Breaking the Bottleneck

Automating the Reconciliation of Named Entities to the Library of Congress Name Authority File

Jenn Wright and Matt Carruthers

University of Michigan Library

LIBRARY

Authority Control Overview

- What it is:
  - We send Backstage lists of headings connected to new records we acquire
    - About 5,000 records on average sent to Backstage per week
  - They run comparisons against their authority files and make as many changes as they can
    - They match about 80% of the records we send them
  - They send those without matches (the leftovers) back to us
- Questions we ask:
  - Are there any authority records for this heading?
  - If so, do we have that authority record?
    - If not, import it
  - Do any of our headings need to be changed (conformed) to match the heading established as the preferred heading in the authority record?

How It Works: Backstage

Backstage Authority Control Report: **R07 - Unmatched Primary Headings**

Report Type: **Names**

Created for: ***University of Michigan***

Created on: ***January 29 2013***

The fields in this report include a primary heading which did not match an established heading or cross reference in any authority record. The *primary heading* in a field includes different subfields depending on the heading type. For example, the primary heading in a Personal Name (x00) heading includes the $a subfield plus any $c, $q, or $d subfields present. For subject (6xx) headings, only authority records in the related national authority file (and internal BSLW authority files) are checked for matching records. For example, a Topical Subject (650) heading with a second indicator zero is only matched against the LC Subject Authority File unless special processing is requested.

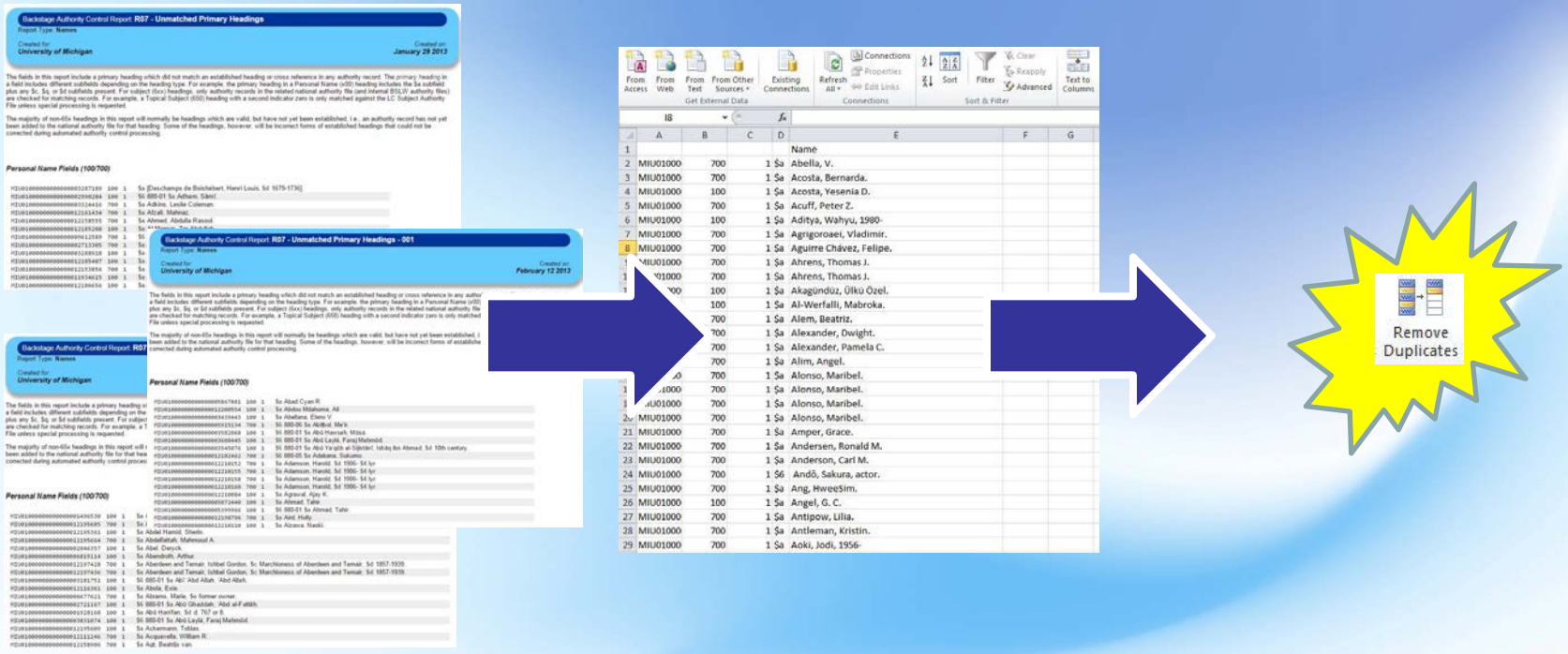
The majority of non-65x headings in this report will normally be headings which are valid, but have not yet been established, i.e., an authority record has not yet been added to the national authority file for that heading. Some of the headings, however, will be incorrect forms of established headings that could not be corrected during automated authority control processing.

Personal Name Fields (100/700)

MIU01000000000000003287189	100	1	$a [Deschamps de Boishébert, Henri Louis, $d 1679-1736].
MIU01000000000000002990284	100	1	$6 880-01 $a Adham, Sāmī.
MIU01000000000000003324416	700	1	$a Adkins, Leslie Coleman.
MIU01000000000000012161434	700	1	$a Afzali, Mahnaz.
MIU01000000000000012158555	700	1	$a Ahmed, Abdulla Rasool.
MIU01000000000000012185208	100	1	$a Al Mamun, Tm Abdullah.
MIU01000000000000009612589	700	1	$6 880-05 $a Aldakaeva, N. G.
MIU01000000000000002713305	700	1	$a Alessandri, Luigi.
MIU01000000000000003288928	100	1	$a Almada, Andreas d'.
MIU01000000000000012185407	100	1	$a Anderson, Latham.
MIU01000000000000012153856	700	1	$a Andrade, Fernando Grostein. $4 drt $4 cng $4 pro
MIU01000000000000011934615	100	1	$a Ansaloni, Enrico, $d 1977-
MIU01000000000000012186656	100	1	$a Arifin, Syarifuddin, $d 1956-

LIBRARY

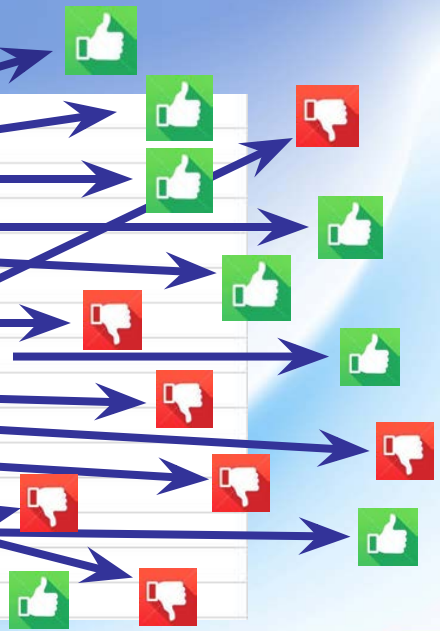
How It Was: Combine Files



How It Was: Problems

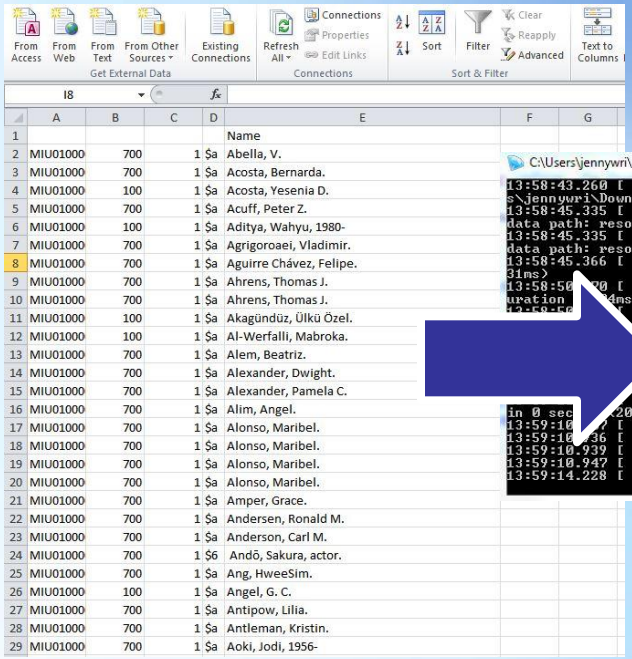
Is there an authority record?

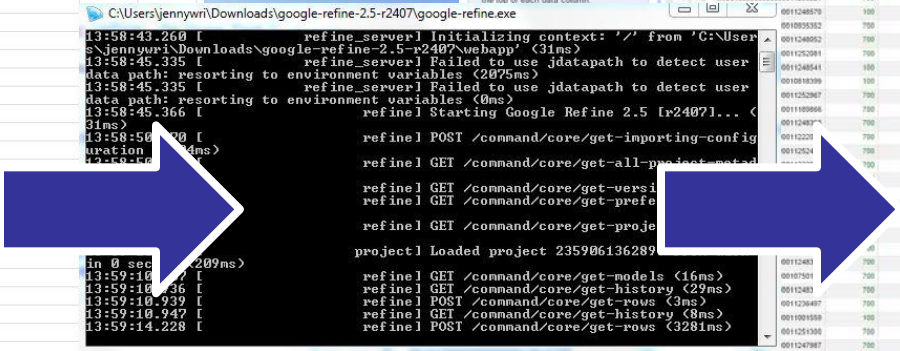
MIU01000000000000009355132	700	1	$a	Beklemishev, Lev.
MIU01000000000000009187801	100	1	$a	Benjamin, Richard.
MIU01000000000000009356462	100	1	$a	Beraud, Pierre.
MIU01000000000000009459420	100	1	$a	Bergheaud, Lise.
MIU01000000000000005237905	100	1	$a	Bergner, Thomas.
MIU01000000000000009613856	700	1	$a	Bermúdez, Daniel, 1950-
MIU01000000000000009617467	700	1	$a	Berquin, Armand.
MIU01000000000000009612951	100	1	$a	Berthier de Grandry, Frédéric de.
MIU01000000000000009476003	700	1	$a	Bettella, Mario.
MIU01000000000000009617734	700	1	$a	Bibke, John,$eformer owner.
MIU01000000000000009523988	700	1	$a	Bickeem, Ramdasha.
MIU01000000000000009460249	100	1	$a	Bierlein, Ulrike
MIU01000000000000007095104	700	1	$a	Bijelić, Severin, 1921-1972.
MIU01000000000000005642913	700	1	$a	Bišrī, Salīm al-
MIU01000000000000009620381	700	1	$a	Bizer, Christian.

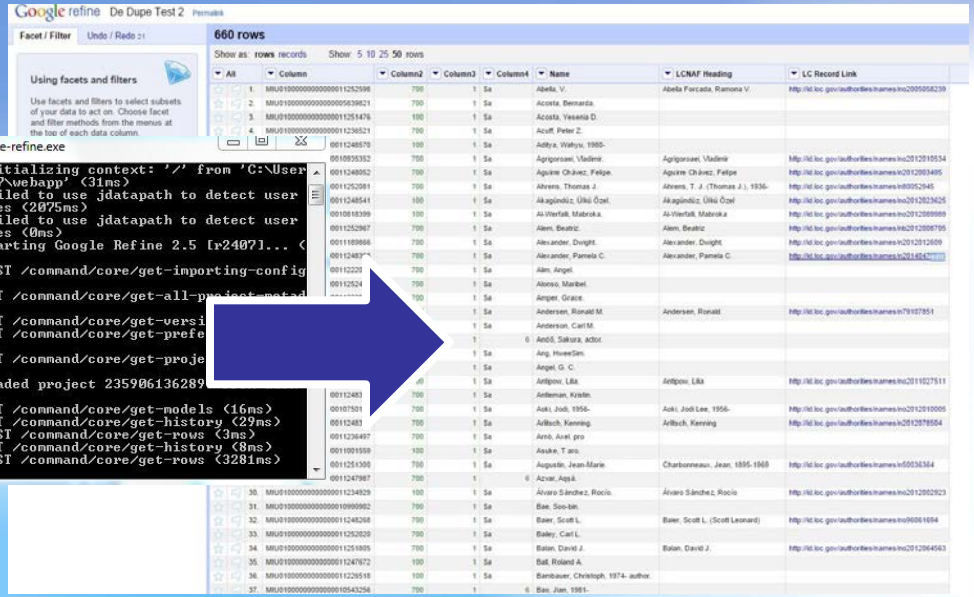


LIBRARY

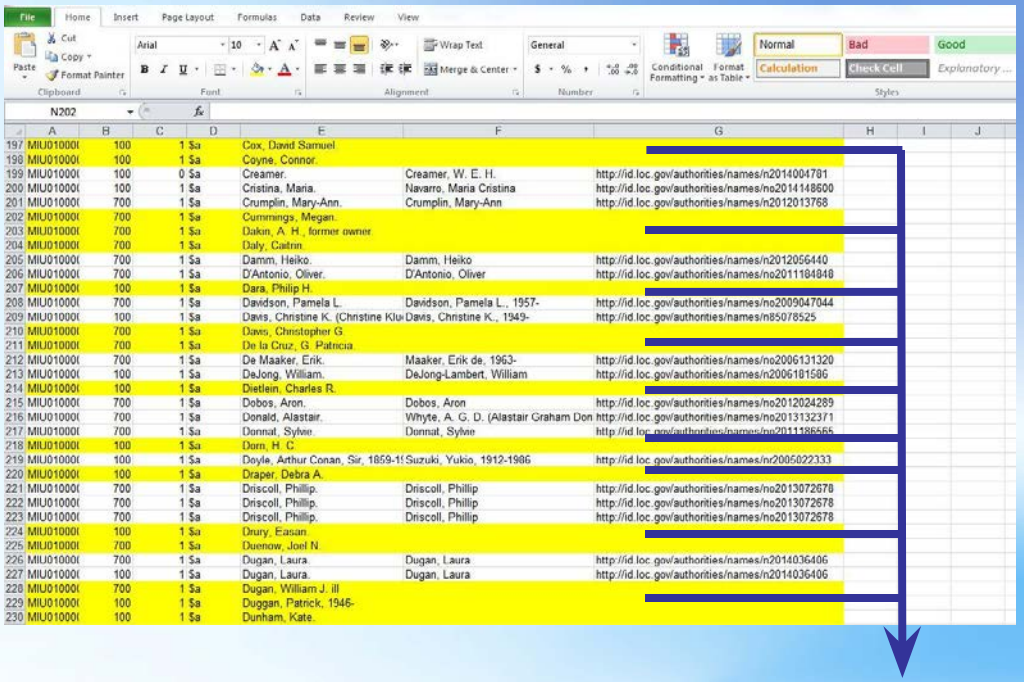
How It Works: GoogleRefine







How It Works: Efficiencies



	A	B	C	D	E	F	G
197	MIU01000(	100	1	$a	Cox, David Samuel.		
198	MIU01000(	100	1	$a	Coyne, Connor.		
199	MIU01000(	100	0	$a	Creamer.	Creamer, W. E. H.	http://id.loc.gov/authorities/names/n2014004781
200	MIU01000(	100	1	$a	Cristina, Maria.	Navarro, Maria Cristina	http://id.loc.gov/authorities/names/no2014148600
201	MIU01000(	700	1	$a	Crumplin, Mary-Ann.	Crumplin, Mary-Ann	http://id.loc.gov/authorities/names/n2012013768
202	MIU01000(	700	1	$a	Cummings, Megan.		
203	MIU01000(	700	1	$a	Dakin, A. H., former owner		
204	MIU01000(	700	1	$a	Daly, Caitrin.		
205	MIU01000(	700	1	$a	Damm, Heiko.	Damm, Heiko	http://id.loc.gov/authorities/names/n2012056440
206	MIU01000(	700	1	$a	D'Antonio, Oliver.	D'Antonio, Oliver	http://id.loc.gov/authorities/names/no2011184848
207	MIU01000(	100	1	$a	Dara, Philip H.		
208	MIU01000(	700	1	$a	Davidson, Pamela L.	Davidson, Pamela L., 1957-	http://id.loc.gov/authorities/names/no2009047044
209	MIU01000(	100	1	$a	Davis, Christine K. (Christine Klui	Davis, Christine K., 1949-	http://id.loc.gov/authorities/names/n85078525
210	MIU01000(	700	1	$a	Davis, Christopher G.		
211	MIU01000(	700	1	$a	De la Cruz, G. Patricia.		
212	MIU01000(	700	1	$a	De Maaker, Erik.	Maaker, Erik de, 1963-	http://id.loc.gov/authorities/names/no2006131320
213	MIU01000(	100	1	$a	DeJong, William.	DeJong-Lambert, William	http://id.loc.gov/authorities/names/n2006101586
214	MIU01000(	100	1	$a	Dietlein, Charles R.		
215	MIU01000(	700	1	$a	Dobos, Aron.	Dobos, Aron	http://id.loc.gov/authorities/names/no2012024289
216	MIU01000(	700	1	$a	Donald, Alastair.	Whyte, A. G. D. (Alastair Graham Don	http://id.loc.gov/authorities/names/no2013132371
217	MIU01000(	700	1	$a	Donnat, Sylvie.	Donnat, Sylvie	http://id.loc.gov/authorities/names/no2011186565
218	MIU01000(	100	1	$a	Dorn, H. C.		
219	MIU01000(	100	1	$a	Doyle, Arthur Conan, Sir, 1859-19	Suzuki, Yukio, 1912-1986	http://id.loc.gov/authorities/names/nr2005022333
220	MIU01000(	100	1	$a	Draper, Debra A.		
221	MIU01000(	700	1	$a	Driscoll, Phillip.	Driscoll, Phillip	http://id.loc.gov/authorities/names/no2013072678
222	MIU01000(	700	1	$a	Driscoll, Phillip.	Driscoll, Phillip	http://id.loc.gov/authorities/names/no2013072678
223	MIU01000(	700	1	$a	Driscoll, Phillip.	Driscoll, Phillip	http://id.loc.gov/authorities/names/no2013072678
224	MIU01000(	100	1	$a	Drury, Easan.		
225	MIU01000(	700	1	$a	Duenow, Joel N.		
226	MIU01000(	700	1	$a	Dugan, Laura.	Dugan, Laura	http://id.loc.gov/authorities/names/n2014036406
227	MIU01000(	100	1	$a	Dugan, Laura.	Dugan, Laura	http://id.loc.gov/authorities/names/n2014036406
228	MIU01000(	700	1	$a	Dugan, William J. III		
229	MIU01000(	100	1	$a	Duggan, Patrick, 1946-		
230	MIU01000(	100	1	$a	Dunham, Kate.		

No authorities?

No problem!

Just move them on down to the end of the spreadsheet.

They'll contribute to a pool of candidates waiting for NACO authority record creation.

LIBRARY

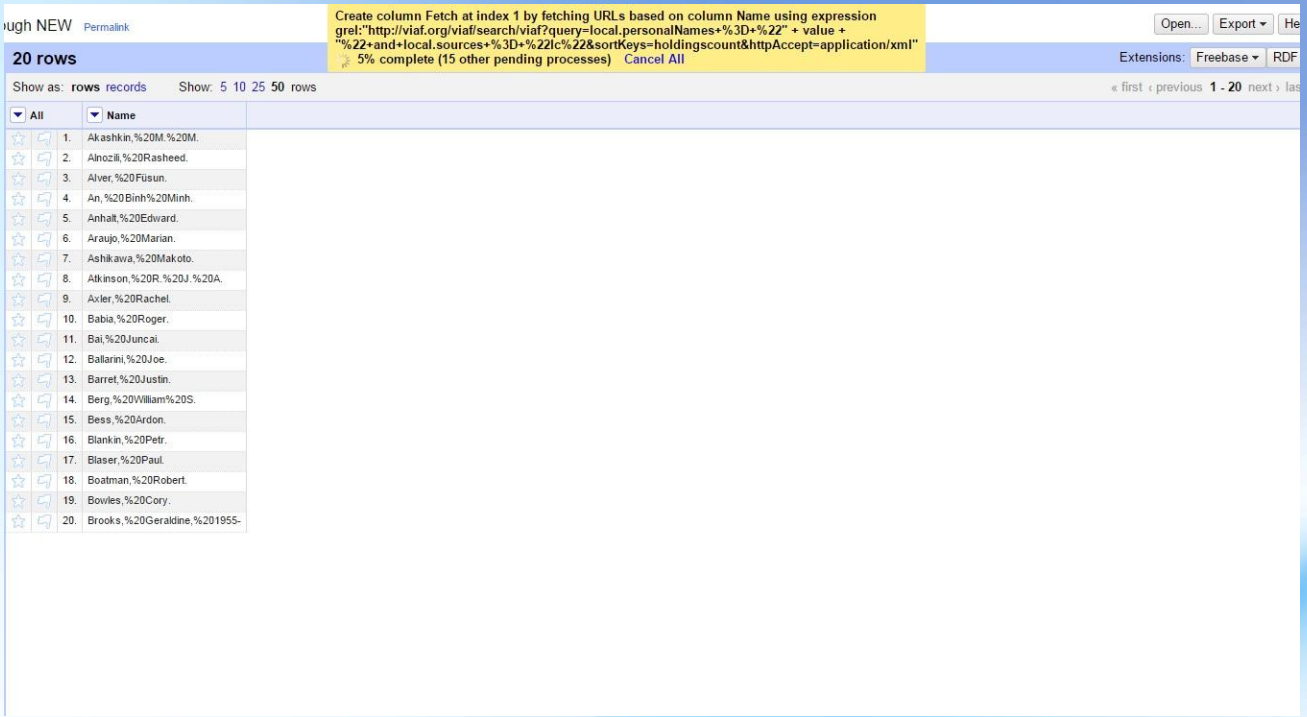
Technical Details

Behind the Scenes:

- Google Refine automatically searches the Virtual International Authority File (VIAF) for matching headings from the Library of Congress Name Authority File

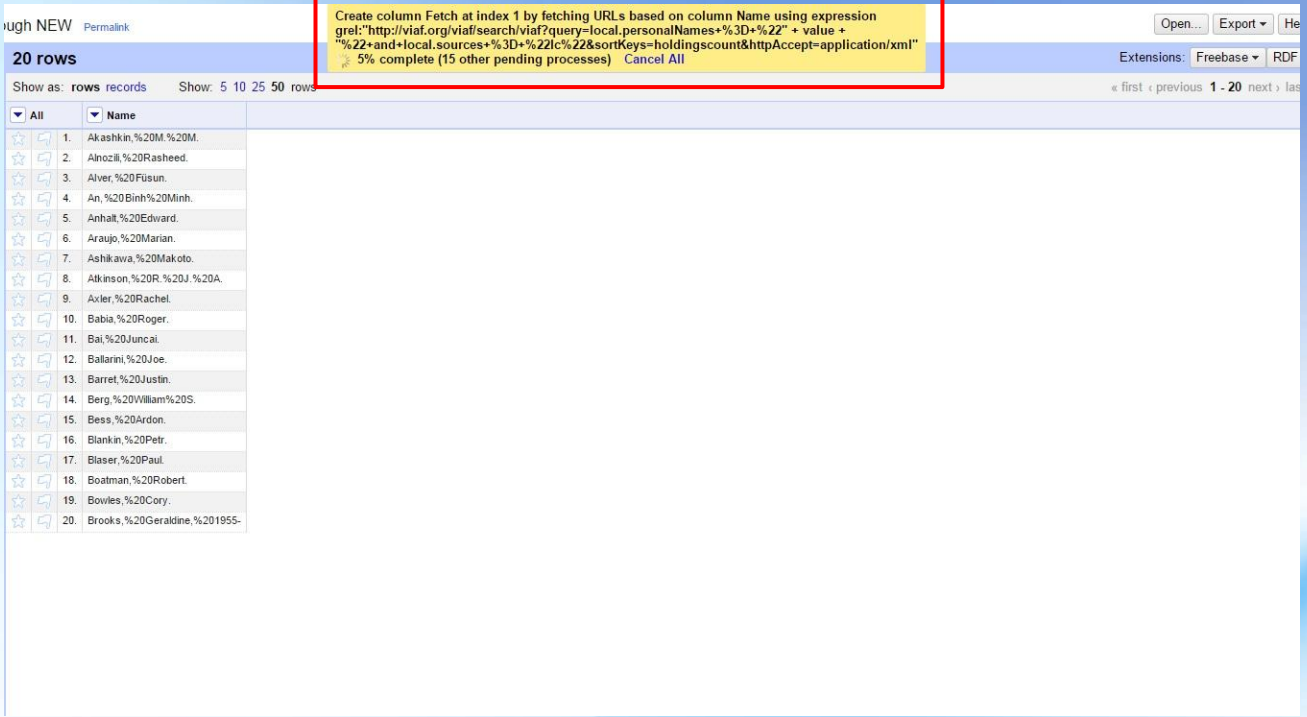
- It then returns any matching headings and links to corresponding records

Technical Details



LIBRARY

Technical Details

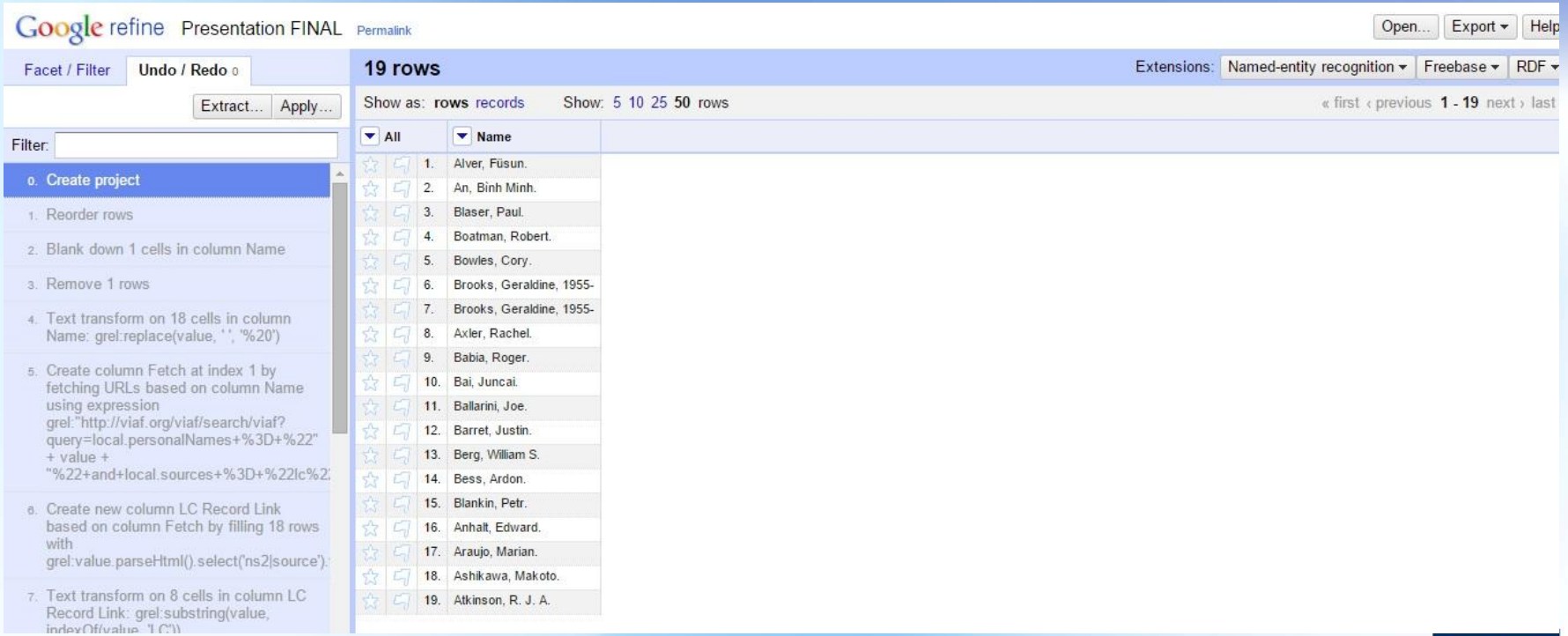


LIBRARY

Technical Details

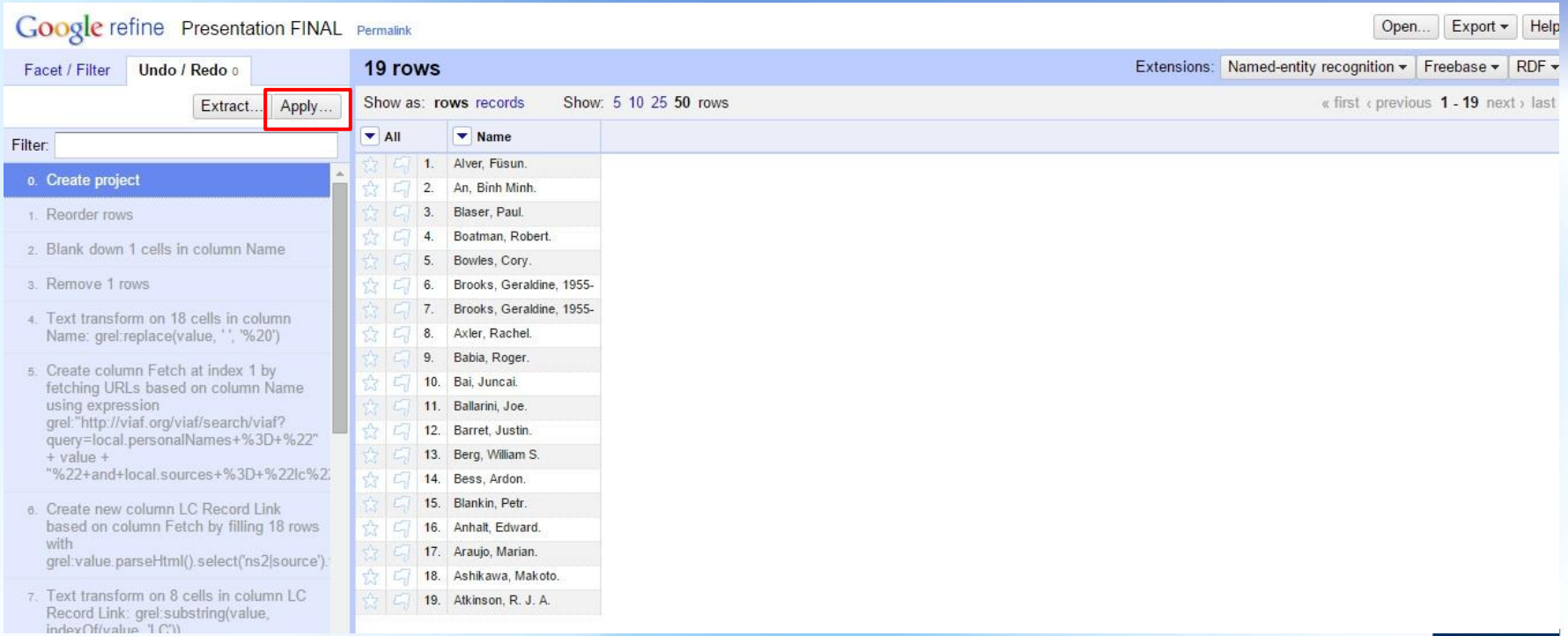
LIBRARY

Technical Details



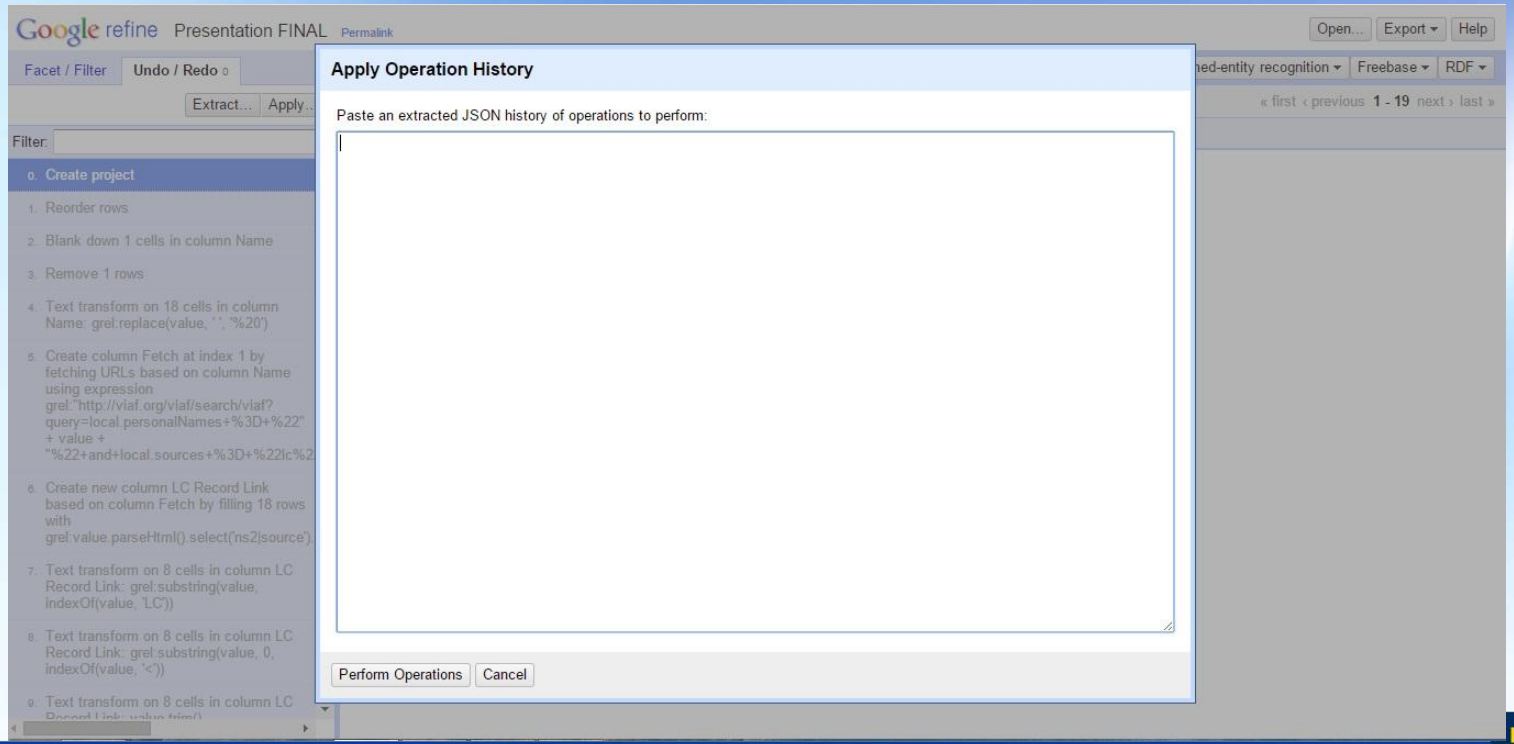
LIBRARY

Technical Details

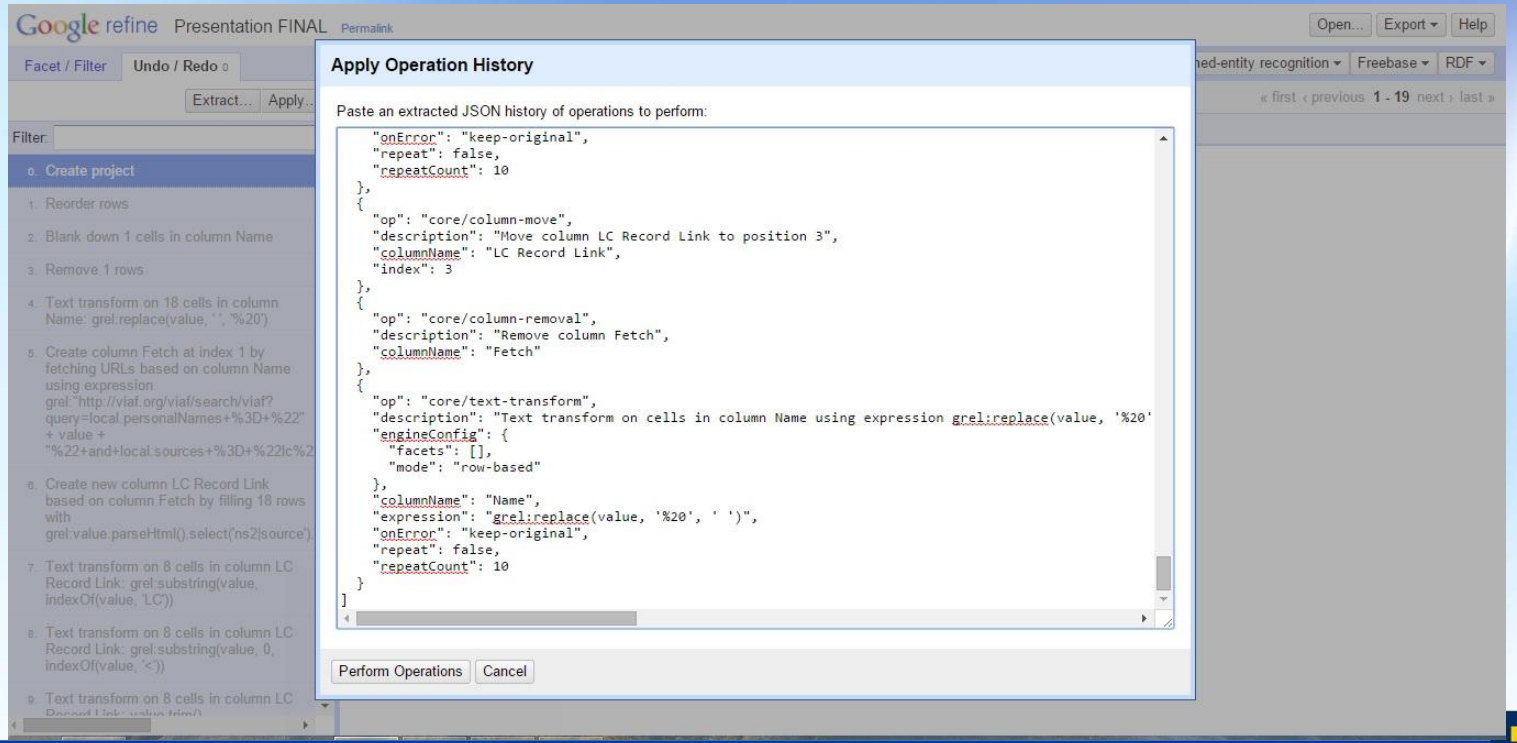


LIBRARY

Technical Details

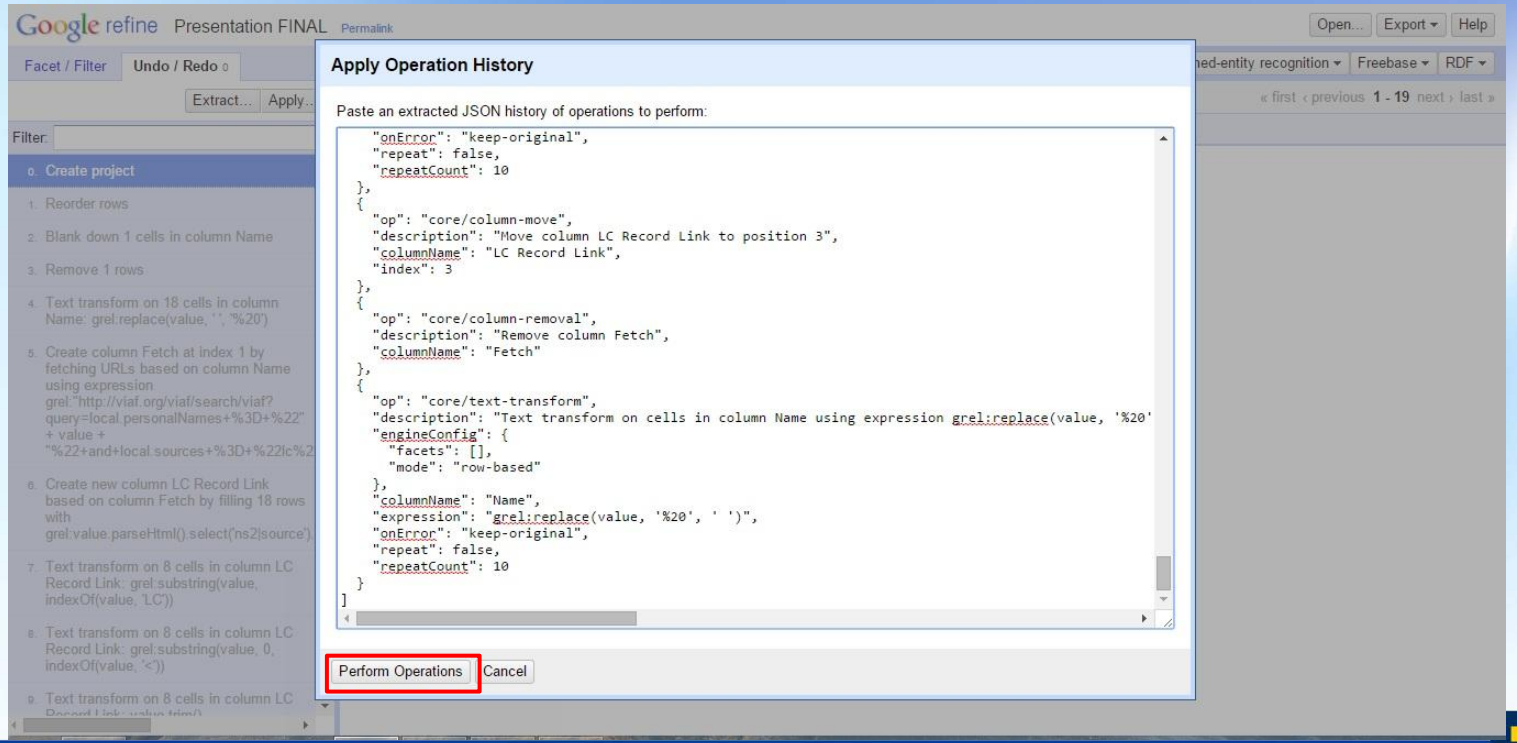


Technical Details



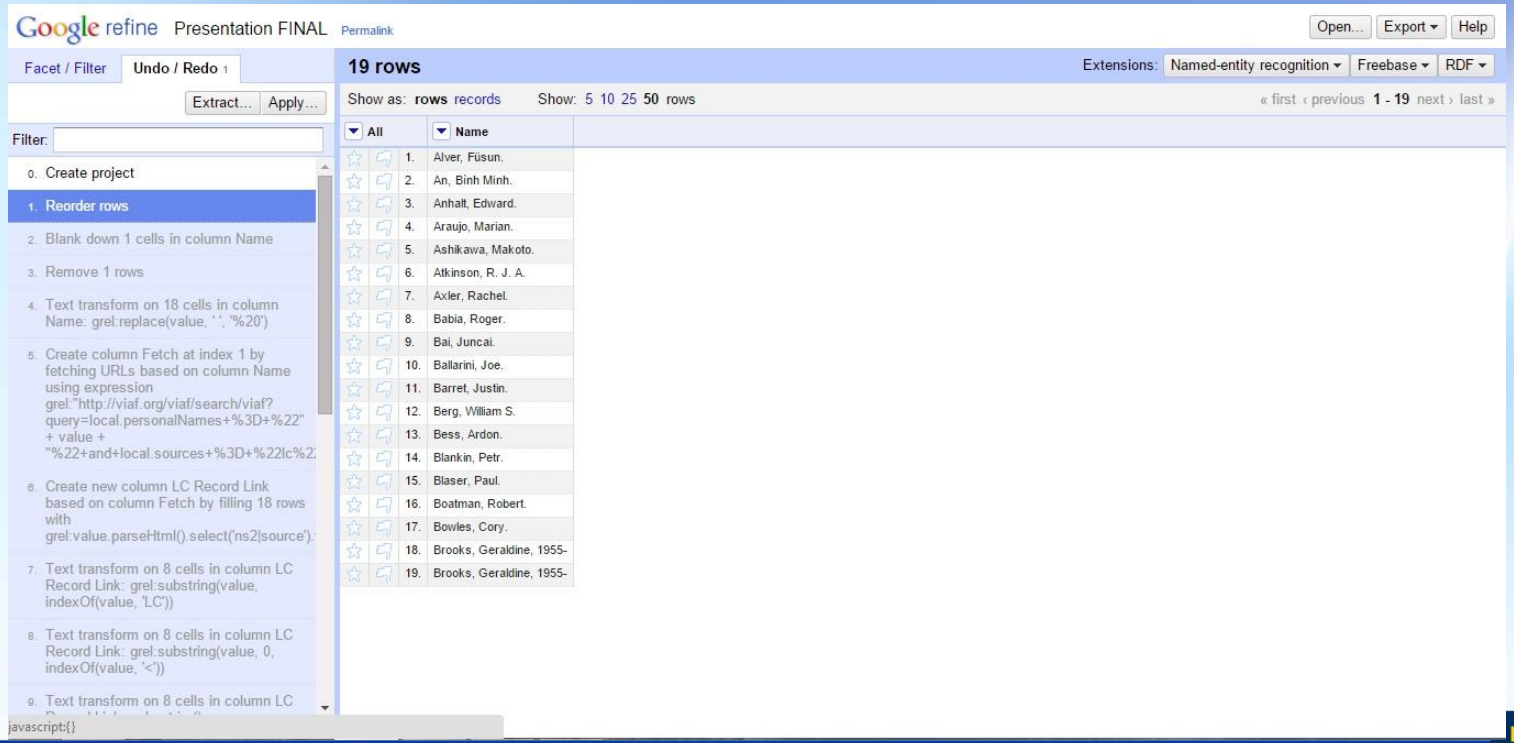
LIBRARY

Technical Details



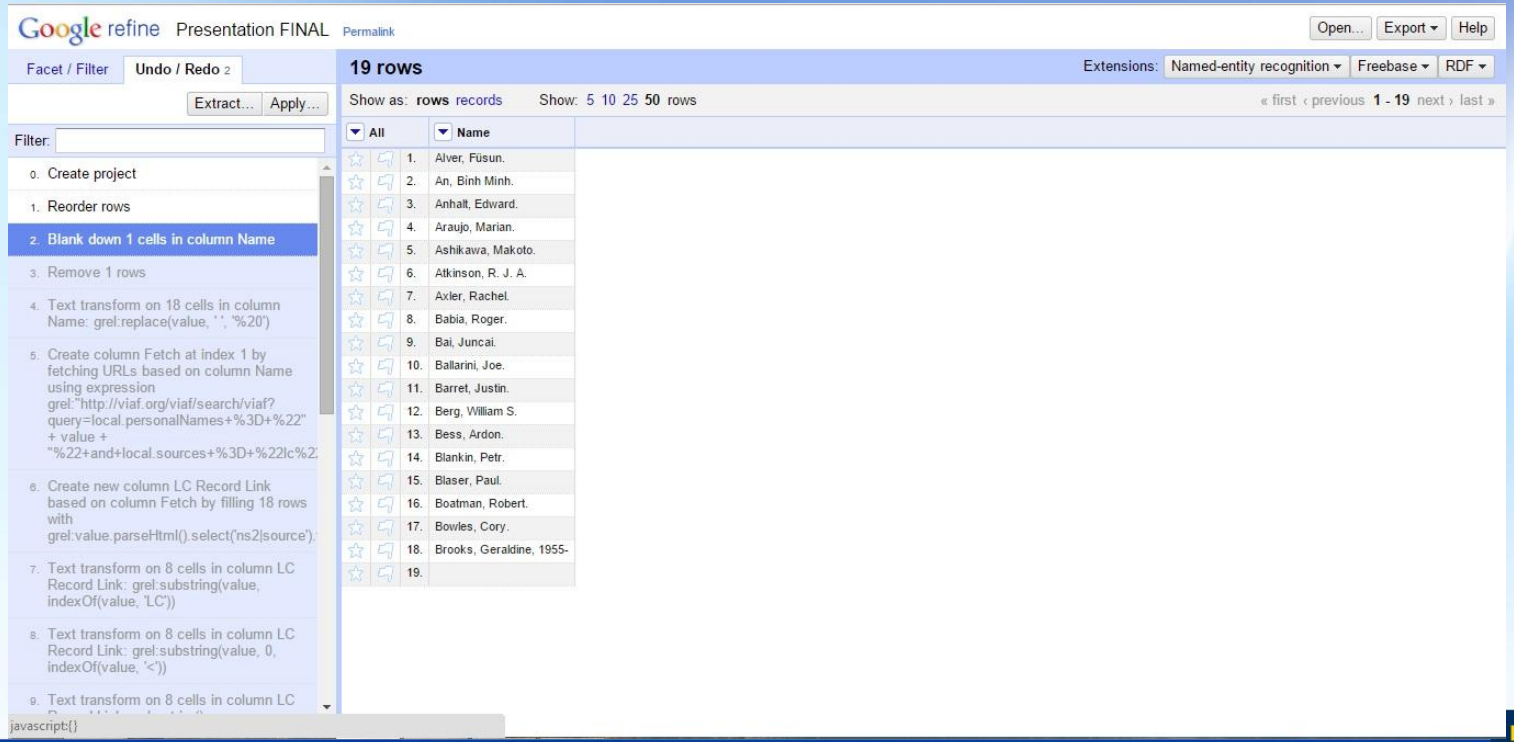
LIBRARY

Technical Details



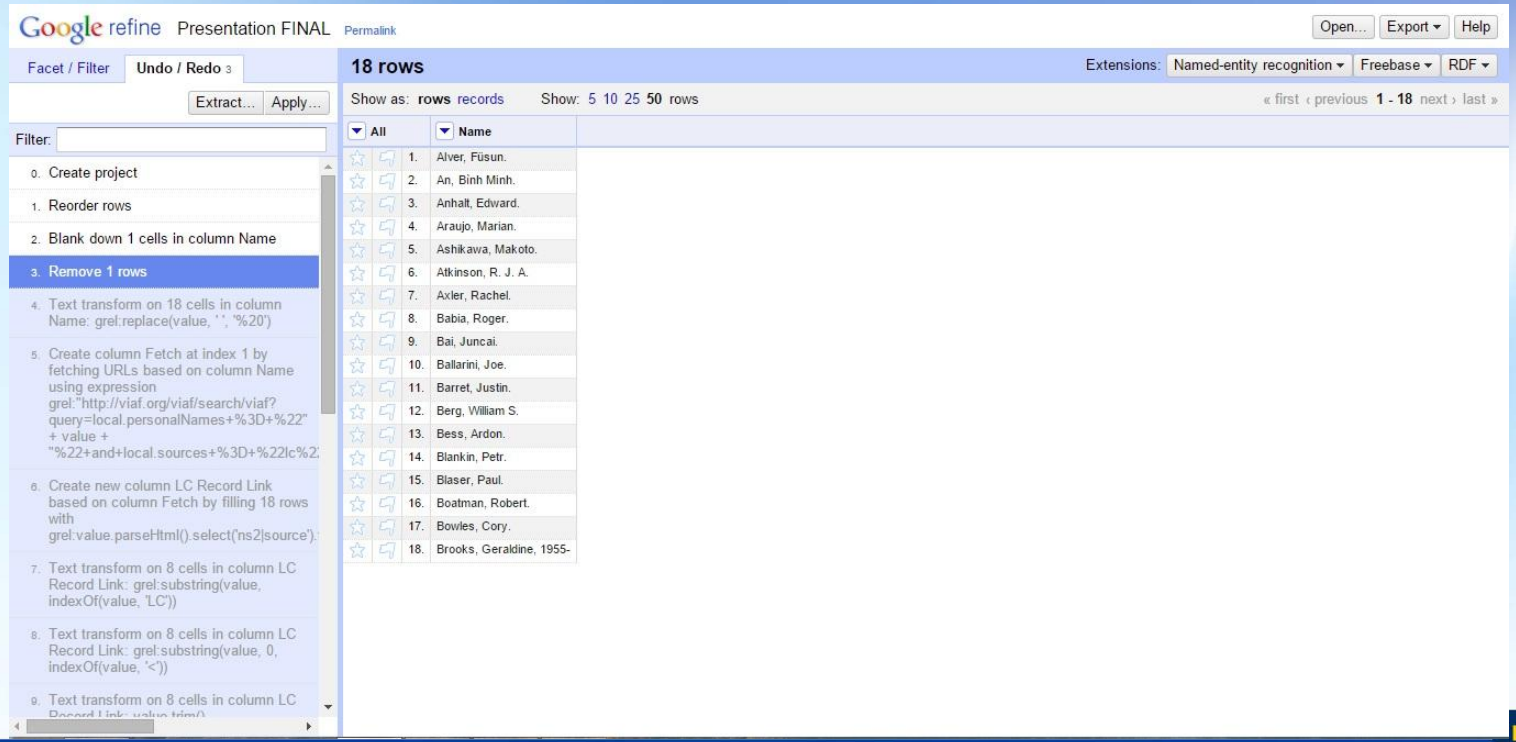
LIBRARY

Technical Details



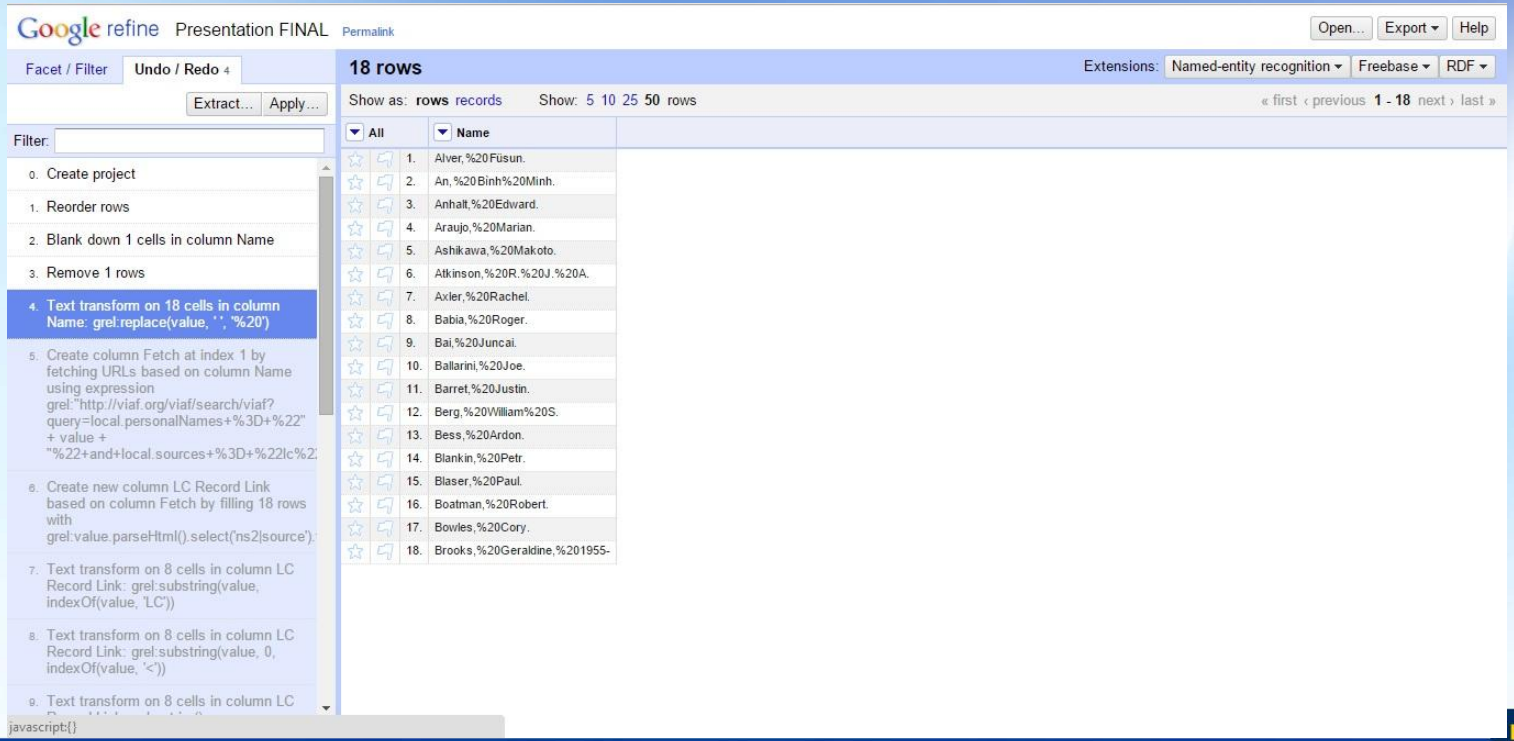
LIBRARY

Technical Details



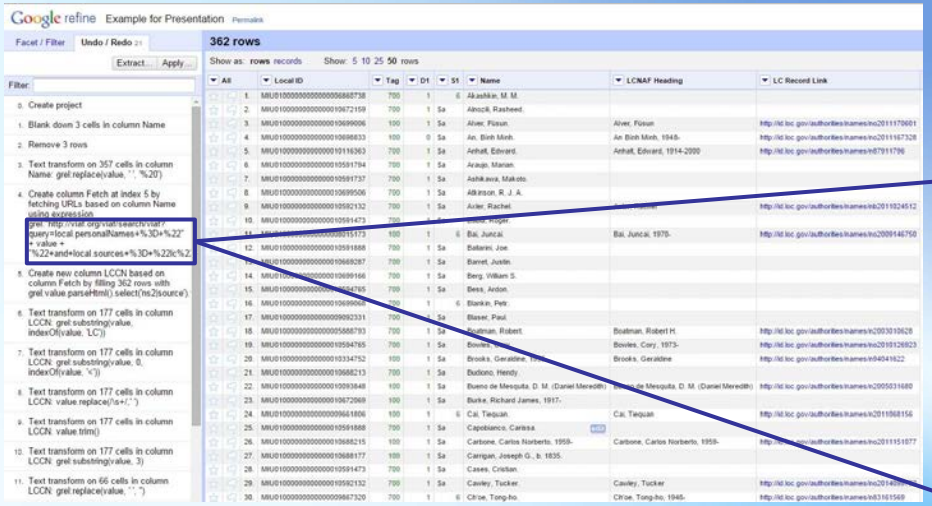
LIBRARY

Technical Details



LIBRARY

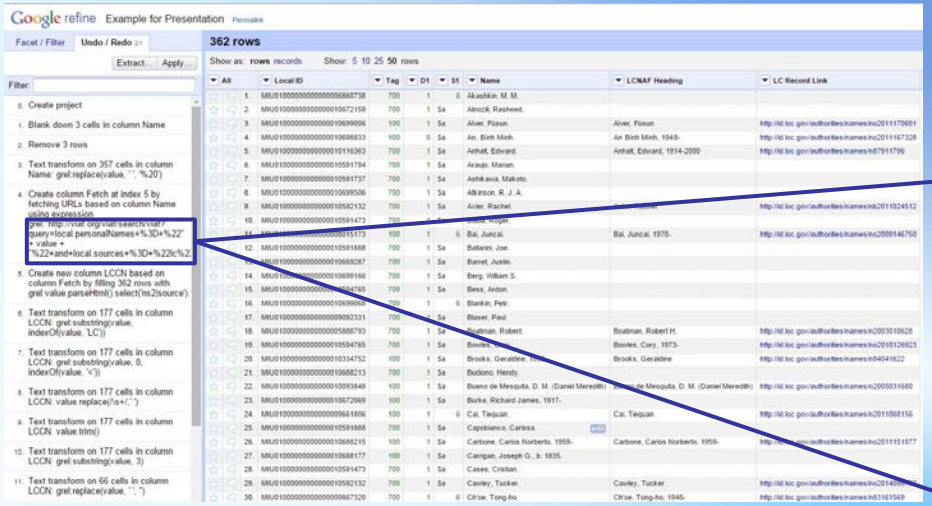
Technical Details



First Query for VIAF API:

"http://viaf.org/viaf/search/viaf?
query=local.personalNames+%
3D+%22" + value + "%
22+and+local.sources+%3D+%
22lc%
22&httpAccept=application/xml"

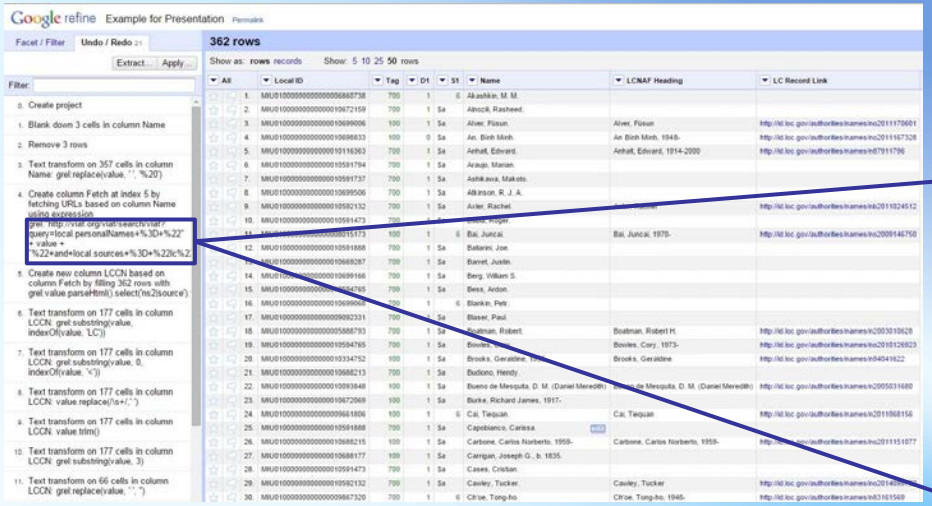
Technical Details



First Query for VIAF API:

"http://viaf.org/viaf/search/viaf?
query=local.personalNames+%
3D+%22" + value + "%
22+and+local.sources+%3D+%
22lc%
22&httpAccept=application/xml"

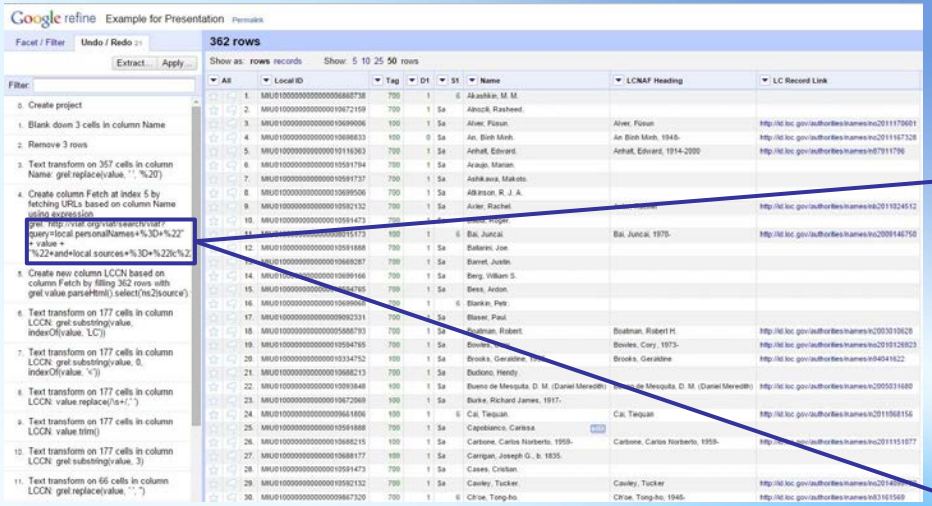
Technical Details



First Query for VIAF API:

"http://viaf.org/viaf/search/viaf?
query=local.personalNames+%
3D+%22" + value + "%
22+and+local.sources+%3D+%
22lc%
22&httpAccept=application/xml"

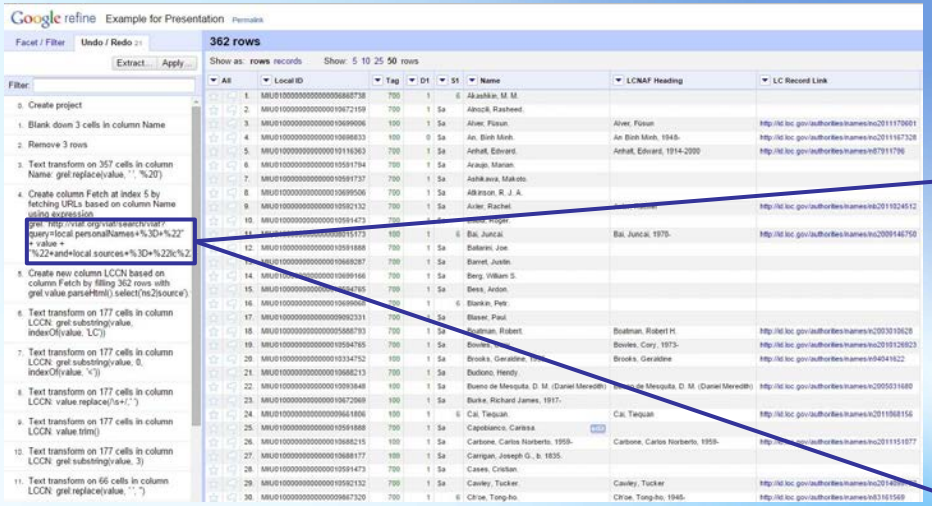
Technical Details



First Query for VIAF API:

"http://viaf.org/viaf/search/viaf?
query=local.personalNames+%
3D+%22" + value + "%
22+and+local.sources+%3D+%
22lc%
22&httpAccept=application/xml"

Technical Details

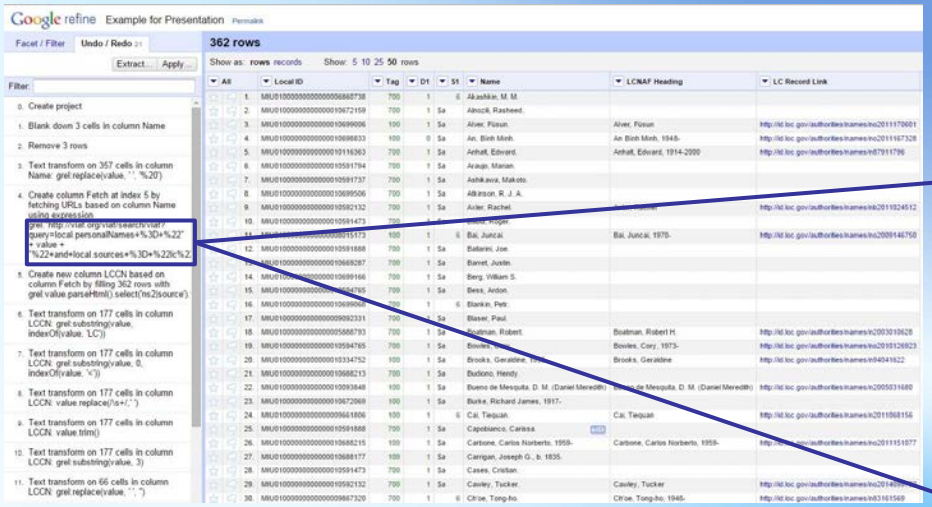


First Query for VIAF API:

"http://viaf.org/viaf/search/viaf?
query=local.personalNames+%
3D+%22" + value + "%
22+and+local.sources+%3D+%
22lc%
22&httpAccept=application/xml"

LIBRARY

Technical Details

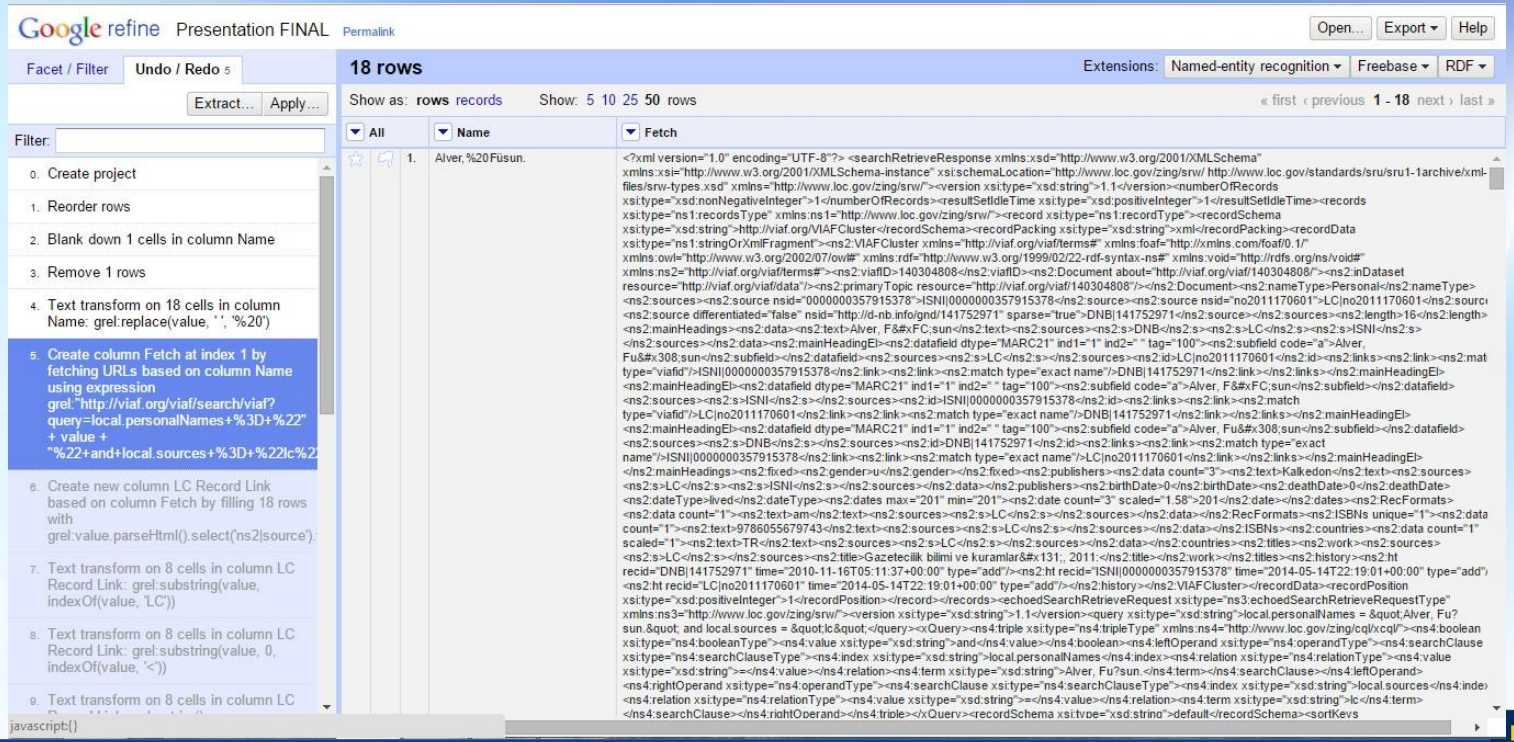


First Query for VIAF API:

"http://viaf.org/viaf/search/viaf?
query=local.personalNames+%
3D+%22" + value + "%
22+and+local.sources+%3D+%
22lc%
22&httpAccept=application/xml"

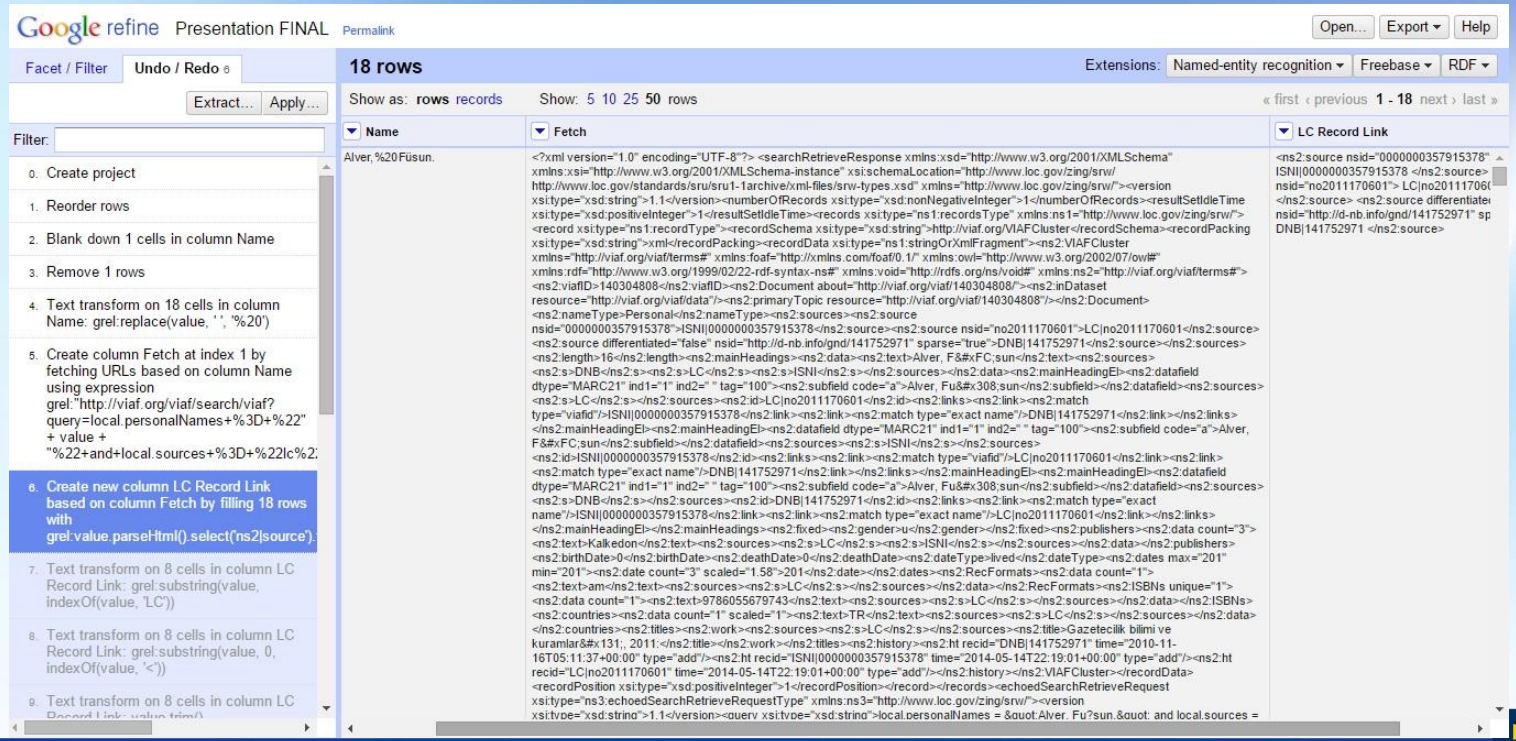
LIBRARY

Technical Details

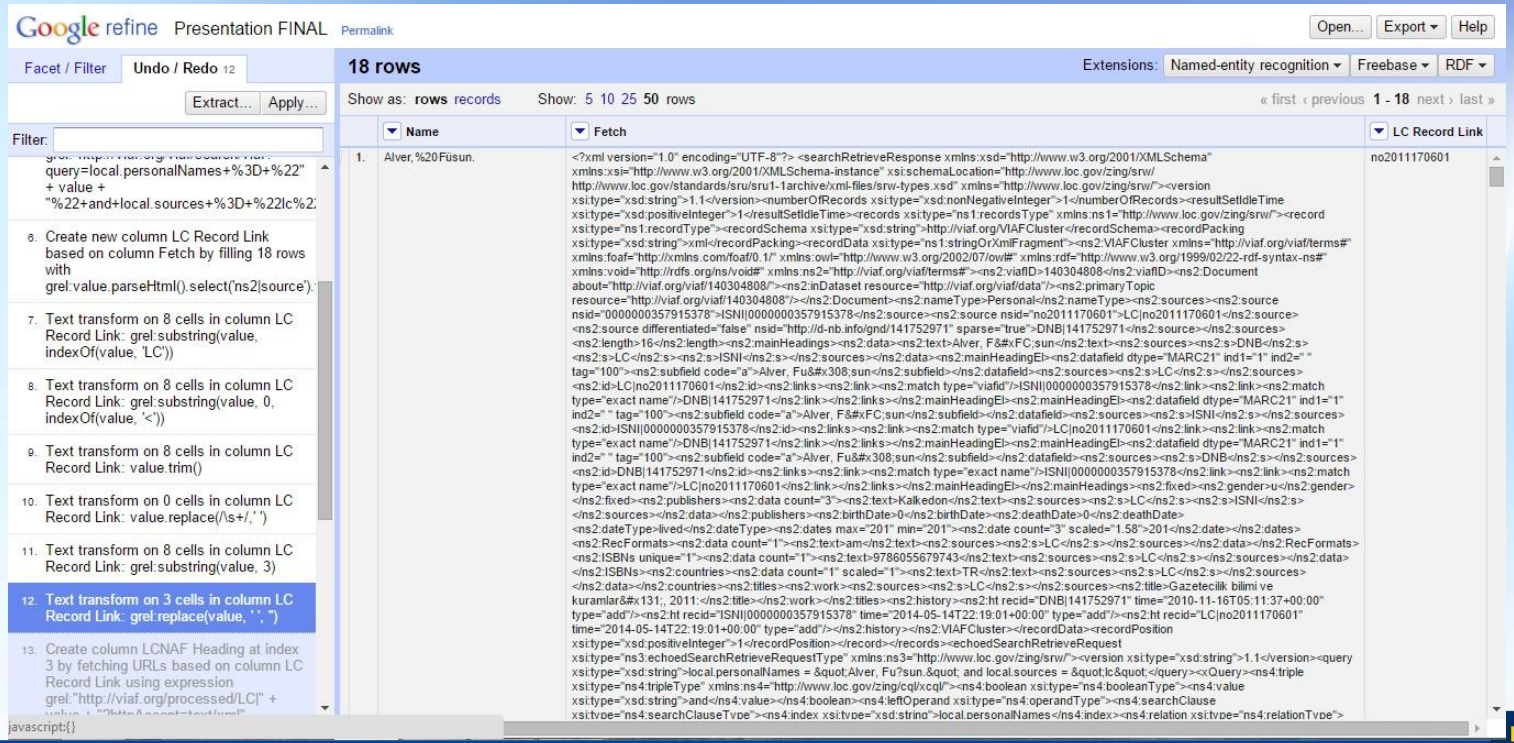


LIBRARY

Technical Details

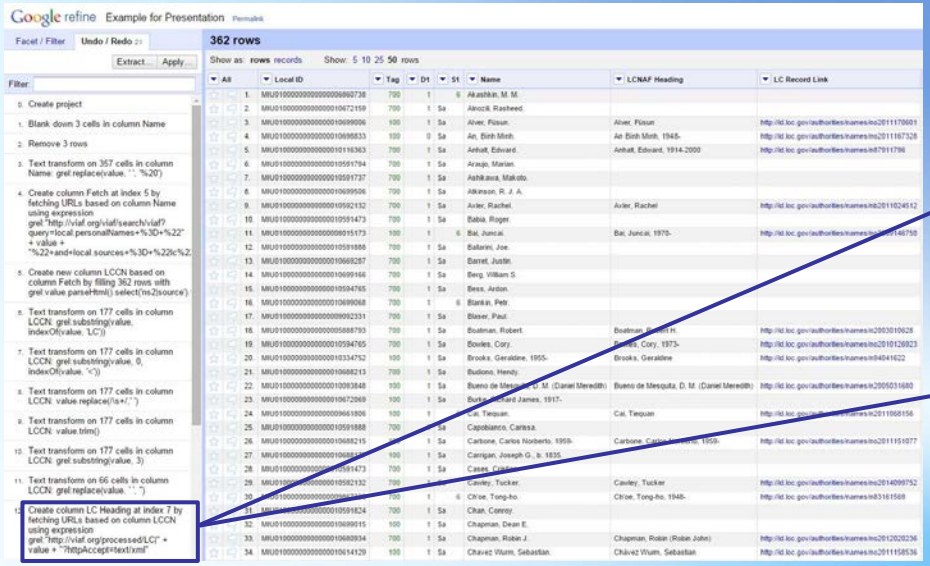


Technical Details



LIBRARY

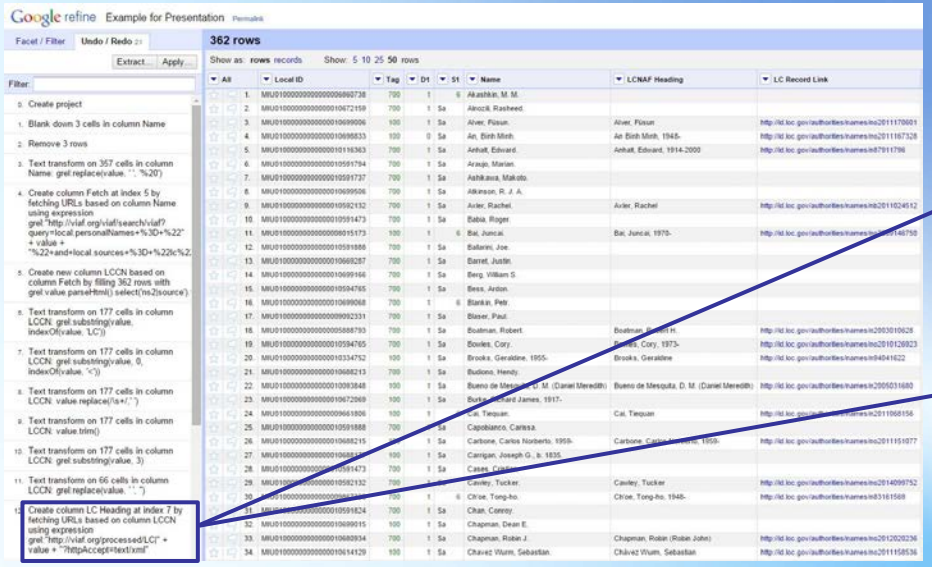
Technical Details



Second Query for VIAF API:

"http://viaf.org/processed/LC|" + value + "?
httpAccept=text/xml"

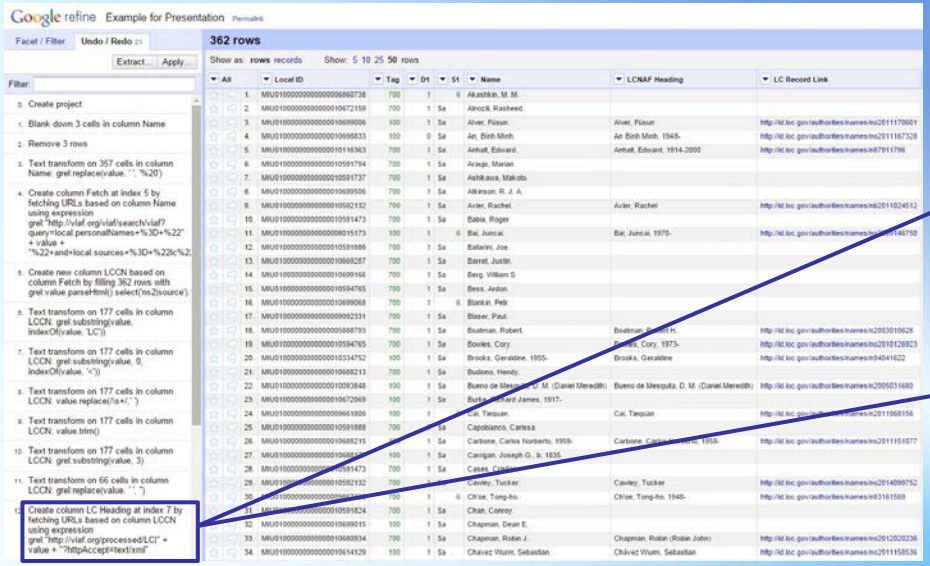
Technical Details



Second Query for VIAF API:

"http://viaf.org/processed/LC|" + value + "?
httpAccept=text/xml"

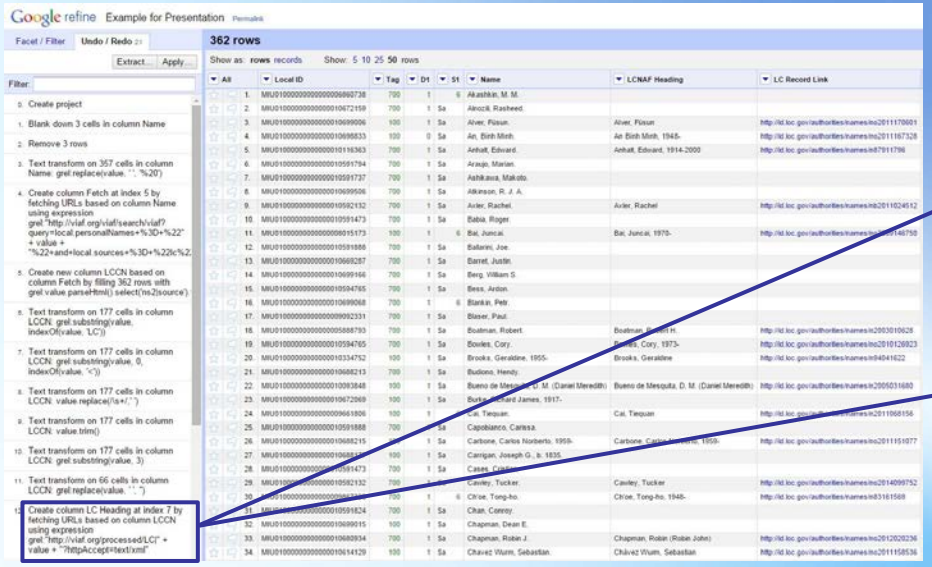
Technical Details



Second Query for VIAF API:

"http://viaf.org/processed/LC|" + value + "?
httpAccept=text/xml"

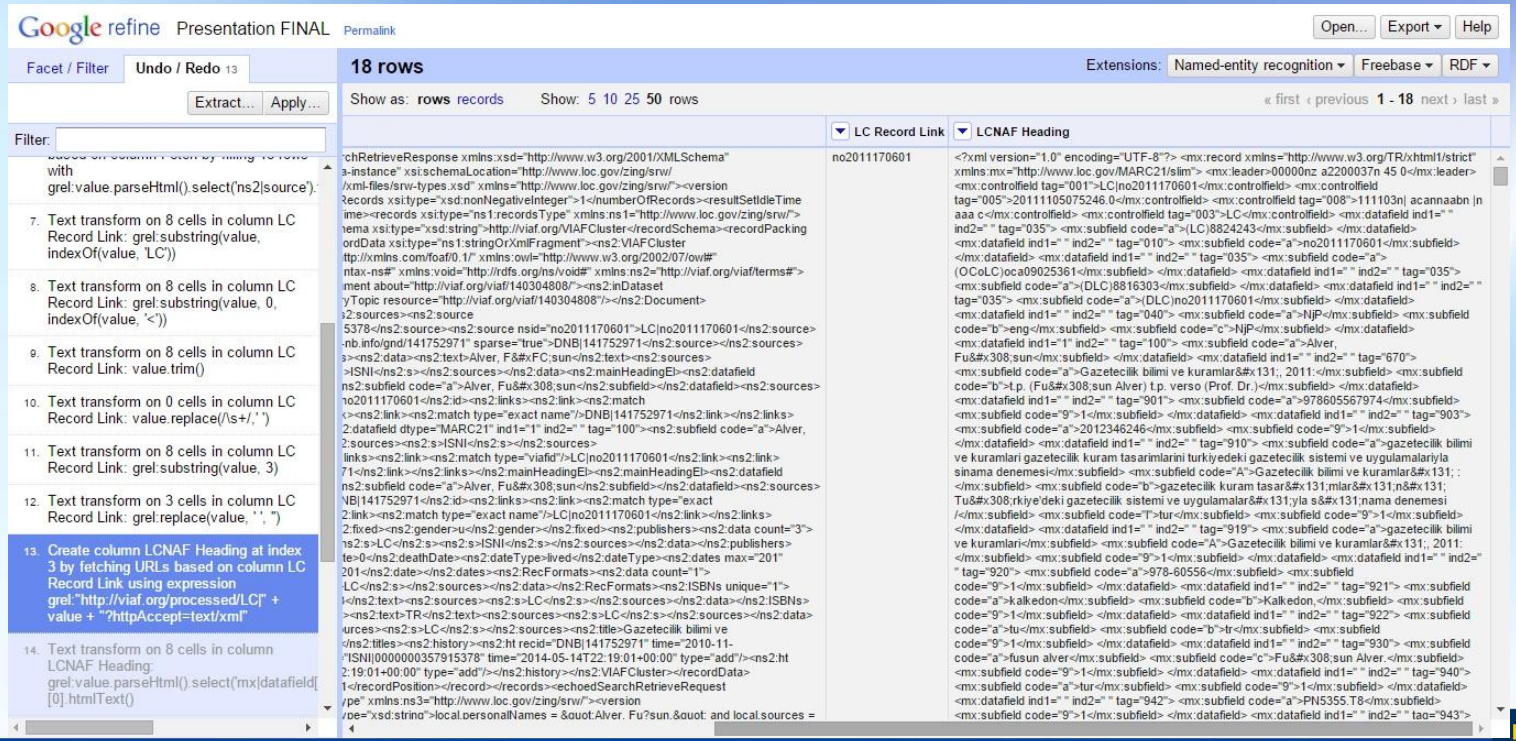
Technical Details



Second Query for VIAF API:

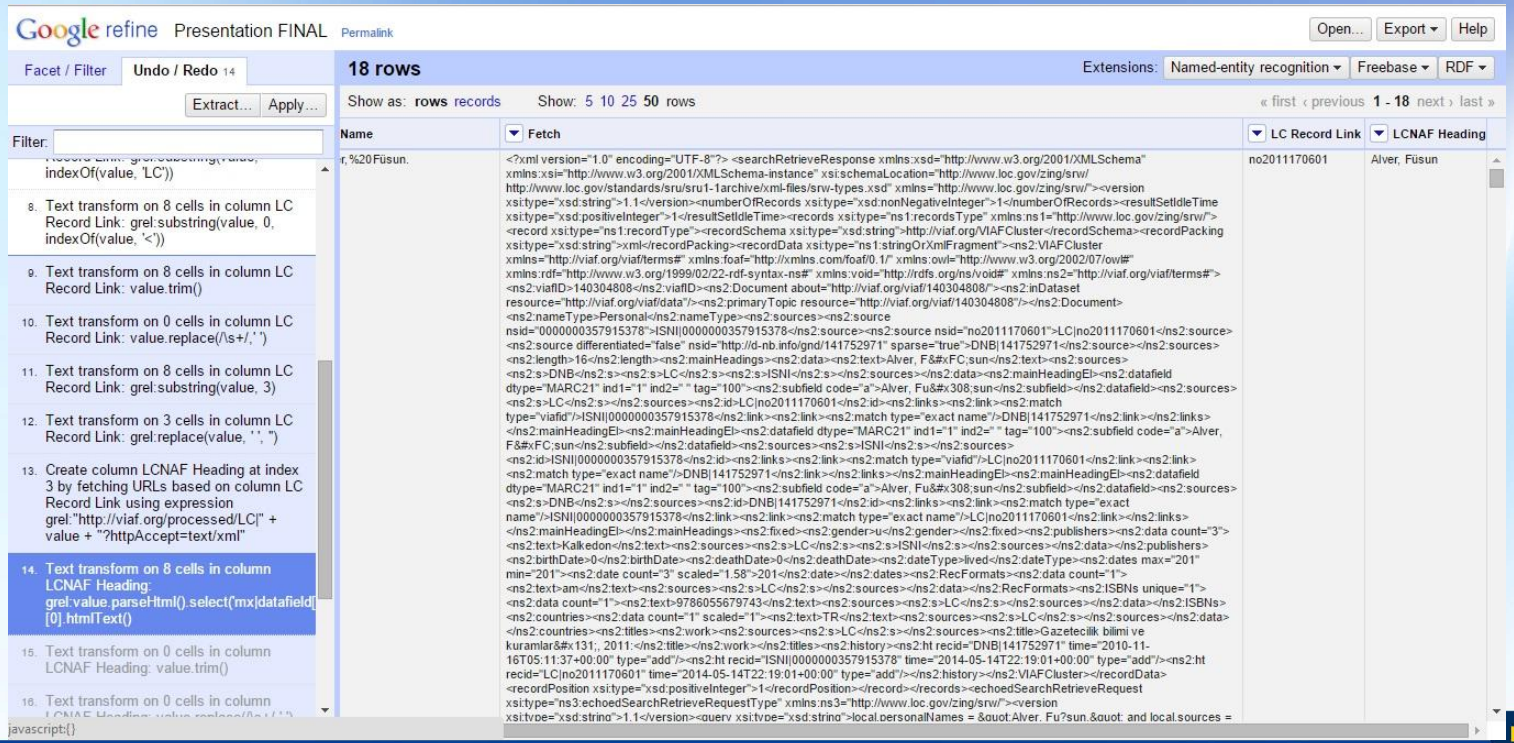
"http://viaf.org/processed/LC|" + value + "?
httpAccept=text/xml"

Technical Details

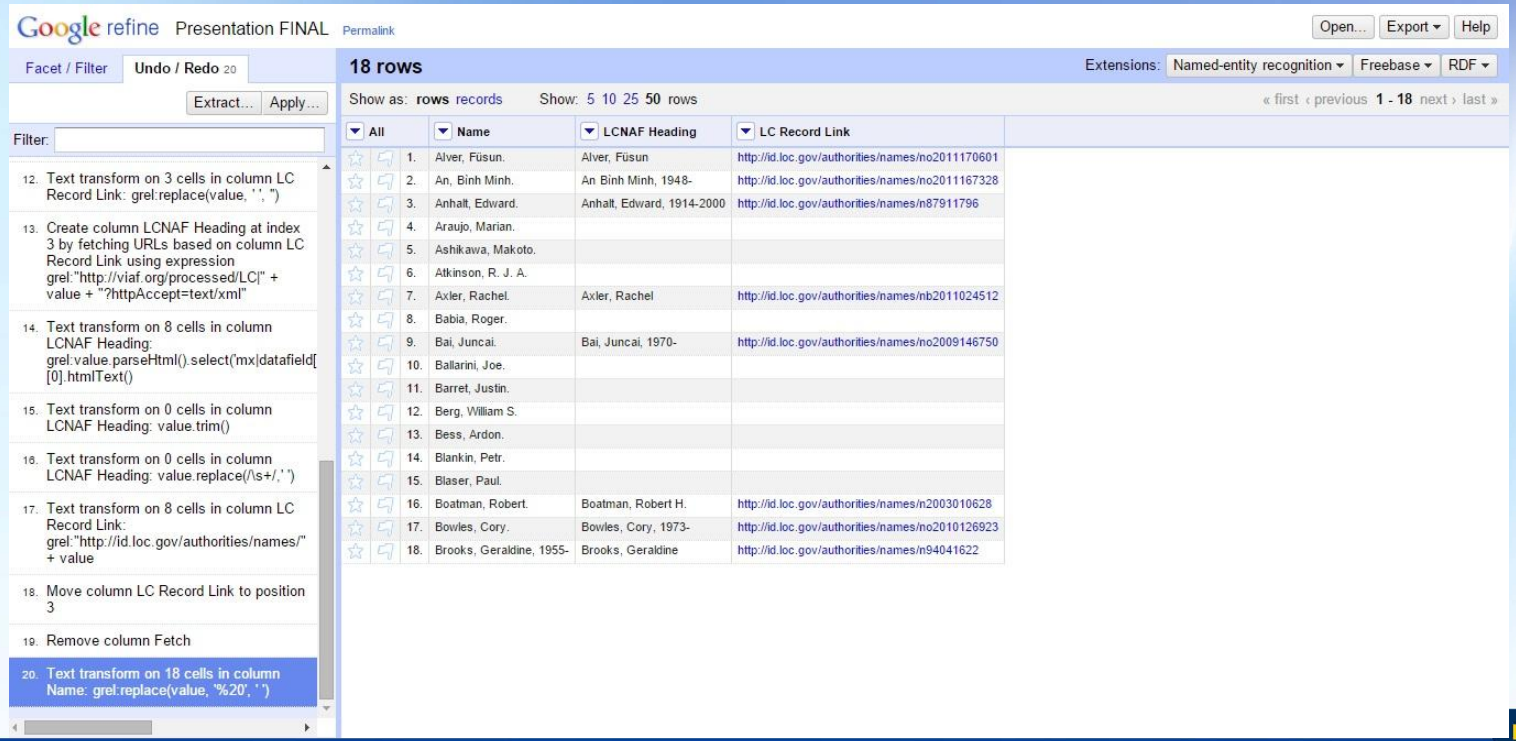


LIBRARY

Technical Details



Technical Details



LIBRARY

Why Use VIAF?

1. More robust dataset to search than other options
2. VIAF algorithm is better at fuzzy name searches
3. Updated weekly

Limitations

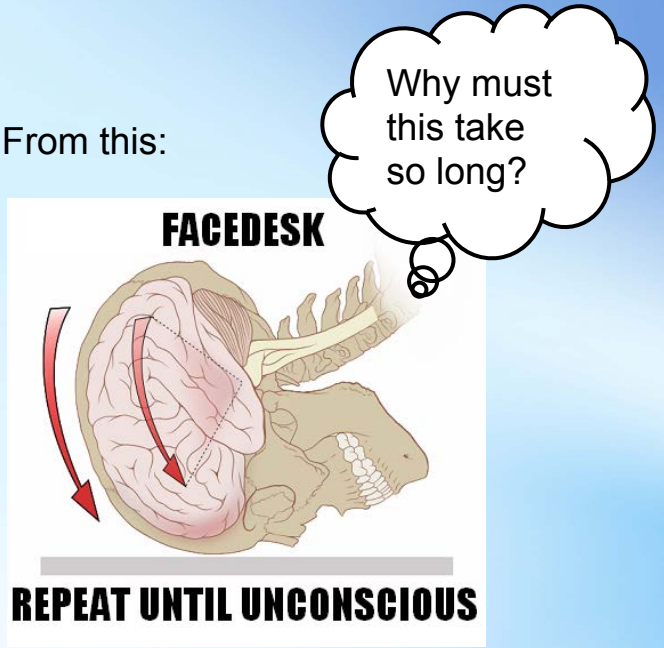
1. There is a time lag if you are searching hundreds or thousands of names at a time.

2. IMPORTANT: It is up to the user to verify the accuracy of the matches found.

3. The process does not currently reconcile geographic name headings.

Impact

From this:



To this:

Impact Details

- With this process, authority control now takes half the time it used to
  - Less time spent looking in vain for non-existent authority files, because the script tells us whether there is a file or not
- More authority control gets done
- More time for other tasks
- Less time wasted = good for morale!
- We contribute to a pool of headings that we can draw from for authority record creation workflows

Conclusion

- This is the sort of thing we always hoped to be possible
  - Has been tremendously helpful to workflows

- We hope it helps others as much as it has us!

Get the files and instructions!

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

Matt Carruthers
mcarruth@umich.edu

Jenn Wright
jennywri@umich.edu

LIBRARY