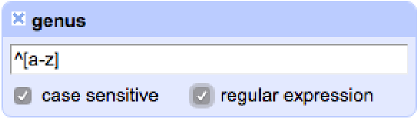
Exercise on the use of Open Refine

Contents

CONVENTIONS

Formulas (copy-paste) Text in blue

Example: ...then paste the expression `^[a-z]`



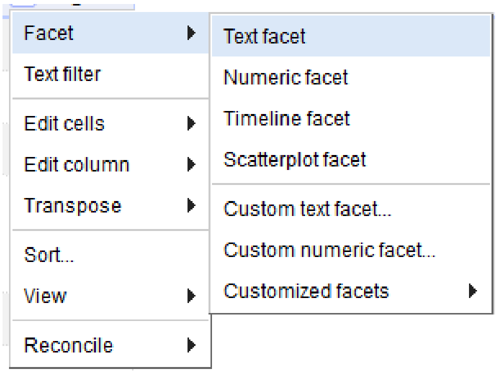
☒ **genus**

^[a-z]

☑ case sensitive ☑ regular expression

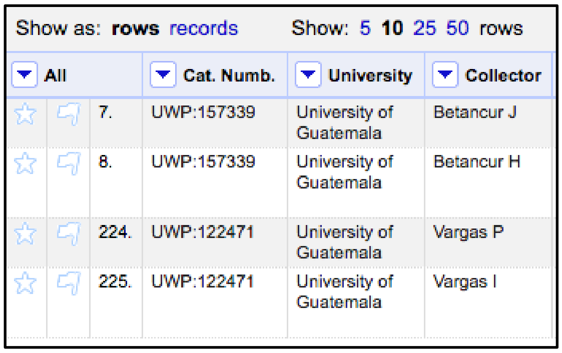
Commands in Refine Text in red

Example: ...and follow the route to Text facet



Column names Text in green

Example: ...go to column Cat. Numb



Show as: **rows** records Show: 5 **10** 25 50 rows

▼ All			▼ Cat. Numb.	▼ University	▼ Collector
☆	⚐	7.	UWP:157339	University of Guatemala	Betancur J
☆	⚐	8.	UWP:157339	University of Guatemala	Betancur H
☆	⚐	224.	UWP:122471	University of Guatemala	Vargas P
☆	⚐	225.	UWP:122471	University of Guatemala	Vargas I

Hyperlinks www.gbif.org

Column menu ▼

2. BASIC USE

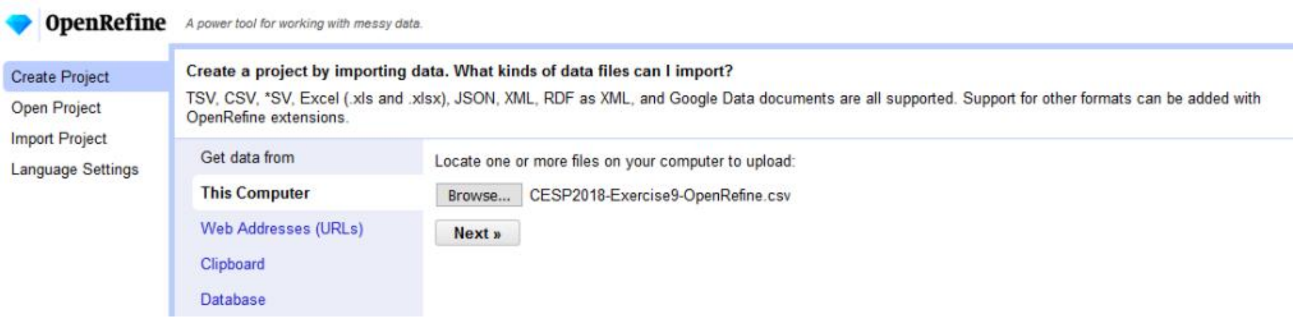
2.1. FILE LOADING AND PROJECTS

2.1.1. Before starting

Data can be loaded from various data sources: TSV, CSV, SV, Excel (.xls and .xlsx), JSON, XML, RDF and XML data as well as Google Docs. Loading data involves two steps: the first is selecting the file, and the second is the creation of the project.

2.1.2. Exercise 1. Create a project

1. Find the file Exercise-OpenRefine.csv in your course folder.
2. Open *OpenRefine* (using openrefine.exe), click on Create Project, and follow the route Get data from > This Computer, then click on Browse. Select the file. Click on Next.



OpenRefine A power tool for working with messy data.

Create Project
Open Project
Import Project
Language Settings

Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can be added with OpenRefine extensions.

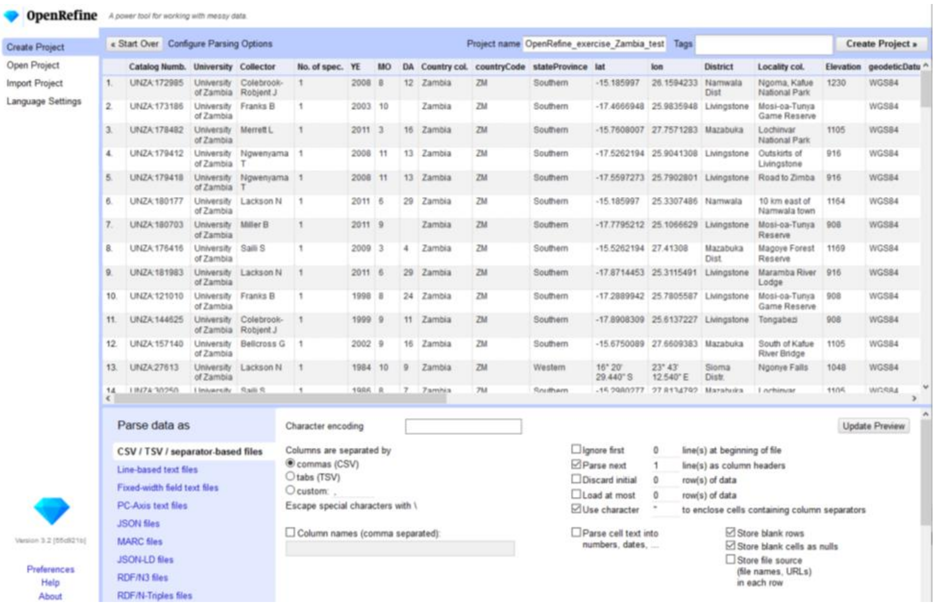
Get data from
This Computer
Web Addresses (URLs)
Clipboard
Database

Locate one or more files on your computer to upload:

Browse... CESP2018-Exercise9-OpenRefine.csv

Next »

3. A parsing options menu will appear. Fill in the options as shown in the picture below. Note that columns may be separated by tabs, commas or semicolons (;), so if you are not sure which delimiter your file uses, experiment with different options, until the data is split into columns correctly, as shown in the image below :



OpenRefine A power tool for working with messy data.

Create Project
Open Project
Import Project
Language Settings

« Start Over Configure Parsing Options

Project name OpenRefine_exercise_Zambia_test Tags

Create Project »

	Catalog Numb.	University	Collector	No. of spec.	YE	MO	DA	Country col.	countryCode	stateProvince	lat	lon	District	Locality col.	Elevation	geodeticDatu
1.	UNZA:172985	University of Zambia	Colebrook-Robjent J	1	2008	8	12	Zambia	ZM	Southern	-15.185997	26.1594233	Namwala Dist	Ngoma, Kafue National Park	1230	WGS84
2.	UNZA:173186	University of Zambia	Franks B	1	2003	10		Zambia	ZM	Southern	-17.4666948	25.9835948	Livingstone	Mosi-oa-Tunya Game Reserve		WGS84
3.	UNZA:178482	University of Zambia	Merrett L	1	2011	3	16	Zambia	ZM	Southern	-15.7608007	27.7571283	Mazabuka	Lochinvar National Park	1105	WGS84
4.	UNZA:179412	University of Zambia	Ngwenyama T	1	2008	11	13	Zambia	ZM	Southern	-17.5262194	25.9041308	Livingstone	Outskirts of Livingstone	916	WGS84
5.	UNZA:179418	University of Zambia	Ngwenyama T	1	2008	11	13	Zambia	ZM	Southern	-17.5597273	25.7902801	Livingstone	Road to Zimba	916	WGS84
6.	UNZA:180177	University of Zambia	Lackson N	1	2011	6	29	Zambia	ZM	Southern	-15.185997	25.3307486	Namwala	10 km east of Namwala town	1164	WGS84
7.	UNZA:180703	University of Zambia	Miller B	1	2011	9		Zambia	ZM	Southern	-17.7795212	25.1066629	Livingstone	Mosi-oa-Tunya Reserve	908	WGS84
8.	UNZA:176416	University of Zambia	Saili S	1	2009	3	4	Zambia	ZM	Southern	-15.5262194	27.41308	Mazabuka Dist	Magoye Forest Reserve	1169	WGS84
9.	UNZA:181983	University of Zambia	Lackson N	1	2011	6	29	Zambia	ZM	Southern	-17.8714453	25.3115491	Livingstone	Maramba River Lodge	916	WGS84
10.	UNZA:121010	University of Zambia	Franks B	1	1998	8	24	Zambia	ZM	Southern	-17.2889942	25.7805587	Livingstone	Mosi-oa-Tunya Game Reserve	908	WGS84
11.	UNZA:144625	University of Zambia	Colebrook-Robjent J	1	1999	9	11	Zambia	ZM	Southern	-17.8908309	25.6137227	Livingstone	Tongabezi	908	WGS84
12.	UNZA:157140	University of Zambia	Bellcross G	1	2002	9	16	Zambia	ZM	Southern	-15.6750089	27.6609383	Mazabuka	South of Kafue River Bridge	1105	WGS84
13.	UNZA:27613	University of Zambia	Lackson N	1	1984	10	9	Zambia	ZM	Western	16° 20' 29.440" S	23° 43' 12.540" E	Sioma Distr.	Ngonye Falls	1048	WGS84
14.	UNZA:30250	University	Saili S	1	1986	8	7	Zambia	ZM	Southern	-15.2980277	27.8134792	Mazabuka	Lochinvar	1105	WGS84

Parse data as

CSV / TSV / separator-based files
Line-based text files
Fixed-width field text files
PC-Axis text files
JSON files
MARC files
JSON-LD files
RDF/N3 files
RDF/N-Triples files

Character encoding

Update Preview

Columns are separated by
● commas (CSV)
○ tabs (TSV)
○ custom: ,
Escape special characters with \

□ Column names (comma separated):

□ 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
☑ Use character " to enclose cells containing column separators

□ Parse cell text into numbers, dates, ...

☑ Store blank rows
☑ Store blank cells as nulls
□ Store file source (file names, URLs) in each row

Version 3.2 [55d921b]

Preferences
Help
About

5. On the top centre you can provide a Project name, and click Create Project and you will be ready to work!

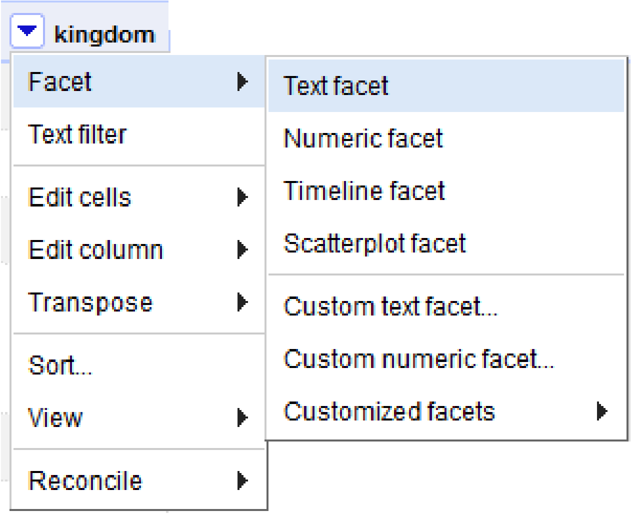
2.2. FACETING

2.2.1. Before starting

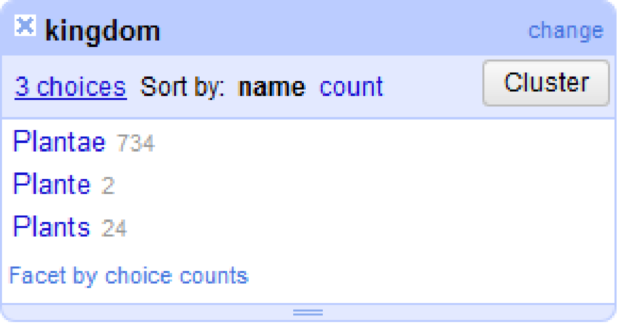
Faceting is a feature that will allow us to get a big picture overview of the data, and to filter down to just the subset of rows that we want to change or view in bulk. It facilitates the use and analysis of data and can be done with cells containing any kind of text, numbers and dates.

2.2.2. Exercise 2. Faceting and mass editing

1. Go to column kingdom, and then click on the column menu ▼ and follow the route to Text facet as shown below:

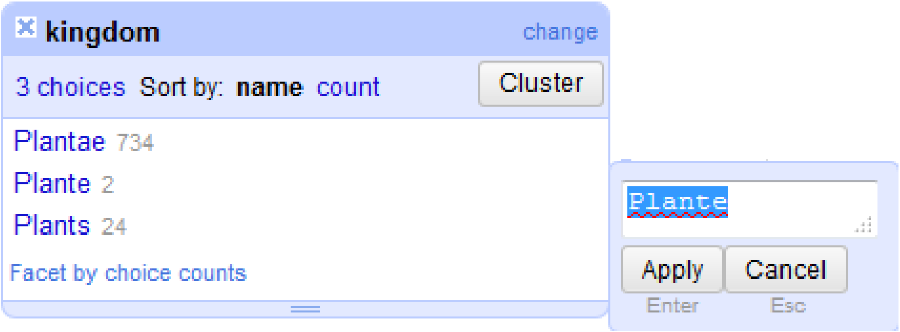


2. On the left a window with the name of the column will appear, which is the facet:



Click on count to sort by count, then click on name to sort alphabetically.

3. To fix the spelling mistakes (all values should be identical), place the cursor over the text in the window and click on Edit, then fix the error in the text box, and to save click on Apply.



Plantae 734
Plante 2
Plants 24

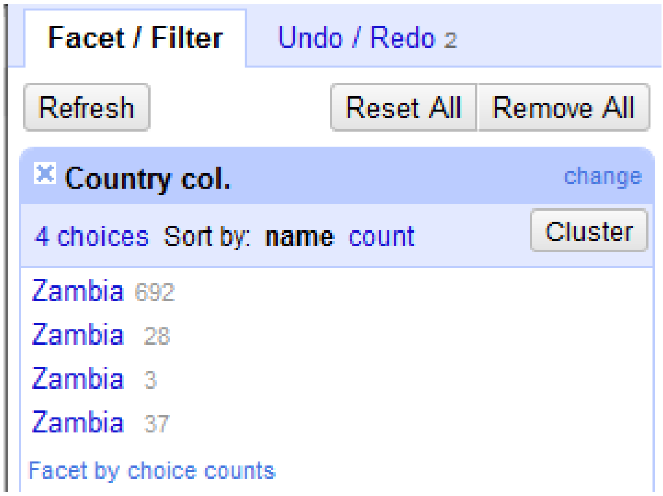
Plante

All the values will be fixed automatically in all records.

4. To close a facet, click on the **x** next to the facet's name.

2.2.3. Exercise 3. Faceting and white spaces I

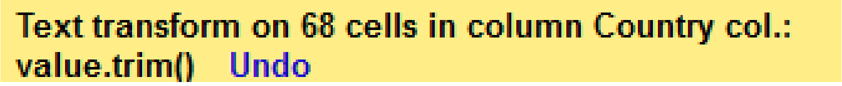
1. Go to Country col. and click on column menu ▼ and perform a Text Facet.



Zambia 692
Zambia 28
Zambia 3
Zambia 37

It appears that the country name is spelled correctly, but the facet shows four different values. This is due to the extra spaces at the end of the text.

2. To fix the error in this column, go to Country col. and click on column menu ▼ > Edit Cells > Common transforms > Trim leading and trailing whitespace. You will see a notification message:

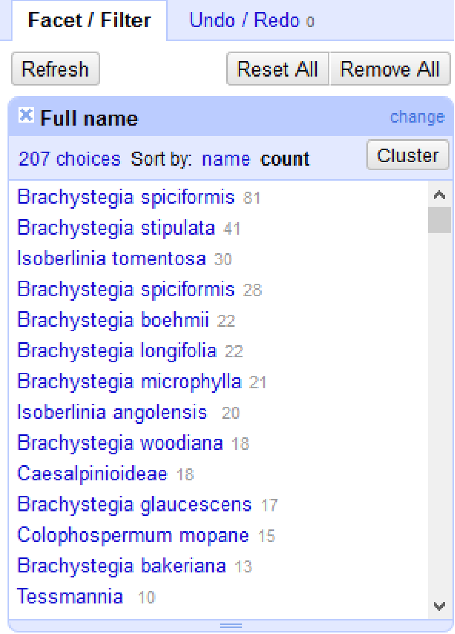


Text transform on 68 cells in column Country col.: value.trim() Undo

3. Now check the facet window: only one value will remain.

2.2.4. Exercise 4. Faceting and white spaces II

1. Go to column Full name and click on ▼ then go to Text facet. Then click on count. The facet will show:



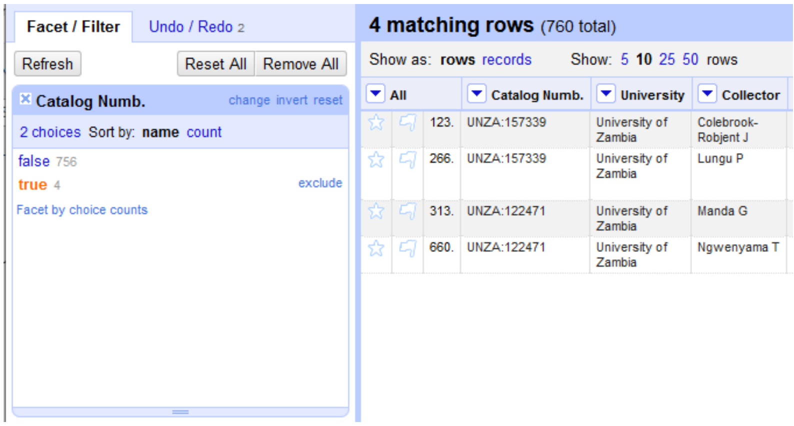
Brachystegia spiciformis	81
Brachystegia stipulata	41
Isoberlinia tomentosa	30
Brachystegia spiciformis	28
Brachystegia boehmii	22
Brachystegia longifolia	22
Brachystegia microphylla	21
Isoberlinia angolensis	20
Brachystegia woodiana	18
Caesalpinioideae	18
Brachystegia glaucescens	17
Colophospermum mopane	15
Brachystegia bakeriana	13
Tessmannia	10

As seen above, *Brachystegia spiciformis* is the first item in the list with 81 specimens, but it is also present in the 4th place with 28 specimens. This is because there are additional white spaces between *Brachystegia* and *spiciformis.*

2. Fix the error from the Full name column menu, Edit Cells > Common transforms > Collapse consecutive whitespaces.
3. Once the whitespaces are removed, *Brachystegia spiciformis* should only appear once in the list with 109 records.

2.2.5. Exercise 5. Faceting and duplicates

1. Go to column catalog in Cat. Numb, and follow the route Facet > Customized facets > Duplicates facet. The facet will show 4 duplicates
2. Click on true, and you'll see the values in the main window:



4 matching rows (760 total)

All	Catalog Numb.	University	Collector
123.	UNZA:157339	University of Zambia	Colebrook-Robjent J
266.	UNZA:157339	University of Zambia	Lungu P
313.	UNZA:122471	University of Zambia	Manda G
660.	UNZA:122471	University of Zambia	Ngwenyama T

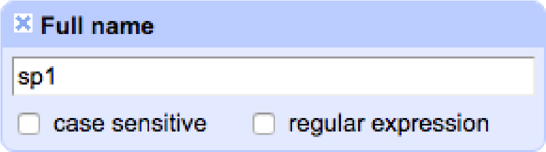
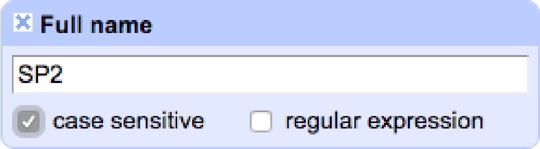
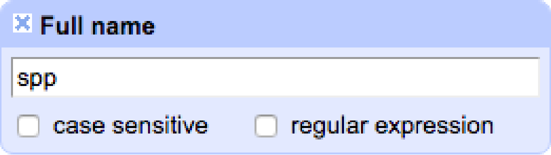
3. Fix the catalog numbers using the correct values shown in the table below, by clicking on edit directly in the relevant cell. A window will open that allows editing of the value. Once it has been corrected, click Apply.

UNZA:122470	Manda G
UNZA:122471	Ngwenyama T
UNZA:157351	Colebrook-Robjent J
UNZA:157339	Lungu P

2.3. FILTERING

2.3.1. Exercise 6. Basic filter

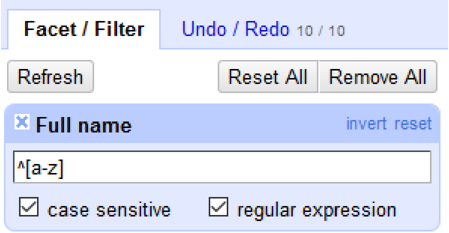
1. Go again to Full name column menu and perform a Text facet to visualize the values, then go again to ▼ and click on Text filter, perform the following filters and fix them as shown below:

Filter	How to fix	Correct value
 Full name sp1 ☐ case sensitive ☐ regular expression	Edit directly in the cell	Eragrostis
 Full name SP2 ☑ case sensitive ☐ regular expression	Edit directly in the cell, check case sensitive	Julbernardia
 Full name spp ☐ case sensitive ☐ regular expression	1. Go to ▼ on Full name, then click Edit cells > Transform… 2. In the text box paste the formula `value.replace(" spp","")` 3. Click OK	Digitaria Diheteropogon Sporobolus Tessmannia

2. Remember to always close filters before continuing to the next exercise.

2.3.2. Exercise 7. Advanced filter I

1. Go to column Full name and perform a Text filter.
2. Check regular expression and case sensitive, then paste the expression `^[a-z]`



Full name

^[a-z]

☑ case sensitive ☑ regular expression

This regular expression filters the strings in which the first letter is lowercase.

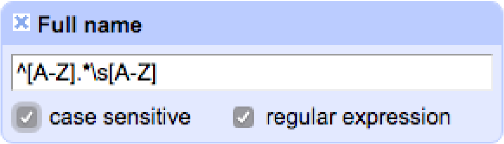
3. **WRITE DOWN THE NUMBER OF RECORDS, AS WELL AS THE NAME IN THE FULL NAME COLUMN, THAT HAVE THE FIRST LETTER AS A LOWERCASE LETTER. YOU WILL REQUIRE THIS INFORMATION TO COMPLETE THE QUIZ.**

4. Perform a correction since the genus should be capitalized.

Note: If you want to know more about regular expressions click here https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Regular-Expressions

2.3.3. Exercise 8. Advanced filter II

1. Go to column Full name and perform a Text filter.
2. Check regular expression and case sensitive, then paste the expression `^[A-Z].*\s[A-Z]`



Full name

^[A-Z].*\s[A-Z]

☑ case sensitive ☑ regular expression

This regular expression filters the strings that start with a capital letter followed by any character, then a space, then a capital letter.

3. **WRITE DOWN THE NUMBER OF RECORDS, AS WELL AS THE NAME IN THE FULL NAME COLUMN, THAT HAVE THE FIRST LETTER OF THE SECOND WORD AS AN UPPERCASE LETTER. YOU WILL REQUIRE THIS INFORMATION TO COMPLETE THE QUIZ.**

4. Perform a correction since the second word of the name should be lowercase.

Note: If you want to know more about regular expressions click here - https://github.com/OpenRefine/OpenRefine/wiki/Understanding-Regular-Expressions

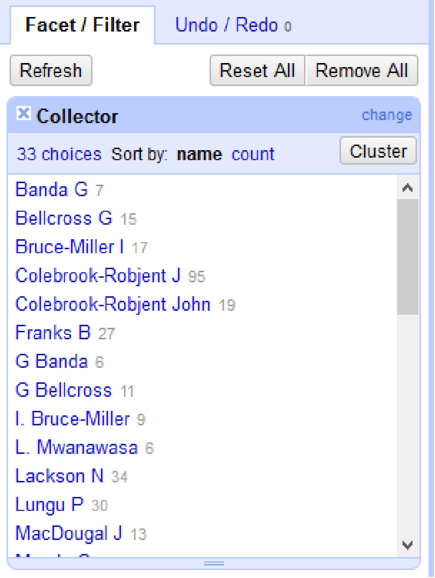
2.3.4. *Exercise 9. Advanced filter – additional exercise on advanced filter 1*

1. Go to column genus and perform a Text filter.
2. Check regular expression and case sensitive, then paste the correct expression to be able to find the strings in which the first letter is lowercase.
3. **WRITE DOWN THE NUMBER OF RECORDS AS WELL AS THE NAME IN THE GENUS COLUMN THAT HAVE THE FIRST LETTER AS A LOWERCASE LETTER. YOU WILL REQUIRE THIS INFORMATION TO COMPLETE THE QUIZ.**
4. Perform a correction since the genus should be capitalized.

2.4. CLUSTERING

2.4.1. Exercise 9. Basic clustering

1. Go to Collector, then in the menu column click Text facet.

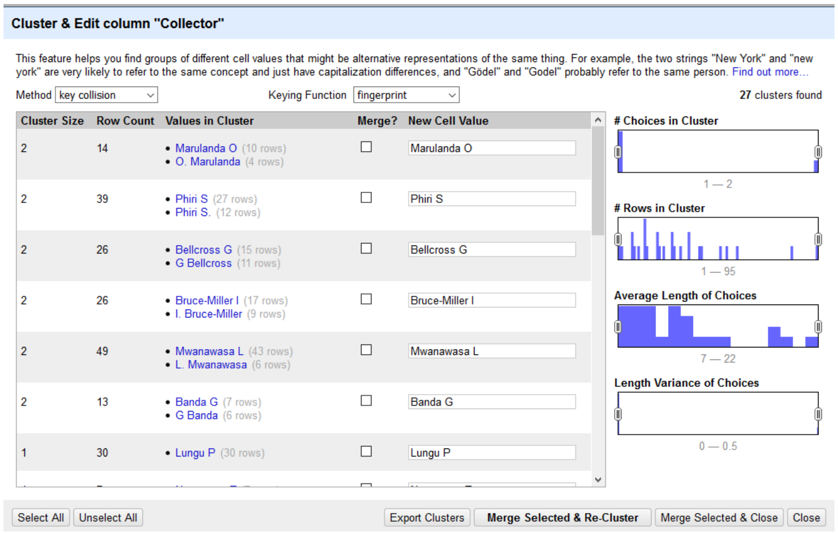


Collector

33 choices Sort by: **name** count

Banda G 7
Bellcross G 15
Bruce-Miller I 17
Colebrook-Robjent J 95
Colebrook-Robjent John 19
Franks B 27
G Banda 6
G Bellcross 11
I. Bruce-Miller 9
L. Mwanawasa 6
Lackson N 34
Lungu P 30
MacDougal J 13

2. On the top right of the facet window click on Cluster, a new window will appear:



Cluster & Edit column "Collector"

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   **27** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	14	• Marulanda O (10 rows) • O. Marulanda (4 rows)	□	Marulanda O
2	39	• Phiri S (27 rows) • Phiri S. (12 rows)	□	Phiri S
2	26	• Bellcross G (15 rows) • G Bellcross (11 rows)	□	Bellcross G
2	26	• Bruce-Miller I (17 rows) • I. Bruce-Miller (9 rows)	□	Bruce-Miller I
2	49	• Mwanawasa L (43 rows) • L. Mwanawasa (6 rows)	□	Mwanawasa L
2	13	• Banda G (7 rows) • G Banda (6 rows)	□	Banda G
1	30	• Lungu P (30 rows)	□	Lungu P

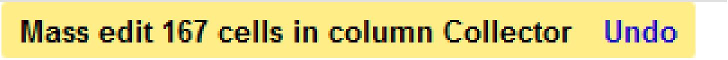
3. Now you can see information about the clusters:
   - Cluster size: the number of different versions that the clustering algorithm believes to be the same.
   - Row count: the number of records with any of the cluster values.
   - Values in cluster: the actual values that the algorithm believes to be the same. There is also the number of records with each particular value, and the possibility to browse the contents of the cluster in a different tab.
   - Merge?: check if values are to be merged into a single standard value.

- New cell value: the value to be applied to every record in the cluster. By default, it is the value with most records. You can also click on any value to apply that to the New cell value.

On the right will be a number of graphs providing additional information about the clusters. The top graph will be #Choices in Cluster. In our example, this will range from 1 to 2. Drag the slider to be greater than 1, so that only those 6 clusters that have more than one choice will show.

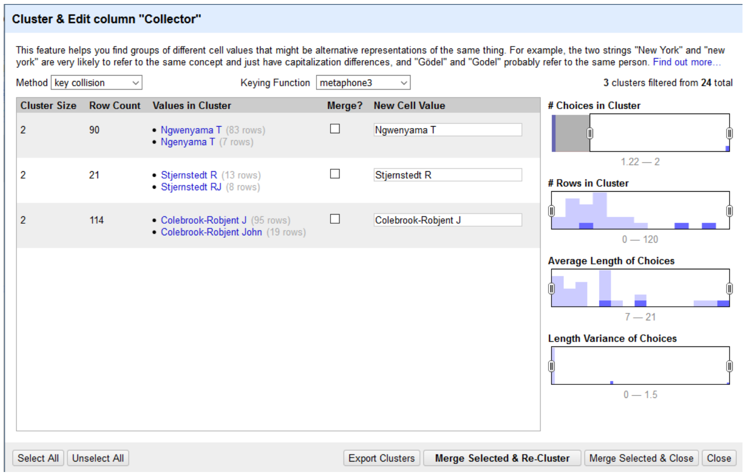
Note: If you want to know more about clustering click here.

4. Click on Select All and then on Merge Selected & close, you will see a notification message:



Mass edit 167 cells in column Collector Undo

5. To fix the remaining collectors go again to Cluster in the facet window of Collector.
6. In the Cluster and edit window, go to Keying Function, then select metaphone3. Again, the top graph on the right will be #Choices in Cluster. Drag the slider to be greater than 1, so that only those clusters that have more than one choice will show. The 3 clusters shown in the image below will be identified.



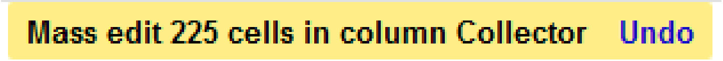
Cluster & Edit column "Collector"

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 3 clusters filtered from 24 total

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	90	• Ngwenyama T (83 rows) • Ngenyama T (7 rows)	□	Ngwenyama T
2	21	• Stjernstedt R (13 rows) • Stjernstedt RJ (8 rows)	□	Stjernstedt R
2	114	• Colebrook-Robjent J (95 rows) • Colebrook-Robjent John (19 rows)	□	Colebrook-Robjent J

7. Click on Select All and then on Merge Selected & close, you will see a notification message:



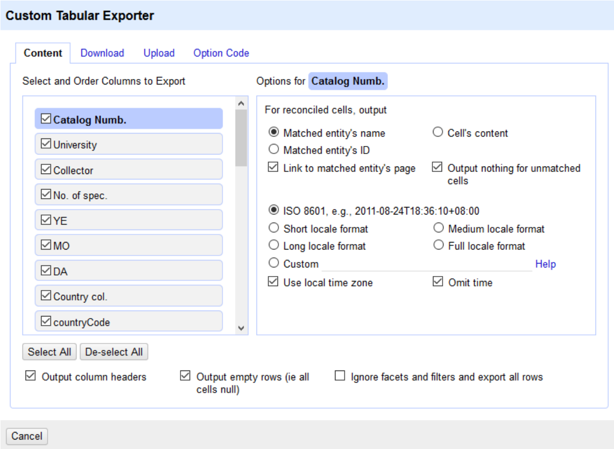
Mass edit 225 cells in column Collector Undo

8. Your collectors are now fixed.

2.5. EXPORTING

You will have several options for exporting your cleaned data, but the following option is useful in most cases.

1. On the upper right corner click on Export and select Custom tabular exporter.
2. You will see the exportation window:



Select and Order Columns to Export

☑ Catalog Numb.
☑ University
☑ Collector
☑ No. of spec.
☑ YE
☑ MO
☑ DA
☑ Country col.
☑ countryCode

Options for Catalog Numb.

For reconciled cells, output

● Matched entity's name ○ Cell's content
○ Matched entity's ID
☑ Link to matched entity's page ☑ Output nothing for unmatched cells

● ISO 8601, e.g., 2011-08-24T18:36:10+08:00
○ Short locale format ○ Medium locale format
○ Long locale format ○ Full locale format
○ Custom
☑ Use local time zone ☑ Omit time

☑ Output column headers ☑ Output empty rows (ie all cells null) □ Ignore facets and filters and export all rows

3. In the content tab you can choose the columns that you want to export, if you select Ignore facets and filters and export all rows, all facets and filterings will be ignored - this is useful if you forget to clear them before exporting.
4. Go to the Download tab and select the separator that you prefer, or whether you would like to download it as an Excel file. Don't modify the other options unless you need to. You can also Upload the data to a Google spreadsheet.

You can also export the whole project to open it in OpenRefine on another computer by following the route Export > Export project. In this case you are not downloading a data file to open in a spreadsheet or text processor, but rather a GZIP file that will only be accessible through OpenRefine.

3. USEFUL LINKS AND REFERENCES

- Documentation https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users
- Resources list for OpenRefine: https://github.com/OpenRefine/OpenRefine/wiki/External-Resource

Exercise concept and content developed by Néstor Beltrán, and adapted for OpenRefine 3.2 and an African example by Lizanne Roxburgh.