peter, 1927-						
		E						
	☆ 대 11.	Abu-Faraj, Ziad U.						
	☆ 듸 12	Aby-Faraj, Ziad U						
	☆ 🤤 15	Acord, Lea						
	숤 디 18	Adams, Gary A.						
	☆ 5 17	Adams, Robert L.						
	☆ 🗐 18	Adibuzzaman, Mohammad						
	☆ 듸 19	Adya, Monica						
	☆ 🧐 20	Afinoguenova, Eugenia						
	☆ 듸 22	Afonso, Meneo, Father						
	☆ 🦪 23	Aftyka, Theresa						
	☆ 듸 24	Agili, Sedig Salem						
	☆ 🦪 25	Aguilu de Murphy, Raquel						
	☆ 뎕 64	7. Aguilu de Murphy, Raquel						
	☆ 🗐 27	Ahamed, Sheikh I.						
	☆ 5기 29	Ahsanullah, M. (Mohammad)						
	☆ 🦪 30	Akbay, Kunter S., 1955-						
	☆ 듸 31	Akers, Michael						
	☆ 🧐 32	Akers, Michael D.						
	☆ 🎝 33	Akhter, Syed H.						
iavascript:{}	\$ 5 35	Albers. Delmar D.						

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 <u>https://github.com/mcarruthers/LCNAF-Named-Entity-Reconciliation</u>. 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")

	-						0					×
Raynor Memorial Libr × Screenshots	- Goo	gie	×	Remove duplicate	row × 🕒 OpenRefine	Tips and ×	Ç) mcarru	thers/LCNAF-	···· × +			_
I27.0.0.1:3333/project?project=191836343372	5				⊽ Cª Q, Sea	rch		₩ 1				=
🗜 Find Music You'll Love 😝 Classification Web Ma	i 🔎	^a My	PBwik	i : Home 尺 RDA Tooll	cit 📋 WeBridge 🎇 SHERPA	/RoMEO - Pub	🚺 CONTEI	NTdm Admini	s 🚺 CONTE	NTdm Admini	s	×
Google refine Screenshots Permalink									Op	en Export	• He	elp
Facet / Fiter Undo / Redo 10	28	75	matc	hing rows (3248 to	tal)				Extensions	Freebase 🔻	RDF	-
Extract Apply	Shr	ow as	: row	s records Show	: 5 10 25 50 rows So	ert 🔻			« first « pre	vious 1 - 50 n e	ext⇒ la:	st »
		A.II		Column 4		Column 2	Column 2	Calumn 4	Column E	Column 6		Celu
Filter:	÷		4	Eacot b	1	· Column 2	- column s	Columna	Columno	Column o	10	COIL
 Create project 	24		2	Taut files								- f
1. Edit single cell on row 13, column Column 1	52		3.	Text filter								
- Education and a solution Optimized	53		4.	Edit cells 🕨								
2. Edit single cell on row 14, column Column 1	컶		5.	Edit column 🕨	Split into several columns							
3. Mass edit 461 cells in column Column 1	গ্ন		6.	Transpose 🕨	Add column based on this colu	umn						
4. Mass edit 4 cells in column Column 1	岔		7.	Sort +	Add column by fetching URLs							
a. Mass edit 6 cells in column Column 1	শ্বি		8.	View •	Add columns from Freebase .							
	. द्रि		10.	Reconcile •	Papama this column							
 Mass edit 290 cells in column Column 1 	23		11.		Remaine this column							
7. Blank down 371 cells in column Column 1	24		12.	Aby-Paraj, 21ad U	Remove this column							
8 Blank down 0 cells in column Column 1	23		16	Adams Gary A	Move column to beginning							
	- 53		17.	Adams, Robert L.	Move column to end							
9. Rename column Column 1 to Name	**		18.	Adibuzzaman, Mohamm	Move column left							
	公		19.	Adya, Monica	Move column right							
	☆		20.	Afinoguenova, Eugenia								
	53		22.	Afonso, Meneo, Father								
	☆		23.	Aftyka, Theresa								
	ជ		24.	Agili, Sedig Salem								
	었		25.	Aguilu de Murphy, Raque	el .							
	23		647.	Aguilu de Murphy, Raque	21							
	24		27.	Anamed, Sneikh I.	and)							
	22 57		30.	Akbay, Kunter S., 1955-								
	숤		31.	Akers, Michael								
	☆		32.	Akers, Michael D.								
	슔		33.	Akhter, Syed H.								
iavaccrint/\	42		35.	Albers. Delmar D.							1	
Jakasenhett	1											P

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

 127.0.0.1:3333/project?project=191836343 	3725	⊽ C'	Q, Search			☆自	J 4 6) ≡
Find Music You'll Love Grace Classification Web	Mai 🔑 My PBv	iki : Home 尺 RDA Toolkit 🗍 WeBridge 🕰	SHERPA/RoMEO - Pu	ь 🚺 солте	NTdm Admini	s 🚺 CONTE	NTdm Admini	s ,
Google refine Screenshots Remain	ink					01	pen Export •	✓ Help
		••				Entracion	En la la	L D D D
Facet / Filter Undo / Redo 11	2875 ma	ching rows (3248 total)				Extension	s: Freebase +	RDF 🕈
Extract Apply.		Show: 5 10 25 50 rows				« first « pre	evious 1 - 50 ne	⊭xt⇒last⇒
iter:		▼ Name	Column 2	Column 3	Column 4	Column 5	Column 6	Colu
0. Create project	☆ 년 1.	Ababei, Cristinel						
 Edit algele call as roug 12, aslume Column 1. 	값 더 2	Abbott, Allison						
1. Eait single ceiron row 13, column column 1	X 4 *	Abbott, Allison L						
2. Edit single cell on row 14, column Column 1		Abbodd, Flandols M.						
3. Mass edit 461 cells in column Column 1	소의*	Abena, Albert						
4 Mass edit 4 cells in column Column 1	* 517.	Abing, Kevin J.						
	☆ 디 8.	Abramoff, Peter, 1927-						
5. Mass edit 6 cells in column Column 1	☆ 🗐 10	Abrams, Richard A.						
e. Mass edit 290 cells in column Column 1	☆ 与 11	Abu-Faraj, Ziad U.						
7 Blank down 371 cells in column Column 1	☆ 🎵 12	Aby-Faraj, Ziad U						
	☆ 与 15	Acord, Lea						
 Blank down 0 cells in column Column 1 	☆ 🗐 18	Adams, Gary A.						
9. Rename column Column 1 to Name	값 되 17	Adams, Robert L.						
10 Rename column Name to Column 1	::::::::::::::::::::::::::::::::::::::	Adibuzzaman, Mohammad						
	었 <i>니</i> 19	Adya, Monica						
11. Rename column Column 1 to Name		Afrance Meneo Eather						
	× ~ 23	Aftyka Theresa						
	\$ 5 5 24	Agili, Sedig Salem						
	* 5 25	Aguilu de Murphy, Raguel						
	☆ 되 27	Ahamed, Sheikh I.						
	☆ 与 29	Ahsanullah, M. (Mohammad)						
	☆ 듸 30	Akbay, Kunter S., 1955-						
	☆ 🎝 31	Akers, Michael						
	☆ 듸 32	Akers, Michael D.						
	☆ 🦪 33	Akhter, Syed H.						
	A (7) 25	Albert, Deleter D						

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

() 🛞 127.	.0.0.1:3333/projec	t?project=1918363433725		⊽ C ⁱ	Q Search	+	俞 ☆	ê 🛡		9 =
Find Musi	c You'll Love 🧲	Classification Web Mai.	🔑 My PBwiki : Home 尺 RD	A Toolkit 🗌 WeBridge 🏔	SHERPA/RoMEO - Pub	ONTENTdm A	dminis 🚺	CONTEN	ITdm Admir	nis
Google	e refine so	reenshots Permalink						Ope	n Export	t 👻 Help
Facet / Filter	Undo / Redo	11	2875 matching rows (3	3248 total)			E	xtensions:	Freebase -	• RDF •
		Apply Operation Hi						; prev	ious 1 - 50 -	next> last »
ilter:		Apply Operation Hi	story					15	Column 6	S 💌 Colu
o. Create p	roiect	Paste an extracted JSON	history of operations to perform:							
 Edit singl Edit singl Edit singl Mass ed 	le cell on row 13 le cell on row 14 it 461 cells in col	"onError": "repeat": f. "repeatCoun" }, { "op": "core	<pre>"keep-original", alse, t": 10 /column-move", " "Your column to be " " "Your column to be " " " ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' ' '</pre>				^			
4. Mass ed	it 4 cells in colum	"columnName "index": 3	": "LC Record Link",	cord Link to posit.	10n 3",					
5. Mass ed	it 6 cells in colum	1								
 Mass ed Blank do Blank do 	it 290 cells in col wn 371 cells in c wn 0 cells in coli	"op": "core. "description "columnName }, { "op": "core.	<pre>/column-removal", n": "Remove column Fet ": "Fetch" /text-transform",</pre>	ch",						
9. Rename	column Column 1	"description "engineConf "facets":	n": "Text transform on ig": { [],	a cells in column N	ame using express	ion grel:repla	ice (valı			
11. Rename	column Column 1	<pre>"mode": ".), "columnName "expression "onError": "repeat": fi "re</pre>	row-based" : "Name", : "gecl:replace(value keep-original", llse, (": 10	:, '%20', ' ')",			E .			
		Perform Operations (Cancel							
			었니 33. Akhter, Syed H. ☆ 디 35. Albers, Delmar D.							

😝 Logout 🛛 🗙 🗓 Rayno	ogout × Raynor Memorial Libraries × 0% complete - Screenshots × CLCNAF-Named-Entity-Rec × + 1270.0.1:3333/project1project1918363433725 ♥ ♥ Cassification Web Main My PBwiki : Home R R0A Tookit : WeBridge A SHERPA/RoMEO - Pub CONTENTION Adminis > Music You'll Love ● Classification Web Main My PBwiki : Home R R0A Tookit :: WeBridge A SHERPA/RoMEO - Pub CONTENTION Adminis > Olic refine Screenshots Pemalint Create column Fetch at index 1 by fetching UBLe based to column Rame using expression graph statescript Contention Adminis > // Ref Undo / Redo 15 2889 r ''''''''''''''''''''''''''''''''''''								
Logout X Raynor Memorial Libraries × Alabel % Create project Left single cell on row 13, column Column 1 Left single cell on row 13, column Column 1 Left single cell on row 13, column Column 1 Left single cell on row 13, column Column 1 Left single cell on row 13, column Column 1 Left single cell on row 13, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 13, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 14, column Column 1 Left single cell on row 14, column Column 1 Left single cell in column Colum 1 Left	▼ C Q Search	ŧ	^ ☆	≙ ♥	4 9	=			
P Find Music You'll Love 😝 Classification Web M	Beynor Memorial Libraris P O Scomplete - Screenhots C C								
Google refine Screenshots Permalini	2869	Cri Nai /via	eate column Fetch at index 1 by fetching URLs based on column me using expression grel:"http://viaf.org/viaf/search 1?query-local.personalNames+%30%228_contRevs.bhd/ingscount& 22andel.ong Lources.%30%226_contRevs.bhd/ingscount&			Open Extensions: Fr	Export 🕶	Help)
Extract Apply	wreject?project=1918363433725 ▼ C Sorch ▲ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠								
Eller	All	. 6		- Screenshots × C LCNAF-Named-Entity-Rec × +					
Filter.	x Image: Raynor Memorial Libraries x 0% complete - Screenshots x C LCNAF-Named-Entity-Rec x + oject?project=1918363433725 v C Q Search + A Q Image: Q								
 Create project 									
1 Edit single cell on row 13 column Column 1									
1. Calculating contained 13, column column 1	2 7	3.	Abbaud %20Erappoin%20M		× + TENTdm Adminis ♥ CONTENTdm Adminis Open Export ▼ Hete Extensions: Freebase ▼ ROF ▼ ← first ← previous 1-50 next) last stumn 2 ♥ Column 3 ♥ Column 4 ♥ Column 5 ♥				
2. Edit single cell on row 14, column Column 1	2 7	5	Abe %20Massinshi%20A				VTENTdm Adminis Open Export • RC shors: Freebase • RC cprevious 1-50 next > Column 4 Column 5		
3. Mass edit 461 cells in column Column 1	x Raynor Memorial Libraris x 0 % complete - Screenshots x + L3333/project/project=191636X433725 C Q. Sarch + A A C B ull Low Classification Web Mai MV P8wiki : Home R RDA Toolkit WeBridge A SHERPA/RoMEO - Pub CONTENTIdm Adminis V CONTENTIdm Adminis V C Demail								
 Mass edit 4 cells in column Column 1 									
5. Mass edit 6 cells in column Column 1		9	Abrams %20Richard%20A						
a Mass adit 290 cells is column Column 1	Logout X Raynor Memorial Libraries X D's complete - Screenshots X CLAAF-Named-Entity-Rec X + 1270.0.1.3333/project:191836433725 C C Search + A <t< td=""><td></td></t<>								
		11	Aby-Farai %20Ziad%20U						
7. Blank down 371 cells in column Column 1	☆ -0 ☆ ⊑7	12	Acord %20Lea%20						
8 Blank down 0 cells in column Column 1	- ~ ~ ~	13	Adams %20Gao/%20A	• O's complete - Screenshots × • • CLAF-Named-Entity-Rec × • • • C C LAF-Named-Entity-Rec × • • • C C C LAF-Named-Entity-Rec × • • • C COUNTENTIon Adminis * • CONTENTIon Adminis * • CONTENTIon Adminis * • CONTENTIon Adminis * • CONTENTIon Adminis * • Contention C C CONTENTIon Adminis * • • Columnaters*\30.94-%221-%224-%210+%222-%alue + • • •					
-		14	Adams %20Robert%20L						
 Rename column Column 1 to Name 	5 57	15	Adibuzzaman %20Mohammad%20						
10. Rename column Name to Column 1		16	Adva %20Monica%20						
Deserve asking Only of the News	57 57	17.	Afinoquenova %20Eugenia%20						
11. Rename column Column 1 to Name	- 57 GJ	18.	Afonso.%20Meneo.%20Father				Column		
12. Reorder rows	* - * G	19.	Aftvka %20Theresa						
12 Black down 6 cells in column Name	1	20.	Agili %20Sedig%20Salem						
13. Diank down o cella in column Name		21	Aquilu%20de%20Murohy%20Raquel%20						
14. Remove 379 rows	☆ -0 ☆ ⊑7	22	Ahamed %20Sheikh%20I						
15. Text transform on 2867 cells in column Name:	+ 51	23.	Ahsanullah.%20M.%20(Mohammad)						
	52 53	24.	Akbay,%20Kunter%20S.,%201955-						
	52 61	25	Akers %20Michael%20						
	52 57	26.	Akers,%20Michael%20D.						
	53 51	27.	Akhter,%20Syed%20H.						
	12 EI	28.	Albers,%20Delmar%20D.						
	* G	29.	Albert %20Carolyne%20						ŀ
	•		W					F.	

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.

127.0.0.1:3333/project?proj	ject=1918363433725						ት ሰ ጎ	1 🖻 🛡	1 😕	Ξ
Find Music You'll Love 😝 Cla	ssification Web Mai	0	Му	PBwi	iki : Home 尺 RDA Toolkit 🗍 Wel	Bridge 🏔 SHERPA/RoMEO - Pub 🕻	CONTENTdm Adminis	CONTENT	m Adminis	
Coogle refine screen								Open	Export +	Hel
Joogle ronno ocieci	ISTICIS Permannik									
Facet / Filter Undo / Redo 31		28	69 I	OW	5			Extensions:	reebase 👻 🛛 RL	JF 4
E	dract Apply	Sho	w as	rov	vs records Show: 5 10 25	50 rows		« first < previou	s 1 - 50 next >	las
er:			All		▼ Name	LCNAF Heading	LC Record Link	Column 2	Column 3	
0. Text transform on 1647 cells in cr	olumn LC Record			1.	Ababei, Cristinel	edit				
Link: value.trim()				2.	Abbott, Allison					
Text transform on 0 cells in colum	in I.C. Record			3.	Abbott, Allison L					
Link: value.replace(/\s+/,' ')				4.	Abboud, Francois M.	Abboud, François M.	http://id.loc.gov/authorities /names/n81065705			
2. Text transform on 1647 cells in cr	olumn LC Record			5.	Abe, Masatoshi A.	Abe, Masatoshi A., 1939-	http://id.loc.gov/authorities /names/n86860726			
LINK: grei:substring(Value, 3)				6.	Abena, Albert					
Text transform on 1467 cells in ce	olumn LC Record			7.	Abing, Kevin J.					
Link: grel:replace(value, ' ', ")				8.	Abramoff, Peter, 1927-	Abramoff, Peter, 1927-	http://id.loc.gov/authorities /nemes/n85813093			
 Create column LCNAF Heading at fetching URLs based on column L 	index 3 by C Record Link			9.	Abrams, Richard A.	Abrams, Richard A., 1951-	http://id.loc.gov/authorities /names/no98106011			
using expression grel:"http://viaf.	org			10.	Abu-Faraj, Ziad U.					
/processed/L/C/" + value + "/nttp/	Accept=text/xmr			11.	Aby-Faraj, Ziad U					
Text transform on 1647 cells in cr	olumn LCNAF			12.	Acord, Lea					
Heading: grativalue parceHtml() select/my	datafield/ten~=10/			13.	Adams, Gary A.	Adams, Gary A.	http://id.loc.gov/authorities			
[0].htmlText()	datanenajtag = rot			14.	Adams, Robert L.	Adams, Robert Lang	http://id.loc.gov/authorities /names/n80080062			
Text transform on 0 cells in colum	n LCNAF			15.	Adibuzzaman, Mohammad					
Heading: value.trim()				16.	Adya, Monica					
 Text transform on 0 cells in colum Heading: value replace(As+/'') 	IN LCNAF			17.	Afinoguenova, Eugenia	Afinoguénova, Eugenia	http://id.loc.gov/authorities /names/no2003094174			
				18.	Afonso, Meneo, Father					
Text transform on 1647 cells in ce	olumn LC Record			19.	Aftyka, Theresa					
value	E			20.	Agili, Sedig Salem					
Move column LC Record Link to p	osition 3			21.	Aguilu de Murphy, Raquel	Aguilú de Murphy, Raquel	http://id.loc.gov/authorities /names/n93097553			
				22.	Ahamed, Sheikh I.					
Remove column Fetch				23.	Ahsanullah, M. (Mohammad)					
Text transform on 2867 cells in ce ore!replace(value, '%20', ' ')	olumn Name:			24.	Akbay, Kunter S., 1955-	Akbay, Kunter S., 1955-	http://id.loc.gov/authorities /names/n86800859			
				25.	Akers, Michael	Akers, Michael J.	http://id.loc.gov/authorities			

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

Find Music You'll Love (Google refine s Facet / Filter Undo / Red Refresh X Name 1 choices Sort by: name count	Classification Web Mai creenshots Permalink o 31 Reset Al Remove Al change Invert reset	My PBwil 1222 mato Show as: row All	ki : Home 🤻 RDA Toolkit 🗋 WeBridge 🏠 SHERPA/Ro ching rows (2869 total) rs records Show: 5 10 25 50 rows	oMEO - Pub 🚺 C	ONTENTdm Admini	🚺 CONTEI Op Extensions	NTdm Adminis en Export • Export •	s x. • Help RDF •			
Facet / Filter Undo / Red Refresh X Name 1 choices Sort by: name count	creenshots Permalink o 31 Reset All Remove All change invert reset	1222 mate	thing rows (2869 total) records Show: 5 10 25 50 rows			Op	en Export •	r Help			
Facet / Filter Undo / Red Refresh X Name 1 cholces Sort by: name count	o 31 Reset All Remove All change invert reset	1222 mate Show as: row	ching rows (2869 total) rs records Show: 5 10 25 50 rows			Extensions	Freebase 👻	RDF -			
Facet / Filter Undo / Red Refresh X Name 1 choices Sort by: name count	Reset All Remove All change invert reset	Show as: row	records Show: 5 10 25 50 rows			Extensions	Freebase 👻	RDF 🔻			
Refresh X Name 1 choices Sort by: name count	Reset All Remove All change invert reset	Show as: row	rs records Show: 5 10 25 50 rows		1222 matching rows (2869 total)						
X Name	change invert reset	💌 All		Show as: rows records Show: 5 10 25 50 rows @ first < previous 1 - 50 next > last >							
1 choices Sort by: name count			▼ Name	LCNAF Heading	LC Record Link	Column 2	Column 3	Colur			
		숤 듸 1.	Ababei, Cristinel								
alse 1222	exclude	☆ 듸 2.	Abbott, Allison								
aget by choice counts		☆ 坷 3.	Abbott, Allison L								
		☆ 듸 6.	Abena, Albert								
		☆ 🧐 7.	Abing, Kevin J.								
		☆ 뎍 10.	Abu-Faraj, Ziad U.								
		숬 듸 11.	Aby-Faraj, Ziad U								
		☆ 덕 12.	Acord, Lea								
		☆ 드/ 15.	Adibuzzaman, Mohammad								
		없 더 18.	Adya, Monica								
		었 딕 18.	Afonso, Meneo, Father								
LC Record Link	change invert reset	었 딕 19.	Aftyka, Theresa								
choices Sort by: name count		었 딕 20.	Agili, Sedig Salem								
choices don by. name count		었 딕 22.	Ahamed, Sheikh I.								
alse 1647	abuhya	었 년 23.	Ahsanullah, M. (Mohammad)								
aget by choice counts	chuduc	없 니 28.	Akers, Michael D.								
for by choice counts		☆ <i>L</i> ₁ 28.	Albers, Delmar D.								
		었 <i>니</i> 29.	Albert, Carolyne								
		었 <i>니</i> 31.	Albright, M. Arline								
		X 4 41.	Alumkai, A.								
		X Y 43.	American Association of Critical-Care Nurses, issuing body								
		X 4 44.	American Association of Pathologists and Bacteriologists								
		X Y 40.	American Catholic Psychological Association								
		ム J 50.	American Society for Personnel Administration								
		↔ √ 57. 57.	Anderson Fric F. Entrepreneur and new enterprise formation								
		× ~ 65.	Anderson, Rebecca Coowell, 1948-								
		A [] 70	Antony Edwin								

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

GitHub:

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