

Use of OpenRefine

Biodiversity Data Mobilization - Data Cleaning - OpenRefine Exercise



CONTENTS

[CONTENTS](#)

[CONVENTIONS](#)

[2. BASIC USE](#)

[2.1. FILE LOADING AND PROJECTS](#)

[2.1.1. Before starting](#)

[2.1.2. Exercise 1. Create a project](#)

[2.2. FACETING](#)

[2.2.1. Before starting](#)

[2.2.2. Exercise 2. Faceting and mass editing](#)

[2.2.3. Exercise 3. Faceting and white spaces I](#)

[2.2.4. Exercise 4. Faceting and white spaces II](#)

[2.2.5. Exercise 5. Faceting and duplicates](#)

[2.3. FILTERING](#)

[2.3.1. Exercise 6. Basic filter](#)

[2.3.2. Exercise 7. Advanced filter I](#)

[2.3.3. Exercise 8. Advanced filter II](#)

[2.4. CLUSTERING](#)

[2.4.1. Exercise 9. Basic clustering](#)

[2.5. EXPORTING](#)

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

[3. BASIC API USE](#)

[3.1. Before starting](#)

[3.2. Exercise 1. Higher taxonomy](#)

[4. USEFUL LINKS AND REFERENCES](#)

CONVENTIONS

Formulas (copy-paste)

Text in blue

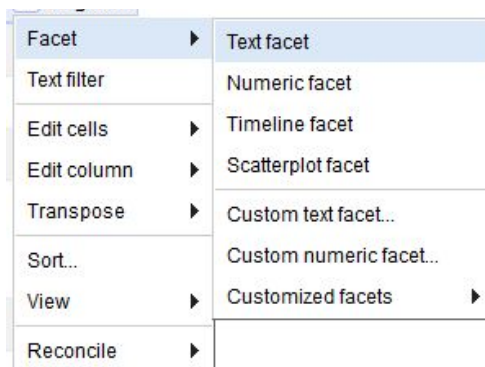
Example: ...then paste the expression `^[a-z]`



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

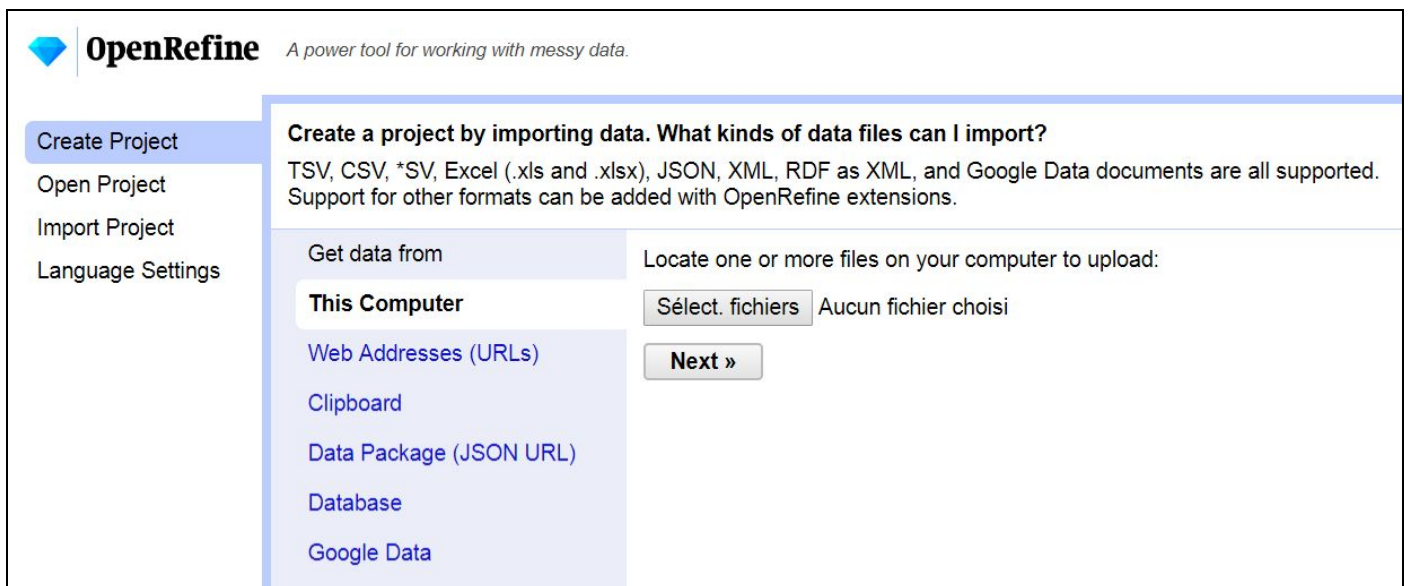
2.1. FILE LOADING AND PROJECTS

2.1.1. Before starting

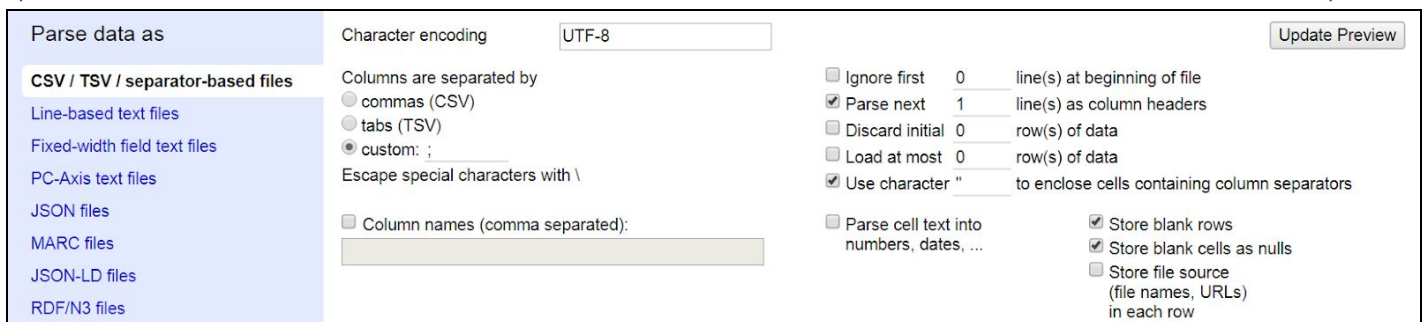
Data loading can be done from various data sources: TSV, CSV, SV, Excel (.xls and .xlsx), JSON, XML, RDF and XML data as Google Docs. Loading data involves two stages: the first is the file load and the second is the creation of the project.

2.1.2. Exercise 1. Create a project

1. Load the base data file from the indicated link in the the e-learning platform.
2. Open *OpenRefine* (in the latest version of OpenRefine, you can select your preferred language in [Language Settings](#)), click on [Create Project](#), and follow the route [Get data from](#) > [This Computer](#), then click on [Choose Files](#). Select the file. Click on [Next](#).



3. OpenRefine will automatically detect the format of your file as csv; a parsing options menu will appear. Choose the options as shown in the picture:




4. On the top right in the Project Name field, rename your file as [YourName]UseCase1OpenRefine and click [Create Project](#) and you will be ready to work!

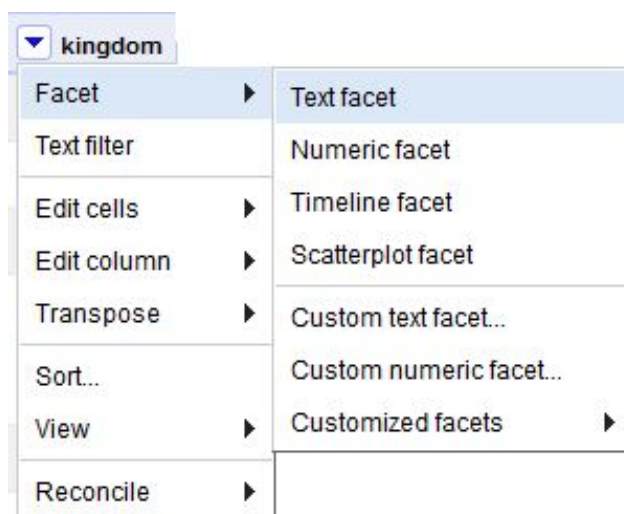
2.2. FACETING

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:

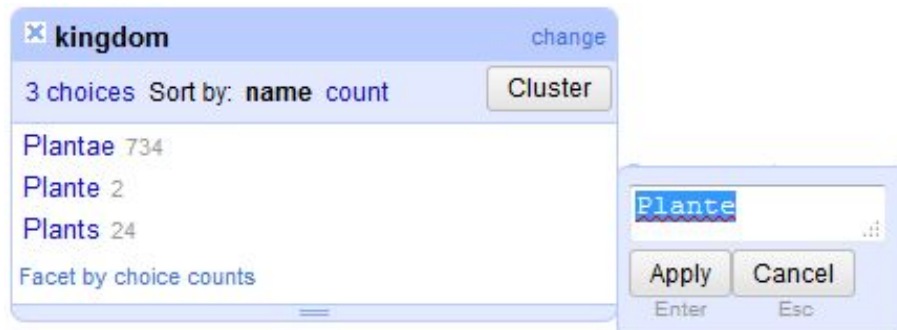


2. On the left a window with the name of the column will appear, that is the facet:




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

- Fix the spelling mistakes. 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**.



All the values will be fixed automatically.

2.2.3. EXERCISE 3. Faceting and white spaces I

- Go to **Country col.** and click on column menu  and perform a **Text Facet**.




On a quick view, the country appears to be spelled correctly, but the facet shows three different values due to the extra spaces at the end of the text.

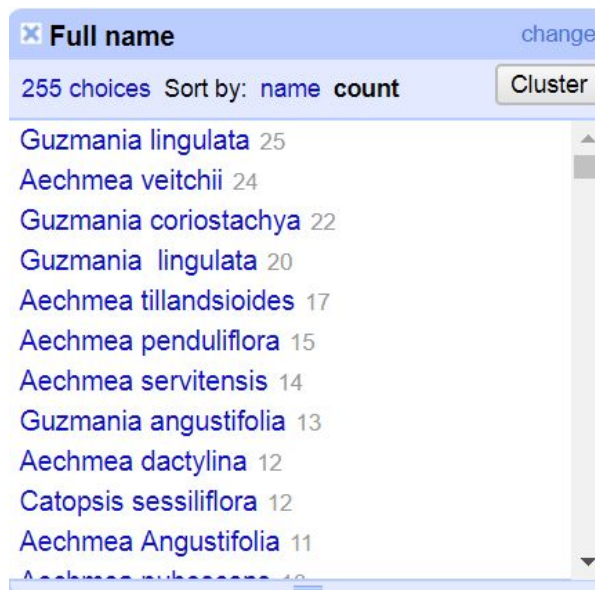
- Fix the error from the column menu on **Country col.**, following the route **Edit Cells > Common transforms > Trim leading and trailing whitespace**. You will see a notification message:

Text transform on 38 cells in column Country col.: value.trim() Undo

- Now check the facet window; only one value will remain.

2.2.4. EXERCISE 4. Faceting and white spaces II

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



As seen above, *Guzmania lingulata* is the first item in the list with 25 specimens, but it is also present in the 4th place with 20 specimens

- Fix the error from the **Full name** column menu, **Edit Cells** > **Common transforms** > **Collapse consecutive whitespaces**.
- Once the white-spaces are removed, *Guzmania lingulata* should only appear in the list with 45 records.

2.2.5. EXERCISE 5. Faceting and duplicates

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


Show as: rows records		Show: 5 10 25 50 rows			
<input type="checkbox"/> All	<input type="checkbox"/> Cat. Numb.	<input type="checkbox"/> University	<input type="checkbox"/> Collector		
<input type="checkbox"/> <input type="checkbox"/>	49.	UWP:122471	University of Guatemala	Vargas P	
<input type="checkbox"/> <input type="checkbox"/>	50.	UWP:122471	University of Guatemala	Vargas I	
<input type="checkbox"/> <input type="checkbox"/>	117.	UWP:157339	University of Guatemala	Betancur J	
<input type="checkbox"/> <input type="checkbox"/>	118.	UWP:157339	University of Guatemala	Betancur H	

After a check with the specimens labels, fix the values clicking edit directly on the cell with the correct catalogue numbers:

UWP:122470 Vargas P
 UWP:122471 Vargas I
 UWP:157351 Betancur H
 UWP:157339 Betancur J

2.3. FILTERING

2.3.1. EXERCISE 6. Basic filter

- 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
	<p>Edit directly in the cell</p>	<p>Cyperus</p>
	<p>Edit directly in the cell, check case sensitive</p>	<p>Cyperus</p>
	<ol style="list-style-type: none"> Go to  on Full name, then click Edit cells > Transform... In the text box paste the formula <code>value.replace(" spp.", "")</code> Click OK 	<p>Aechmea Chusquea Eleocharis Greigia Navia Neurolepis Rhynchospora Tillandsia Xyris</p>

2.3.2. EXERCISE 7. Advanced filter I

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



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

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

Note: If you want to know more about regular expressions click [here](#).

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]`



This regular expression filters the strings that start with a capital letter followed by 0 or more characters, then a space, then a capital letter.

3. 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](#).

2.4. CLUSTERING

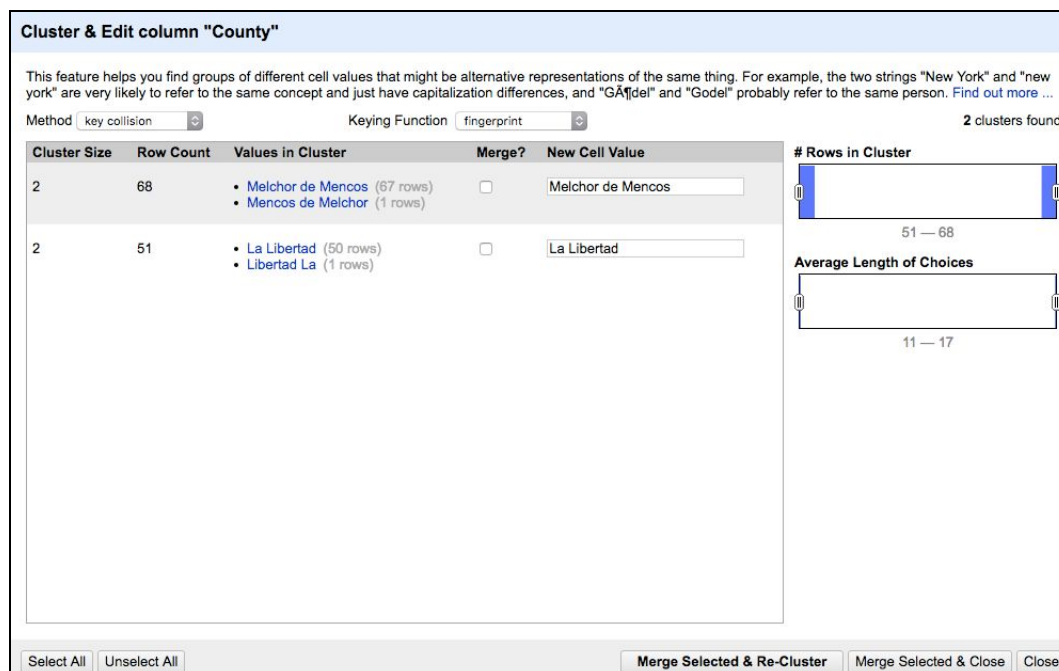
2.4.1. EXERCISE 9. Basic clustering

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



Keep in mind that the correct counties are: Flores, La Libertad, Melchor de Mencos, San Andres and San Jose.

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



Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	# Rows in Cluster
2	68	<ul style="list-style-type: none"> Melchor de Mencos (67 rows) Mencos de Melchor (1 rows) 	<input type="checkbox"/>	Melchor de Mencos	51 — 68
2	51	<ul style="list-style-type: none"> La Libertad (50 rows) Libertad La (1 rows) 	<input type="checkbox"/>	La Libertad	11 — 17

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**.

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 119 cells in column County Undo

5. To fix the remaining counties go again to **Cluster** in the facet window of **County**.
6. In the Cluster and edit window, go to **Keying Function**, then select **ngram-fingerprint**, and set **1** as the value in **Ngram Size**. Press the enter key.
7. Click on **Select All** and then on **Merge Selected & close**, you will see a notification message:

Mass edit 360 cells in column County Undo

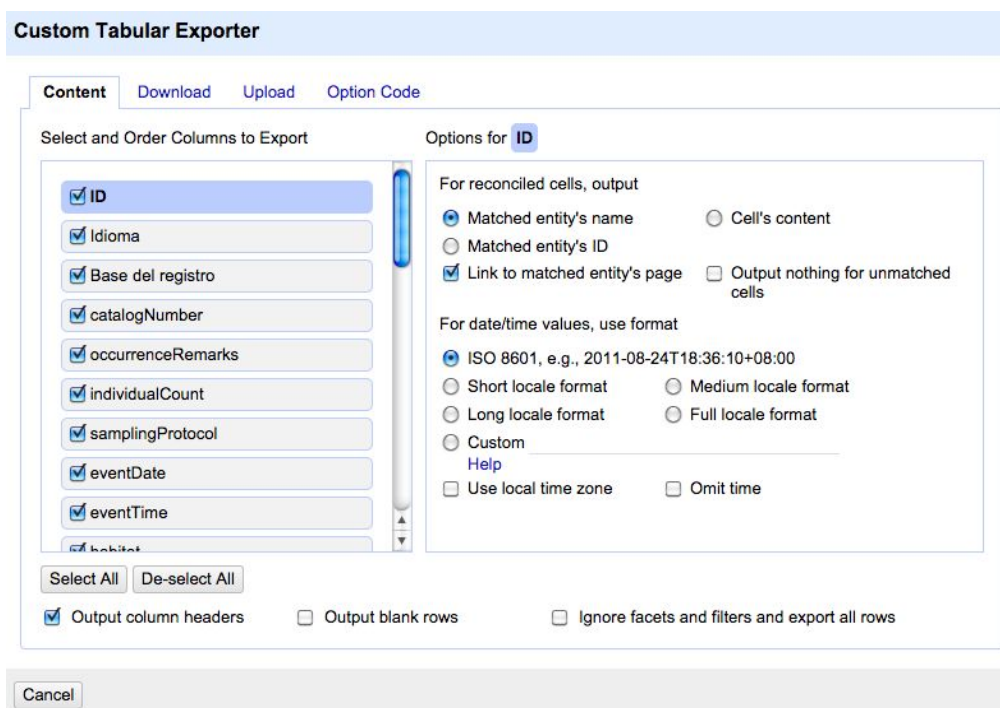
8. Your counties are now fixed and your window should look like the image below:



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:



Custom Tabular Exporter

Content Download Upload Option Code

Select and Order Columns to Export

- ID
- Idioma
- Base del registro
- catalogNumber
- occurrenceRemarks
- individualCount
- samplingProtocol
- eventDate
- eventTime
- habitat

Options for ID

For reconciled cells, output

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

For date/time values, use format

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

Help

Select All De-select All

Output column headers Output blank rows Ignore facets and filters and export all rows

Cancel

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. Don't modify the other options unless you need to.

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're 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. BASIC API USE

3.1. Before starting

Reconciliation matches the information in one of your columns to an outside database. This is particularly helpful when it comes to name validation, as it proves the name you have exists somewhere else. This is a really useful service, but can be time consuming. In this case we will go through the process with only three records using the API from GBIF. Internet connection is required.

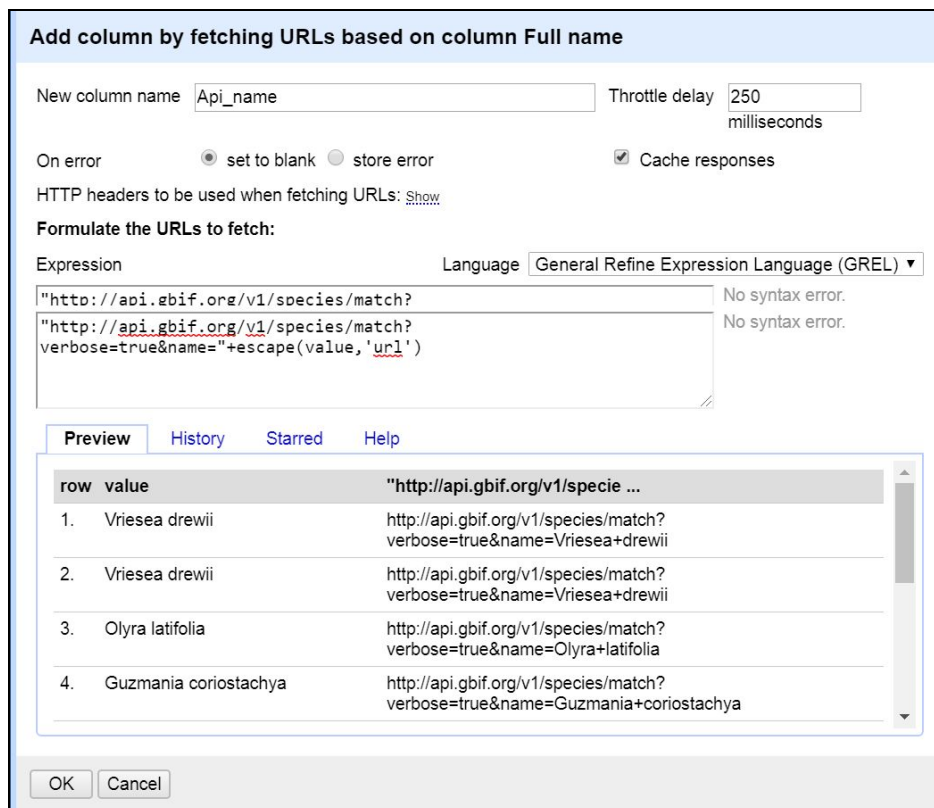
3.2. Exercise 1. Higher taxonomy

1. Go to **Collector**, then make a **Text facet**. Select the collector Elsa P



2. Under **Full name**, click on column menu and then **Edit column** > **Add column by fetching URLs...**, call the new column **Api_name**
3. Change the **Throttle Delay** to **250** and paste the expression:

`"http://api.gbif.org/v1/species/match?verbose=true&name="+escape(value, 'ur1')`



- Click **ok** and wait, this might take some time depending on your internet connection and the number of taxa.
- Go to **Api_name**, click on column menu and then **Edit column > Add column based on this column....** Call the new column **higherClassification** and paste the expression:

```
value.parseJson().get("kingdom")+
", "+value.parseJson().get("phylum")+
", "+value.parseJson().get("class")+
", "+value.parseJson().get("order")+
", "+value.parseJson().get("family")
```

You will see the Kingdom, Phylum, Class, Order and family of each taxon.

Add column based on column Api_name

New column name

core-views/addasdasd set to blank store error copy value from original column

Expression Language

```
value.parseJson().get("kingdom")+
", "+value.parseJson().get("phylum")+
", "+value.parseJson().get("class")+
", "+value.parseJson().get("order")+
", "+value.parseJson().get("family")
```

No syntax error.

Preview
History
Starred
Help

row	value	value.parseJson().get("kingdom ...
1.	{ "usageKey": 5288980, "scientificName": "Vriesea drewii L.B.Sm.", "canonicalName": "Vriesea drewii", "rank": "SPECIES", "status": "ACCEPTED", "name=110; authorship=0; classification=-2; rank=5; status=1; singleMatch=5", "matchType": "EXACT", "kingdom": "drewii", "kingdomKey": 6, "phylumKey": 7707728, "cla	Plantae, Tracheophyta, Liliopsida, Poales, Bromeliaceae
2.	{ "usageKey": 5288980, "scientificName": "Vriesea drewii L.B.Sm.", "canonicalName": "Vriesea	Plantae, Tracheophyta, Liliopsida, Poales, Bromeliaceae

6. Under **higherClassification** follow the route **Edit column > Split into several columns...**, leave the settings as shown:

Split column higherClassification into several columns

How to Split Column

by separator

Separator regular expression

Split into columns at most (leave blank for no limit)

by field lengths

List of integers separated by commas, e.g., 5, 7, 15

After Splitting

Guess cell type

Remove this column

7. Now you know how to obtain the taxonomic categories of a given taxon if this is available in the GBIF API. Column names can be edited in **Edit column > Rename this column**.
8. For the purpose of the workshop, the columns created in this exercise (Higher taxonomy) must be deleted. Under **All**, which is the first column, go to **Edit columns > Re-order / remove columns...**
9. Drop the columns as shown and click **OK**:

Re-order / Remove Columns

Drag columns to re-order

coordinate Uncertainty

identifiedBy

typeStatus

kingdom

phylum

class

order

family

genus

specificEpithet

infraspecificEpithet

Full name

taxonRank

Authorship

Other name

Drop columns here to remove

Api_name

higherClassification 1

higherClassification 2

higherClassification 3

higherClassification 4

higherClassification 5

4. USEFUL LINKS AND REFERENCES

- Name validation Tutorial:
https://docs.google.com/document/d/1tkDRXIYhmassYAk5T4v5oac5prF0jAiSMr_JEGTvhRo/edit
- Higher Taxonomy Tutorial:
https://docs.google.com/document/d/1XZ_pM9gldQzHzl8wfUCVea-52yub5T_3tc-snBgPRa0/edit
- 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.

Updated: 03 July 2019. Sophie Pamerlon

Updated: 11 July 2019. Laura Russell, Sophie Pamerlon.