QUEEN'S UNIVERSITY LIBRARY

Queen's UNIVERSITY

OpenRefine

'a free, open source, powerful tool for working with messy data'

ACCOLEDS, University of Calgary, December, 2015

Jeff Moon
Data Librarian &
Academic Director, Queen's Research Data Centre
Queen's University Library

http://upload.wikimedia.org/wikipedia/commons/thumb/7/76/Google-refine-logo.svg/2000px-Google-refine-logo.svg.png

What is OpenRefine?

- Formerly known as “Google Refine”
- Free & open-source
- Tool for cleaning large datasets
- Can create links between metadata sets
- Desktop application – run locally

http://openrefine.org

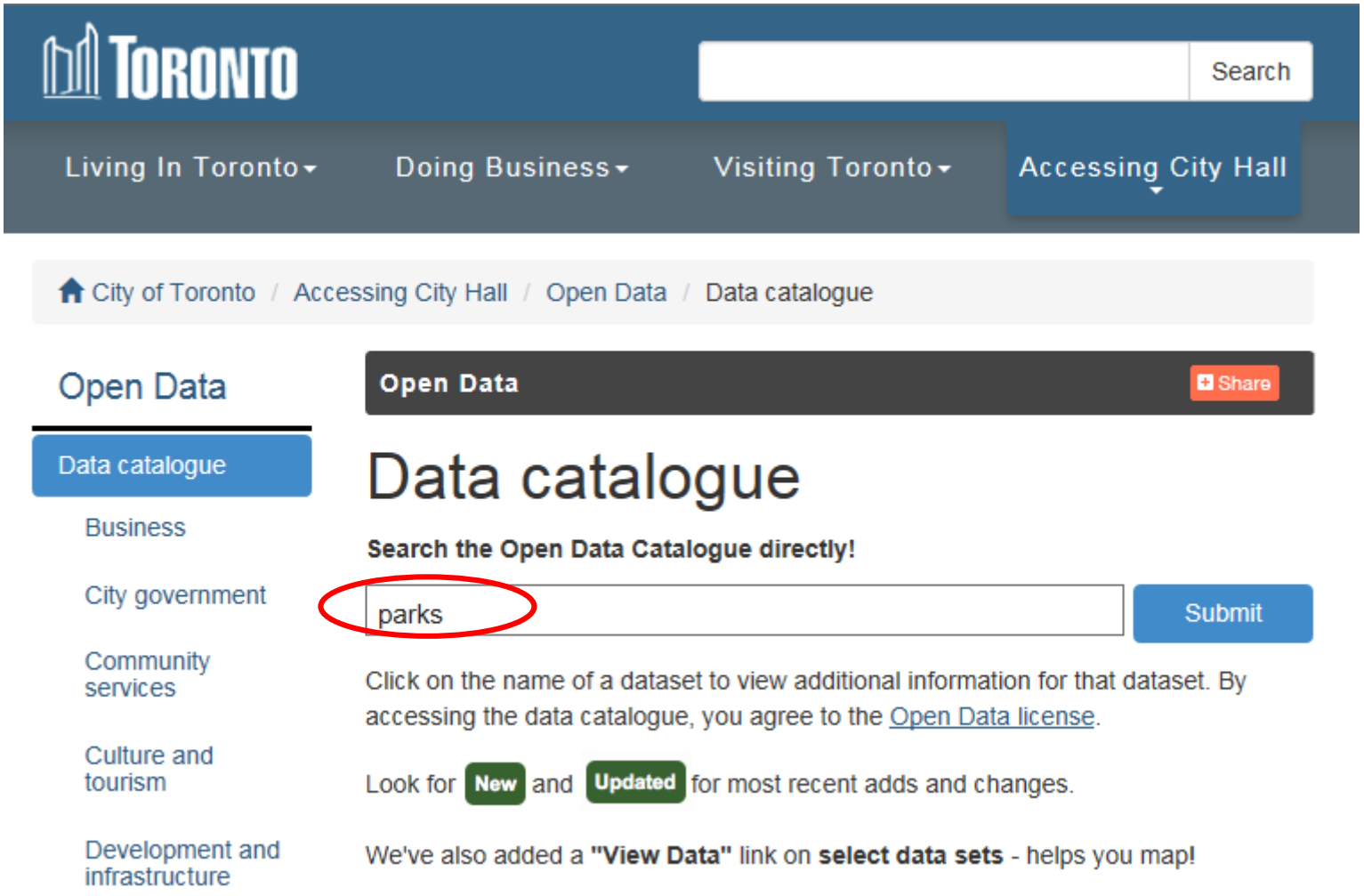
openrefine.org

OpenRefine Core

Google Refine 2.5 - Stable version

- **Windows kit**, Download, unzip, and double-click on *google-refine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it. **NOTE:** If you have issues installing Refine on Mac, please refer to issue 590
- **Linux kit**, Download, extract, then type *./refine* to start.

Real-world question…



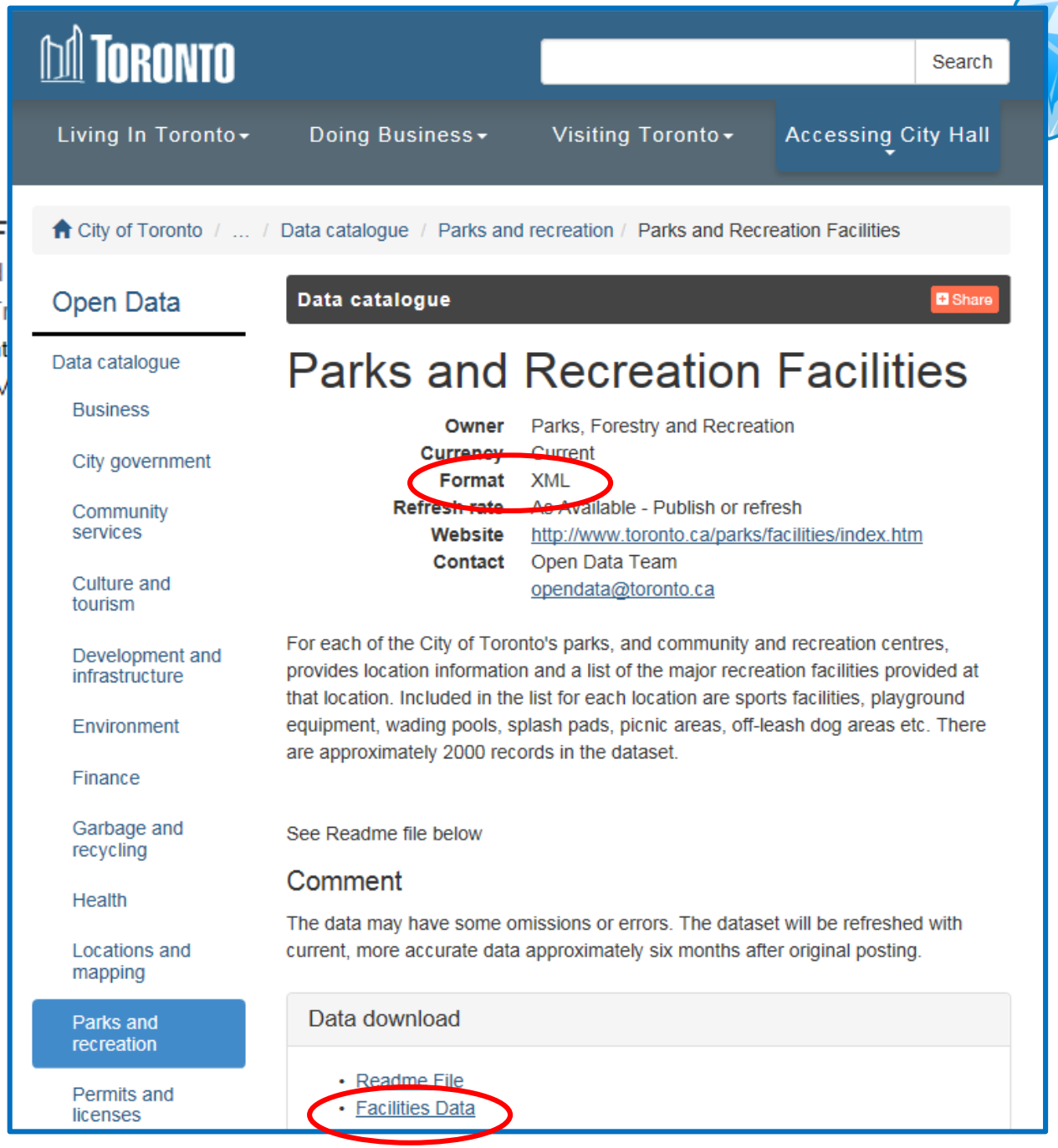
Parks and Recreation F

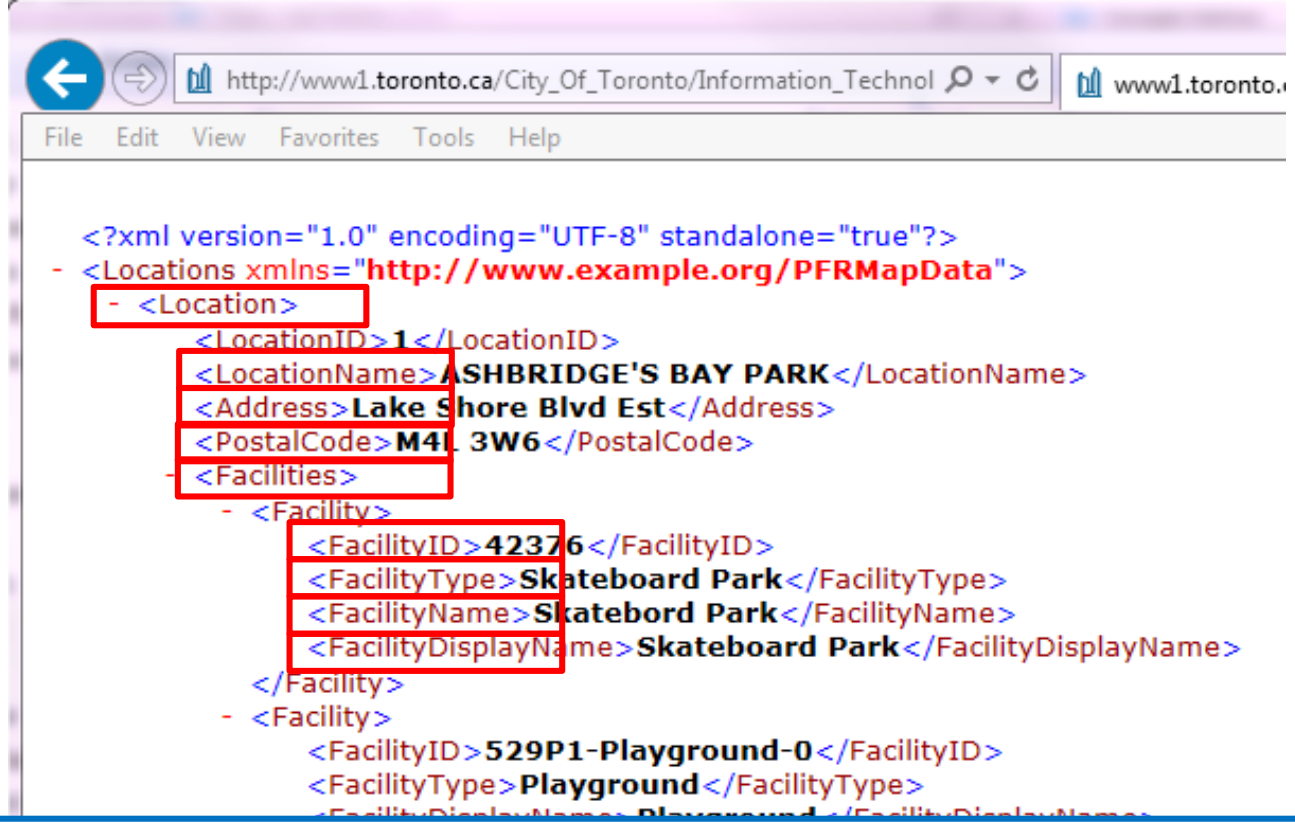
... Live With Culture; Parks and

Toronto Transit Commission; Tr

www1.toronto.ca/wps/portal/content

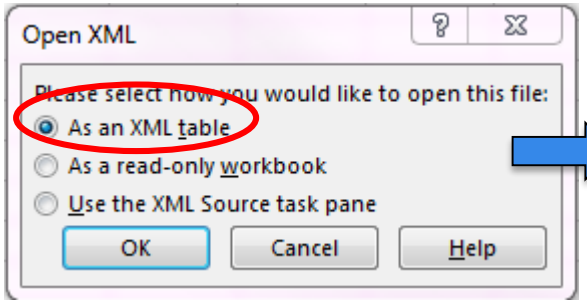
vgnextoid=d28b12f464151310VgnV

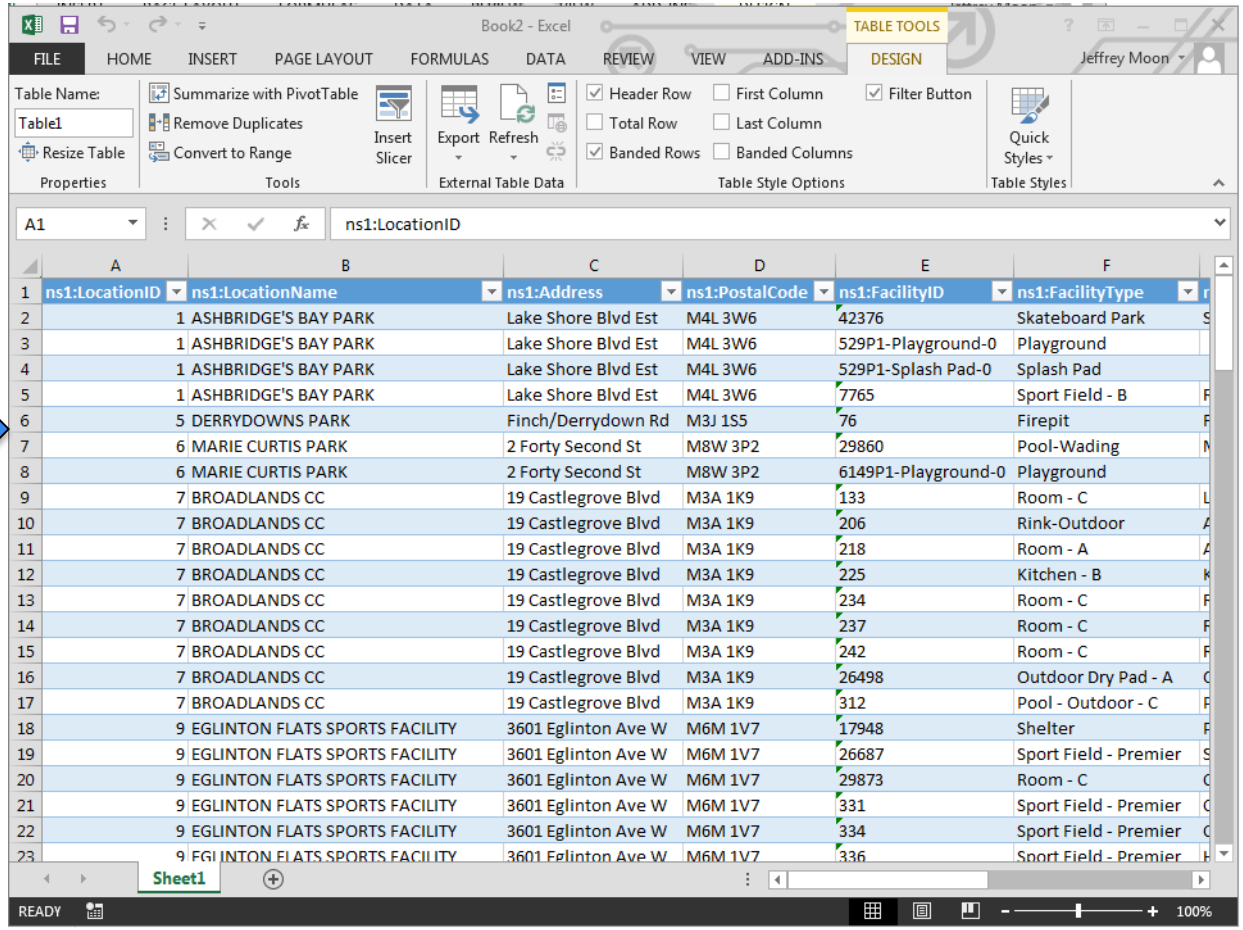




So, how do we get these data into a **spreadsheet** and manipulate them?

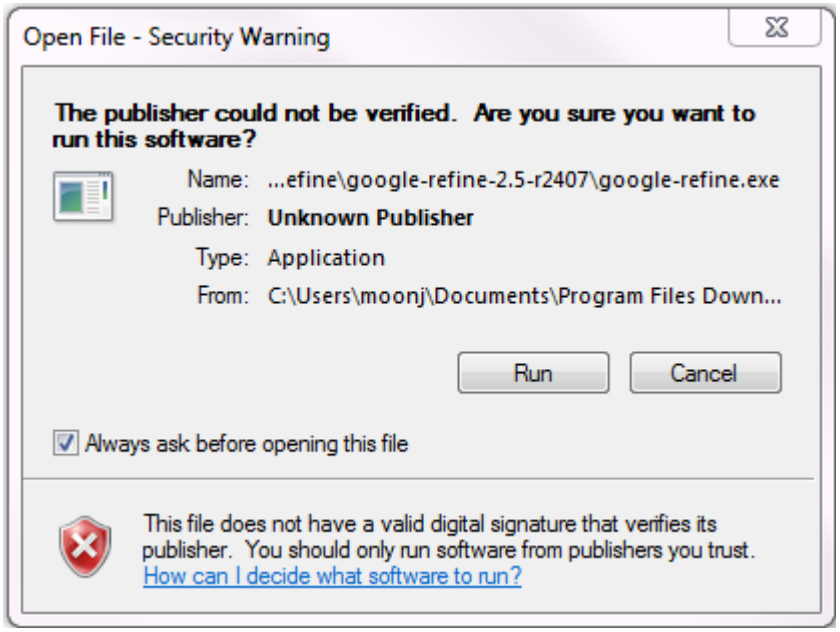
You could use Excel…

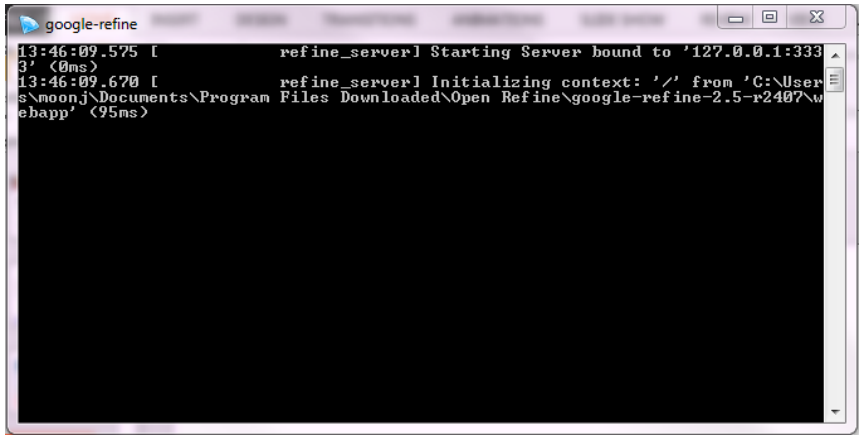




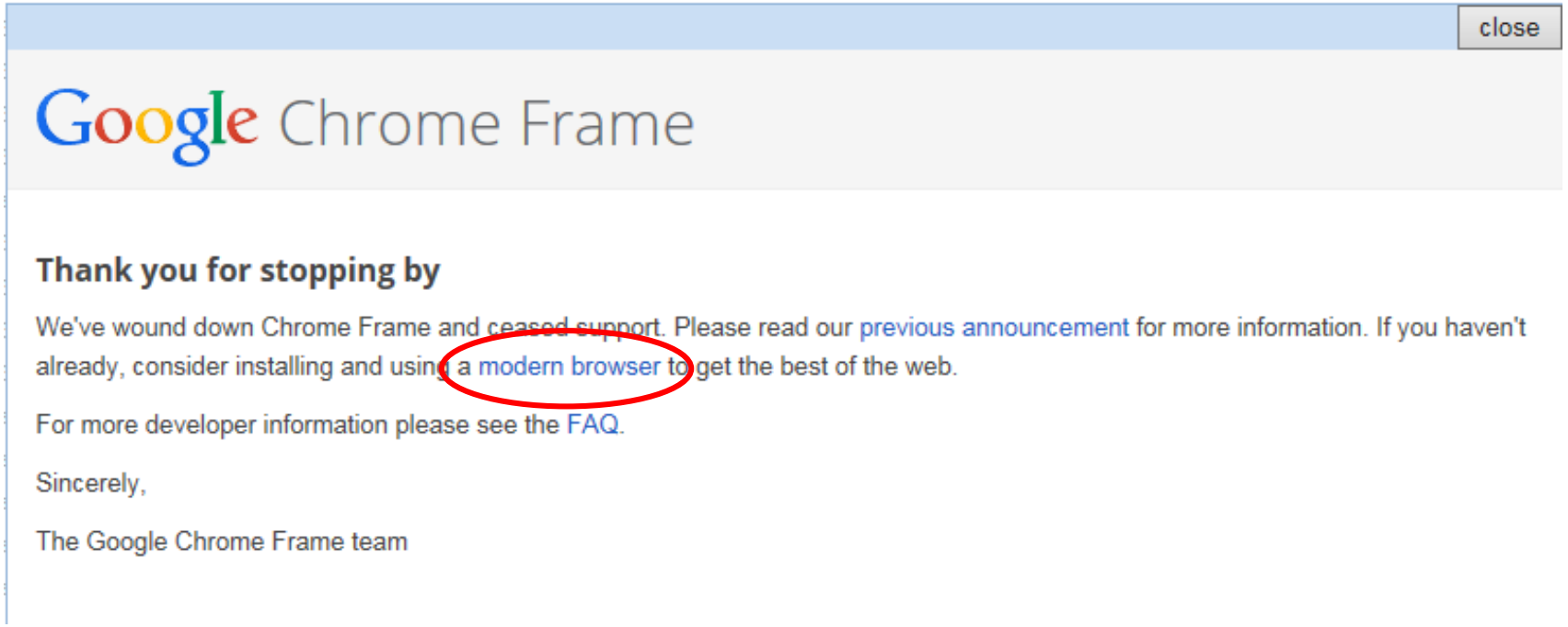
But let's try OpenRefine

- Start the program → Still called 'google-refine'
- You'll see:



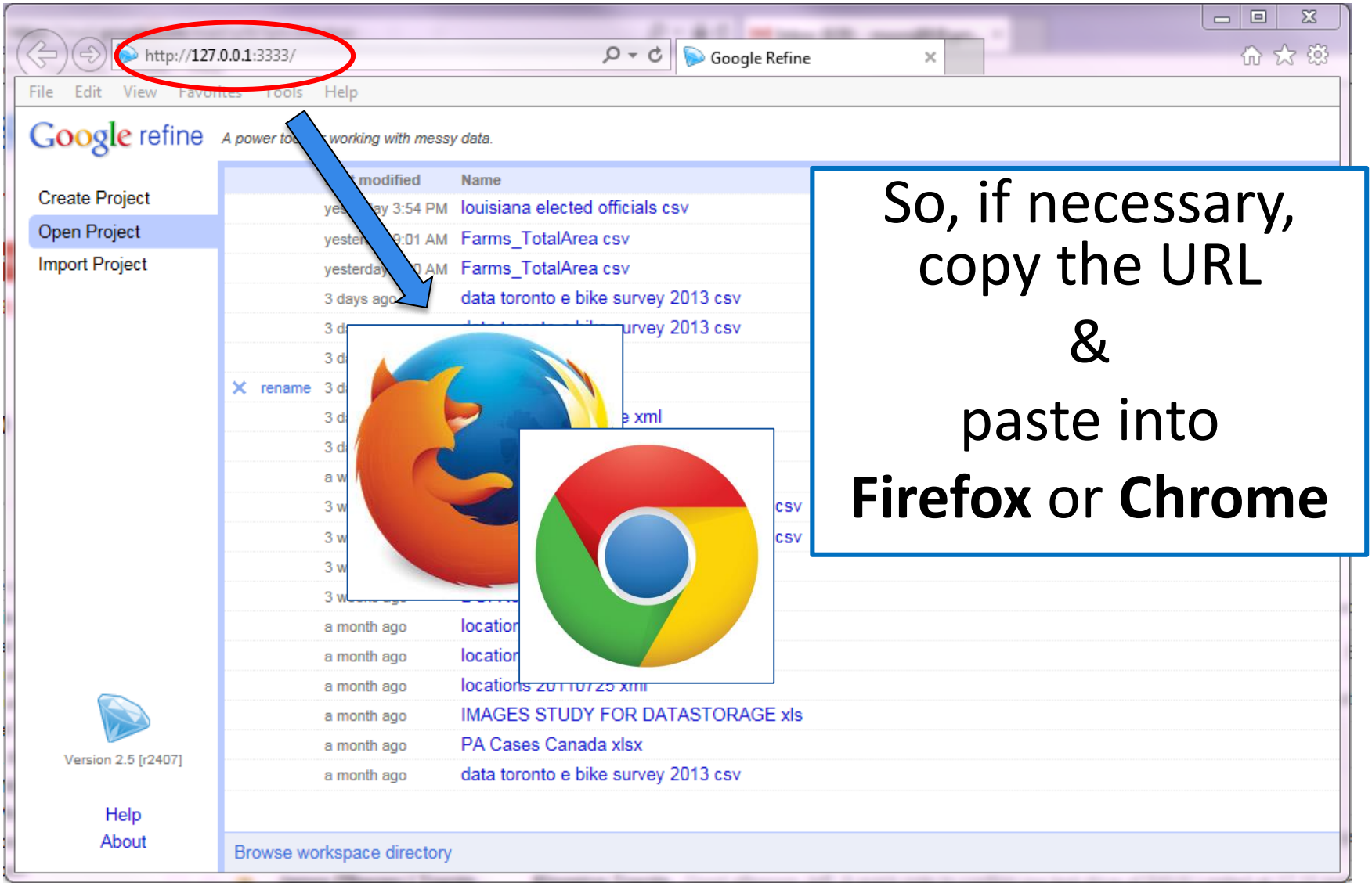


Warning to use a ‘modern browser’



On my PC, Internet Explorer is the default,

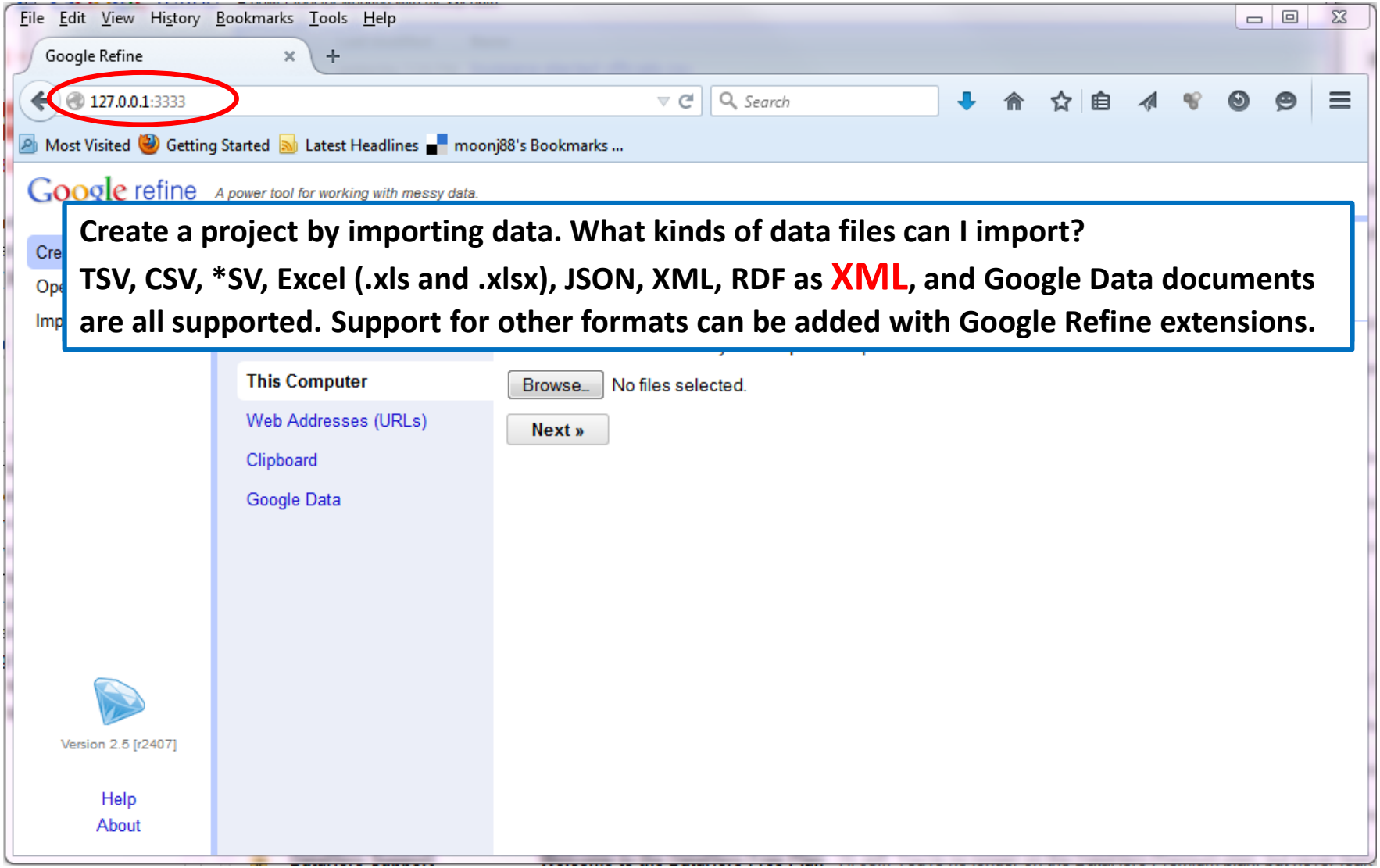
BUT, IE is not the best choice for OpenRefine



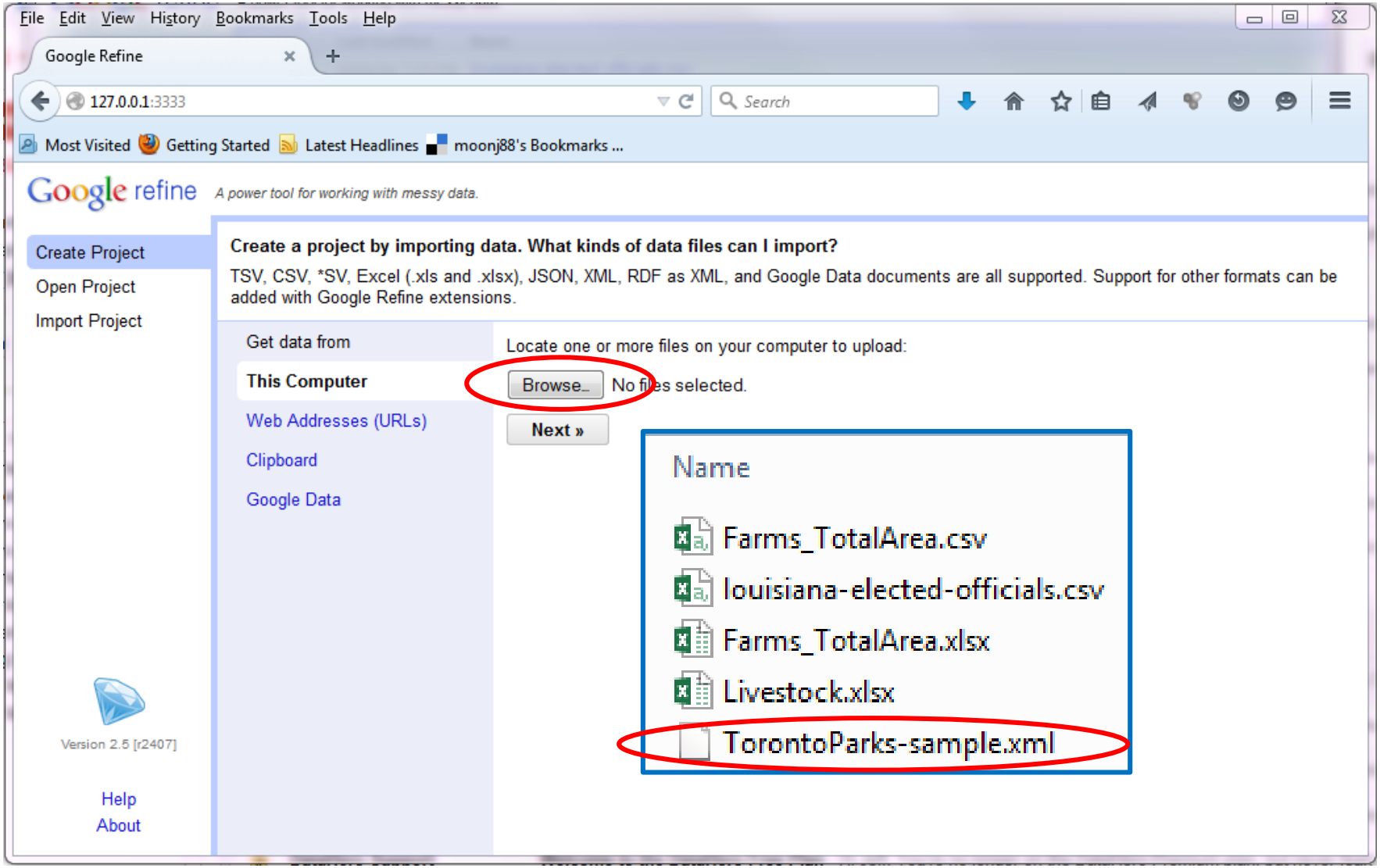
So, if necessary,
copy the URL
&
paste into
Firefox or **Chrome**

https://mozorg.cdn.mozilla.net/media/img/firefox/firefox-256.e2c1fc556816.jpg

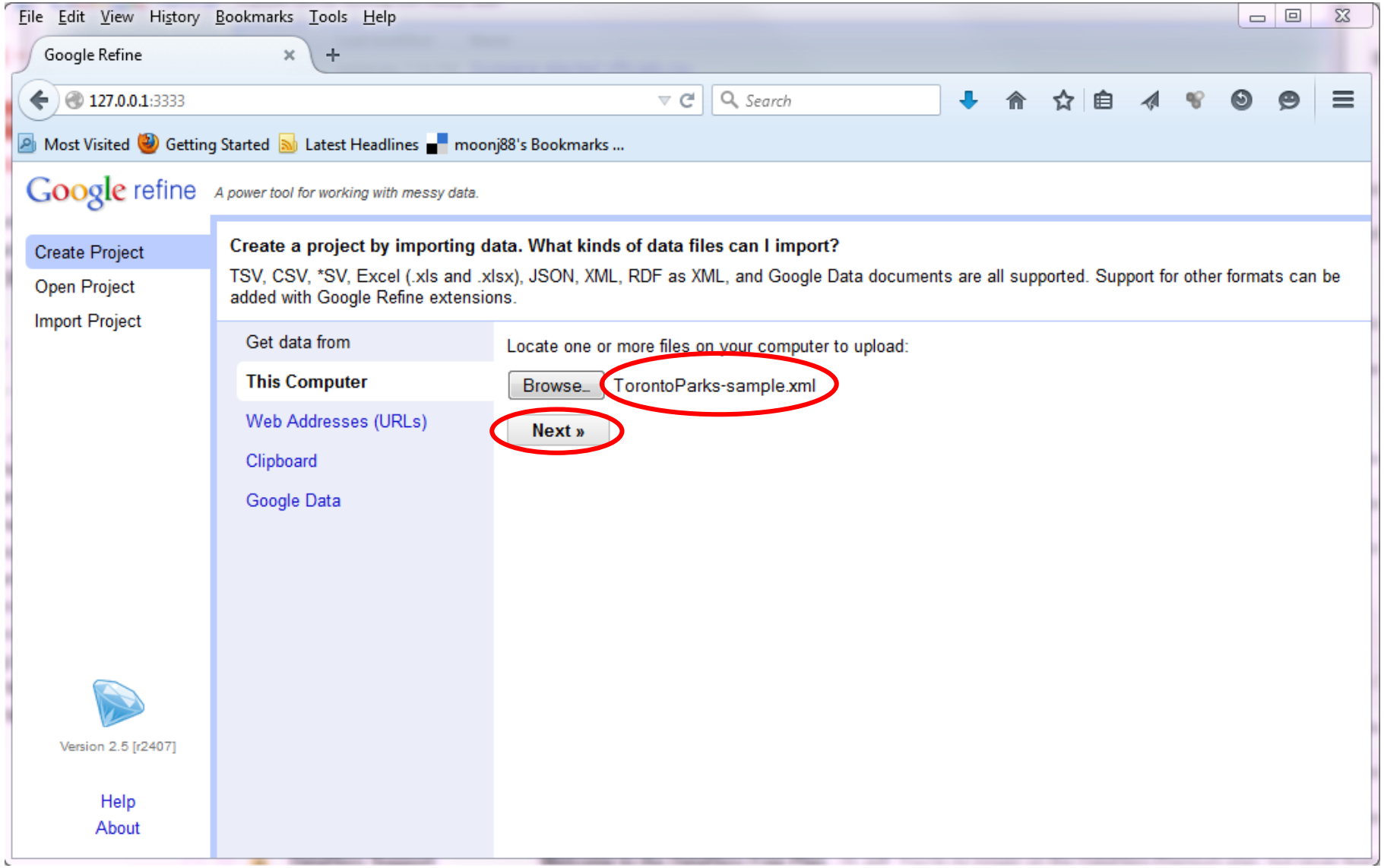
OpenRefine in Firefox



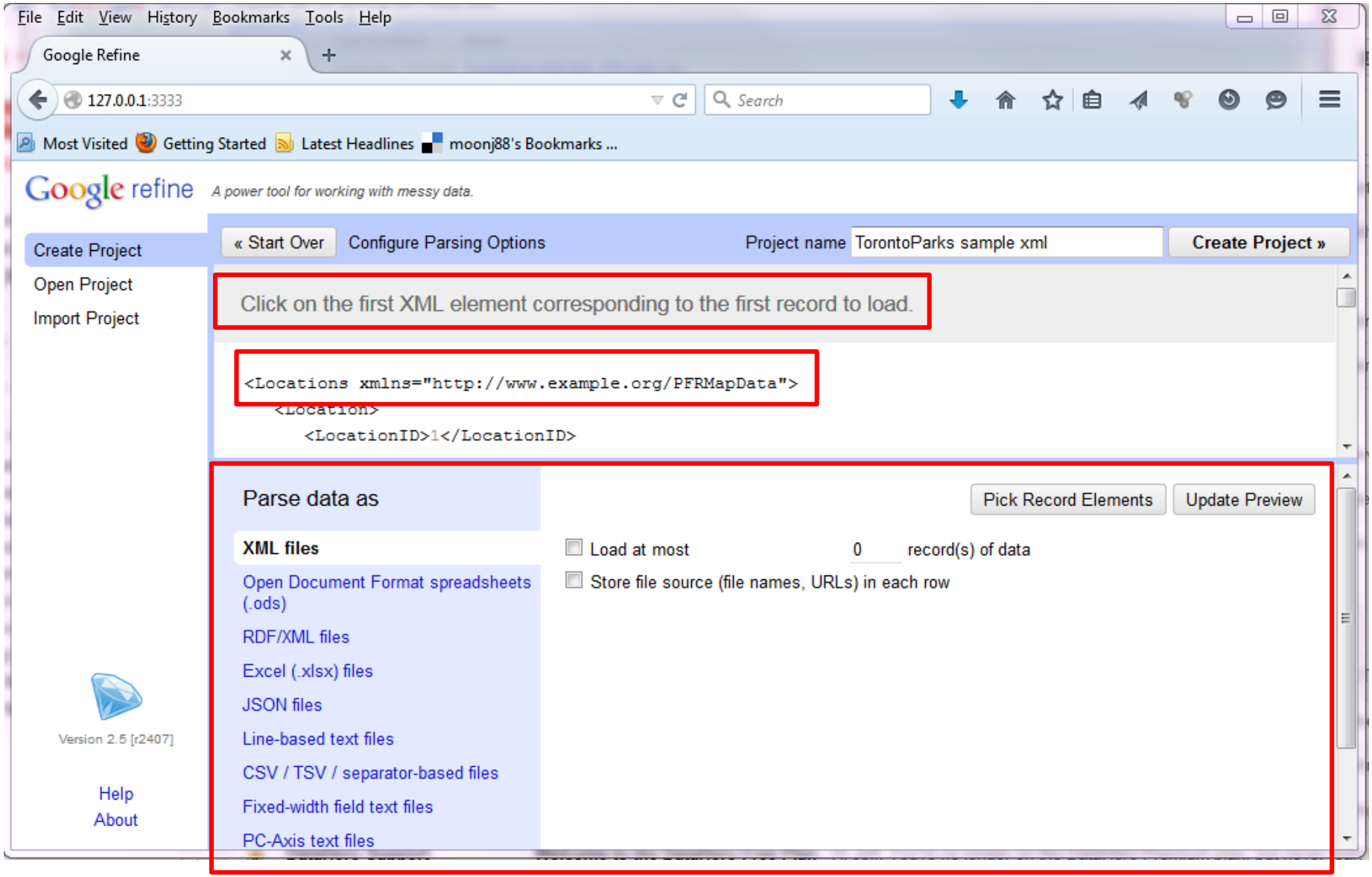
Browse to the XML file



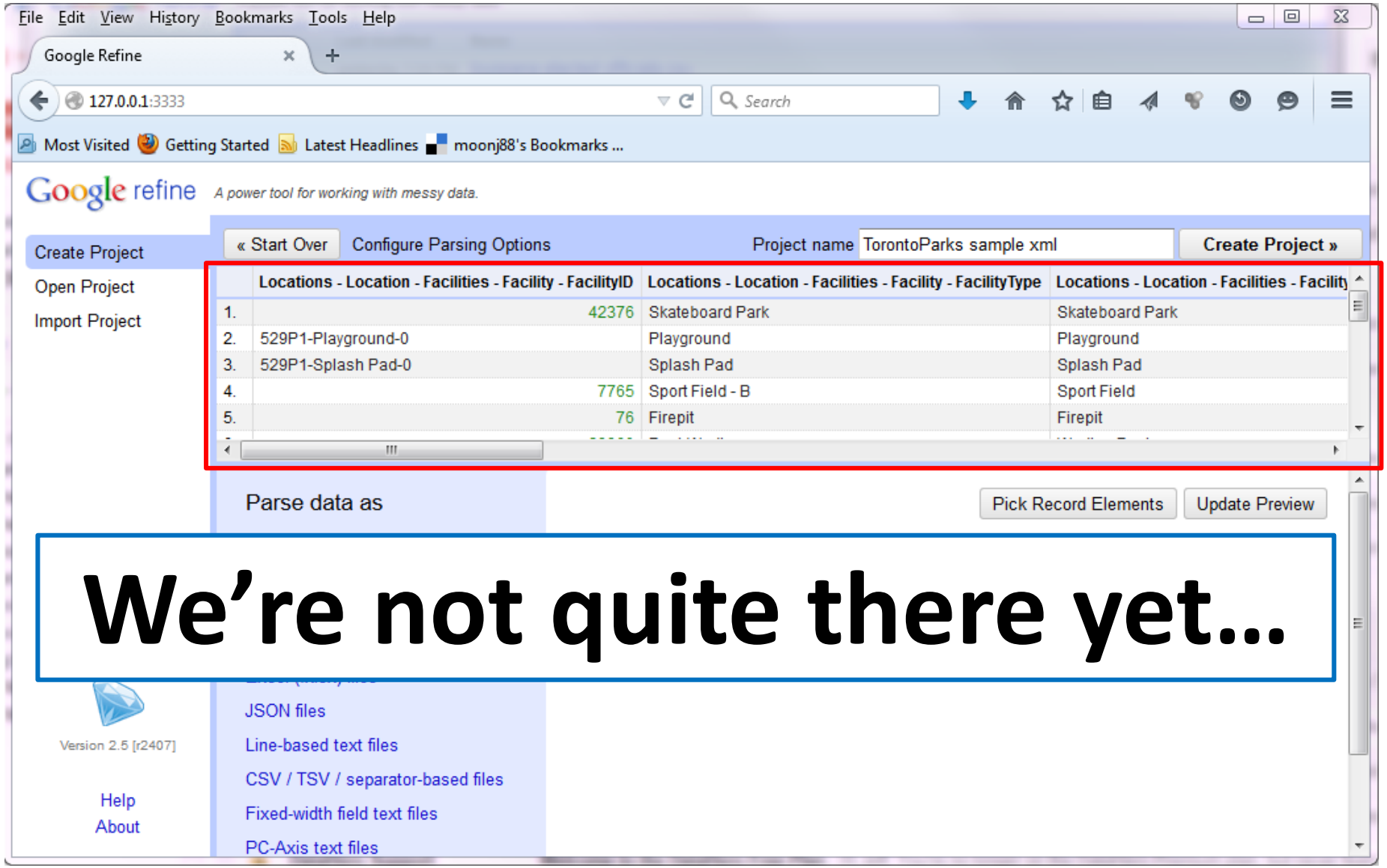
Click ‘Next’ when ready…



OpenRefine recognizes your file

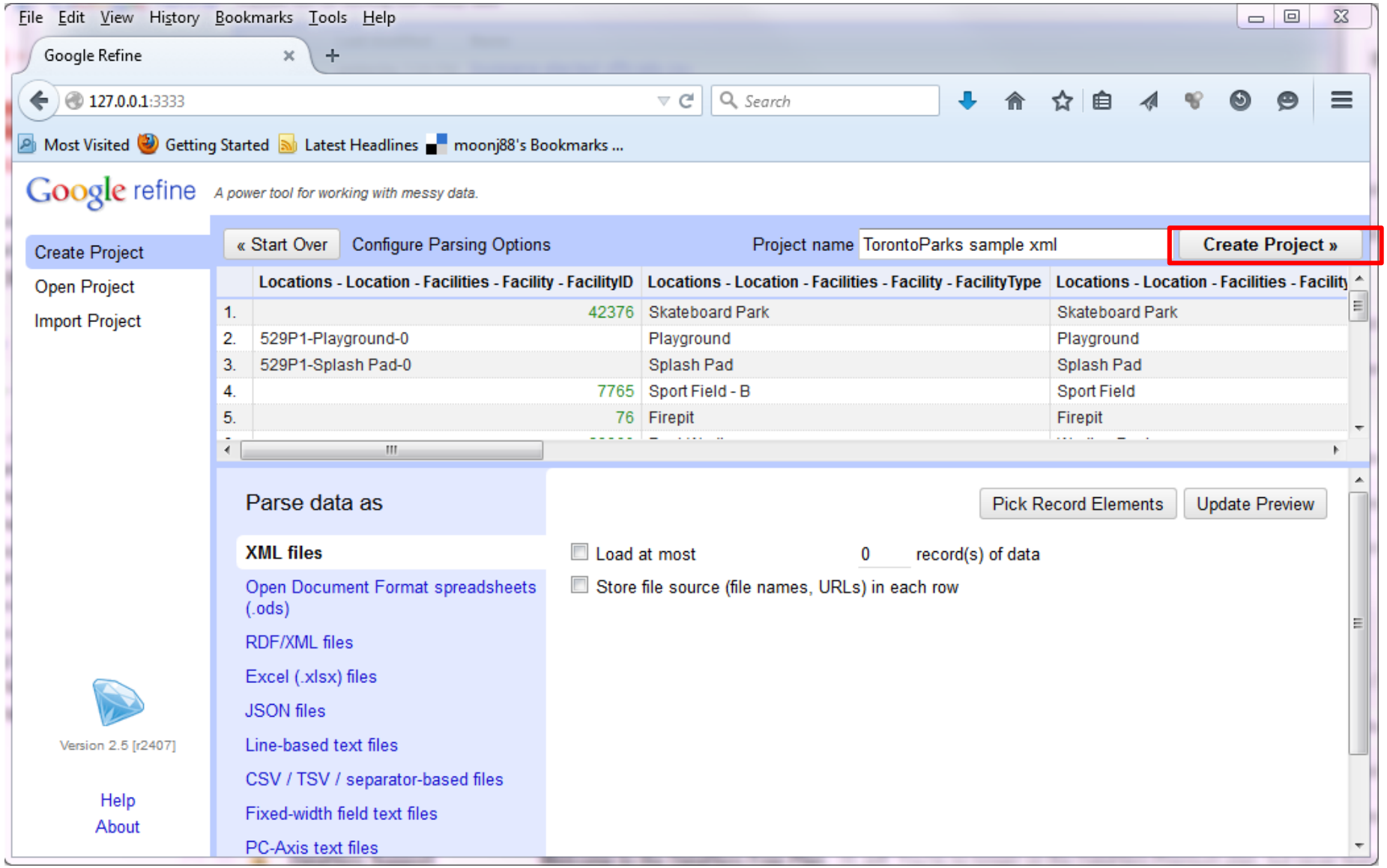


We get a preview of the data…

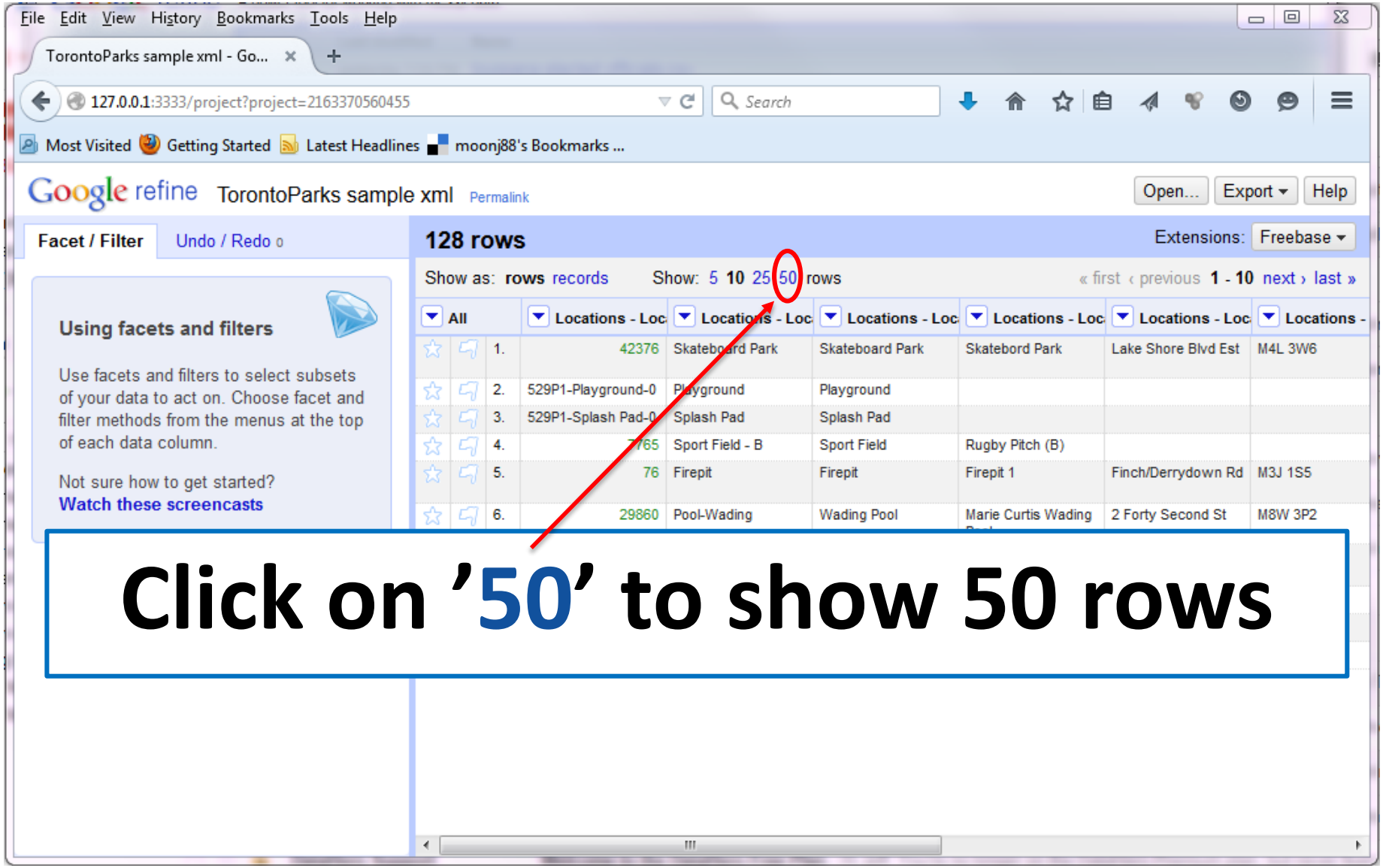


We're not quite there yet…

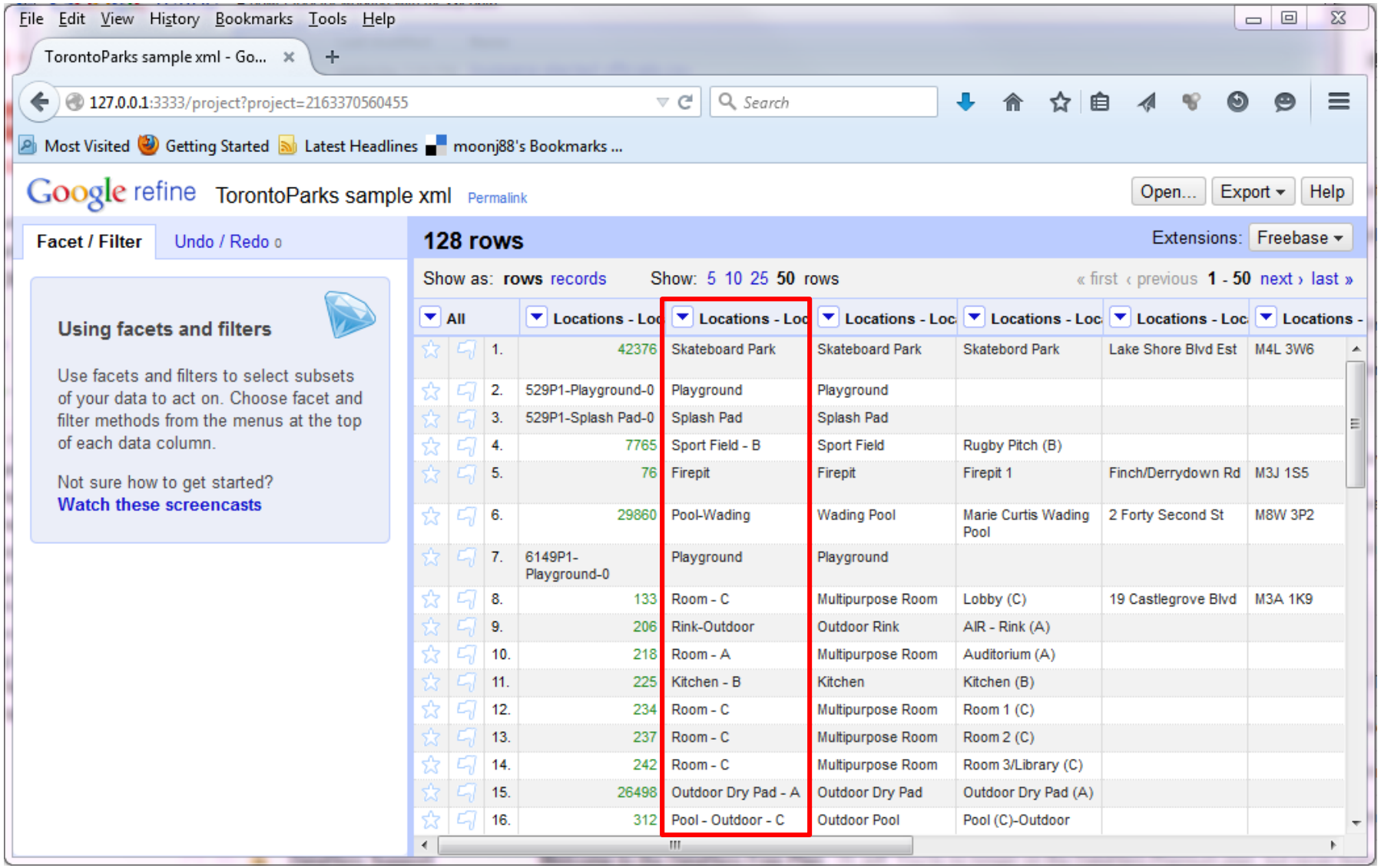
We’re ready to ‘create’ our project…



Data is now in OpenRefine

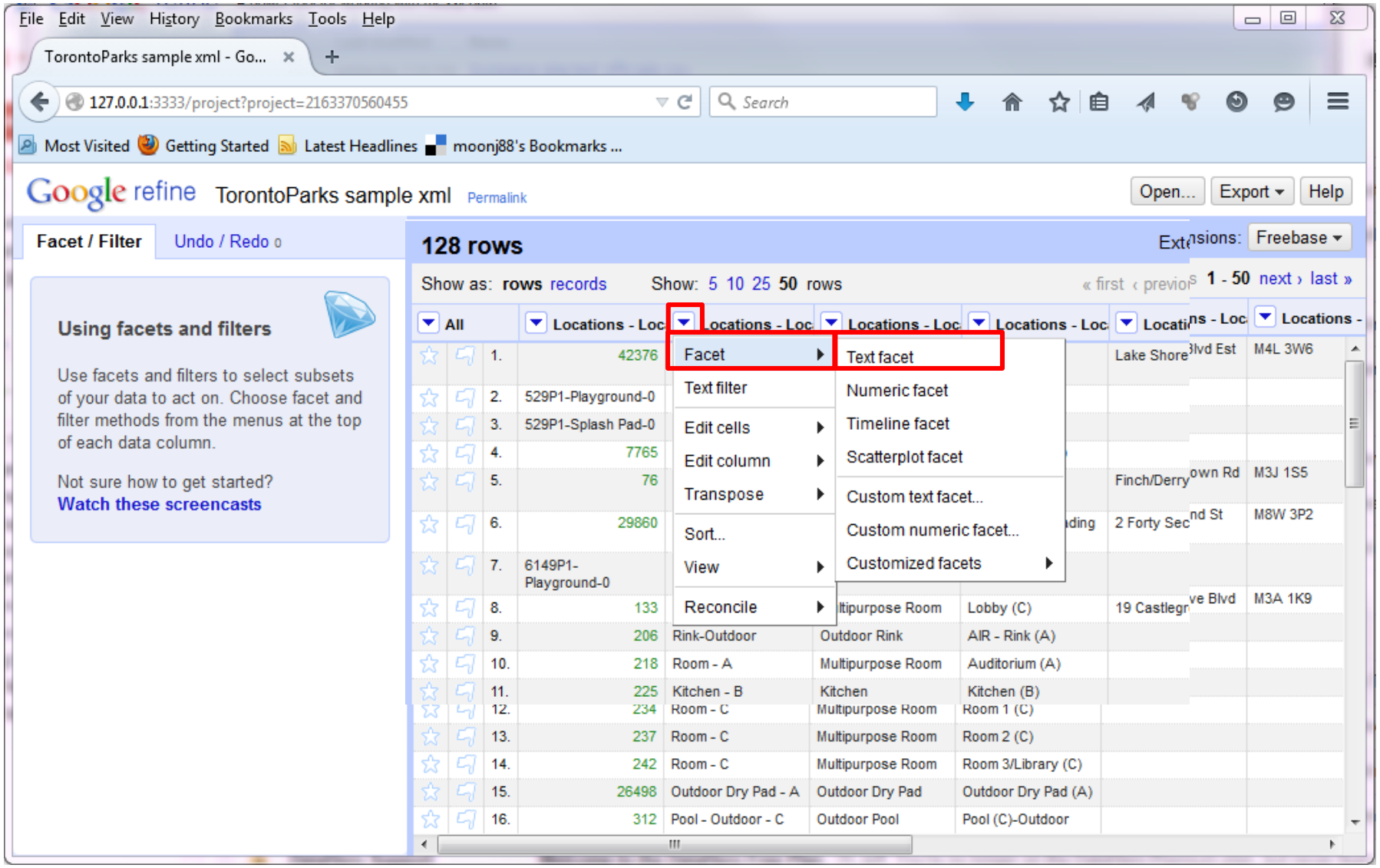


Text facets – let's look at the 2nd 'Location' column…

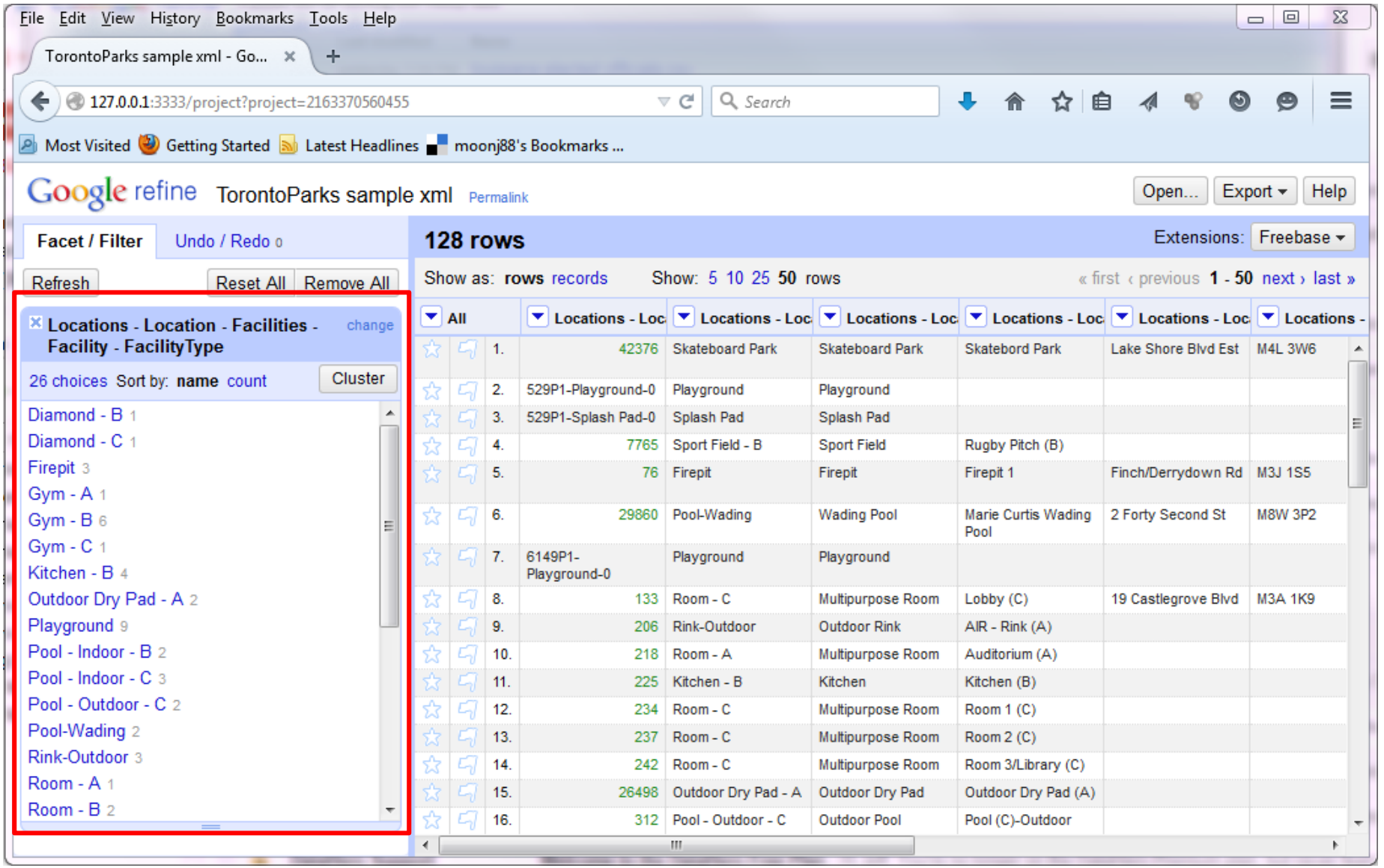


All		Locations - Loc	Locations - Loc	Locations - Loc	Locations - Loc	Locations - Loc	Locations -
1.	42376	Skateboard Park	Skateboard Park	Skatebord Park	Lake Shore Blvd Est	M4L 3W6	
2.	529P1-Playground-0	Playground	Playground				
3.	529P1-Splash Pad-0	Splash Pad	Splash Pad				
4.	7765	Sport Field - B	Sport Field	Rugby Pitch (B)			
5.	76	Firepit	Firepit	Firepit 1	Finch/Derrydown Rd	M3J 1S5	
6.	29860	Pool-Wading	Wading Pool	Marie Curtis Wading Pool	2 Forty Second St	M8W 3P2	
7.	6149P1-Playground-0	Playground	Playground				
8.	133	Room - C	Multipurpose Room	Lobby (C)	19 Castlegrove Blvd	M3A 1K9	
9.	206	Rink-Outdoor	Outdoor Rink	AIR - Rink (A)			
10.	218	Room - A	Multipurpose Room	Auditorium (A)			
11.	225	Kitchen - B	Kitchen	Kitchen (B)			
12.	234	Room - C	Multipurpose Room	Room 1 (C)			
13.	237	Room - C	Multipurpose Room	Room 2 (C)			
14.	242	Room - C	Multipurpose Room	Room 3/Library (C)			
15.	26498	Outdoor Dry Pad - A	Outdoor Dry Pad	Outdoor Dry Pad (A)			
16.	312	Pool - Outdoor - C	Outdoor Pool	Pool (C)-Outdoor			

Dropdown: Facet → Text facet



Facets show all unique entries, with frequencies



Facet / Filter | Undo / Redo 0

Refresh | Reset All | Remove All

Locations - Location - Facilities - Facility - FacilityType change

26 choices Sort by: **name** count | Cluster

- Diamond - B 1
- Diamond - C 1
- Firepit 3
- Gym - A 1
- Gym - B 6
- Gym - C 1
- Kitchen - B 4
- Outdoor Dry Pad - A 2
- Playground 9
- Pool - Indoor - B 2
- Pool - Indoor - C 3
- Pool - Outdoor - C 2
- Pool-Wading 2
- Rink-Outdoor 3
- Room - A 1
- Room - B 2

Unique values & Frequencies

Let's look at another example

Canadian Century Research Infrastructure
Infrastructure de recherche sur le Canada au 20e siècle

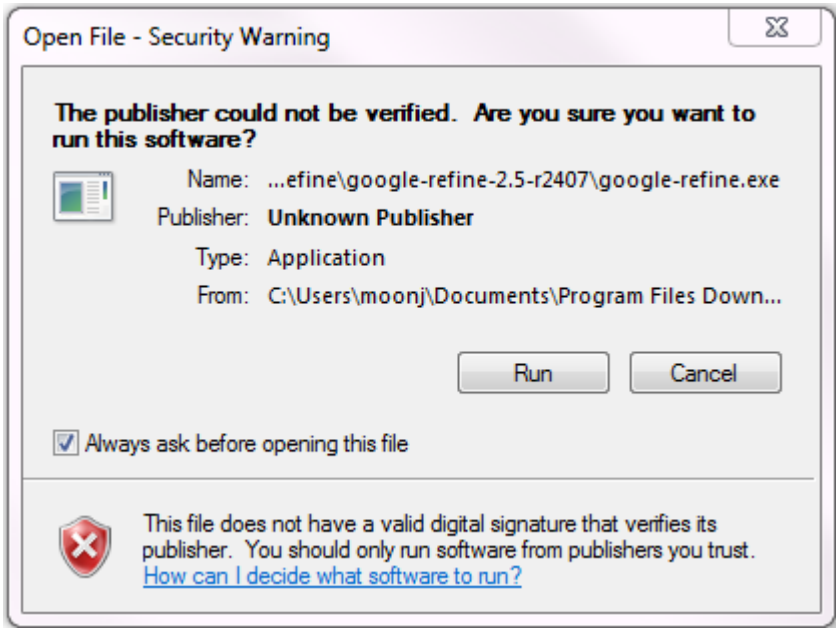
1911 Census Microdata

For this workshop I created a subset of ~100 records from each Province/Territory

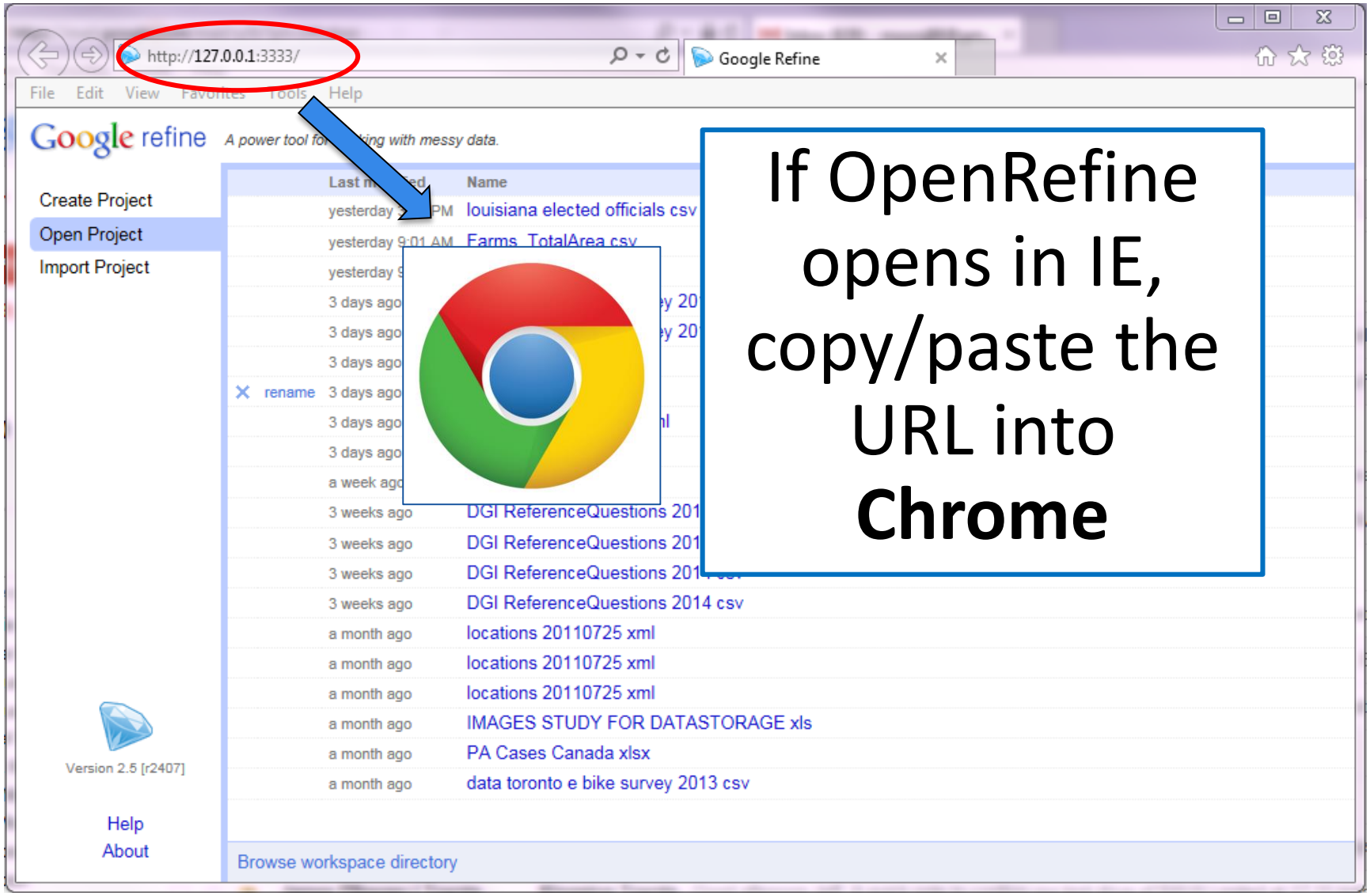
http://127.0.0.1:3333/

Start OpenRefine

- Start the program → Still called ‘google-refine’
- You’ll see:



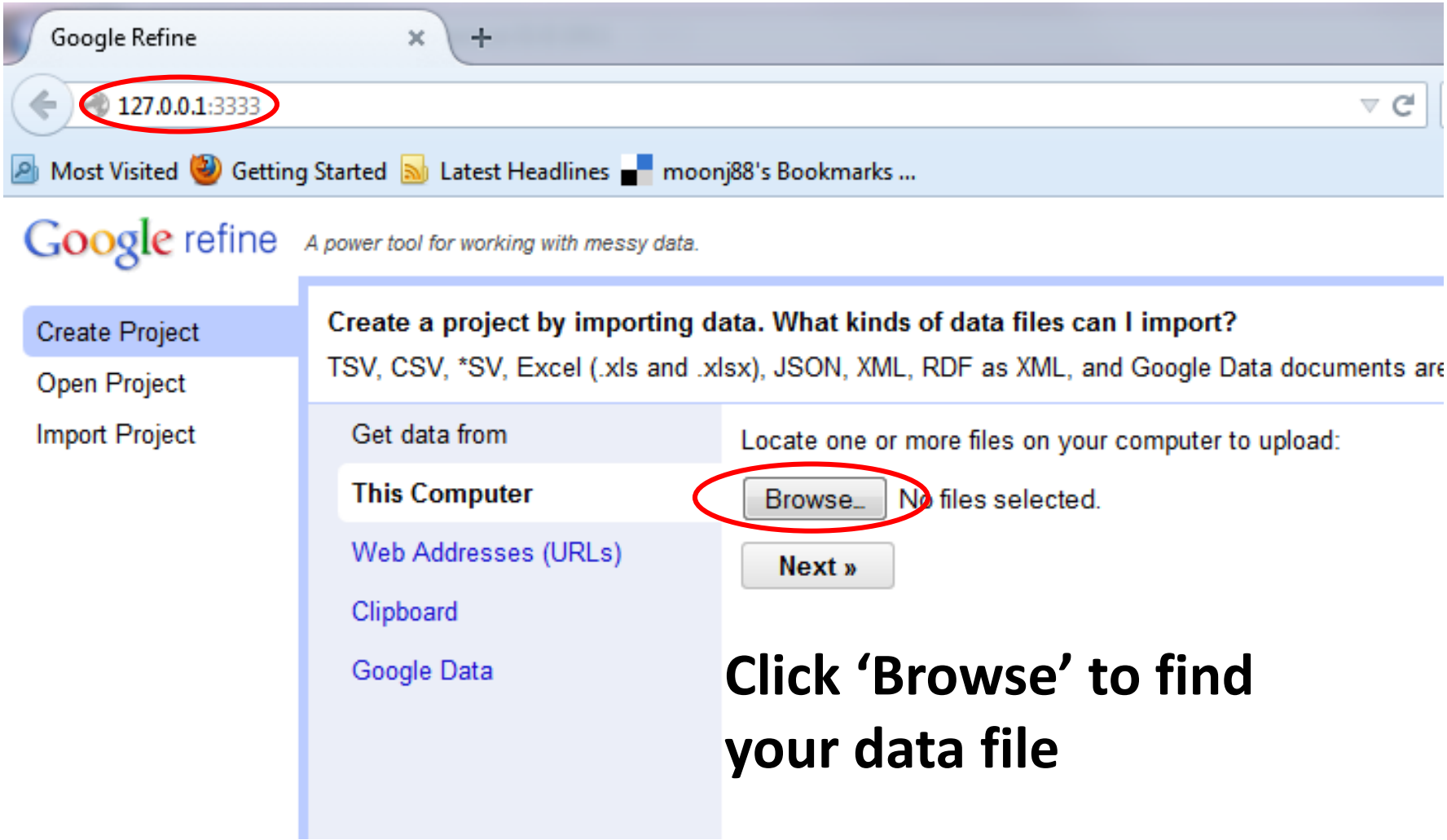
If needed, copy the URL to Chrome…



If OpenRefine opens in IE, copy/paste the URL into **Chrome**

https://mozorg.cdn.mozilla.net/media/img/firefox/firefox-256.e2c1fc556816.jpg

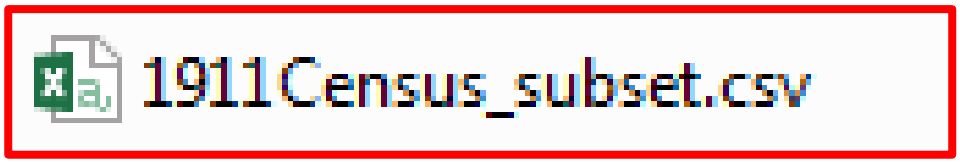
OpenRefine start screen
→ Click Browse



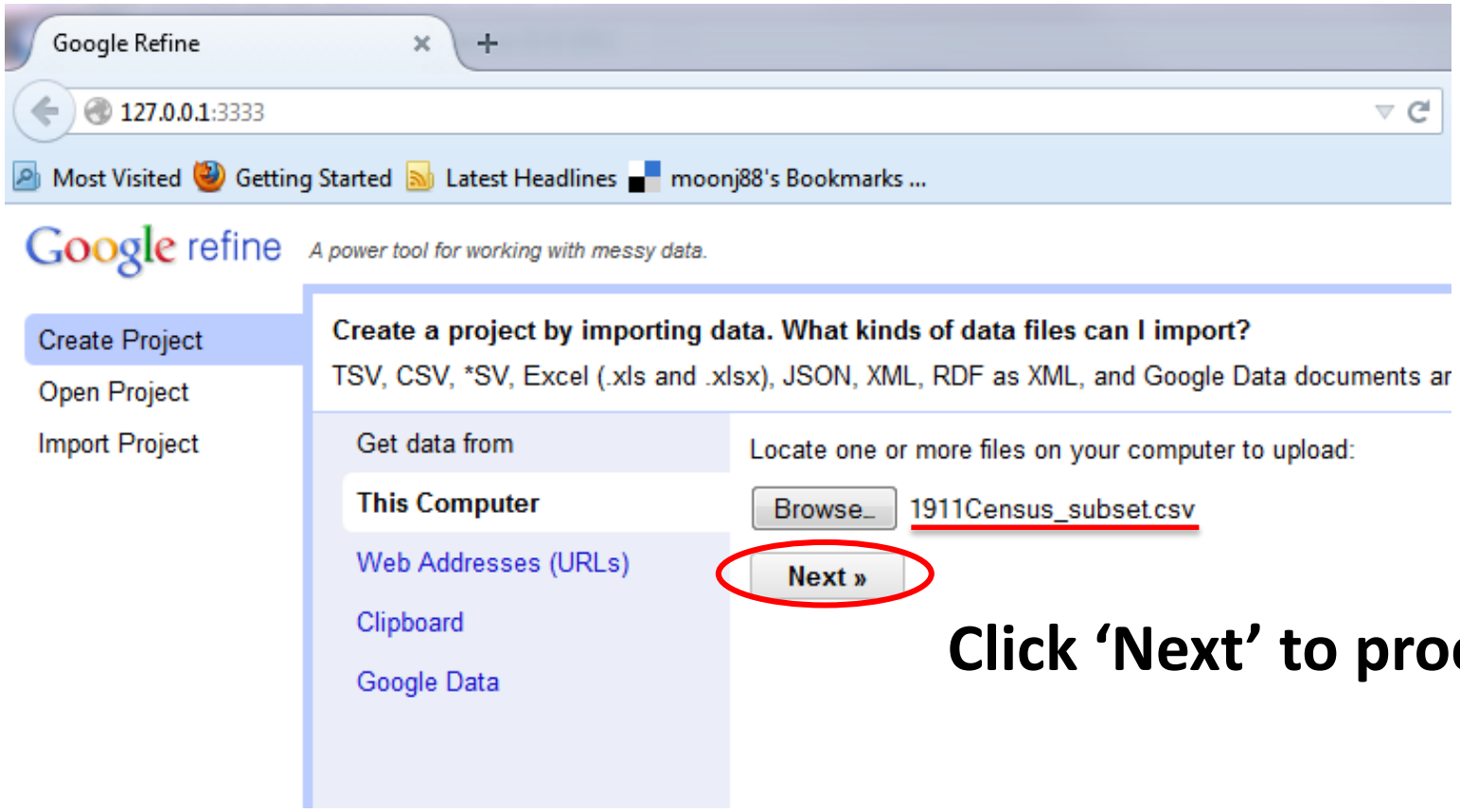
Click 'Browse' to find your data file

Navigate to the '.csv' file

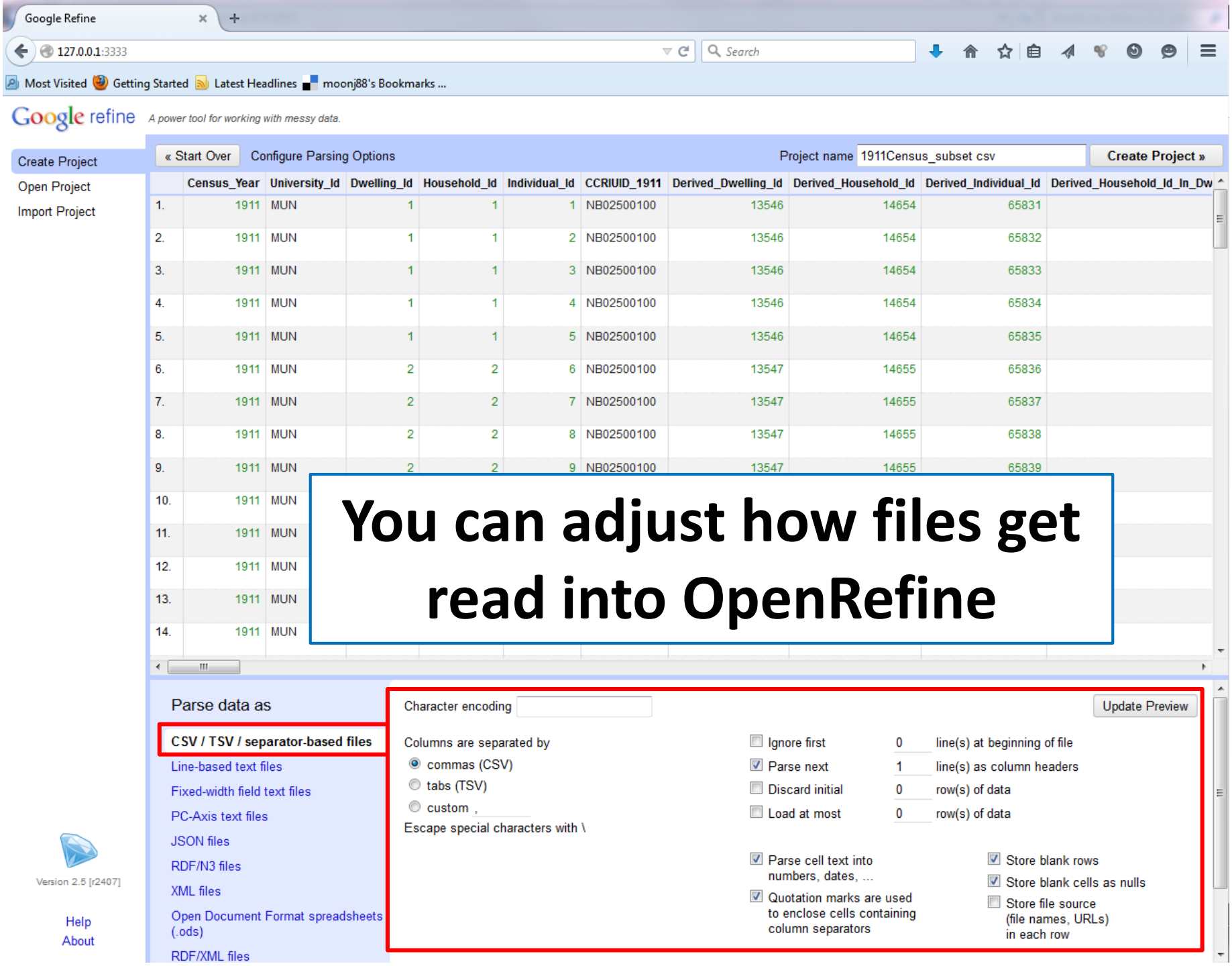
D:\pccf\1911Census_subset.csv

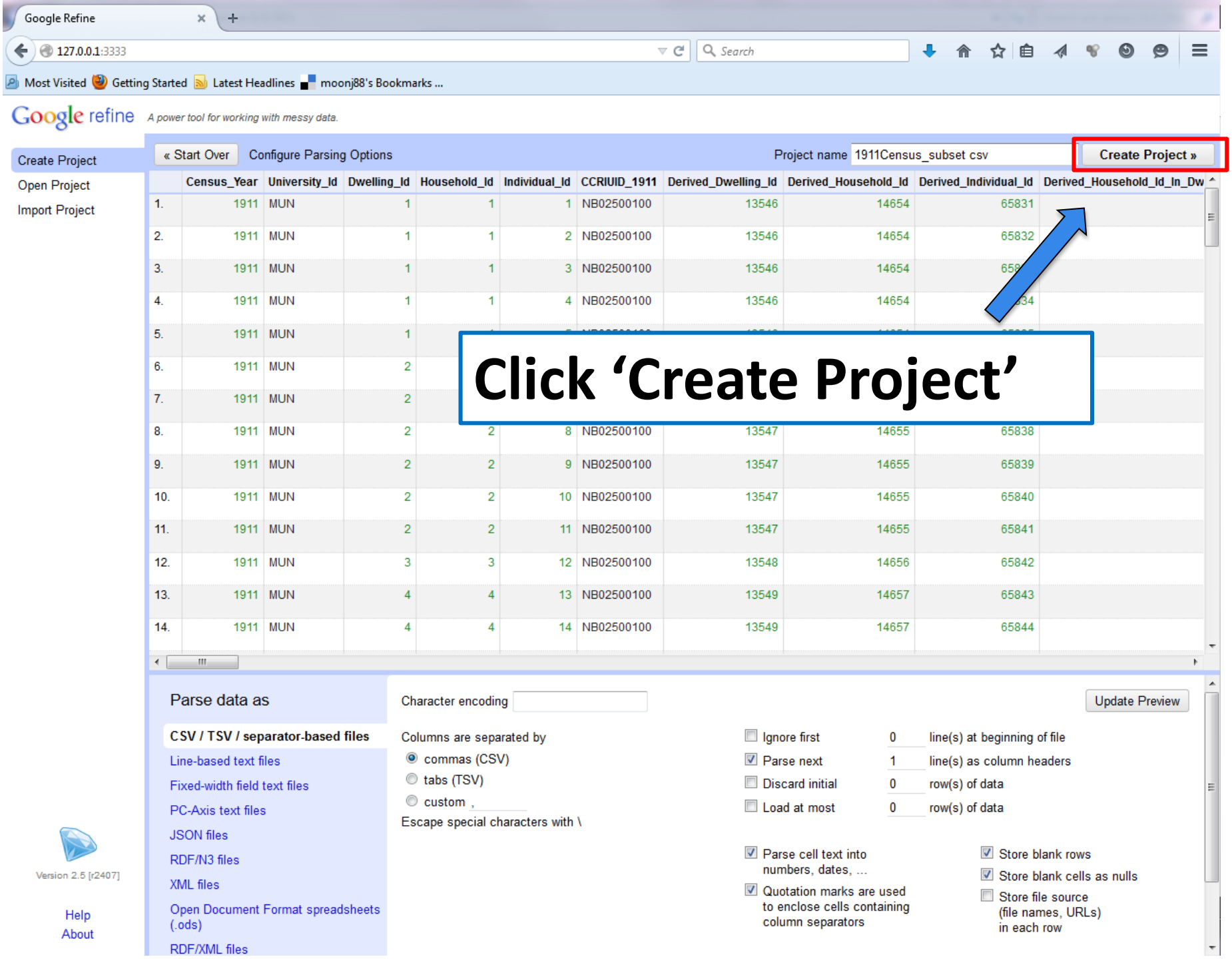


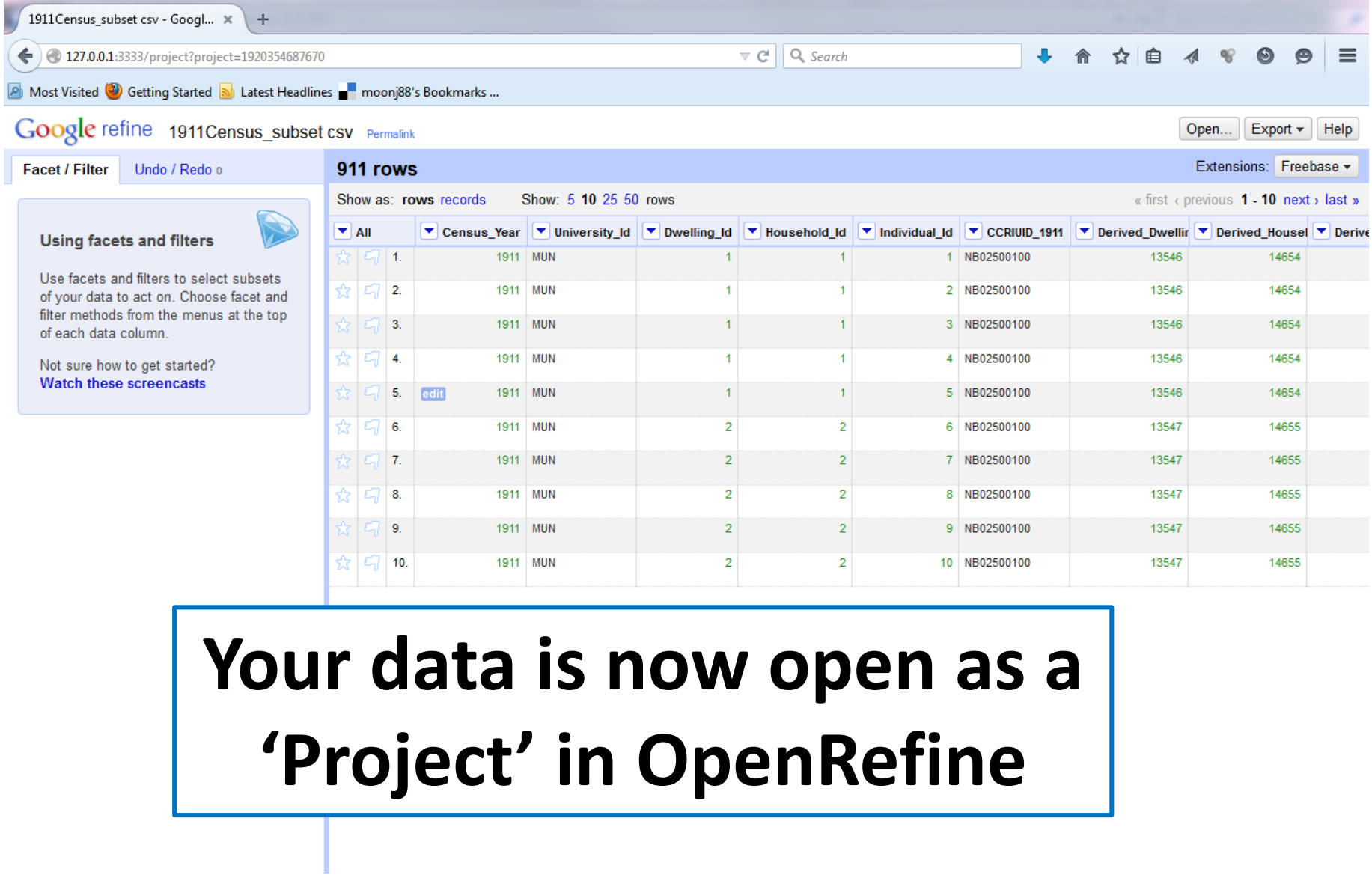
And click 'Open'



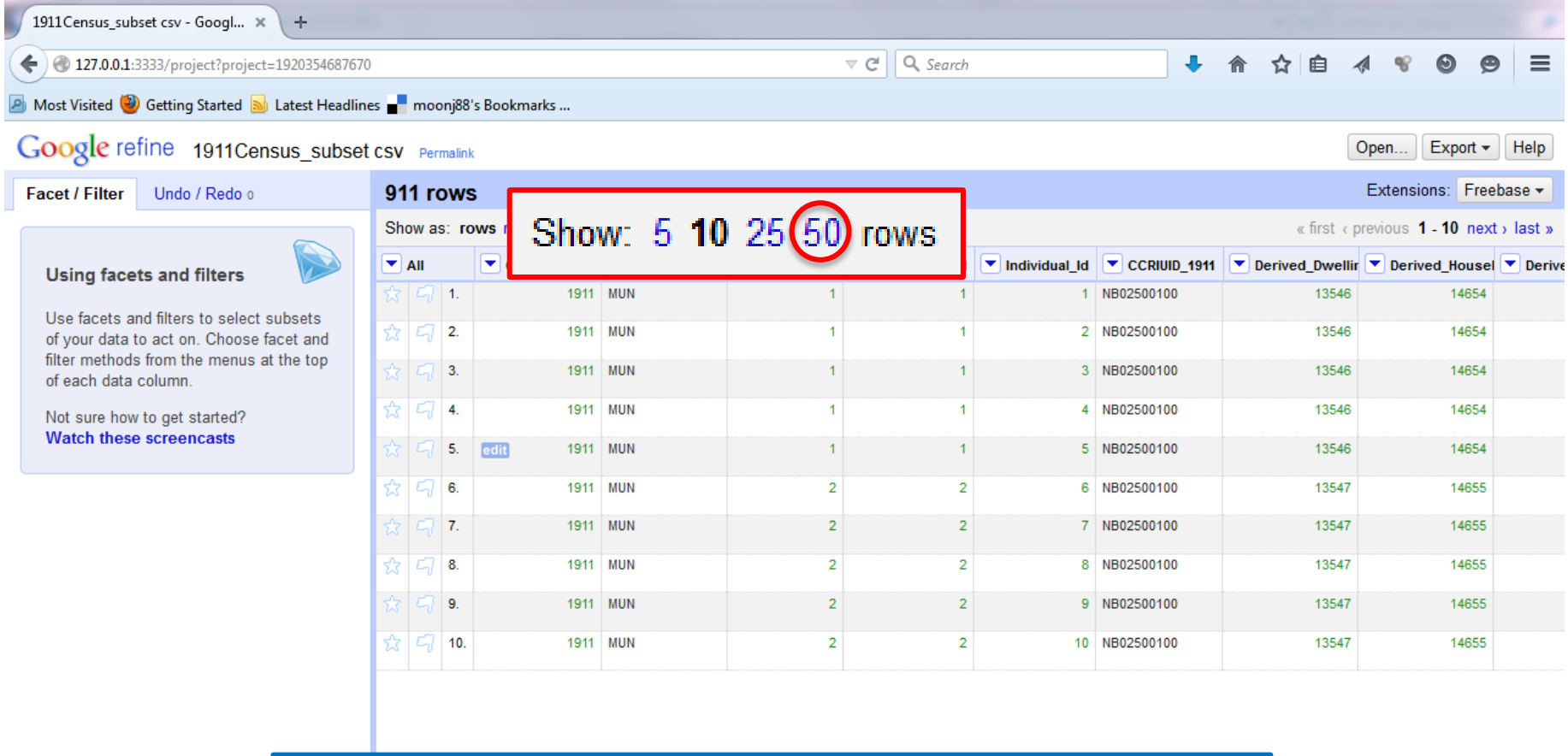
Click 'Next' to proceed


Google Refine
127.0.0.1:3333
Search
Most Visited
Getting Started
Latest Headlines
moonj88's Bookmarks ...
Google refine
A power tool for working with messy data.
Create Project
Open Project
Import Project
« Start Over
Configure Parsing Options
Project name
1911Census_subset csv
Create Project »
Census_Year University_Id Dwelling_Id Household_Id Individual_Id CCRIUID_1911 Derived_Dwelling_Id Derived_Household_Id Derived_Individual_Id Derived_Household_Id_In_Dw
1. 1911 MUN 1 1 1 NB02500100 13546 14654 65831
2. 1911 MUN 1 1 2 NB02500100 13546 14654 65832
3. 1911 MUN 1 1 3 NB02500100 13546 14654 65833
4. 1911 MUN 1 1 4 NB02500100 13546 14654 65834
5. 1911 MUN 1 1 5 NB02500100 13546 14654 65835
6. 1911 MUN 2 2 6 NB02500100 13547 14655 65836
7. 1911 MUN 2 2 7 NB02500100 13547 14655 65837
8. 1911 MUN 2 2 8 NB02500100 13547 14655 65838
9. 1911 MUN 2 2 9 NB02500100 13547 14655 65839
10. 1911 MUN
11. 1911 MUN
12. 1911 MUN
13. 1911 MUN
14. 1911 MUN
You can adjust how files get read into OpenRefine
Parse data as
CSV / TSV / separator-based files
Line-based text files
Fixed-width field text files
PC-Axis text files
JSON files
RDF/N3 files
XML files
Open Document Format spreadsheets (.ods)
RDF/XML files
Character encoding
Update Preview
Columns are separated by
commas (CSV)
tabs (TSV)
custom ,
Escape special characters with \
Ignore first 0 line(s) at beginning of file
Parse next 1 line(s) as column headers
Discard initial 0 row(s) of data
Load at most 0 row(s) of data
Parse cell text into numbers, dates, ...
Quotation marks are used to enclose cells containing column separators
Store blank rows
Store blank cells as nulls
Store file source (file names, URLs) in each row
Version 2.5 [r2407]
Help
About

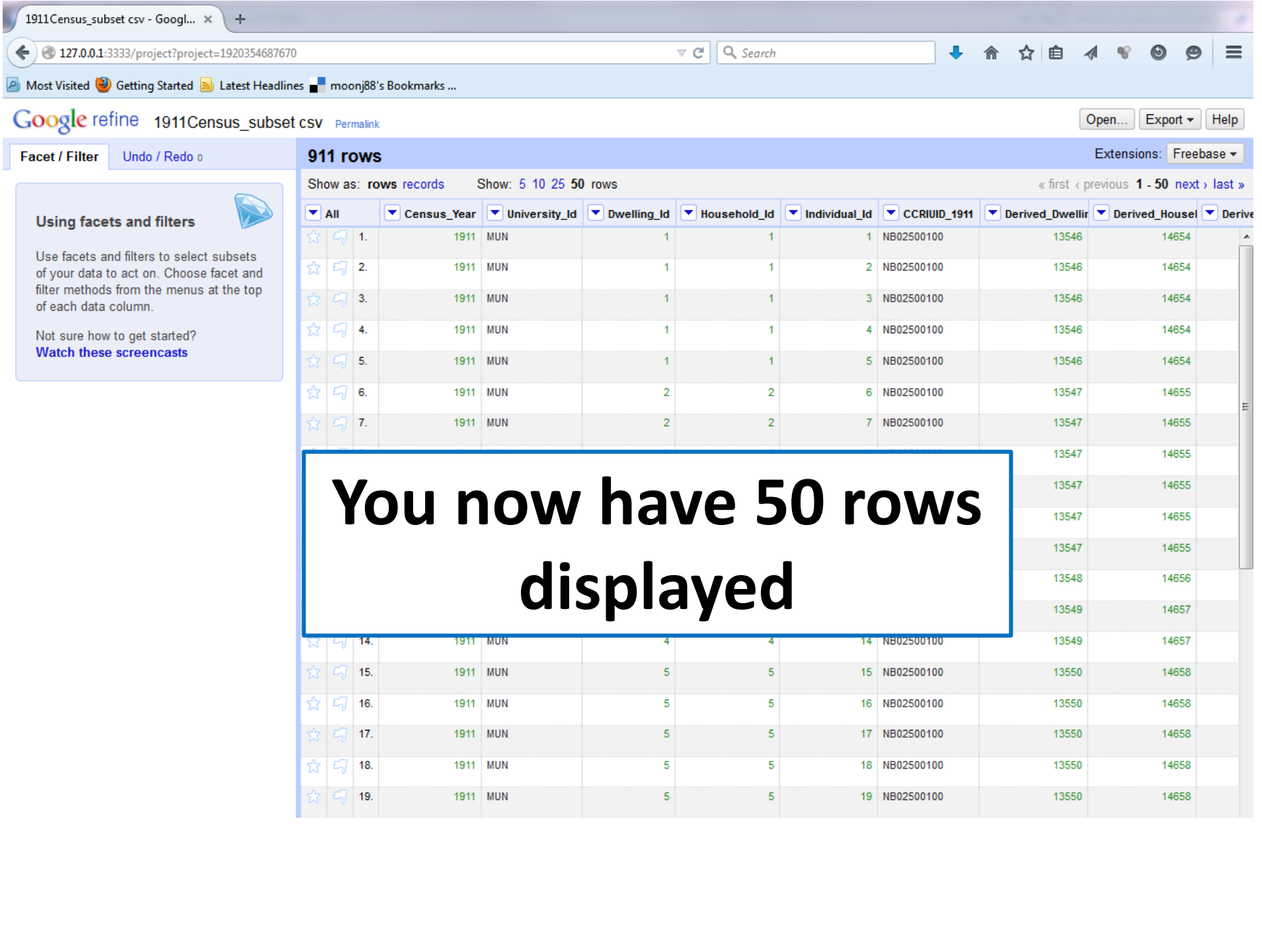

Google Refine
127.0.0.1:3333
Search
Most Visited
Getting Started
Latest Headlines
moonj88's Bookmarks ...
Google refine
A power tool for working with messy data.
Create Project
Open Project
Import Project
« Start Over
Configure Parsing Options
Project name
1911Census_subset csv
Create Project »
Census_Year
University_Id
Dwelling_Id
Household_Id
Individual_Id
CCRIUID_1911
Derived_Dwelling_Id
Derived_Household_Id
Derived_Individual_Id
Derived_Household_Id_In_Dw
1. 1911 MUN 1 1 1 NB02500100 13546 14654 65831
2. 1911 MUN 1 1 2 NB02500100 13546 14654 65832
3. 1911 MUN 1 1 3 NB02500100 13546 14654
4. 1911 MUN 1 1 4 NB02500100 13546 14654
5. 1911 MUN 1
6. 1911 MUN 2
7. 1911 MUN 2
8. 1911 MUN 2 2 8 NB02500100 13547 14655 65838
9. 1911 MUN 2 2 9 NB02500100 13547 14655 65839
10. 1911 MUN 2 2 10 NB02500100 13547 14655 65840
11. 1911 MUN 2 2 11 NB02500100 13547 14655 65841
12. 1911 MUN 3 3 12 NB02500100 13548 14656 65842
13. 1911 MUN 4 4 13 NB02500100 13549 14657 65843
14. 1911 MUN 4 4 14 NB02500100 13549 14657 65844
Click ‘Create Project’
Parse data as
CSV / TSV / separator-based files
Line-based text files
Fixed-width field text files
PC-Axis text files
JSON files
RDF/N3 files
XML files
Open Document Format spreadsheets (.ods)
RDF/XML files
Character encoding
Update Preview
Columns are separated by
commas (CSV)
tabs (TSV)
custom ,
Escape special characters with \
Ignore first 0 line(s) at beginning of file
Parse next 1 line(s) as column headers
Discard initial 0 row(s) of data
Load at most 0 row(s) of data
Parse cell text into numbers, dates, ...
Quotation marks are used to enclose cells containing column separators
Store blank rows
Store blank cells as nulls
Store file source (file names, URLs) in each row
Version 2.5 [r2407]
Help
About



Your data is now open as a ‘Project’ in OpenRefine

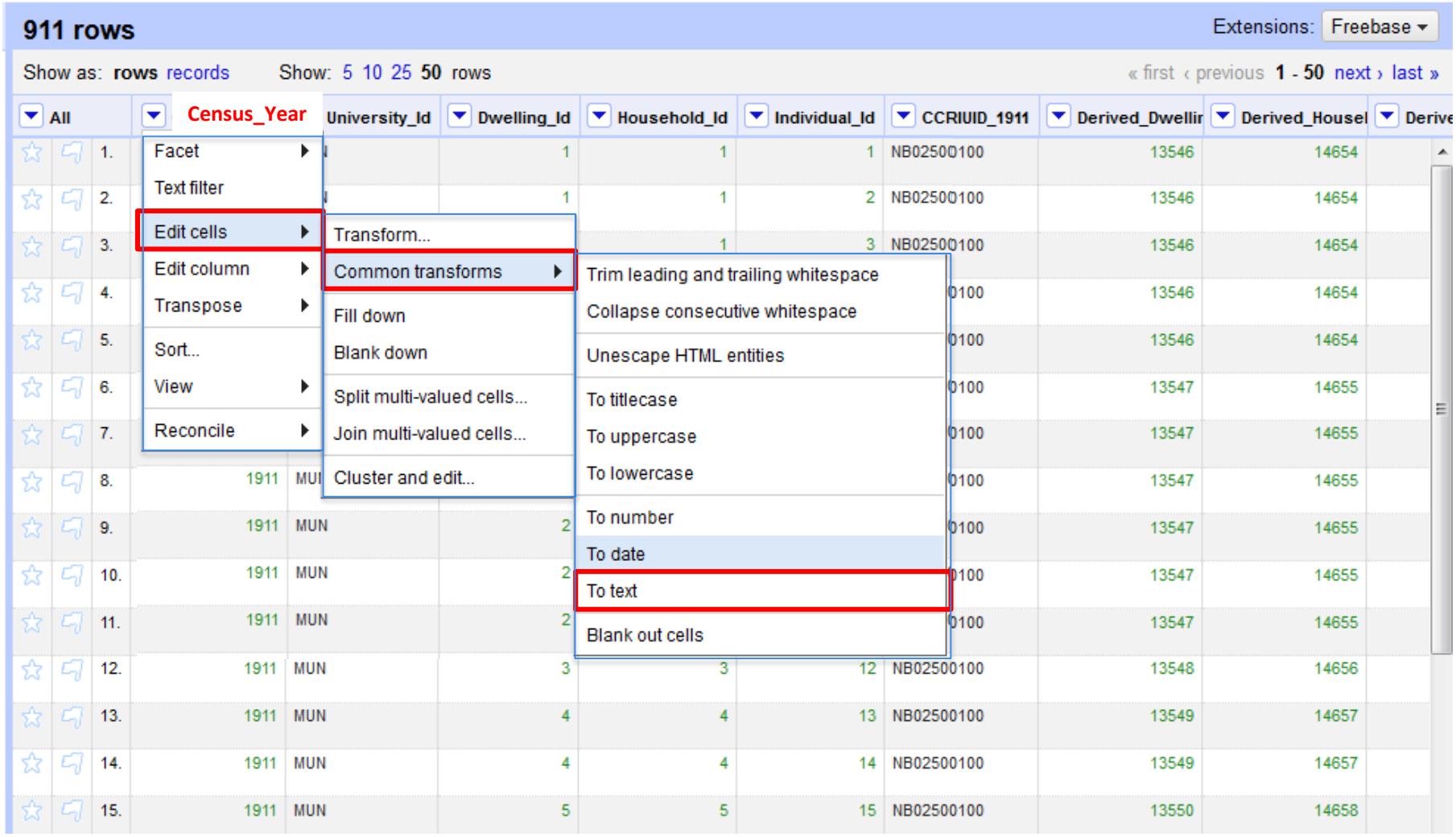


You can specify how many rows to show


1911Census_subset csv - Googl...
127.0.0.1:3333/project?project=1920354687670
Search
Most Visited
Getting Started
Latest Headlines
moonj88's Bookmarks ...
Google refine
1911Census_subset csv
Permalink
Open...
Export
Help
Facet / Filter
Undo / Redo 0
Using facets and filters
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.
Not sure how to get started?
Watch these screencasts
911 rows
Extensions:
Freebase
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 - 50 next › last »
All
Census_Year
University_Id
Dwelling_Id
Household_Id
Individual_Id
CCRIUID_1911
Derived_Dwellir
Derived_Housel
Derive
1. 1911 MUN 1 1 1 NB02500100 13546 14654
2. 1911 MUN 1 1 2 NB02500100 13546 14654
3. 1911 MUN 1 1 3 NB02500100 13546 14654
4. 1911 MUN 1 1 4 NB02500100 13546 14654
5. 1911 MUN 1 1 5 NB02500100 13546 14654
6. 1911 MUN 2 2 6 NB02500100 13547 14655
7. 1911 MUN 2 2 7 NB02500100 13547 14655
13547 14655
13547 14655
13547 14655
13547 14655
13548 14656
13549 14657
14. 1911 MUN 4 4 14 NB02500100 13549 14657
15. 1911 MUN 5 5 15 NB02500100 13550 14658
16. 1911 MUN 5 5 16 NB02500100 13550 14658
17. 1911 MUN 5 5 17 NB02500100 13550 14658
18. 1911 MUN 5 5 18 NB02500100 13550 14658
19. 1911 MUN 5 5 19 NB02500100 13550 14658
You now have 50 rows displayed

Exercise 1. Transform data

Census Year: Edit cells → Common transforms → To text



Exercise 1. Transform data

Census Year: Edit cells → Common transforms → To text

911 rows

Extensions: Freebase

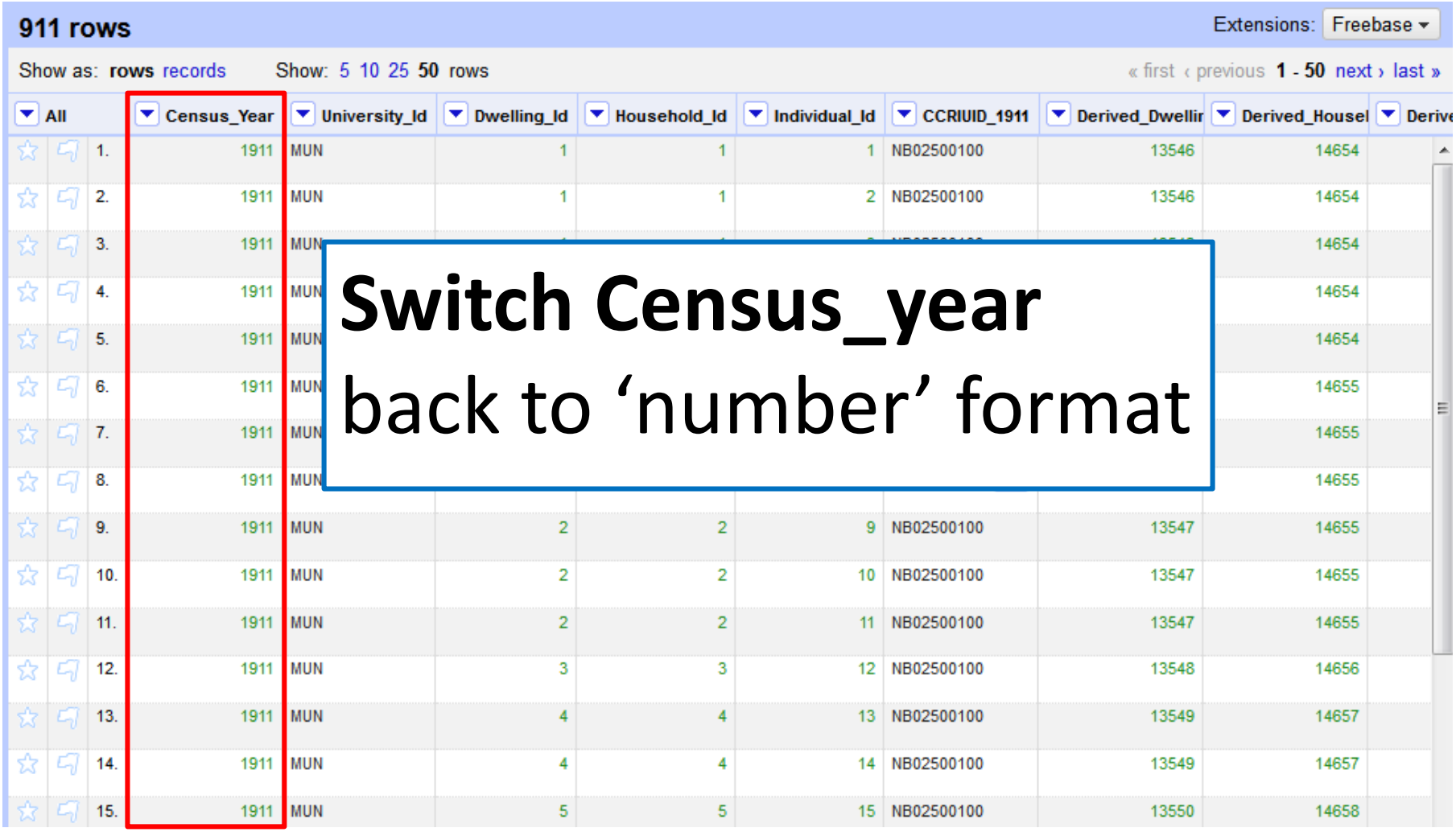
Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 50 next › last »

All		Census_Year	University_Id	Dwelling_Id	Household_Id	Individual_Id	CCRIUID_1911	Derived_Dwellir	Derived_Housel	Derive
1.	1911	MUN	1	1	1	NB02500100	13546	14654		
2.	1911	MUN	1	1	2	NB02500100	13546	14654		
3.	1911	MUN					13546	14654		
4.	1911	MUN					13546	14654		
5.	1911	MUN					13546	14654		
6.	1911	MUN					13547	14655		
7.	1911	MUN					13547	14655		
8.	1911	MUN	2	2	8	NB02500100	13547	14655		
9.	1911	MUN	2	edit 2	9	NB02500100	13547	14655		
10.	1911	MUN	2	2	10	NB02500100	13547	14655		
11.	1911	MUN	2	2	11	NB02500100	13547	14655		
12.	1911	MUN	3	3	12	NB02500100	13548	14656		
13.	1911	MUN	4	4	13	NB02500100	13549	14657		
14.	1911	MUN	4	4	14	NB02500100	13549	14657		
15.	1911	MUN	5	5	15	NB02500100	13550	14658		

Census_year is now in ‘Text’ format

Exercise 2. Transform back to number

Census_Year: Edit cells → Common transforms → To number



911 rows

Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 50 next › last »

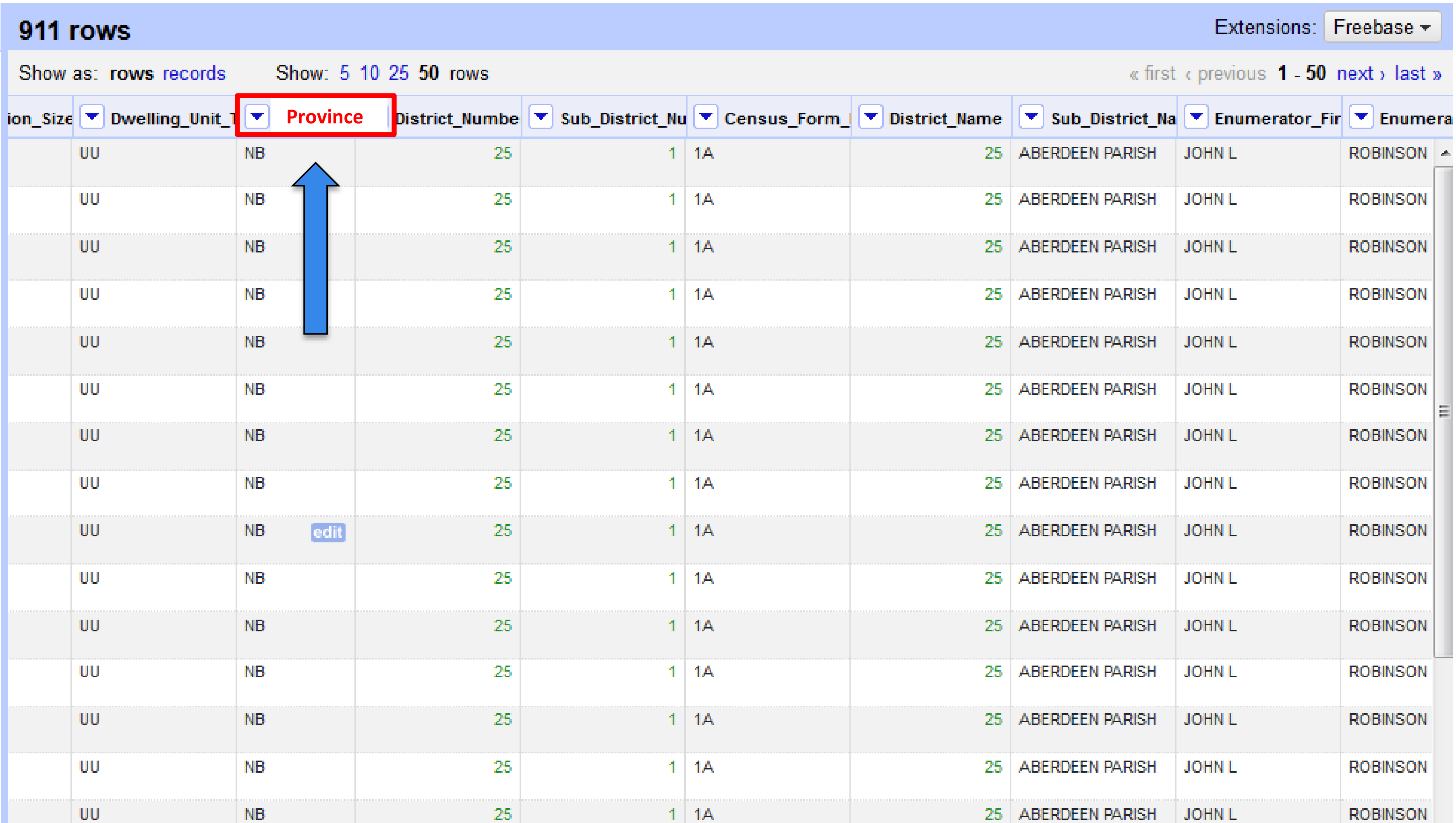
	All	Census_Year	University_Id	Dwelling_Id	Household_Id	Individual_Id	CCRIUID_1911	Derived_Dwellir	Derived_Housel	Derive
1.		1911	MUN	1	1	1	NB02500100	13546	14654	
2.		1911	MUN	1	1	2	NB02500100	13546	14654	
3.		1911	MUN						14654	
4.		1911	MUN						14654	
5.		1911	MUN						14654	
6.		1911	MUN						14655	
7.		1911	MUN						14655	
8.		1911	MUN						14655	
9.		1911	MUN	2	2	9	NB02500100	13547	14655	
10.		1911	MUN	2	2	10	NB02500100	13547	14655	
11.		1911	MUN	2	2	11	NB02500100	13547	14655	
12.		1911	MUN	3	3	12	NB02500100	13548	14656	
13.		1911	MUN	4	4	13	NB02500100	13549	14657	
14.		1911	MUN	4	4	14	NB02500100	13549	14657	
15.		1911	MUN	5	5	15	NB02500100	13550	14658	

Switch Census_year

back to ‘number’ format

Exercise 3. Text facet

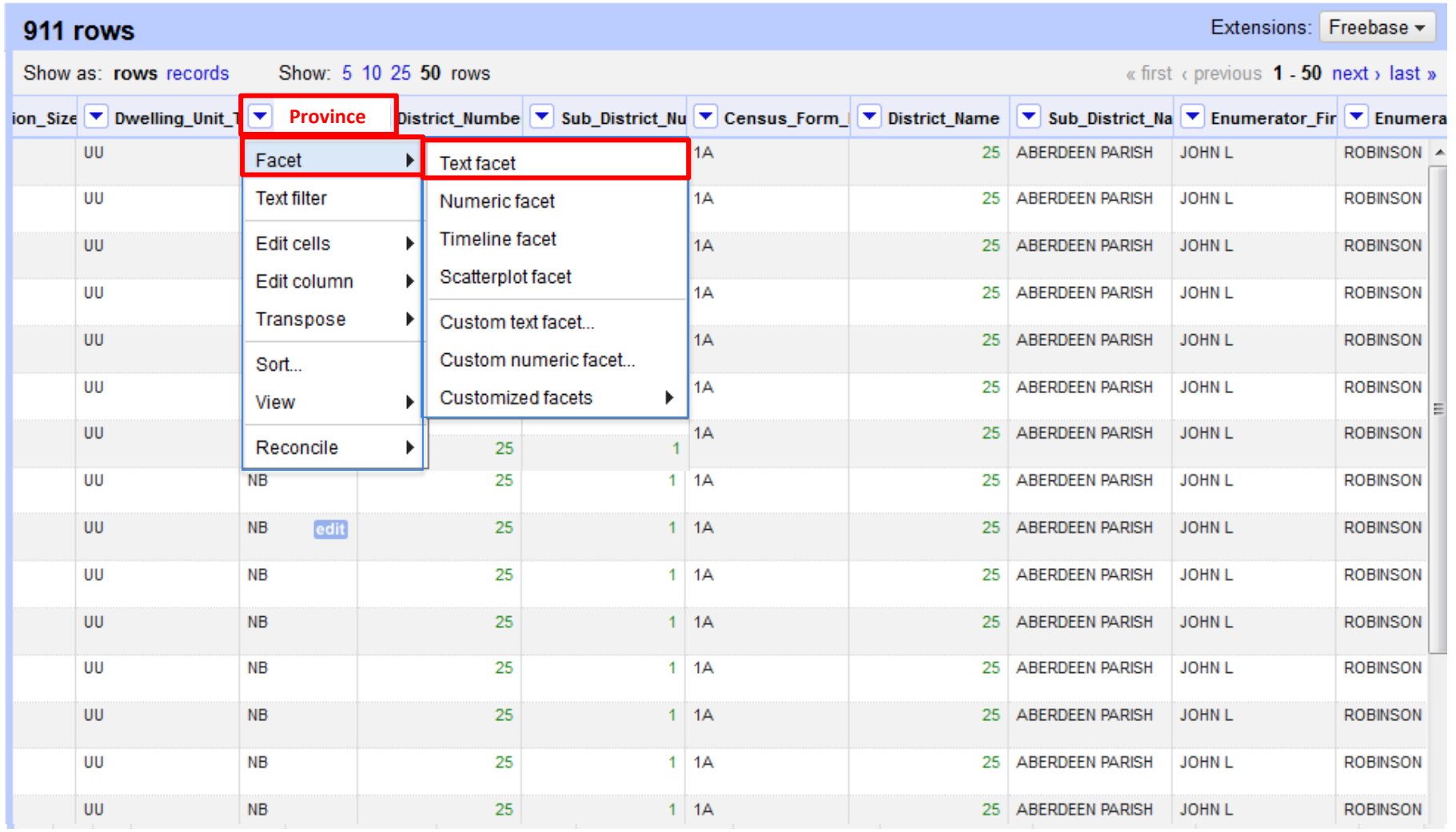
First, SCROLL to Province column



ion_Size	Dwelling_Unit_T	Province	District_Numbe	Sub_District_Nu	Census_Form_	District_Name	Sub_District_Na	Enumerator_Fir	Enumera
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB edit	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON
	UU	NB	25	1	1A	25	ABERDEEN PARISH	JOHN L	ROBINSON

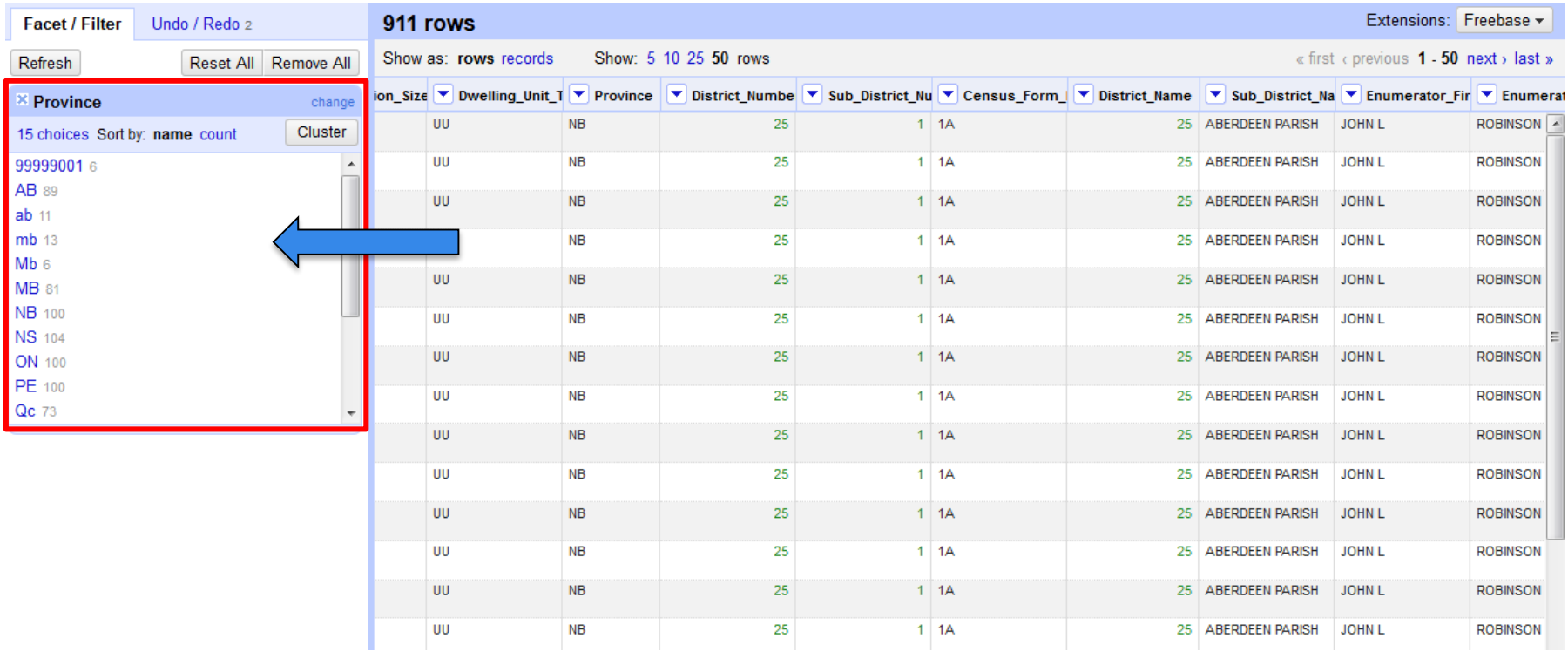
Exercise 3. Text facet

Province: Facet → Text facet



Exercise 3. Facets

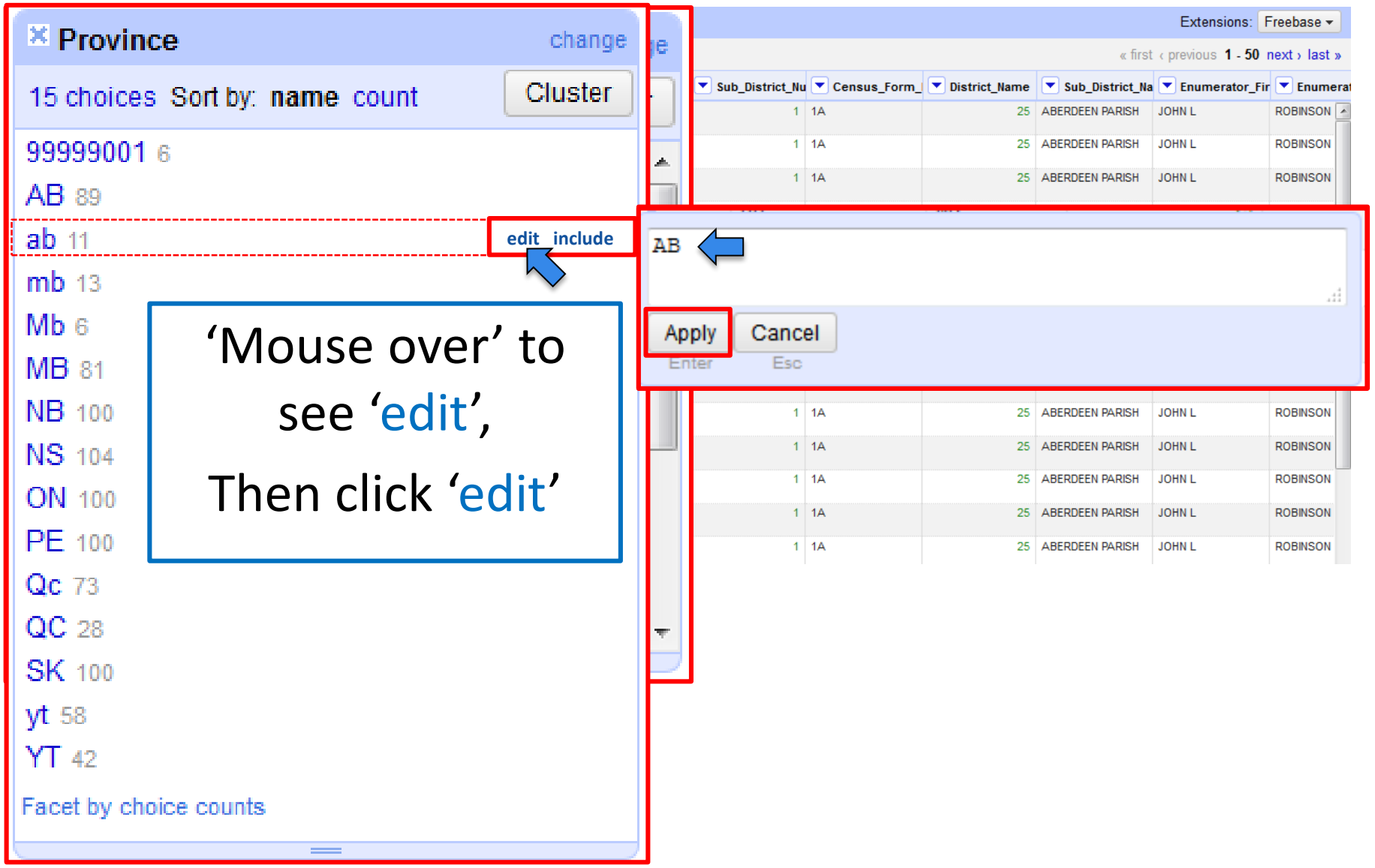
Facets appear at the left of the screen



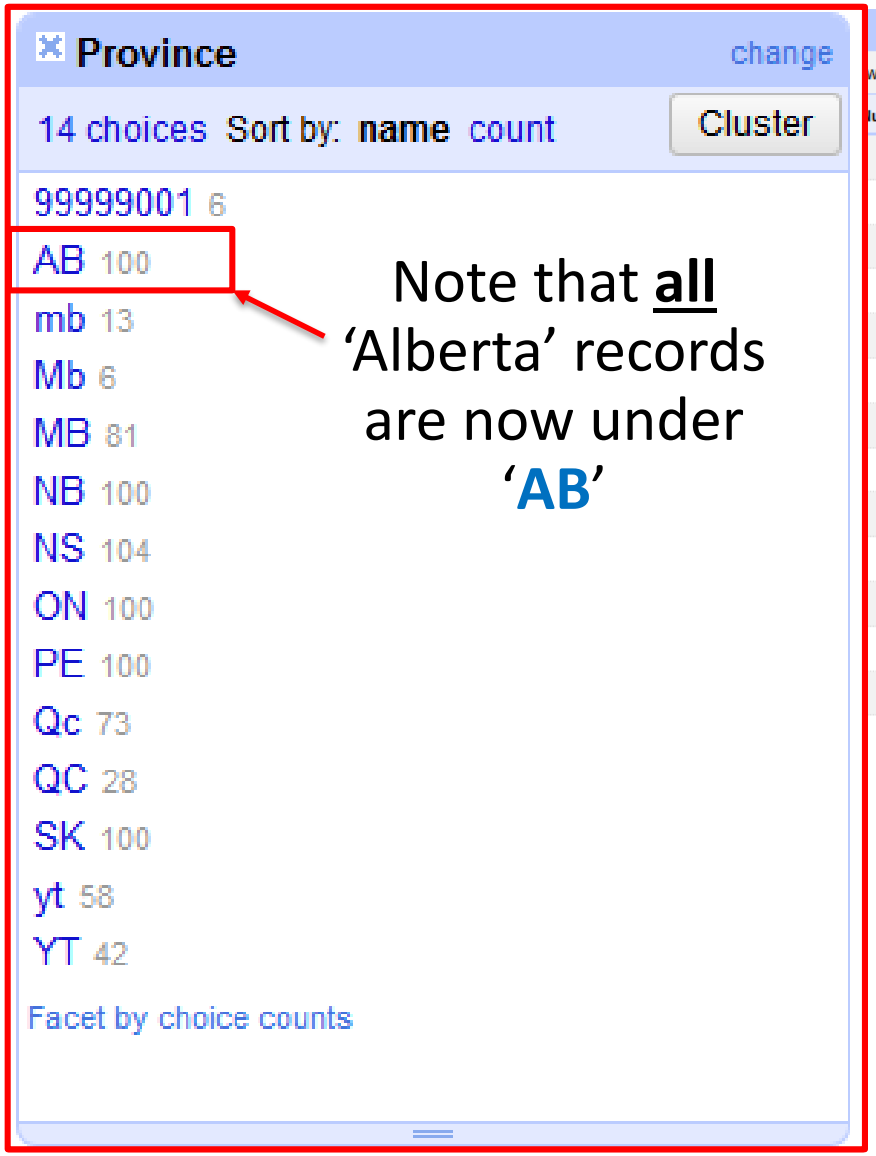
Exercise 3. Facets

Note frequencies for each value in the column…

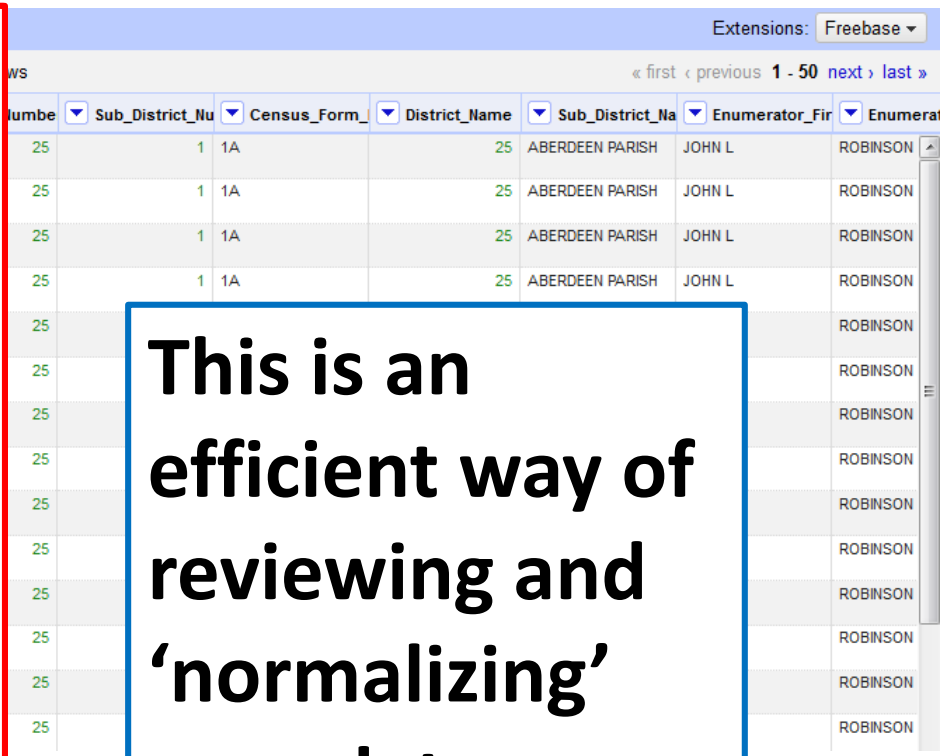
Exercise 4. EDIT entry for Alberta [ab]



Exercise 4. EDIT entry for Alberta [ab]



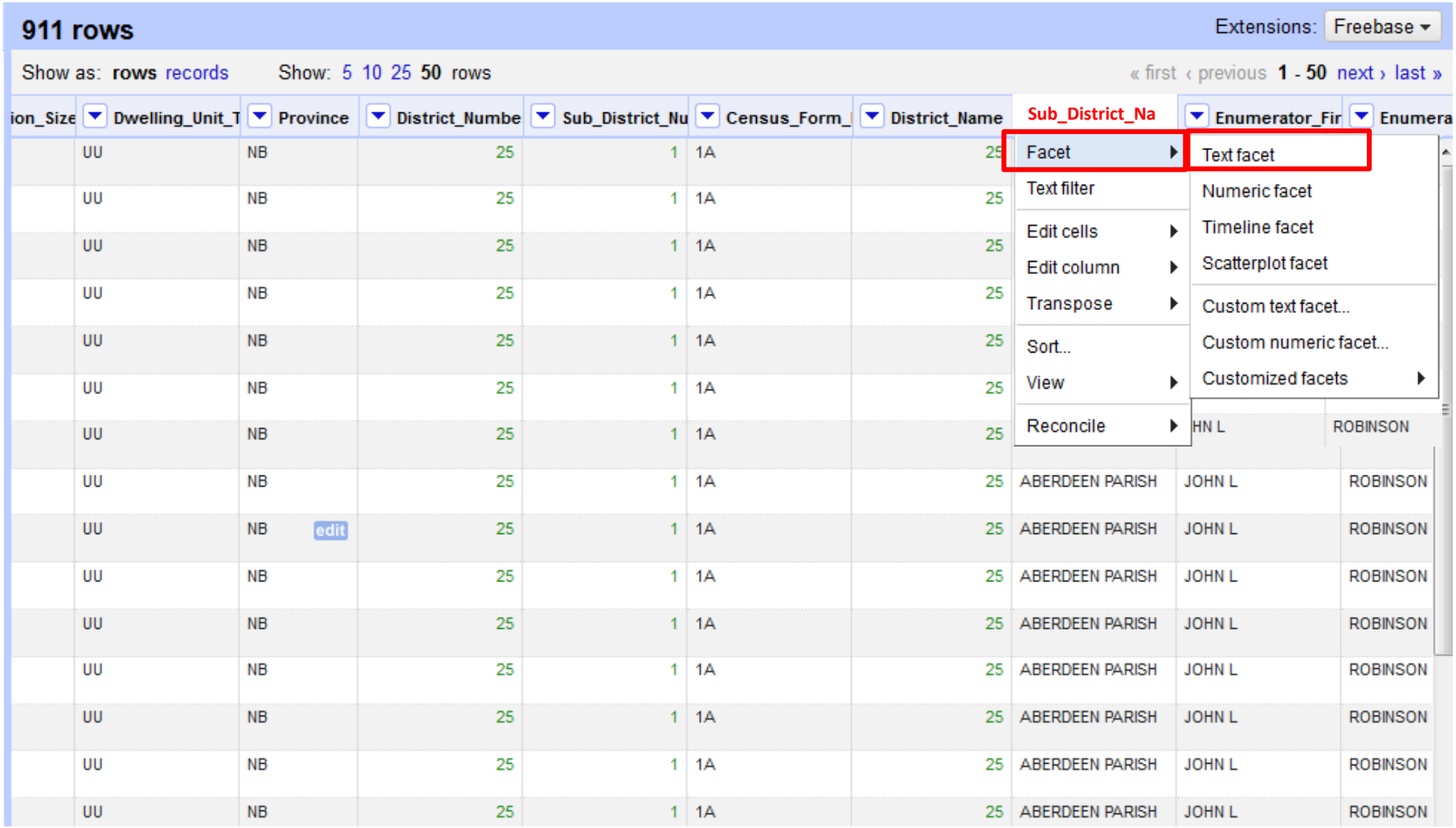
Note that **all** 'Alberta' records are now under '**AB**'



This is an efficient way of reviewing and 'normalizing' your data…

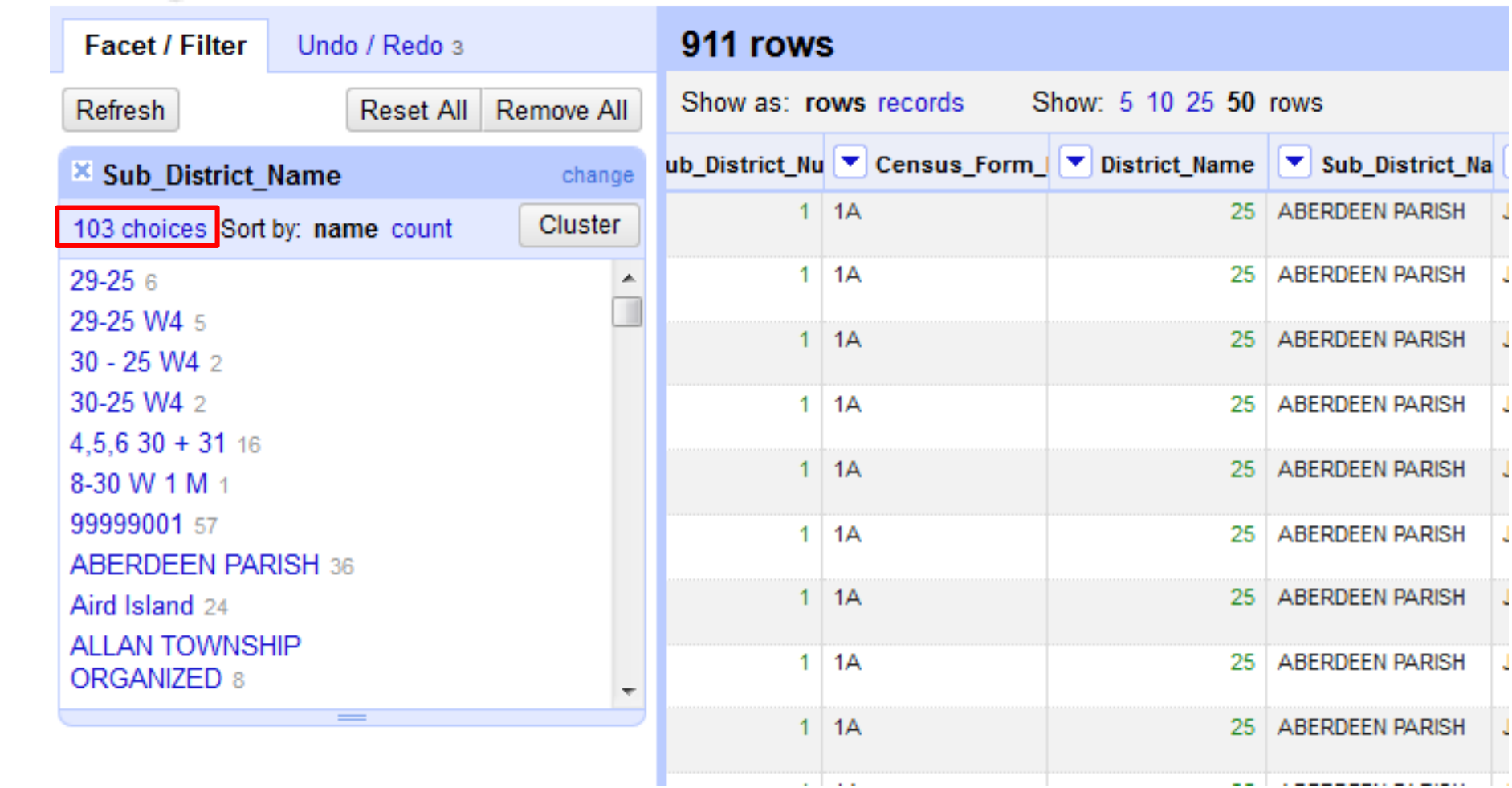
Exercise 5. Text facets & Clustering

Sub_District_Name: Facet → Text facet



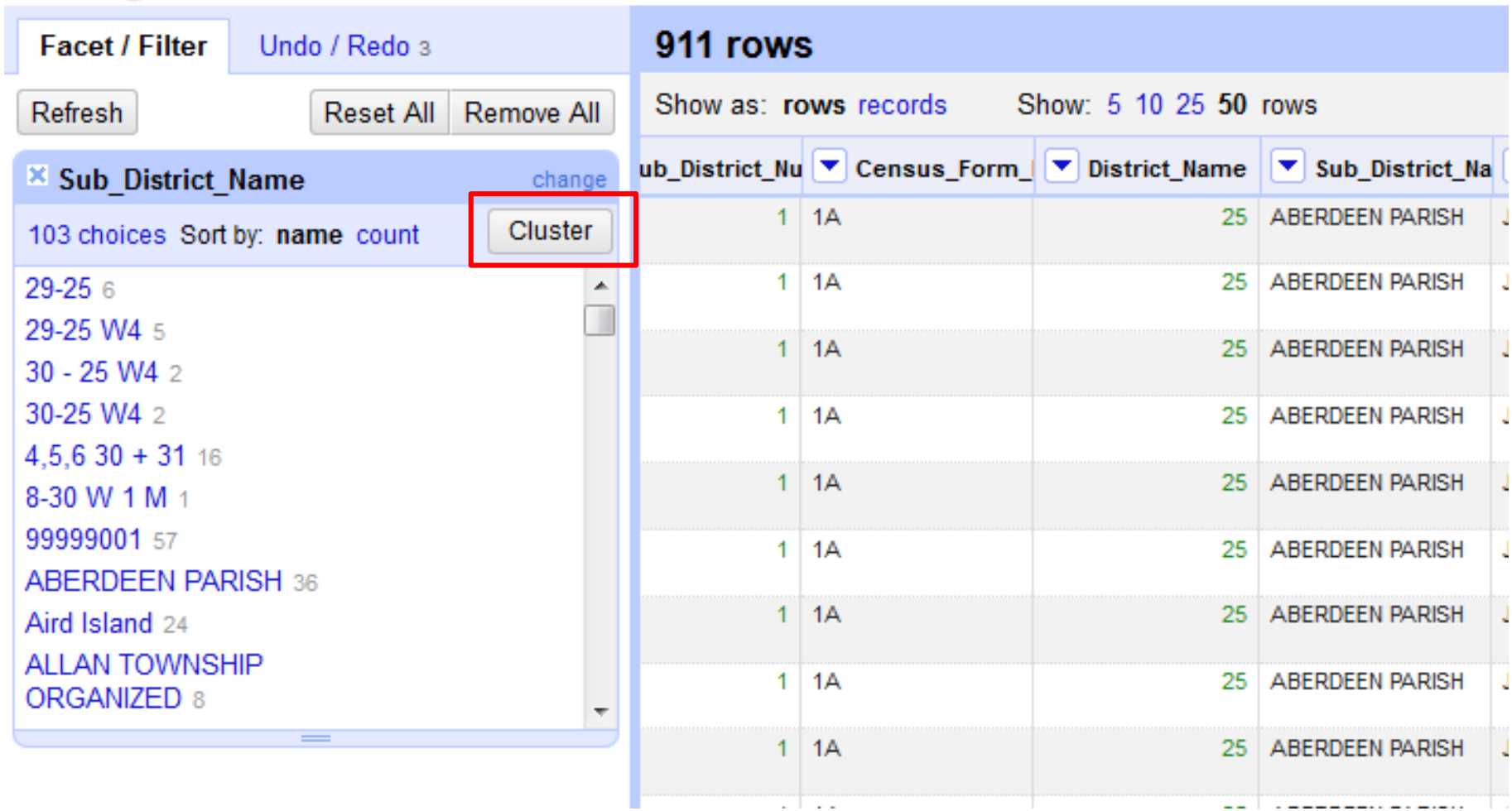
Exercise 5. Text facets & Clustering

Sub_District_Name: Again, facets shown on left side



Exercise 5. Text facets & Clustering

Sub_District_Name: Click on Cluster



Exercise 5. Text facets & Clustering

Note different ‘methods’ and ‘keying

Cluster & Edit column "Sub_District_Name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision Keying Function: fingerprint 6 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	7	• Bonanza Creek, Division no 10b, district no 4 (4 rows) • Bonanza Creek, Division no 10b District no 4 (1 rows) • Bonanza Creek, Division no 10b district 4 (1 rows) • Bonanza Creek, division no 10b District no 4 (1 rows)	☐	Bonanza Creek, Division no
2	3	• Hunker Road Bear Creek (2 rows) • Bear Creek Hunker Road (1 rows)	☐	Hunker Road Bear Creek
2	4	• Township 3 Range 31 West 1st PM (3 rows) • Township 3 Range 31 West 1st P.M. (1 rows)	☐	Township 3 Range 31 West
2	15	• Township 11 & 12 Range 17 (13 rows) • Township 11 +12 Range 17 (2 rows)	☐	Township 11 & 12 Range 17
2	4	• Grand Forks, Bonanza Creek, Division no 10 district no 4 (3 rows) • Grand Forks, Bonanza Creek, Division no 10	☐	Grand Forks, Bonanza Cree

 # Choices in Cluster

2 — 4

 # Rows in Cluster

3 — 15

Average Length of Choices

22 — 56

Length Variance of Choices

0 — 1.5

Select All | Deselect All | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

Exercise 5. Text facets & Clustering

Note suggested ‘clusters’…

Cluster & Edit column "Sub_District_Name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision | Keying Function: fingerprint | 6 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	7	• Bonanza Creek, Division no 10b, district no 4 (4 rows) • Bonanza Creek, Division no 10b District no 4 (1 rows) • Bonanza Creek, Division no 10b district 4 (1 rows) • Bonanza Creek, division no 10b District no 4 (1 rows)	☐	Bonanza Creek, Division no
2	3	• Hunker Road Bear Creek (2 rows) • Bear Creek Hunker Road (1 rows)	☐	Hunker Road Bear Creek
2	4	• Township 3 Range 31 West 1st PM (3 rows) • Township 3 Range 31 West 1st P.M. (1 rows)	☐	Township 3 Range 31 West
2	15	• Township 11 & 12 Range 17 (13 rows) • Township 11 +12 Range 17 (2 rows)	☐	Township 11 & 12 Range 17
2	4	• Grand Forks, Bonanza Creek, Division no 10 district no 4 (3 rows) • Grand Forks, Bonanza Creek, Division no 10	☐	Grand Forks, Bonanza Cree

 # Choices in Cluster

2 — 4

 # Rows in Cluster

3 — 15

Average Length of Choices

22 — 56

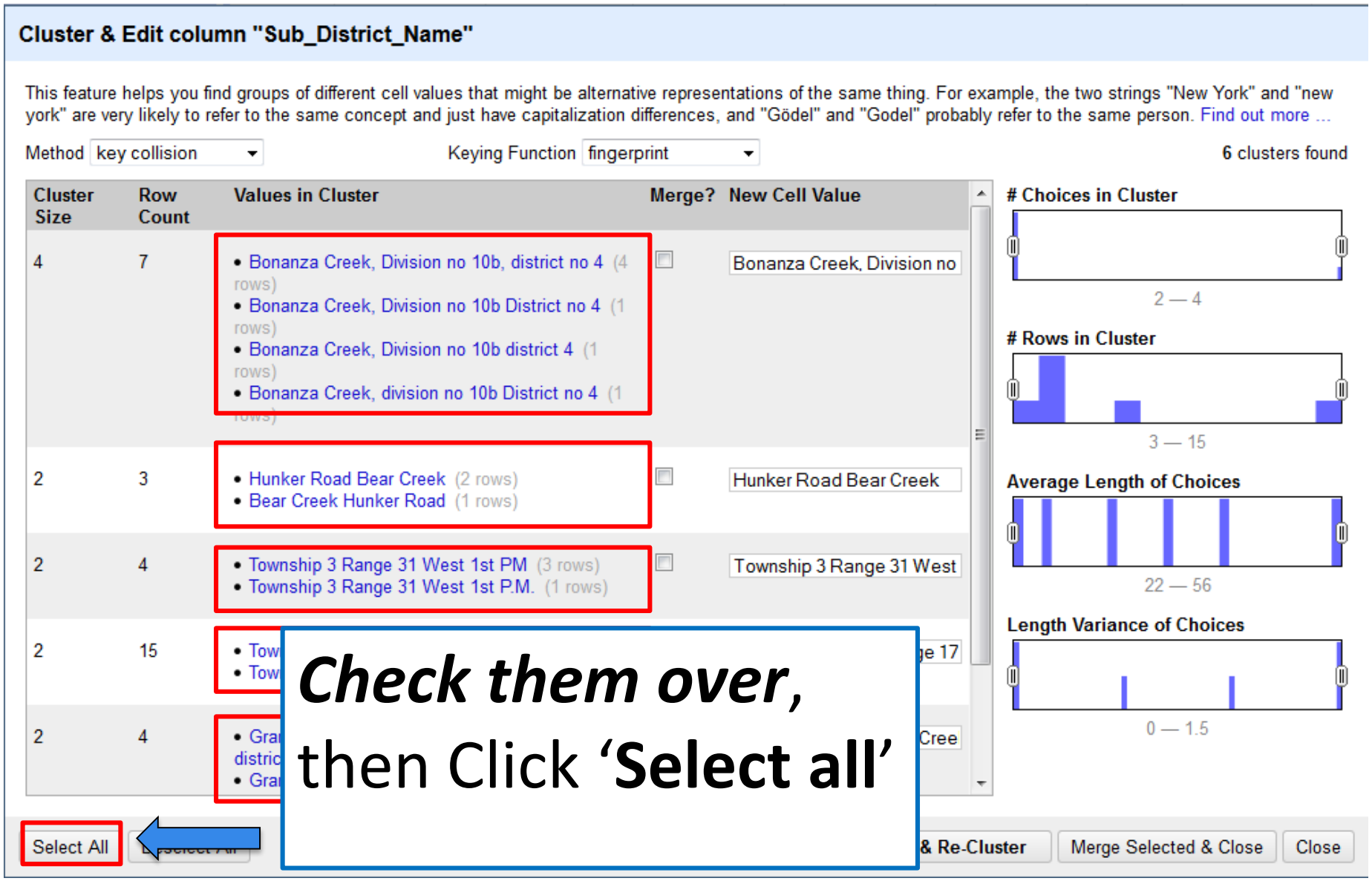
Length Variance of Choices

0 — 1.5

Select All | Deselect All | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

Exercise 5. Text facets & Clustering

Note suggested ‘clusters’…



Check them over, then Click ‘**Select all**’

Exercise 5. Text facets & Clustering

Now we're ready to 'Re-Cluster'

Cluster & Edit column "Sub_District_Name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision | Keying Function: fingerprint | 6 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
4	7	• Bonanza Creek, Division no 10b, district no 4 (4 rows) • Bonanza Creek, Division no 10b District no 4 (1 rows) • Bonanza Creek, Division no 10b district 4 (1 rows) • Bonanza Creek, division no 10b District no 4 (1 rows)	☑	Bonanza Creek, Division no
2				eek
2				West
2	15	• Township 11 & 12 Range 17 (13 rows) • Township 11 +12 Range 17 (2 rows)	☑	Township 11 & 12 Range 17
2	4	• Grand Forks, Bonanza Creek, Division no 10 district no 4 (3 rows) • Grand Forks, Bonanza Creek, Division no 10	☑	Grand Forks, Bonanza Cree

 # Choices in Cluster

2 — 4

 # Rows in Cluster

3 — 15

Average Length of Choices

22 — 56

Length Variance of Choices

0 — 1.5

Click

'Merge Selected & Re-Cluster

Select All | Deselect All | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

Exercise 5. Text facets & Clustering

‘fingerprint’ clustering done!

Cluster & Edit column "Sub_District_Name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method key collision ▾ Keying Function fingerprint ▾

No clusters were found with the selected method

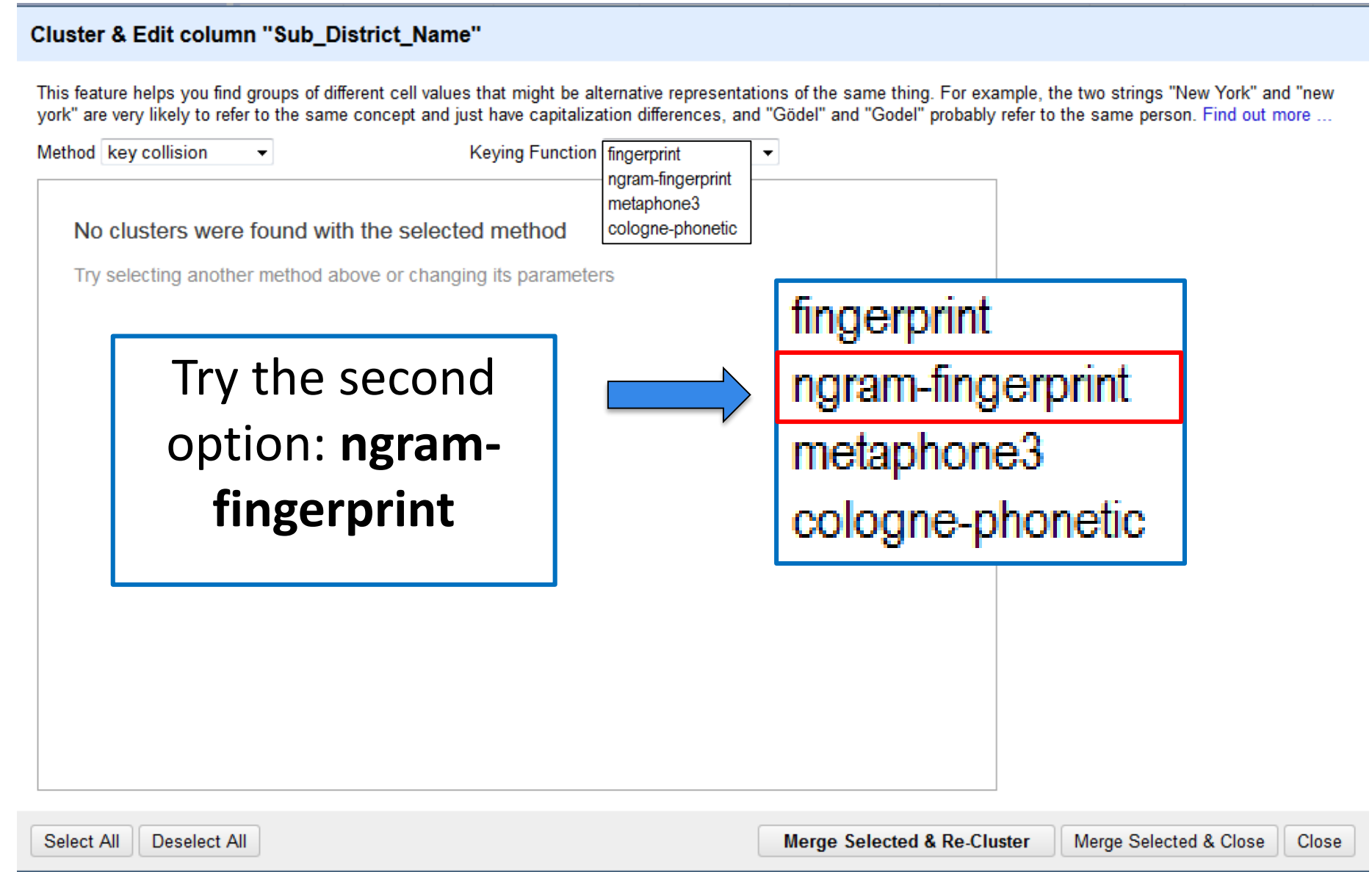
Try selecting another method above or changing its parameters

DONE!

Select All | Deselect All | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

Exercise 5. Text facets & Clustering

Now try the next function, ‘ngram-fingerprint’



Exercise 5. Text facets & Clustering

Results for ‘ngram-fingerprint’

Cluster & Edit column "Sub_District_Name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision | Keying Function: ngram-fingerprint | Ngram Size: 2 | 2 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	36	• Township 11+12 Range 17 (21 rows) • Township 11 & 12 Range 17 (15 rows)	☐	Township 11+12 Range 17
2	4	• 30 - 25 W4 (2 rows) • 30-25 W4 (2 rows)	☐	30 - 25 W4

 # Rows in Cluster

4 — 36

Average Length of Choices

9 — 24

Review,
Select All,
then Click
Merge Selected & Re-Cluster

Two more ‘clusters’

Select All | Deselect All | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

Exercise 5. Text facets & Clustering

‘ngram-fingerprint’ clustering done!

Cluster & Edit column "Sub_District_Name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method key collision Keying Function ngram-fingerprint

No clusters were found with the selected method

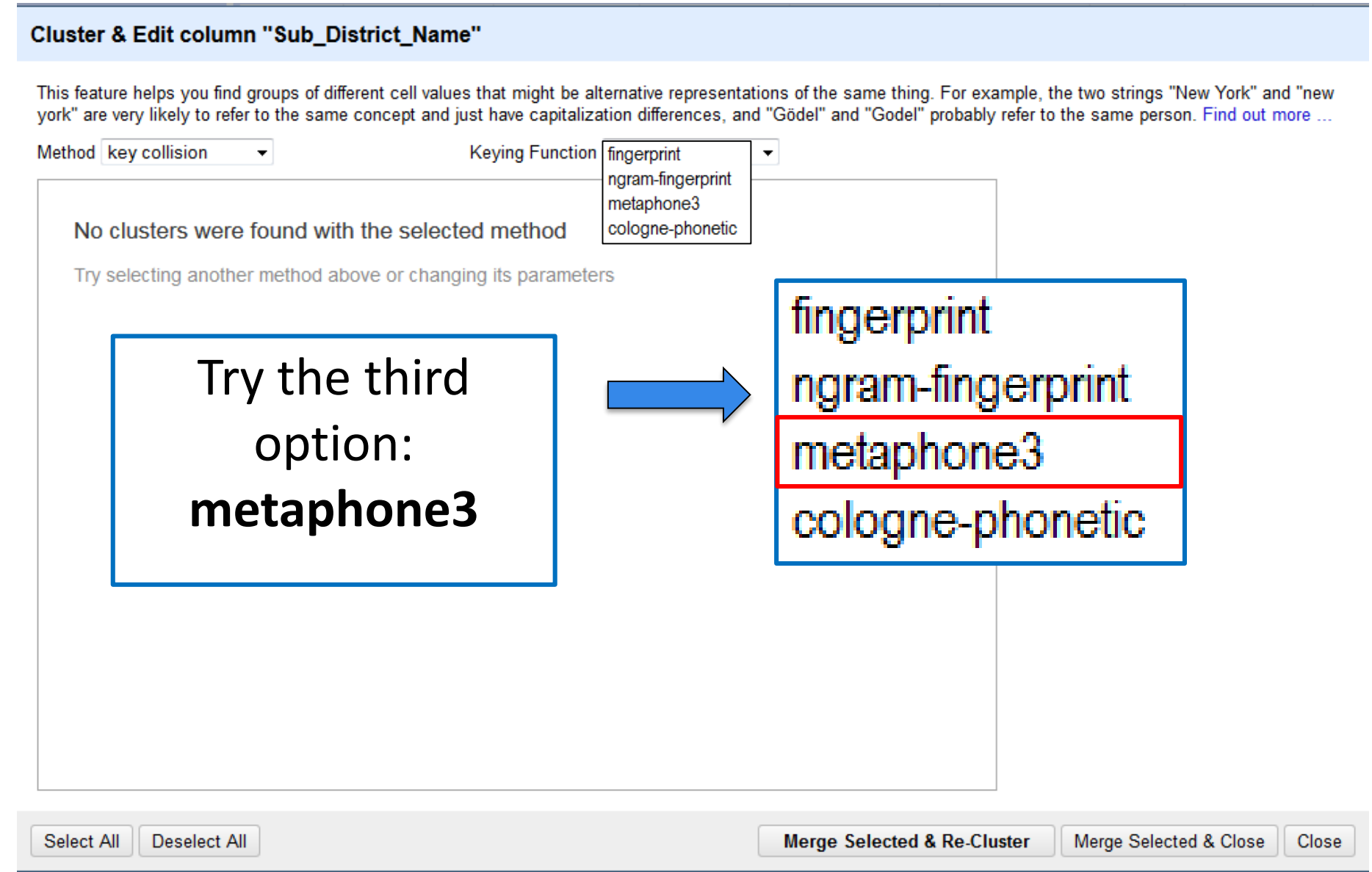
Try selecting another method above or changing its parameters

DONE!

Select All | Deselect All | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

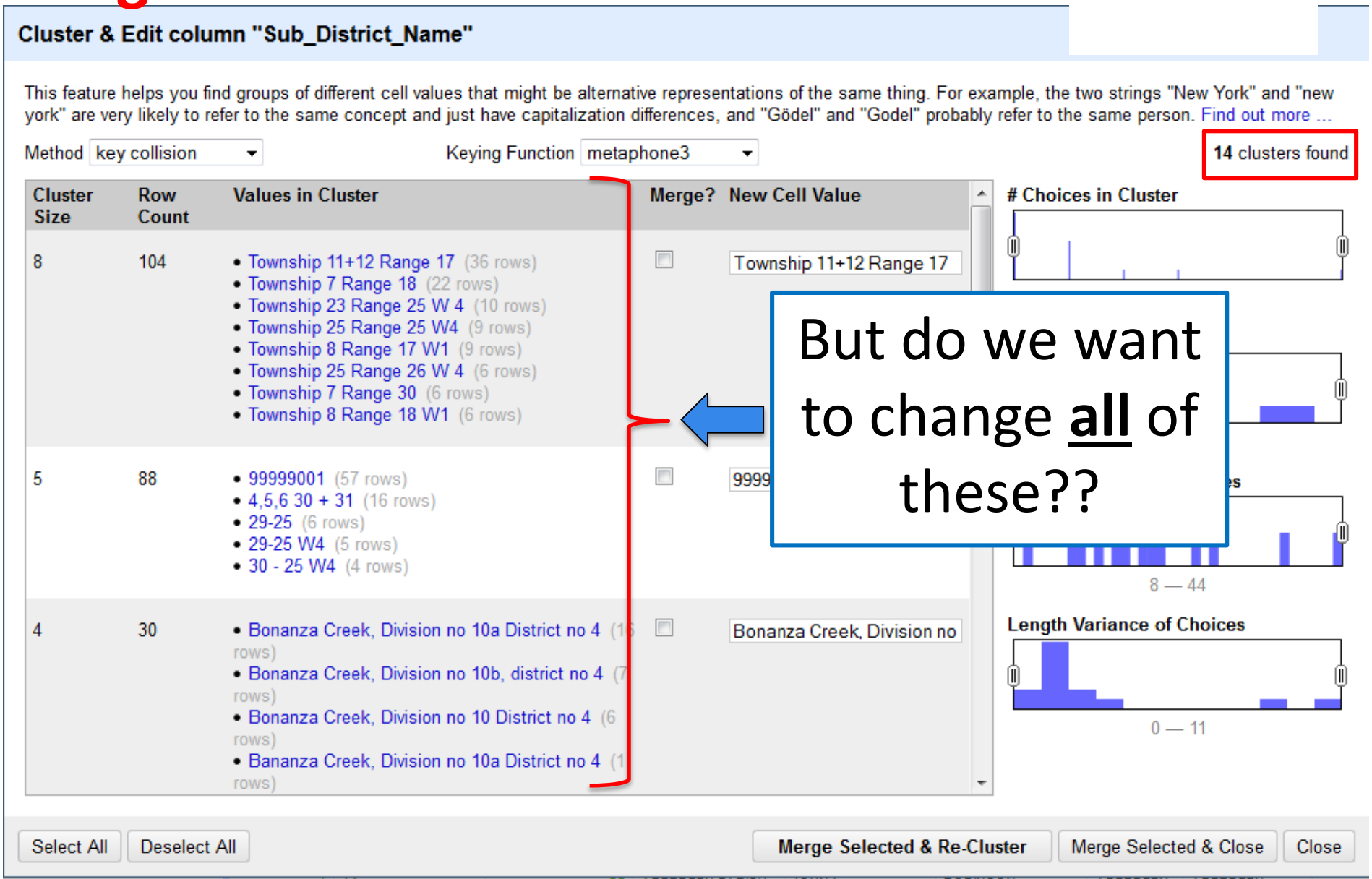
Exercise 5. Text facets & Clustering

Now try the next keying function, ‘metaphone3’



Exercise 5. Text facets & Clustering

We get a lot of new clusters…



Exercise 5. Text facets & Clustering

We get a lot of new clusters… scroll through the list

Cluster & Edit column "Sub_District_Name"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision | Keying Function: metaphone3 | **14** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
8	104	• Township 11+12 Range 17 (36 rows) • Township 7 Range 18 (22 rows) • Township 23 Range 25 W 4 (10 rows) • Township 25 Range 25 W4 (9 rows) • Township 8 Range 17 W1 (9 rows) • Township 25 Range 26 W 4 (6 rows) • Township 7 Range 30 (6 rows) • Township 8 Range 18 W1 (6 rows)	☐	Township 11+12 Range 17
5	88	• 99999001 (57 rows) • 4,5,6 30 + 31 (16 rows) • 29-25 (6 rows) • 29-25 W4 (5 rows) • 30 - 25 W4 (4 rows)	☐	99999001

 # Choices in Cluster

2 — 8

Are these really the same?

Length Variance of Choices

0 — 11

Go through the list and decide which clusters to include (or not), then **Merge Selected & Re-Cluster**

Select All | Deselect All | **Merge Selected & Re-Cluster** | Merge Selected & Close | Close

Exercise 5. Text facets & Clustering

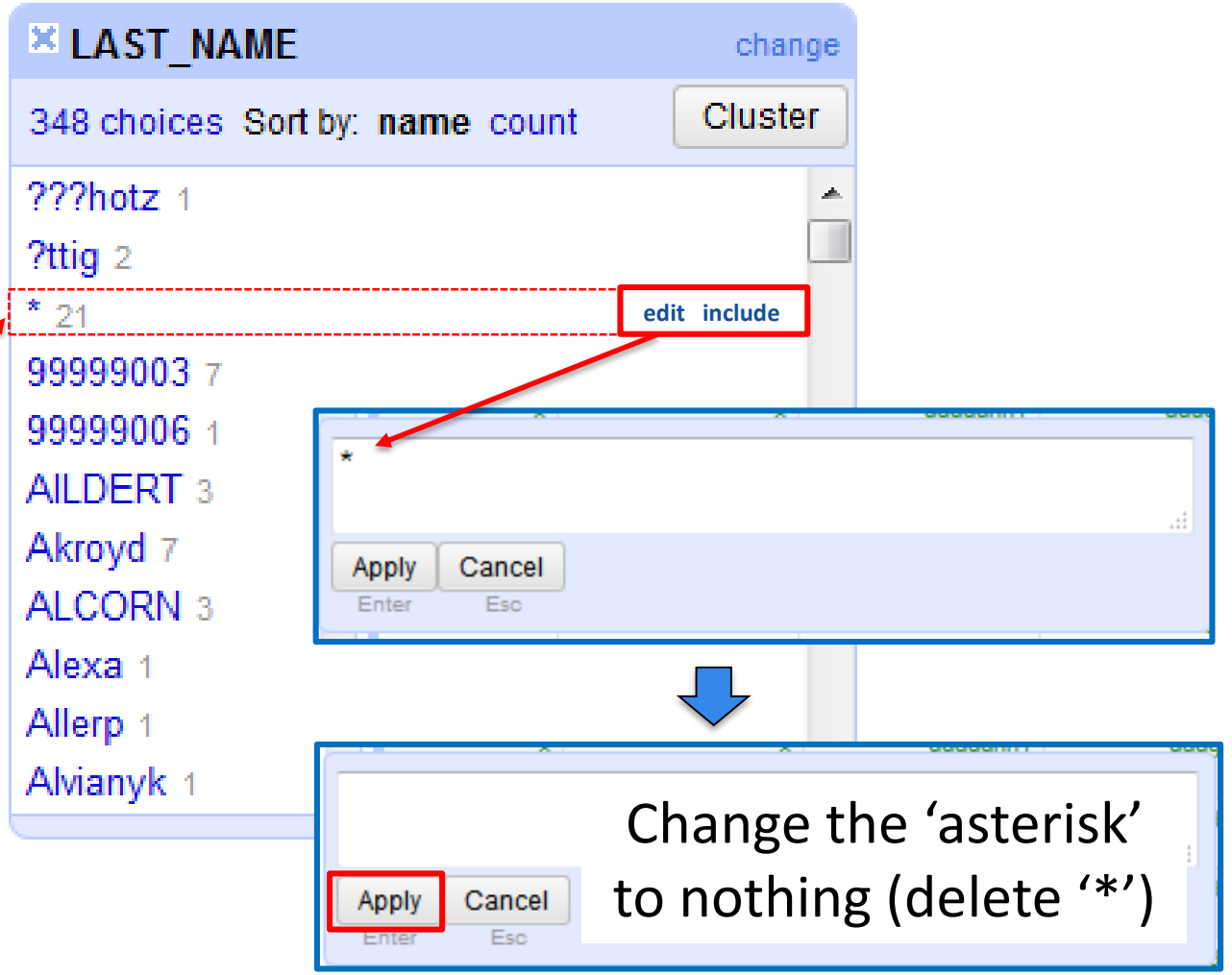
Clustering… a powerful clean-up tool

You can use, and re-use clustering techniques until your data is ‘clean’

Exercise 6. Editing data in columns

LAST_NAME: Facet → Text facet

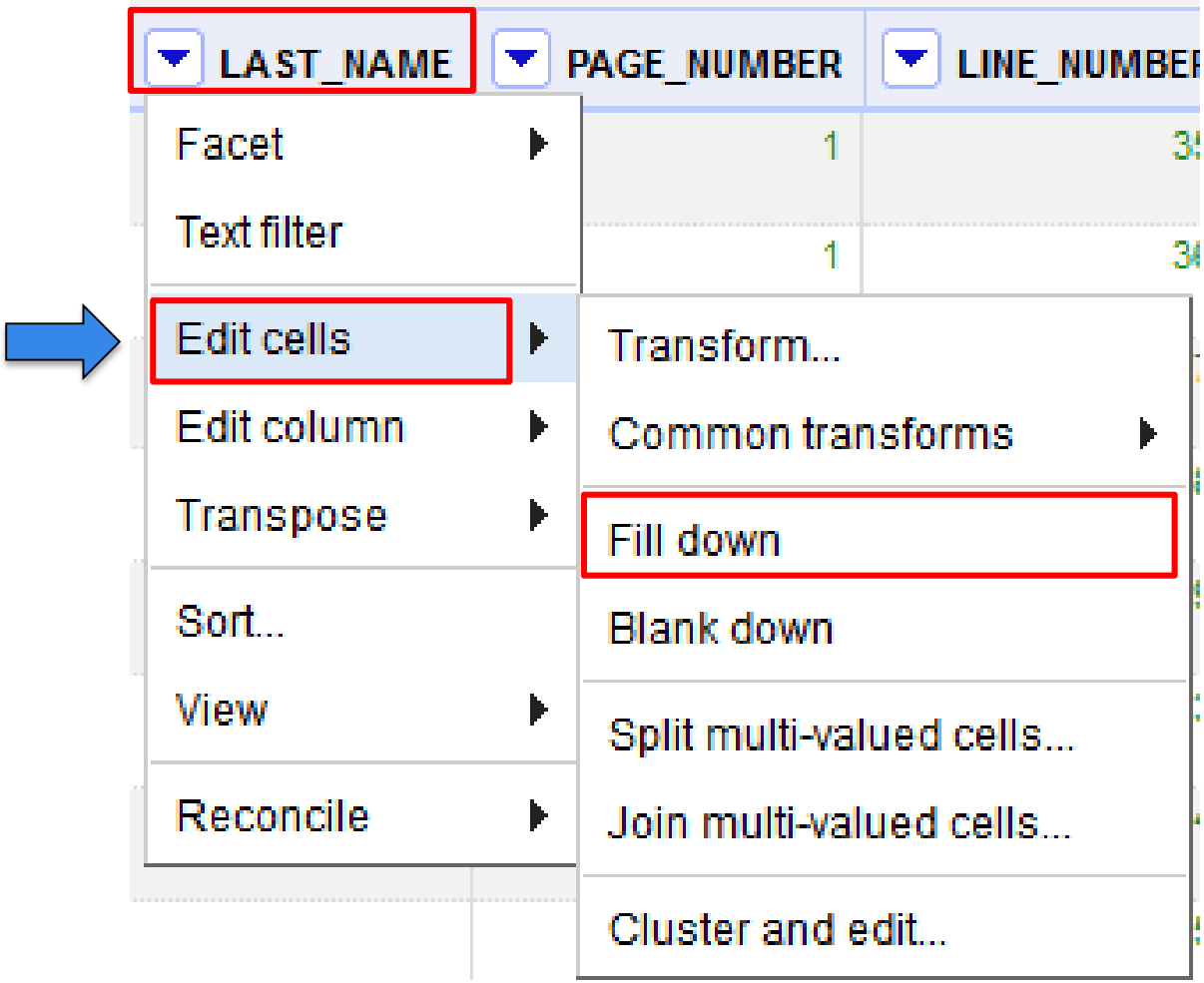
FIRST_NAME	LAST_NAME
RICHARD	BROMLEY
LIZZIE	BROMLEY
ARTHUR	BROMLEY
BARBARA	NIXON
FRANK	NIXON
RALPH	GREGORY
CECLIA	*
RALPH	*
RAY	*
R?LLIA	*
ANNIE	FREMAN
GUY	JONES
ANNIE	SNIDOR
TUVA E	*



Exercise 6. Editing data in columns

LAST_NAME: Edit cells → Fill down

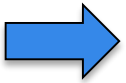
FIRST_NAME	LAST_NAME
RICHARD	BROMLEY
LIZZIE	BROMLEY
ARTHUR	BROMLEY
BARBARA	NIXON
FRANK	NIXON
RALPH	GREGORY
CECLIA	
RALPH	
RAY	
R?LLIA	
ANNIE	FREMAN
GUY	JONES
ANNIE	SNIDOR
TUVA E	



Exercise 6. Editing data in columns

LAST_NAME: Edit cells → Fill down

FIRST_NAME	LAST_NAME
RICHARD	BROMLEY
LIZZIE	BROMLEY
ARTHUR	BROMLEY
BARBARA	NIXON
FRANK	NIXON
RALPH	GREGORY
CECLIA	
RALPH	
RAY	
R?LLIA	
ANNIE	FREMAN
GUY	JONES
ANNIE	SNIDOR
TUVA E	



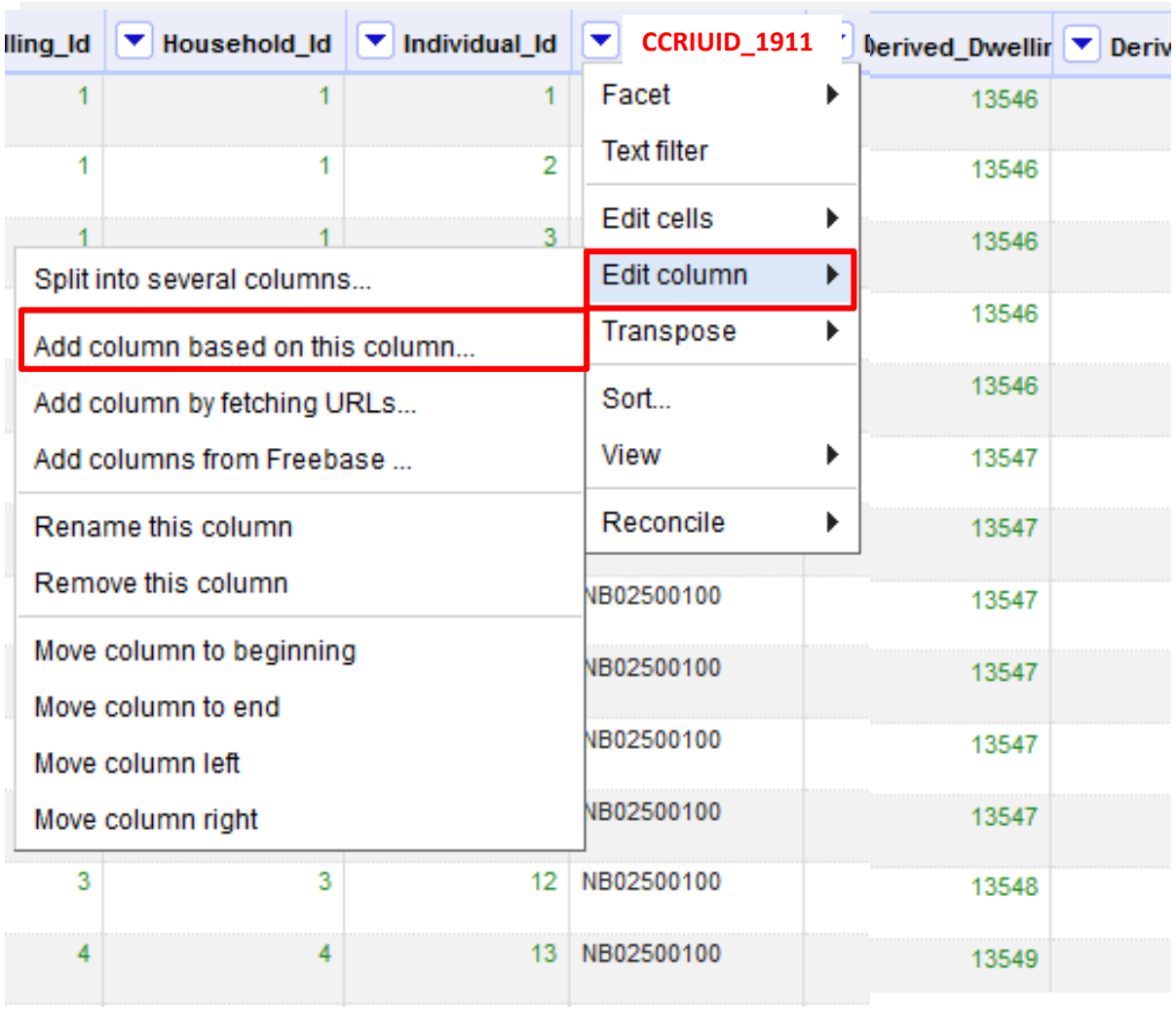
FIRST_NAME	LAST_NAME
RICHARD	BROMLEY
LIZZIE	BROMLEY
ARTHUR	BROMLEY
BARBARA	NIXON
FRANK	NIXON
RALPH	GREGORY
CECLIA	GREGORY
RALPH	GREGORY
RAY	GREGORY
R?LLIA	GREGORY
ANNIE	FREMAN
GUY	JONES
ANNIE	SNIDOR
TUVA E	SNIDOR

Not so easy to do in Excel…

Exercise 7. Creating new columns

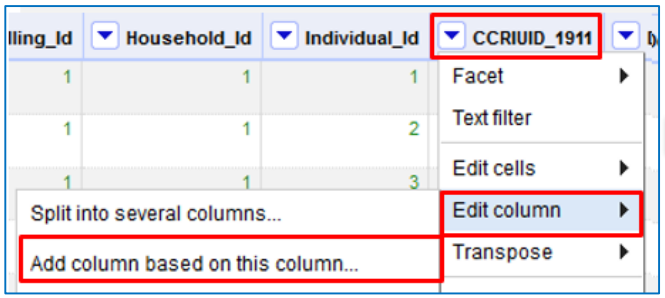
CCRIUID_1911: Edit column

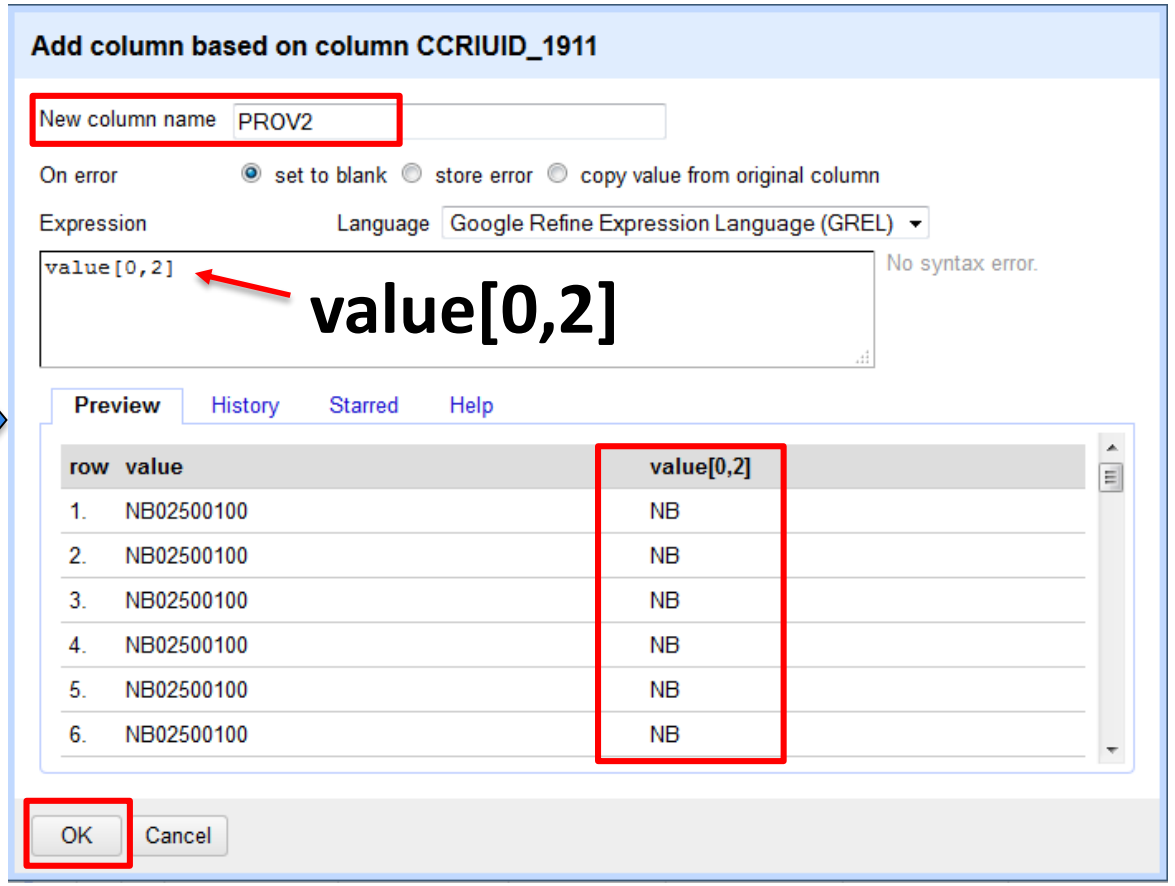
→ Add column based on this column



Exercise 7. Creating new columns

CCRIUID_1911: Define values for new column using ‘regular expression’





Regular Expressions

A very powerful tool!

Create new column PROV2 based on column CCRIUID_1911 by filling 911 rows with grel:value[0,2] Undo

‘grel’

OpenRefine [*formerly* ***G****oogle*]
Regular **E**xpression **L**anguage

‘A *regular expression* is a string that describes a text pattern occurring in other strings.’

CCRIUID_1911	PROV2
NB02500100	NB
NB02500100	NB
NB02500100	NB
NB02500100	NB
NB02500100	NB
NB02500100	NB
NB02500100	NB
NB02500100	NB
NB02500100	NB
NB02500100	NB

https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Regular-Expressions

Regular Expressions

In Transformations:

Edit cells → Transform:

- value.toNumber() ← change to 'number'
- value.toString() ← change to 'string'
- value.replace("+", "") ← replace one thing
- value.replace("~", "").replace(",","") ← replace more than 1 thing
- value.unescape('url') ← remove 'odd' characters
- value.toUppercase() ← change to Uppercase

In Facets

Find entries based on what they contain:

Facet → Custom text facet:

- value.contains("million") ← Facet based on presence/absence of a "string": returns 'True' or 'False'

Source: http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

Regular Expressions

value.replace("F", "f").replace("a","A")

Custom text transform on column OCCUPATION_CHIEF_OCC_IND_CL

Expression Language Google Refine Expression Language (GREL)

```
value.replace("F", "f").replace("a","A")
```

No syntax error.

Preview | History | Starred | Help

row	value	value.replace("F", "f").replace("a","A")
1.	Farmer	fArmer
2.	None	None
3.	null	Error: replace expects 3 strings, or 1 string, 1 regex, and 1 string
4.	None	None
5.	None	None
6.	Farmer	fArmer

On error: keep original / set to blank / store error

Re-transform up to 10 times until no change

OK Cancel

Source: http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

Regular Expressions

value.contains(“Farmer”)

Custom Facet on column OCCUPATION_CHIEF_OCC_IND_CL

Expression Language Google Refine Expression Language (GREL)

```
value.contains("Farmer")
```

No syntax error.

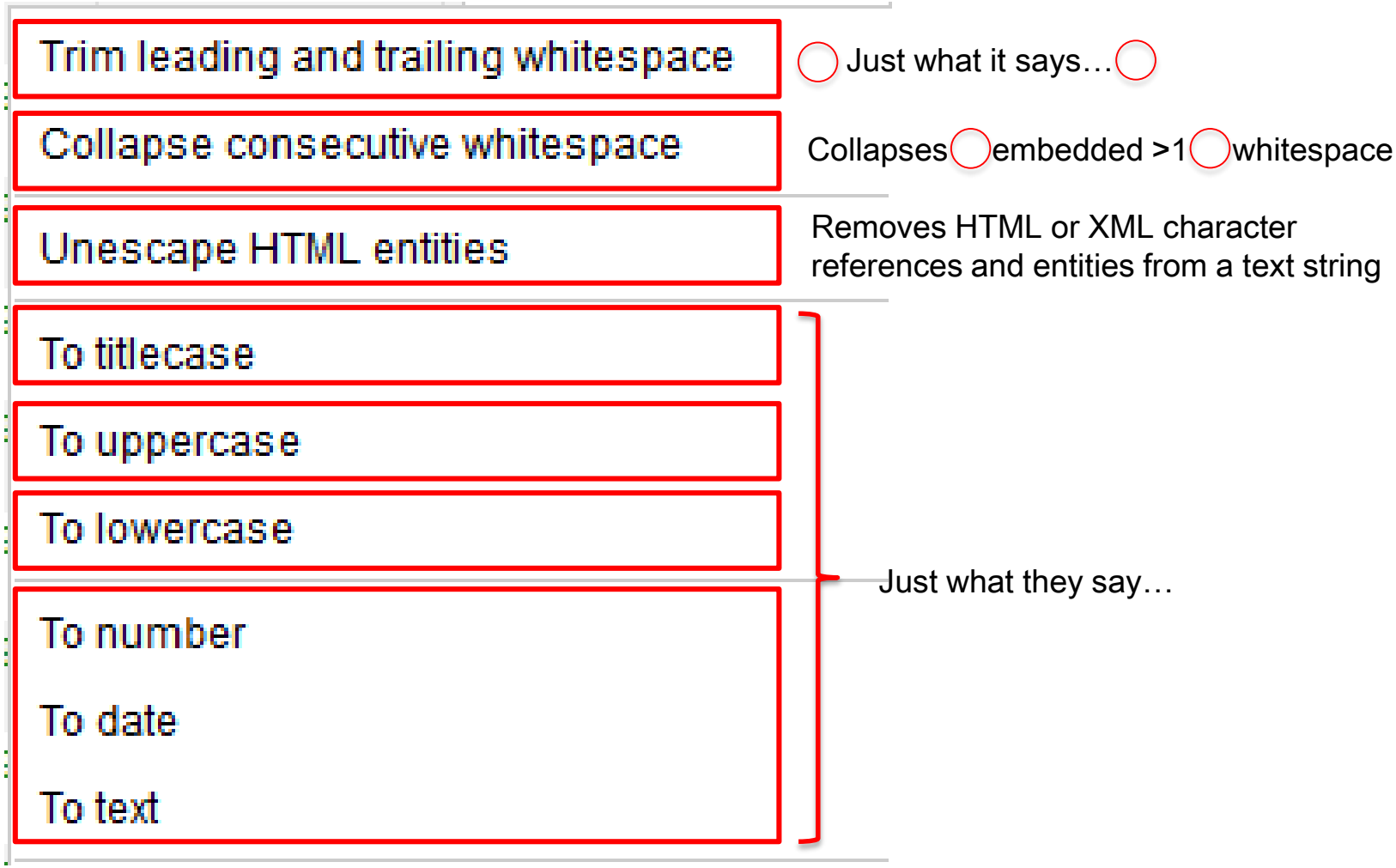
Preview History Starred Help

row	value	value.contains("Farmer")
1.	Farmer	true
2.	None	false
3.	null	null
4.	None	false
5.	None	false
6.	Farmer	true

OK Cancel

Source: http://enipedia.tudelft.nl/wiki/OpenRefine_Tutorial

There is so much more…
Edit cells → Common transforms

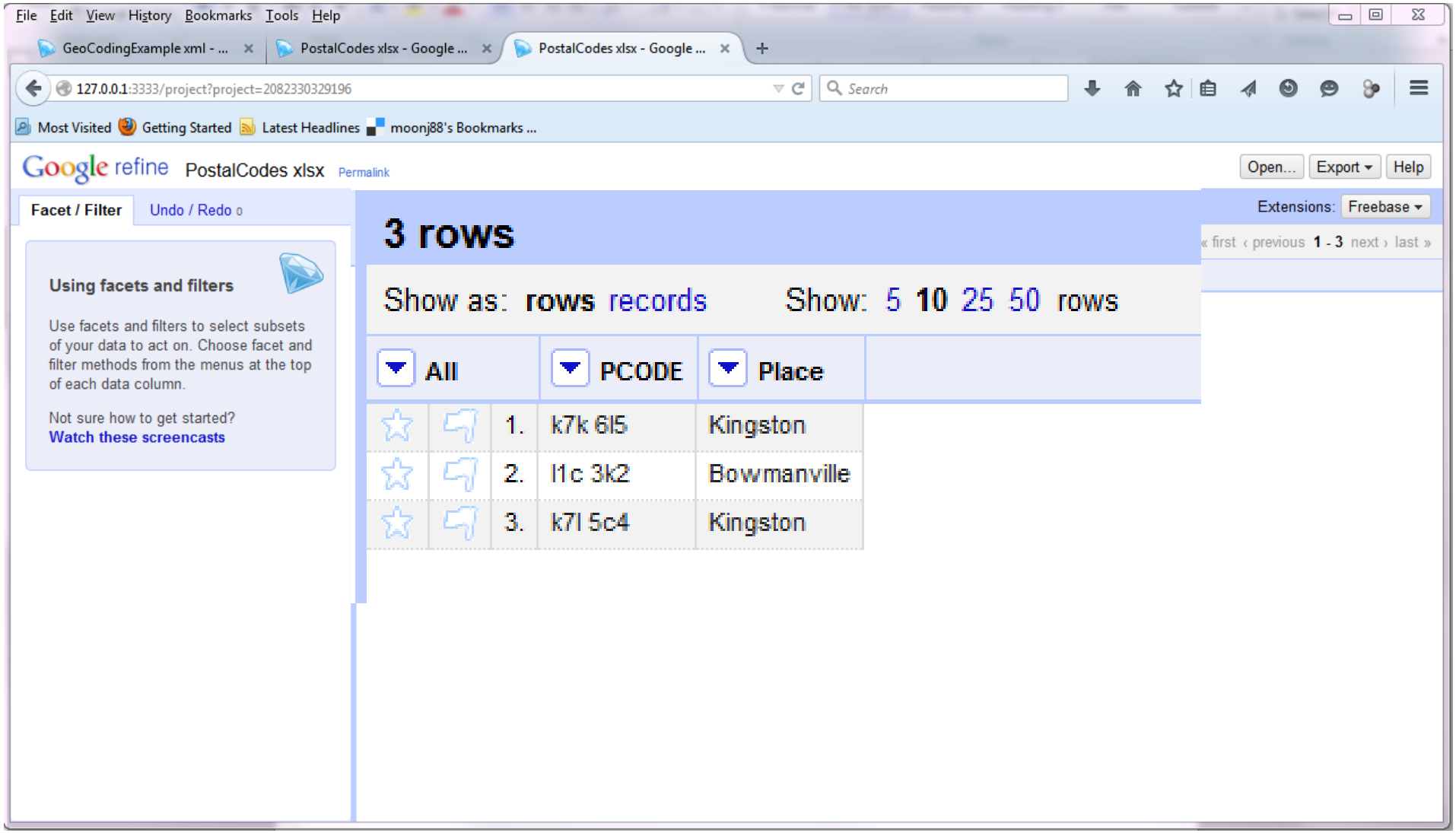


There is so much more…

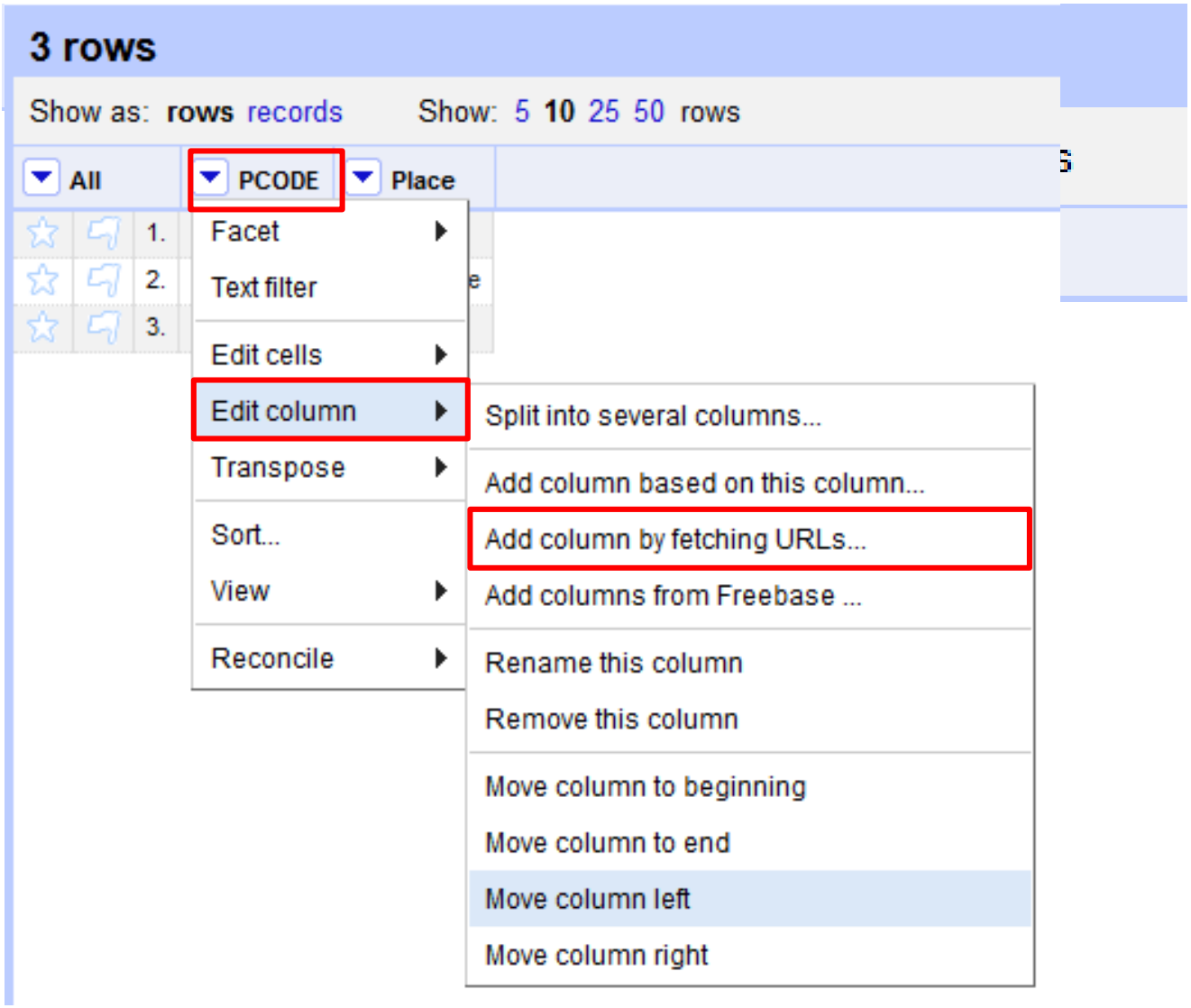
Use Google Regular Expression Language (GREL) to query external data sources

For example, using Postal Codes to extract information from Google Maps

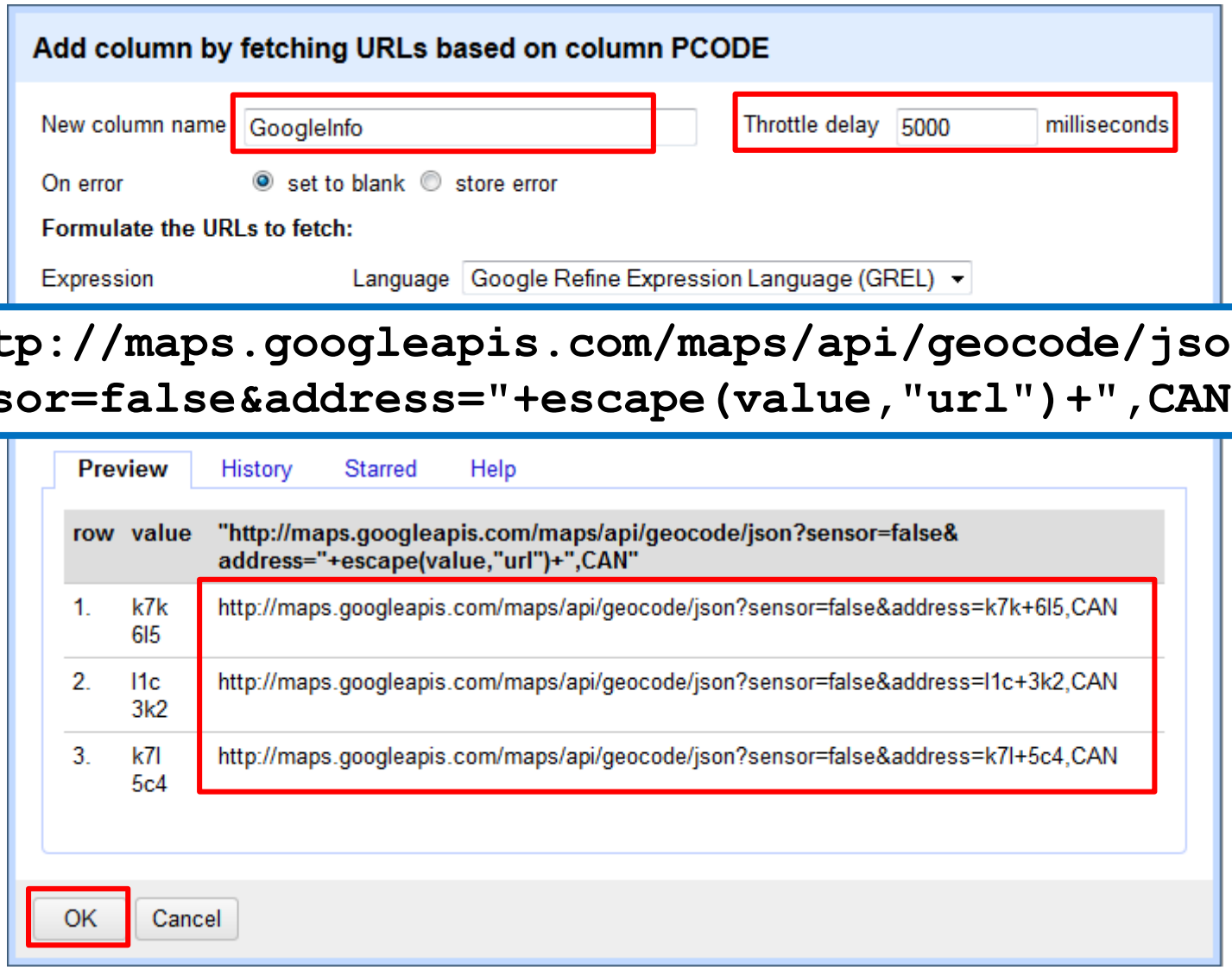
Here's a very simple example…



Use PCODE to retrieve external data…



Add GREL: Google Regular Expression Language



http://datadrivenjournalism.net/resources/geocoding_using_the_google_maps_geocoder_via_openrefine

Google Map API is queried and sends back results…

3 rows Extensions: Freebase

Show as: **rows** records Show: 5 **10** 25 50 rows « first ‹ previous **1 - 3** next › last »

All		PCODE	GoogleInfo	Place
1.		k7k 6l5	{ "results" : [{ "address_components" : [{ "long_name" : "K7K 6L5", "short_name" : "K7K 6L5", "types" : ["postal_code"] }, { "long_name" : "Cataraqui River East", "short_name" : "Cataraqui River East", "types" : ["neighborhood", "political"] }, { "long_name" : "Kingston", "short_name" : "Kingston", "types" : ["locality", "political"] }, { "long_name" : "Frontenac County", "short_name" : "Frontenac County", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : ["administrative_area_level_1", "political"] }, { "long_name" : "Canada", "short_name" : "CA", "types" : ["country", "political"] }], "formatted_address" : "Kingston, ON K7K 6L5, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.2550538, "lng" : -76.4626056 }, "southwest" : { "lat" : 44.2535827, "lng" : -76.4638978 } }, "location" : { "lat" : 44.2540515, "lng" : -76.463049 }, "location_type" : "APPROXIMATE", "viewport" : { "northeast" : { "lat" : 44.25566723029149, "lng" : -76.4619027197085 }, "southwest" : { "lat" : 44.2529692697085, "lng" : -76.46460068029151 } } }, "partial_match" : true, "place_id" : "ChIJf723pzCq0kwRvM3KrXzj__0", "types" : ["postal_code"] }], "status" : "OK" }	Kingston
2.		l1c 3k2	{ "results" : [{ "address_components" : [{ "long_name" : "L1C 3K2", "short_name" : "L1C 3K2", "types" : ["postal_code"] }, { "long_name" : "Haydon", "short_name" : "Haydon", "types" : ["locality", "political"] }, { "long_name" : "Durham Regional Municipality", "short_name" : "Durham Regional Municipality", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : ["administrative_area_level_1", "political"] }, { "long_name" : "Canada", "short_name" : "CA", "types" : ["country", "political"] }], "formatted_address" : "Haydon, ON L1C 3K2, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.066854, "lng" : -78.69502659999999 }, "southwest" : { "lat" : 43.9239747, "lng" : -78.779425 } }, "location" : { "lat" : 43.97851499999999, "lng" : -78.7306158 }, "location_type" : "APPROXIMATE", "viewport" : { "northeast" : { "lat" : 44.066854, "lng" : -78.69502659999999 }, "southwest" : { "lat" : 43.9239747, "lng" : -78.779425 } } }, "partial_match" : true, "place_id" : "ChIJp9BwSfcL1YkRr0ANR05zWyE", "postcode_localities" : ["Bowmanville", "Haydon"], "types" : ["postal_code"] }], "status" : "OK" }	Bowmanville
3.		k7l 5c4 edit	{ "results" : [{ "address_components" : [{ "long_name" : "K7L 5C4", "short_name" : "K7L 5C4", "types" : ["postal_code"] }, { "long_name" : "Kingston", "short_name" : "Kingston", "types" : ["locality", "political"] }, { "long_name" : "Frontenac County", "short_name" : "Frontenac County", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : ["administrative_area_level_1", "political"] }, { "long_name" : "Canada", "short_name" : "CA", "types" : ["country", "political"] }], "formatted_address" : "Kingston, ON K7L 5C4, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.2275953, "lng" : -76.4946668 }, "southwest" : { "lat" : 44.2269466, "lng" : -76.4954425 } }, "location" : { "lat" : 44.2272811, "lng" : -76.49506989999999 }, "location_type" : "APPROXIMATE", "viewport" : { "northeast" : { "lat" : 44.22861993029149, "lng" : -76.4937056970849 }, "southwest" : { "lat" : 44.22592196970849, "lng" : -76.4964036302915 } } }, "partial_match" : true, "place_id" : "ChIJo1lXewSr0kwRqBwN8Nxp4HM", "types" : ["postal_code"] }], "status" : "OK" }	Kingston

Embedded in this data, you'll find 'latitude' & 'longitude'

JSON output can be formatted…


```
{ "results" : [ { "address_components" : [ { "long_name" : "K7L 5C4", "short_name" : "K7L 5C4", "types" : [ "postal_code" ] }, { "long_name" : "Kingston", "short_name" : "Kingston", "types" : [ "locality", "political" ] }, { "long_name" : "Frontenac County", "short_name" : "Frontenac County", "types" : [ "administrative_area_level_2", "political" ] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : [ "administrative_area_level_1", "political" ] }, { "long_name" : "Canada", "short_name" : "CA", "types" : [ "country", "political" ] } ], "formatted_address" : "Kingston, ON K7L 5C4, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.2275953, "lng" : -76.4946668 }, "southwest" : { "lat" : 44.2269466, "lng" : -76.4954425 } }, "location" : { "lat" : 44.2272811, "lng" : -76.49506989999999 }, "location_type" : "APPROXIMATE", "viewport" : { "northeast" : { "lat" : 44.22861993029149, "lng" : -76.49370566970849 }, "southwest" : { "lat" : 44.22592196970849, "lng" : -76.4964036302915 } } }, "partial_match" : true, "place_id" : "ChIJo1lXewSr0kwRqBwN8Nxp4HM", "types" : [ "postal_code" ] } ], "status" : "OK" }
```

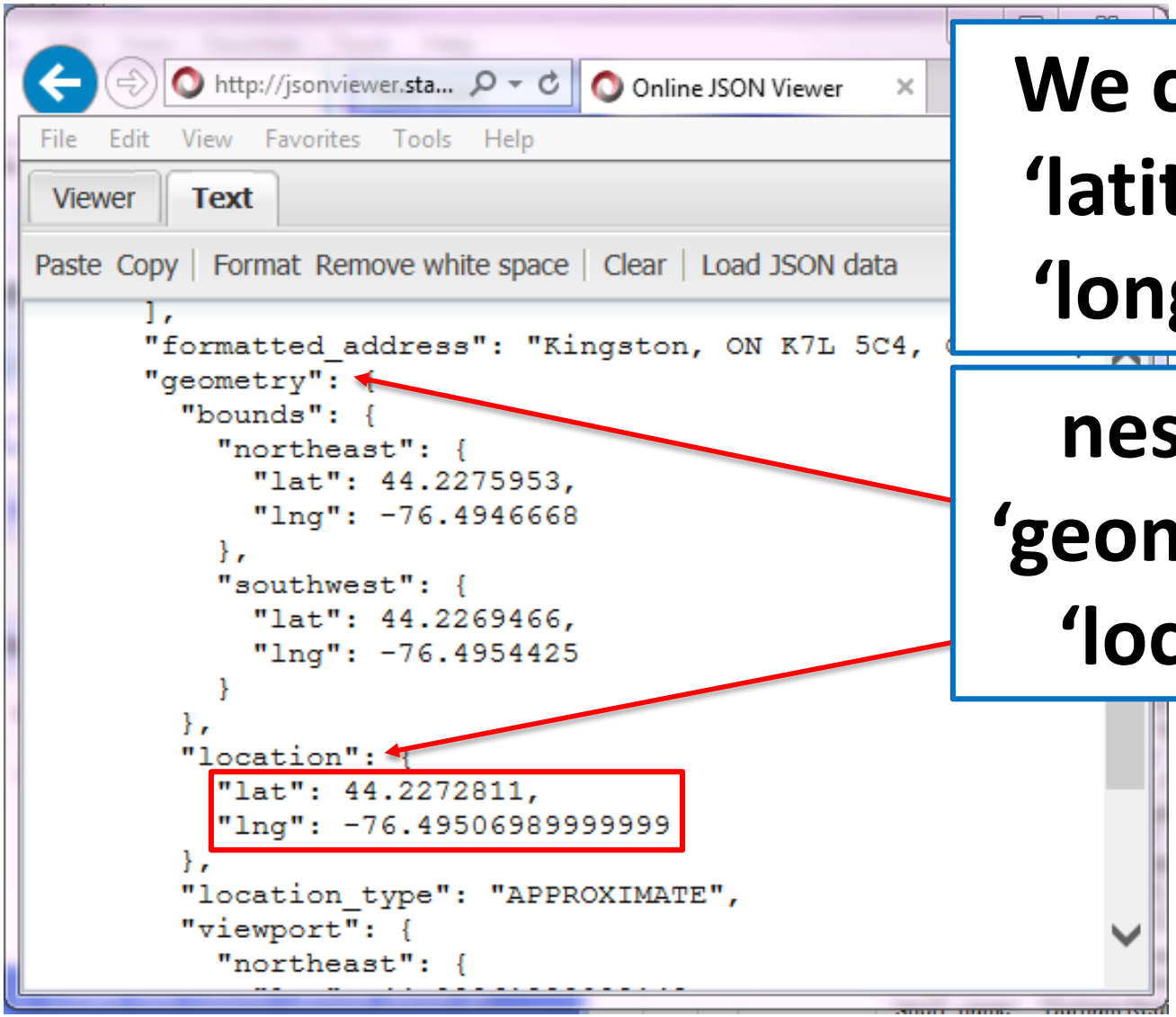

Online JSON Viewer

http://jsonviewer.stack.hu

JSON output can be formatted…

{ "results" : [{ "address_components" : [{ "long_name" : "K7L 5C4", "short_name" : "K7L 5C4", "types" : ["postal_code"] }, { "long_name" : "Kingston", "short_name" : "Kingston", "types" : ["locality", "political"] }, { "long_name" : "Frontenac County", "short_name" : "Frontenac County", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : ["administrative_area_level_1", "political"] }, { "long_name" : "Canada", "short_name" : "CA", "types" : ["country", "political"] }], "formatted_address" : "Kingston, ON K7L 5C4, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.2275953, "lng" : -76.4946668 }, "southwest" : { "lat" : 44.2269466, "lng" : -76.4954425 } }, "location" : { "lat" : 44.2272811, "lng" : -76.49506989999999 }, "location_type" : "APPROXIMATE", "viewport" : { "northeast" : { "lat" : 44.22861993029149, "lng" : -76.49370566970849 }, "southwest" : { "lat" : 44.22592196970849, "lng" : -76.4964036302915 } } }, "partial_match" : true, "place_id" : "ChIJo1lXewSr0kwRqBwN8Nxp4HM", "types" : ["postal_code"] }], "status" : "OK" }

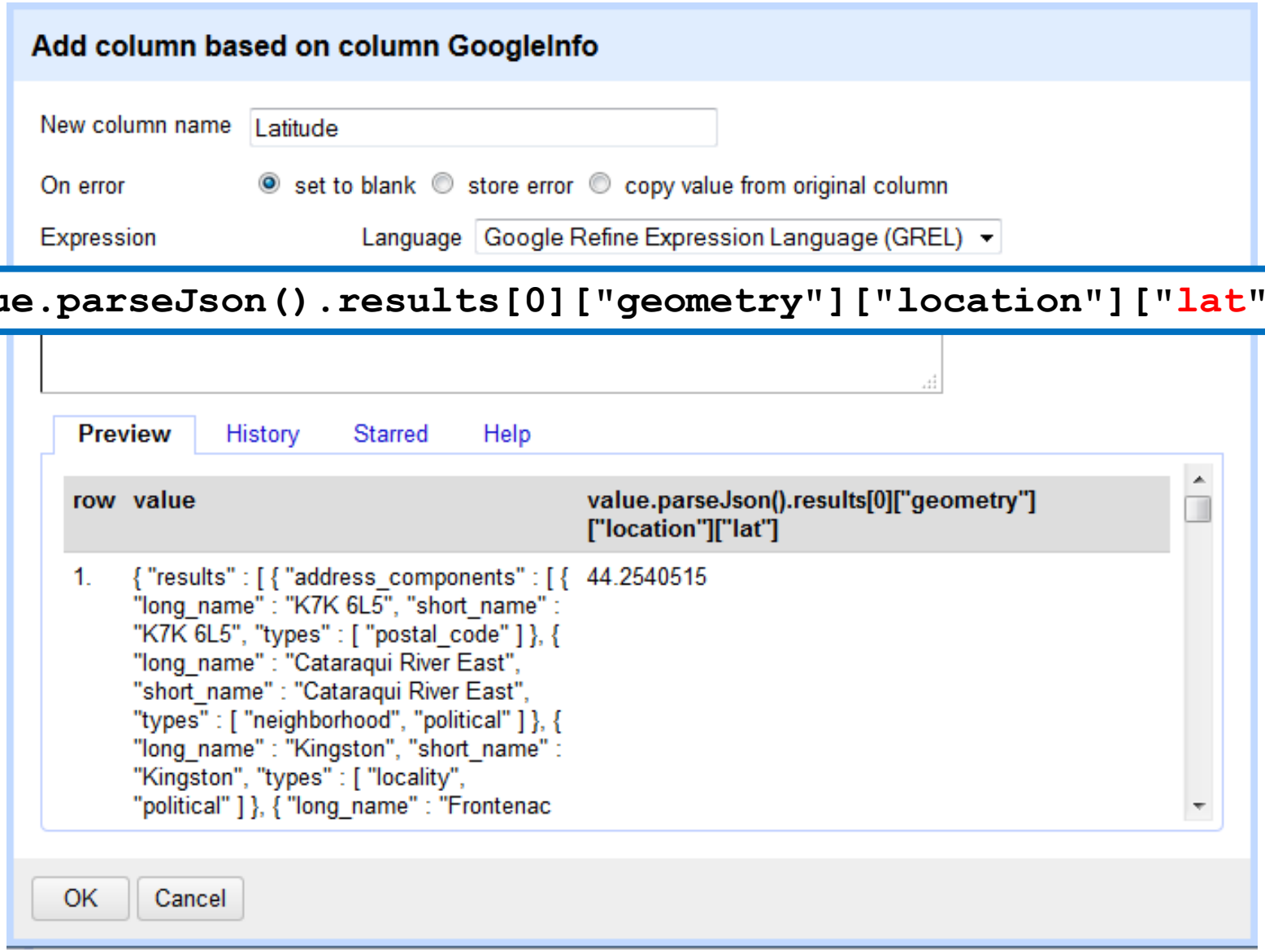
Now it is easier to find elements we might want to use



We can find 'latitude' & 'longitude'

nested in 'geometry' & 'location'

Edit column → Add column based on this column



New ‘Latitude’ column

3 rows Extensions: Freebase

Show as: **rows** records Show: 5 **10** 25 50 rows « first ‹ previous **1 - 3** next › last »

All		PCODE	GoogleInfo	Latitude	Place
1.	k7k 6l5	{ "results" : [{ "address_components" : [{ "long_name" : "K7K 6L5", "short_name" : "K7K 6L5", "types" : ["postal_code"] }, { "long_name" : "Cataraqui River East", "short_name" : "Cataraqui River East", "types" : ["neighborhood", "political"] }, { "long_name" : "Kingston", "short_name" : "Kingston", "types" : ["locality", "political"] }, { "long_name" : "Frontenac County", "short_name" : "Frontenac County", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : ["administrative_area_level_1", "political"] }, { "long_name" : "Canada", "short_name" : "CA", "types" : ["country", "political"] }], "formatted_address" : "Kingston, ON K7K 6L5, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.2550538, "lng" : -76.4626056 }, "southwest" : { "lat" : 44.2535827, "lng" : -76.4638978 } }, "location" : { "lat" : 44.2540515, "lng" : -76.463049 }, "location_type" : "APPROXIMATE", "viewport" : { "northeast" : { "lat" : 44.25566723029149, "lng" : -76.4619027197085 }, "southwest" : { "lat" : 44.2529692697085, "lng" : -76.46460068029151 } } }, "partial_match" : true, "place_id" : "ChIJf723pzCq0kwRvM3KrXzj__0", "types" : ["postal_code"] }], "status" : "OK" }	44.2540515	Kingston
2.	l1c 3k2 edit	{ "results" : [{ "address_components" : [{ "long_name" : "L1C 3K2", "short_name" : "L1C 3K2", "types" : ["postal_code"] }, { "long_name" : "Haydon", "short_name" : "Haydon", "types" : ["locality", "political"] }, { "long_name" : "Durham Regional Municipality", "short_name" : "Durham Regional Municipality", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : ["administrative_area_level_1", "political"] }, { "long_name" : "Canada", "short_name" : "CA", "types" : ["country", "political"] }], "formatted_address" : "Haydon, ON L1C 3K2, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.066854, "lng" : -78.69502659999999 }, "southwest" : { "lat" : 43.9239747, "lng" : -78.779425 } }, "location" : { "lat" : 43.97851499999999, "lng" : -78.7306158 }, "location_type" : "APPROXIMATE", "viewport" : { "northeast" : { "lat" : 44.066854, "lng" : -78.69502659999999 }, "southwest" : { "lat" : 43.9239747, "lng" : -78.779425 } } }, "partial_match" : true, "place_id" : "ChIJp9BwSfcL1YkRr0ANR05zWyE", "postcode_localities" : ["Bowmanville", "Haydon"], "types" : ["postal_code"] }], "status" : "OK" }	43.97851499999999	Bowmanv
3.	k7l 5c4	{ "results" : [{ "address_components" : [{ "long_name" : "K7L 5C4", "short_name" : "K7L 5C4", "types" : ["postal_code"] }, { "long_name" : "Kingston", "short_name" : "Kingston", "types" : ["locality", "political"] }, { "long_name" : "Frontenac County", "short_name" : "Frontenac County", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : ["administrative_area_level_1", "political"] }, { "long_name" : "Canada", "short_name" : "CA", "types" : ["country", "political"] }], "formatted_address" : "Kingston, ON K7L 5C4, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.2275953, "lng" : -76.4946668 }, "southwest" : { "lat" : 44.2269466, "lng" : -76.4954425 } }, "location" : { "lat" : 44.2272811, "lng" : -76.49506989999999 }, "location_type" : "APPROXIMATE", "viewport" : { "northeast" : { "lat" : 44.22861993029149, "lng" : -76.49370566970849 }, "southwest" : { "lat" : 44.22592196970849, "lng" : -76.4964036302915 } } }, "partial_match" : true, "place_id" : "ChIJo1lXewSr0kwRqBwN8Nxp4HM", "types" : ["postal_code"] }], "status" : "OK" }	44.2272811	Kingston

Edit column → Add column based on this column

New ‘Longitude’ column

Extensions: Fr

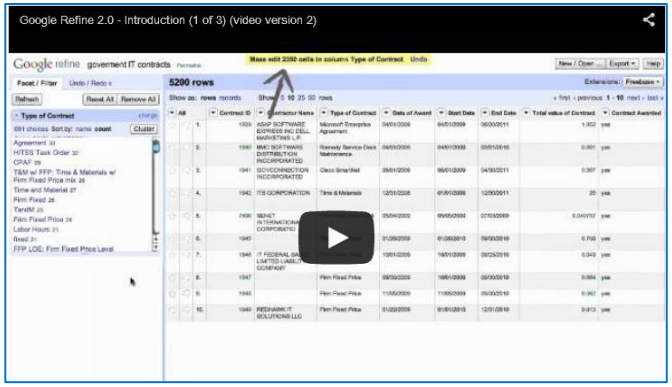
Show as: **rows** records Show: 5 **10** 25 50 rows « first ‹ previous **1 - 3** n

All		PCODE	GoogleInfo	Longitude	Latitude
1.	k7k 6l5	{ "results" : [{ "address_components" : [{ "long_name" : "K7K 6L5", "short_name" : "K7K 6L5", "types" : ["postal_code"] }, { "long_name" : "Cataraqui River East", "short_name" : "Cataraqui River East", "types" : ["neighborhood", "political"] }, { "long_name" : "Kingston", "short_name" : "Kingston", "types" : ["locality", "political"] }, { "long_name" : "Frontenac County", "short_name" : "Frontenac County", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : ["administrative_area_level_1", "political"] }, { "long_name" : "Canada", "short_name" : "CA", "types" : ["country", "political"] }], "formatted_address" : "Kingston, ON K7K 6L5, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.2550538, "lng" : -76.4626056 }, "southwest" : { "lat" : 44.2535827, "lng" : -76.4638978 } }, "location" : { "lat" : 44.2540515, "lng" : -76.463049 }, "location_type" : "APPROXIMATE", "viewport" : { "northeast" : { "lat" : 44.25566723029149, "lng" : -76.4619027197085 }, "southwest" : { "lat" : 44.2529692697085, "lng" : -76.46460068029151 } } }, "partial_match" : true, "place_id" : "ChIJf723pzCq0kwRvM3KrXzj__0", "types" : ["postal_code"] }], "status" : "OK" }	-76.463049	44.2540515
2.	l1c 3k2	{ "results" : [{ "address_components" : [{ "long_name" : "L1C 3K2", "short_name" : "L1C 3K2", "types" : ["postal_code"] }, { "long_name" : "Haydon", "short_name" : "Haydon", "types" : ["locality", "political"] }, { "long_name" : "Durham Regional Municipality", "short_name" : "Durham Regional Municipality", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : ["administrative_area_level_1", "political"] }, { "long_name" : "Canada", "short_name" : "CA", "types" : ["country", "political"] }], "formatted_address" : "Haydon, ON L1C 3K2, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.066854, "lng" : -78.69502659999999 }, "southwest" : { "lat" : 43.9239747, "lng" : -78.779425 } }, "location" : { "lat" : 43.97851499999999, "lng" : -78.7306158 }, "location_type" : "APPROXIMATE", "viewport" : { "northeast" : { "lat" : 44.066854, "lng" : -78.69502659999999 }, "southwest" : { "lat" : 43.9239747, "lng" : -78.779425 } } }, "partial_match" : true, "place_id" : "ChIJp9BwSfcL1YkRr0ANR05zWyE", "postcode_localities" : ["Bowmanville", "Haydon"], "types" : ["postal_code"] }], "status" : "OK" }	-78.7306158	43.97851499999999
3.	k7l 5c4	{ "results" : [{ "address_components" : [{ "long_name" : "K7L 5C4", "short_name" : "K7L 5C4", "types" : ["postal_code"] }, { "long_name" : "Kingston", "short_name" : "Kingston", "types" : ["locality", "political"] }, { "long_name" : "Frontenac County", "short_name" : "Frontenac County", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ontario", "short_name" : "ON", "types" : ["administrative_area_level_1", "political"] }, { "long_name" : "Canada", "short_name" : "CA", "types" : ["country", "political"] }], "formatted_address" : "Kingston, ON K7L 5C4, Canada", "geometry" : { "bounds" : { "northeast" : { "lat" : 44.2275953, "lng" : -76.4946668 }, "southwest" : { "lat" : 44.2269466, "lng" : _type" : 370566970849 },	-76.49506989999999	44.2272811

Closing comments

- Just scratched the surface of what OpenRefine can do
- More resources can be found at OpenRefine.org

Manual



Videos

Regular Expressions Cheat Sheet:
http://arcadiafalcone.net/GoogleRefineCheatSheets.pdf

Using OpenRefine to Update, Clean up, and Link your Metadata to the Wider World, Sarah Weeks, St. Olaf College.

Youtube: https://www.youtube.com/watch?v=E-NbMR3_MRw

Data: https://raw.githubusercontent.com/PaulMakepeace/refine-client-py/master/tests/data/louisiana-elected-officials.csv

Questions