OpenRefine for Data Cleaning

What is OpenRefine?
A data cleaning tool.

Why use OpenRefine?
Tracks changes and is easily reversible.

Reference: Naidoo, H (Hush Naidoo Jade Photography), October 2020, Unsplash License, https://unsplash.com/photos/SutAduhzdRQ

Agenda

What is OpenRefine?	01
Opening OpenRefine	02
Getting Data into OpenRefine	03
Manipulating/Mangling Data in OpenRefine	04
Getting Stuff out of OpenRefine	05

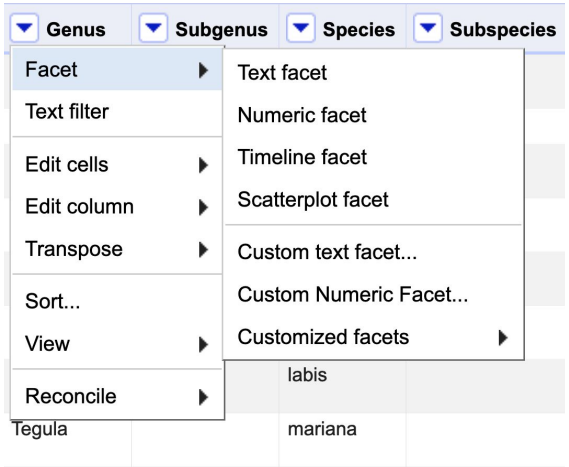
Reference: (c) Field Museum of Natural History - CC BY-NC 4.0 https://mm.fieldmuseum.org/3994ee8a-634d-4e95-bbd7-b008d415fb81 (accessed on 6-15-2022)

Conventions

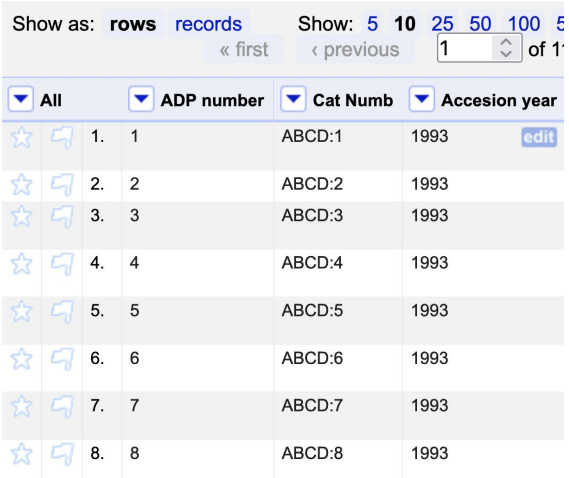
- Formulas (copy and paste)
  - Text in blue
  - Example: …then paste the expression

```
^[a-z]
```

- Column menu



- Commands in Refine
  - Text in magenta
  - Example: …and follow the route to Text facet



- Column names
  - Text in green
  - Example: …go to column **Cat. Numb**

What is OpenRefine?

- The tool formerly known as Google Refine
- Free, open source project
- Stand-alone desktop application
- Uses General Refine Expression Language (GREL)

Reference: openrefine.org, Public domain, via Wikimedia Commons

What Can/Can’t OpenRefine Do?

	OpenRefine
Formulas	☑
Charts & Graphs	
Auto-formats fields	
Allow creation of lookup lists	
Allow character sets besides Latin1 (e.g., UTF-8)	☑
Track changes after closing	☑
Maintain integrity between rows and columns	☑
Enforce referential integrity between different sheets	
Creates a copy of original data	☑
Allow you to share steps	☑

FIELD MUSEUM

Opening OpenRefine

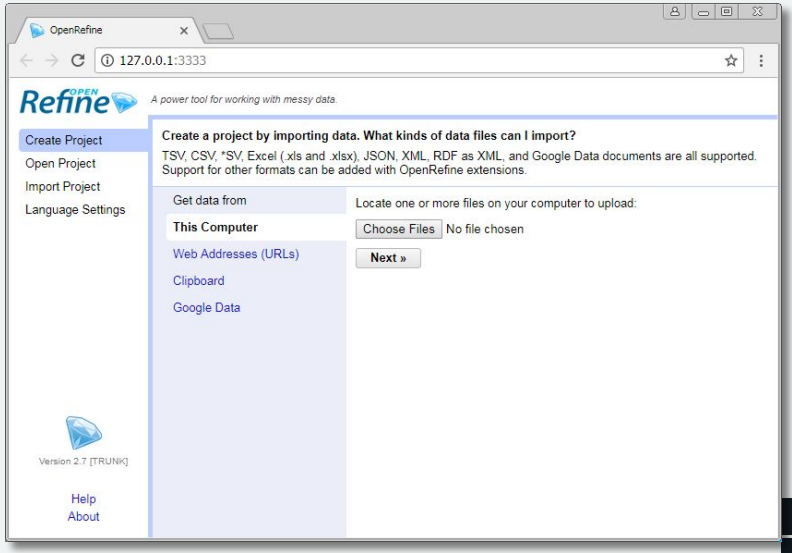
Installation

- Download OpenRefine: http://openrefine.org/download.html
- OpenRefine requires Java to run.
  - Get Java: https://java.com/en/download/
- Extract all files from the .zip folder and run **`openrefine.exe`**
- For detailed instructions see: https://docs.openrefine.org/manual/installing

Opening OpenRefine

Once the program is installed, click **`openrefine.exe`** again to launch it.

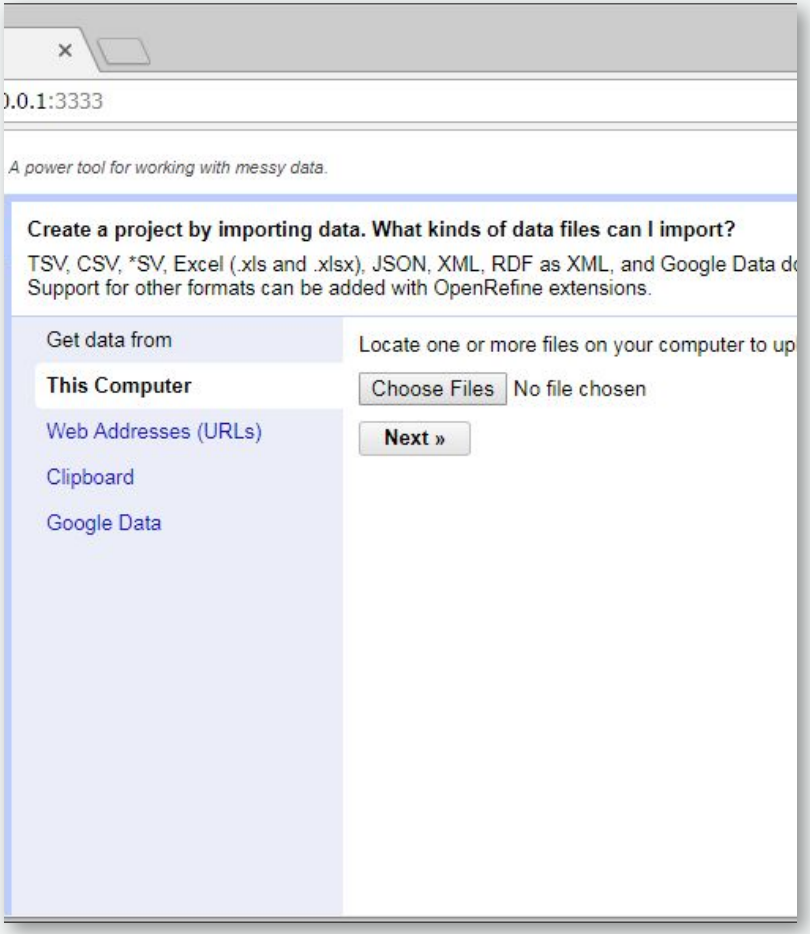
When you open the program, a black screen with code will open and launch OpenRefine in your web browser. Leave the application running to use OpenRefine.



Getting Data into OpenRefine

There are two steps:

1. Loading a file
2. Creating a project

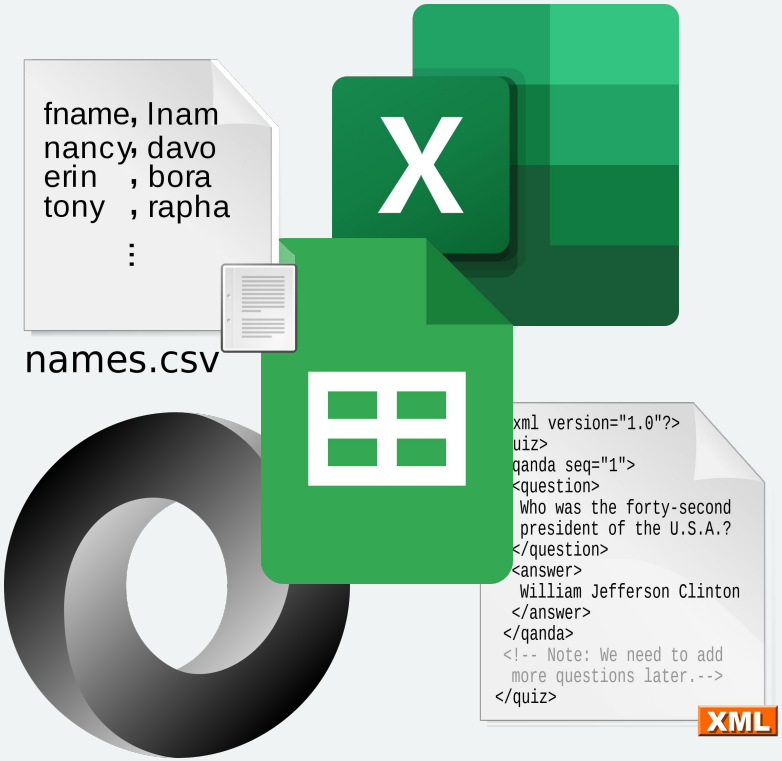


Getting Data into OpenRefine

Load File

OpenRefine can load data in a number of formats:

- Text (tsv, csv)
- Google Sheets
- Excel
- JSON
- XML
- RDF

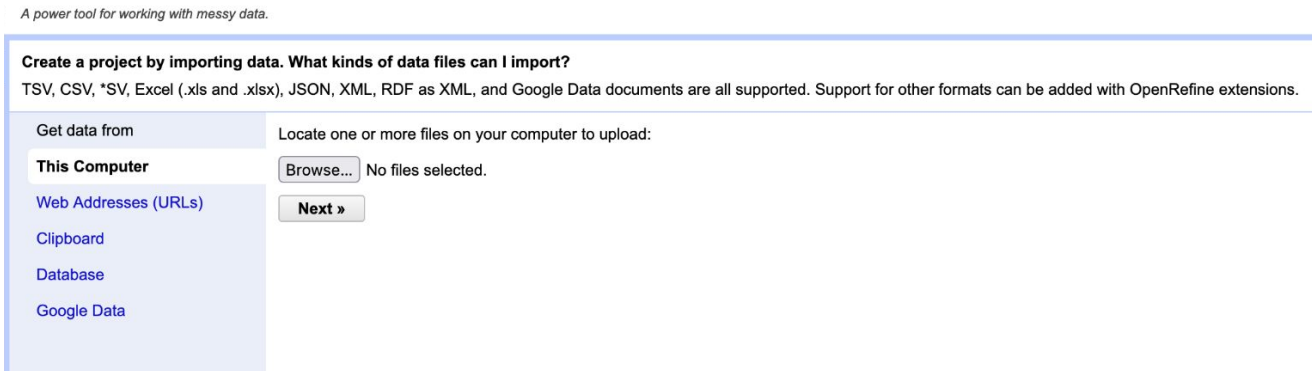


Reference:
User:Dreftymac, CC BY-SA 2.5 <https://creativecommons.org/licenses/by-sa/2.5>, via Wikimedia Commons
Google, Public domain, via Wikimedia Commons
Douglas Crockford, Public domain, via Wikimedia Commons
en:User:Dreftymac, CC BY-SA 2.5 <https://creativecommons.org/licenses/by-sa/2.5>, via Wikimedia Commons
Microsoft Corporation, Public domain, via Wikimedia Commons

Getting Data into OpenRefine

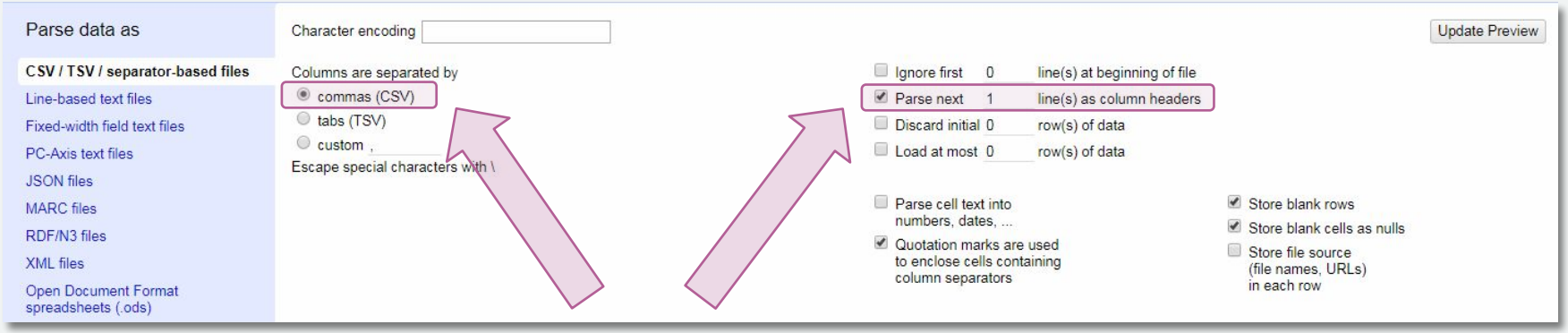
Load File

- Select the appropriate method for importing data from the Get data from menu.
  - For local data files, select This Computer.
  - Click the Browse… button and open the data file using the operating system's file manager.
- Click Next.



Load File

The following dialog will open in the bottom portion of the browser window:



Make sure the appropriate boxes are checked based on the data file's format.

Load File

The top portion of the browser window will contain a preview of the imported data.

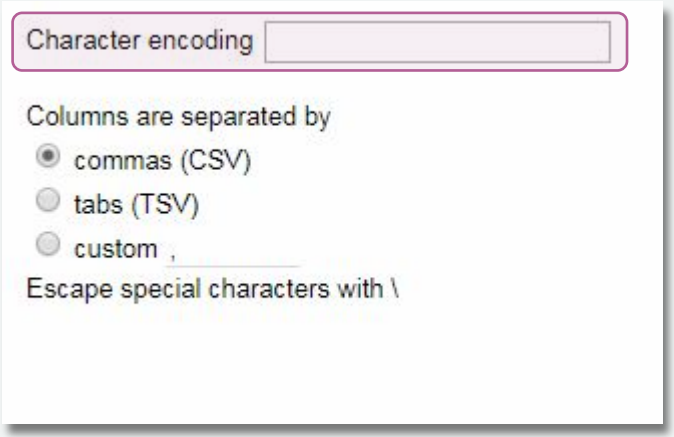
Rename the project in the upper righthand corner.

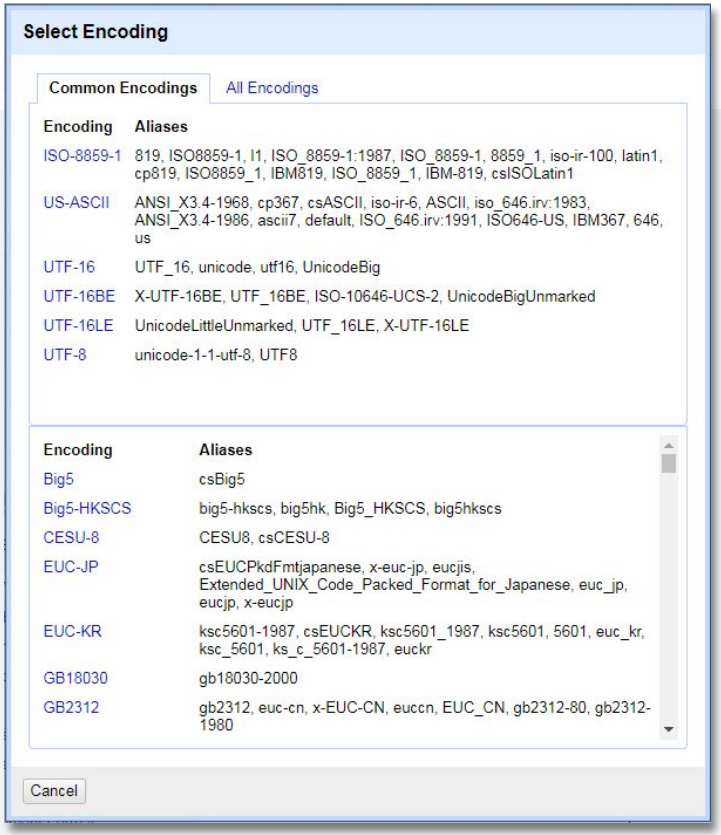
« Start Over | Configure Parsing Options | Project name WorkshopDataset csv | Create Project »

	ADP number	Cat Numb	Accesion year	ACC_N	Former number	count in lot	Specimen identifier's name	Type	Size (mm)	condition	Data	Valves	Kingdom	Superfamily	Family	Subfamily	Genus	Subg
1.	1	ABCD:1	1993	9999		1	Heiser, J., 1993	41	92		None			Pleurotomarioidea	Haliotidae		Haliotis	
2.	2	ABCD:2	1993	9999		1	Heiser, J., 1993	41	34		None		Animalia	Pleurotomarioidea	Haliotidae		Haliotis	
3.	3	ABCD:3	1993	9999		1	Heiser, J., 1993	41	15		None		Animalia	Trochoidea	Trochidae	Trochinae	Clanculus	
4.	4	ABCD:4	1993	9999		2	Heiser, J., 1993	41	15		None		Animalia	Trochoidea	Trochidae	Calliostomatinae	Calliostoma	
5.	5	ABCD:5	1993	9999		1	Heiser, J., 1993	41	20		None		Animalia	Trochoidea			Unidentified	
6.	6	ABCD:6	1993	9999		1	Heiser, J., 1993	41	55		None		Animalia	Trochoidea	Trochidae	Gibbulinae	Cittarium	
7.	7	ABCD:7	1993	9999		3	Per label supplied	41	17		None		Animalia	Trochoidea	Trochidae	Monodontinae	Monodonta	
8.	8	ABCD:8	1993	9999		1	Heiser, J., 1995	41	19	Good	None		Animalia	Trochoidea	Trochidae	Monodontinae	Tegula	
9.	9	ABCD:9	1993	9999		7	Heiser, J., 1995	43	44	Good	Basic		Animalia	Lymnaeoidea	Lymnaeidae		Lymnaea	
10.	10	ABCD:10	1993	9999		1	Heiser, J. 1993	41	28		None		Animalia	Cerithioidea	Vermetidae		Dendropoma	

Load File

Make sure the right character encoding is selected.



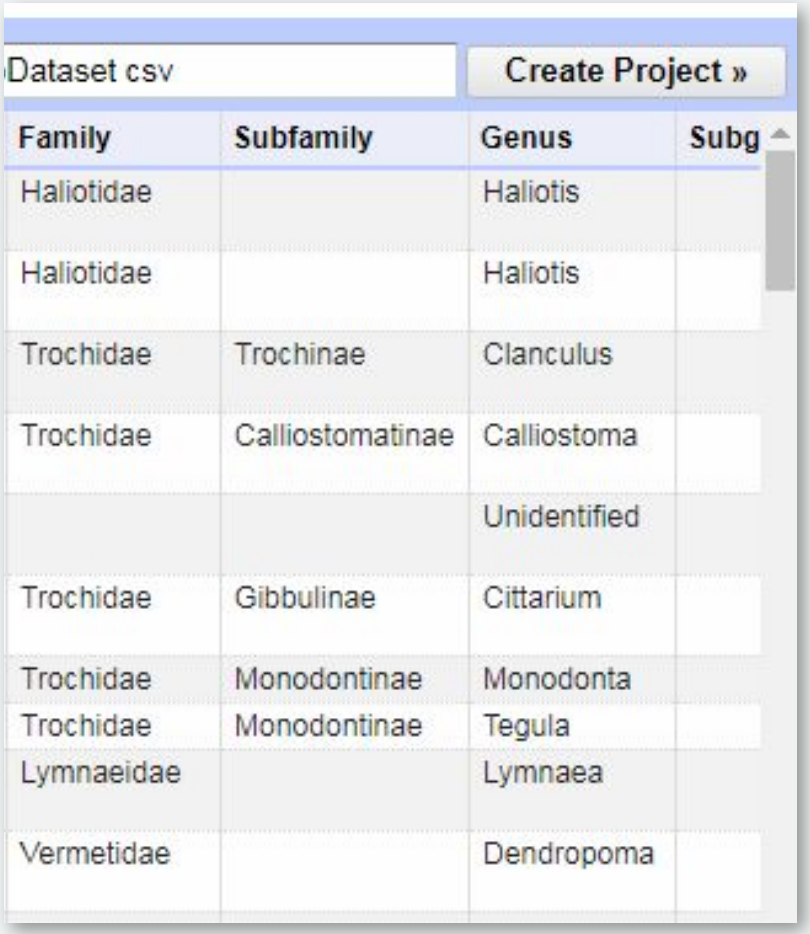


Getting Data into OpenRefine

Create Project

A project is a combination of:

1. The imported data
2. Code that transforms the data



Create Project

The top portion of the browser window will contain a preview of the imported data.

Click Create Project.

« Start Over | Configure Parsing Options | Project name WorkshopDataset csv | Create Project »

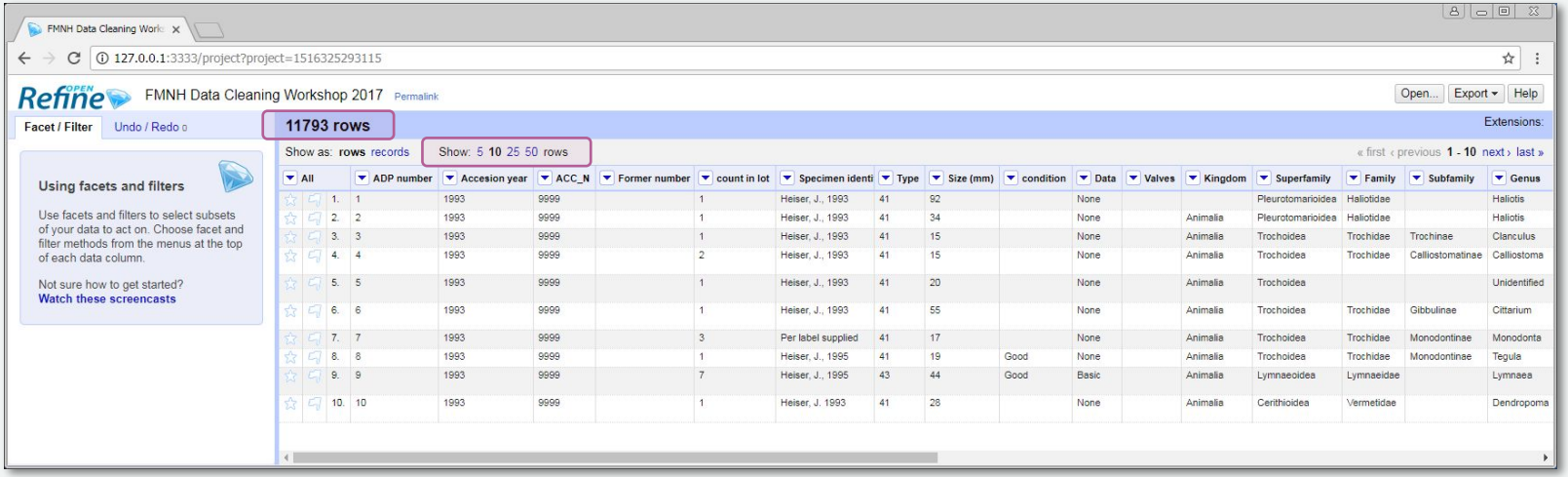
	ADP number	Cat Numb	Accesion year	ACC_N	Former number	count in lot	Specimen identifier's name	Type	Size (mm)	condition	Data	Valves	Kingdom	Superfamily	Family	Subfamily	Genus	Subg
1.	1	ABCD:1	1993	9999		1	Heiser, J., 1993	41	92		None			Pleurotomarioidea	Haliotidae		Haliotis	
2.	2	ABCD:2	1993	9999		1	Heiser, J., 1993	41	34		None		Animalia	Pleurotomarioidea	Haliotidae		Haliotis	
3.	3	ABCD:3	1993	9999		1	Heiser, J., 1993	41	15		None		Animalia	Trochoidea	Trochidae	Trochinae	Clanculus	
4.	4	ABCD:4	1993	9999		2	Heiser, J., 1993	41	15		None		Animalia	Trochoidea	Trochidae	Calliostomatinae	Calliostoma	
5.	5	ABCD:5	1993	9999		1	Heiser, J., 1993	41	20		None		Animalia	Trochoidea			Unidentified	
6.	6	ABCD:6	1993	9999		1	Heiser, J., 1993	41	55		None		Animalia	Trochoidea	Trochidae	Gibbulinae	Cittarium	
7.	7	ABCD:7	1993	9999		3	Per label supplied	41	17		None		Animalia	Trochoidea	Trochidae	Monodontinae	Monodonta	
8.	8	ABCD:8	1993	9999		1	Heiser, J., 1995	41	19	Good	None		Animalia	Trochoidea	Trochidae	Monodontinae	Tegula	
9.	9	ABCD:9	1993	9999		7	Heiser, J., 1995	43	44	Good	Basic		Animalia	Lymnaeoidea	Lymnaeidae		Lymnaea	
10.	10	ABCD:10	1993	9999		1	Heiser, J. 1993	41	28		None		Animalia	Cerithioidea	Vermetidae		Dendropoma	

Create Project

The project will be displayed in a window like the one shown below:

The top bar displays the total number of rows of data.

The program displays ten rows per page by default, but this setting is adjustable.



	All	ADP number	Accesion year	ACC_N	Former number	count in lot	Specimen identi	Type	Size (mm)	condition	Data	Valves	Kingdom	Superfamily	Family	Subfamily	Genus
1.	1	1	1993	9999		1	Heiser, J., 1993	41	92		None			Pleurotomarioidea	Haliotidae		Haliotis
2.	2	2	1993	9999		1	Heiser, J., 1993	41	34		None		Animalia	Pleurotomarioidea	Haliotidae		Haliotis
3.	3	3	1993	9999		1	Heiser, J., 1993	41	15		None		Animalia	Trochoidea	Trochidae	Trochinae	Clanculus
4.	4	4	1993	9999		2	Heiser, J., 1993	41	15		None		Animalia	Trochoidea	Trochidae	Calliostomatinae	Calliostoma
5.	5	5	1993	9999		1	Heiser, J., 1993	41	20		None		Animalia	Trochoidea			Unidentified
6.	6	6	1993	9999		1	Heiser, J., 1993	41	55		None		Animalia	Trochoidea	Trochidae	Gibbulinae	Cittarium
7.	7	7	1993	9999		3	Per label supplied	41	17		None		Animalia	Trochoidea	Trochidae	Monodontinae	Monodonta
8.	8	8	1993	9999		1	Heiser, J., 1995	41	19	Good	None		Animalia	Trochoidea	Trochidae	Monodontinae	Tegula
9.	9	9	1993	9999		7	Heiser, J., 1995	43	44	Good	Basic		Animalia	Lymnaeoidea	Lymnaeidae		Lymnaea
10.	10	10	1993	9999		1	Heiser, J. 1993	41	28		None		Animalia	Cerithioidea	Vermetidae		Dendropoma

Manipulating/ Mangling Data

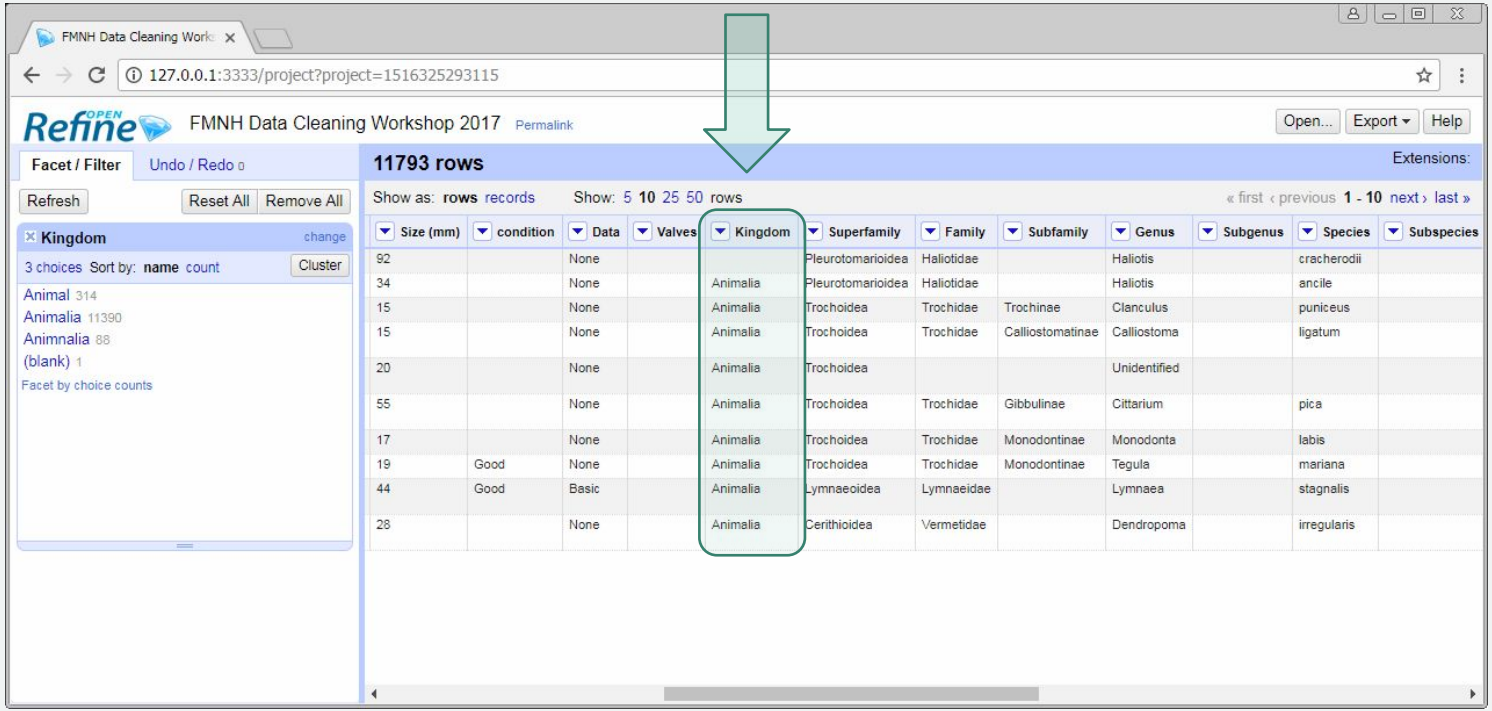
Faceting

Faceting is a feature that allows users to:

1. Find anomalies
2. Group records

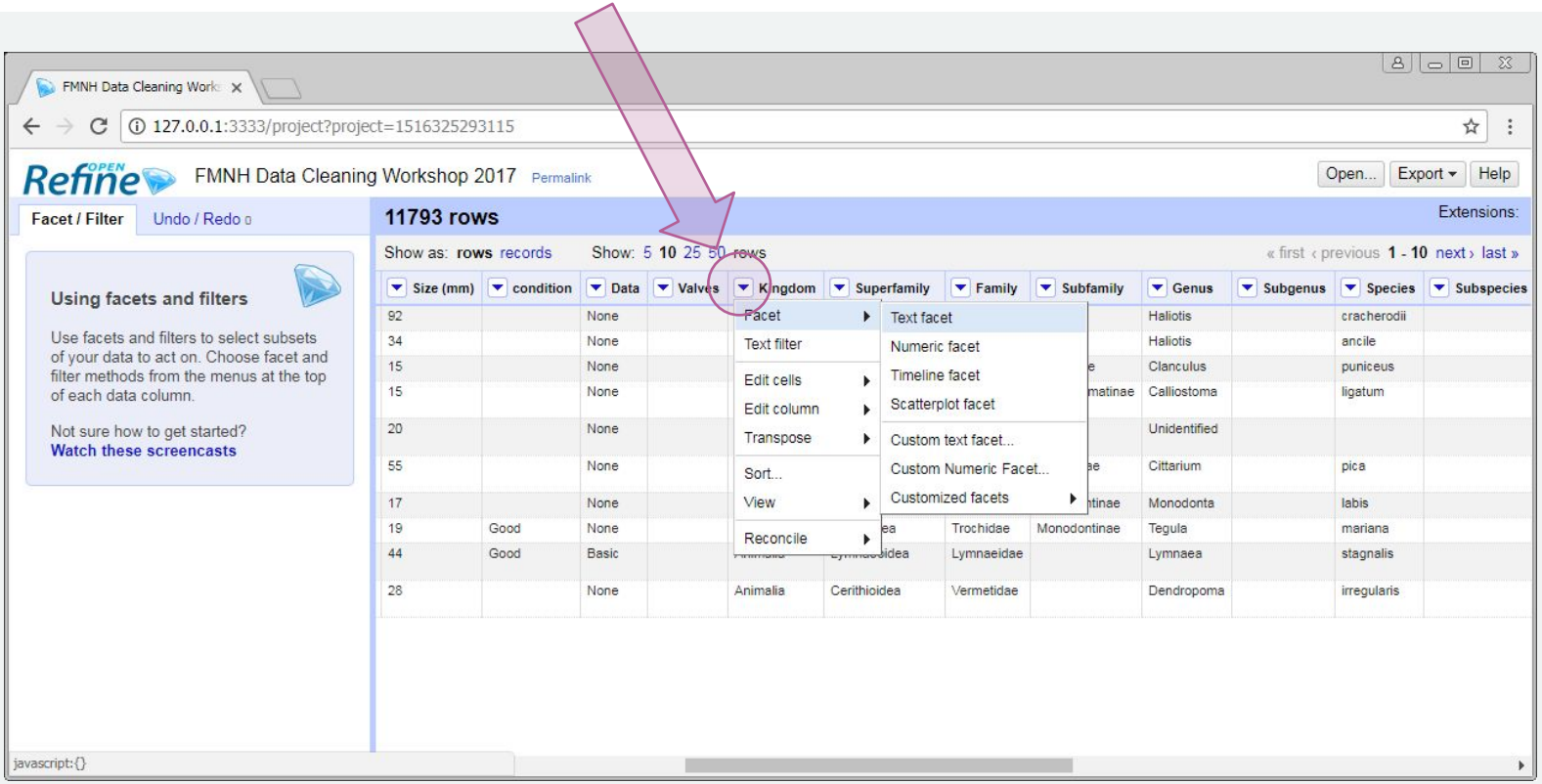
Faceting

Select a column to start working with.



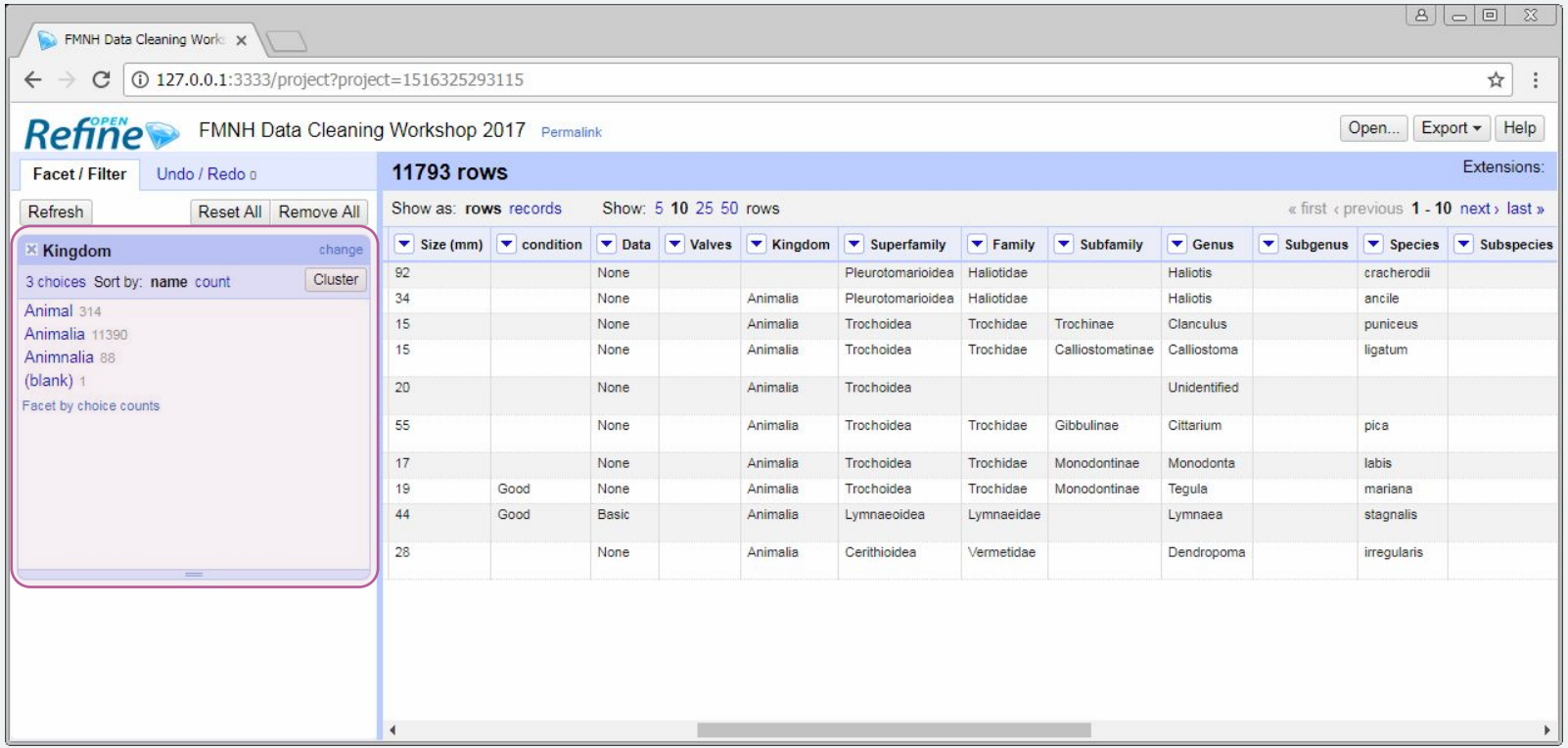
Faceting

Click the drop-down icon in the corner of the column header, and select Text facet from the Facet options.



Faceting

A new panel with the different data values from the column with open in the left-hand sidebar. Users can sort by (name or count) or edit these values.

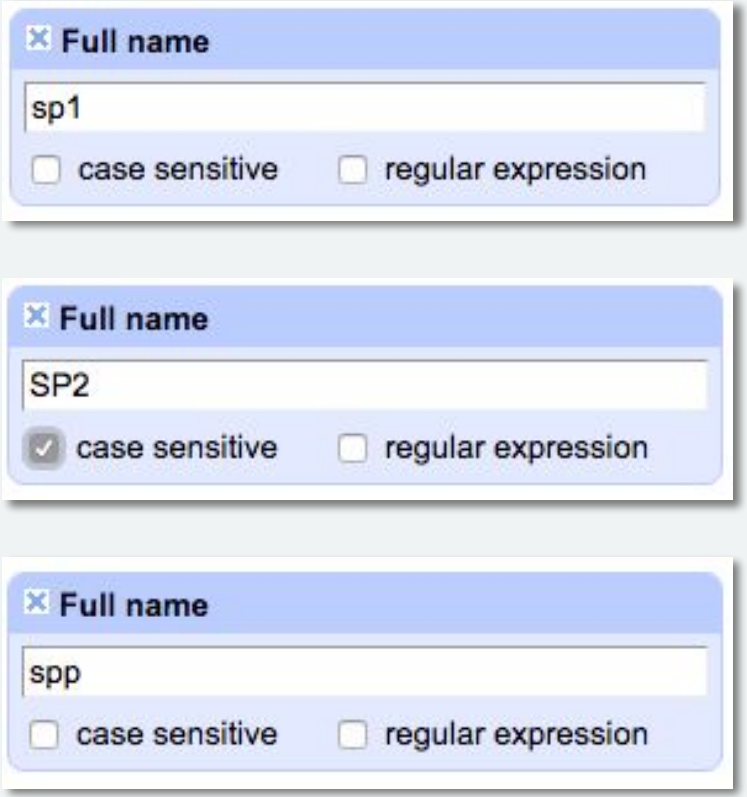


Manipulating/ Mangling Data

Filtering

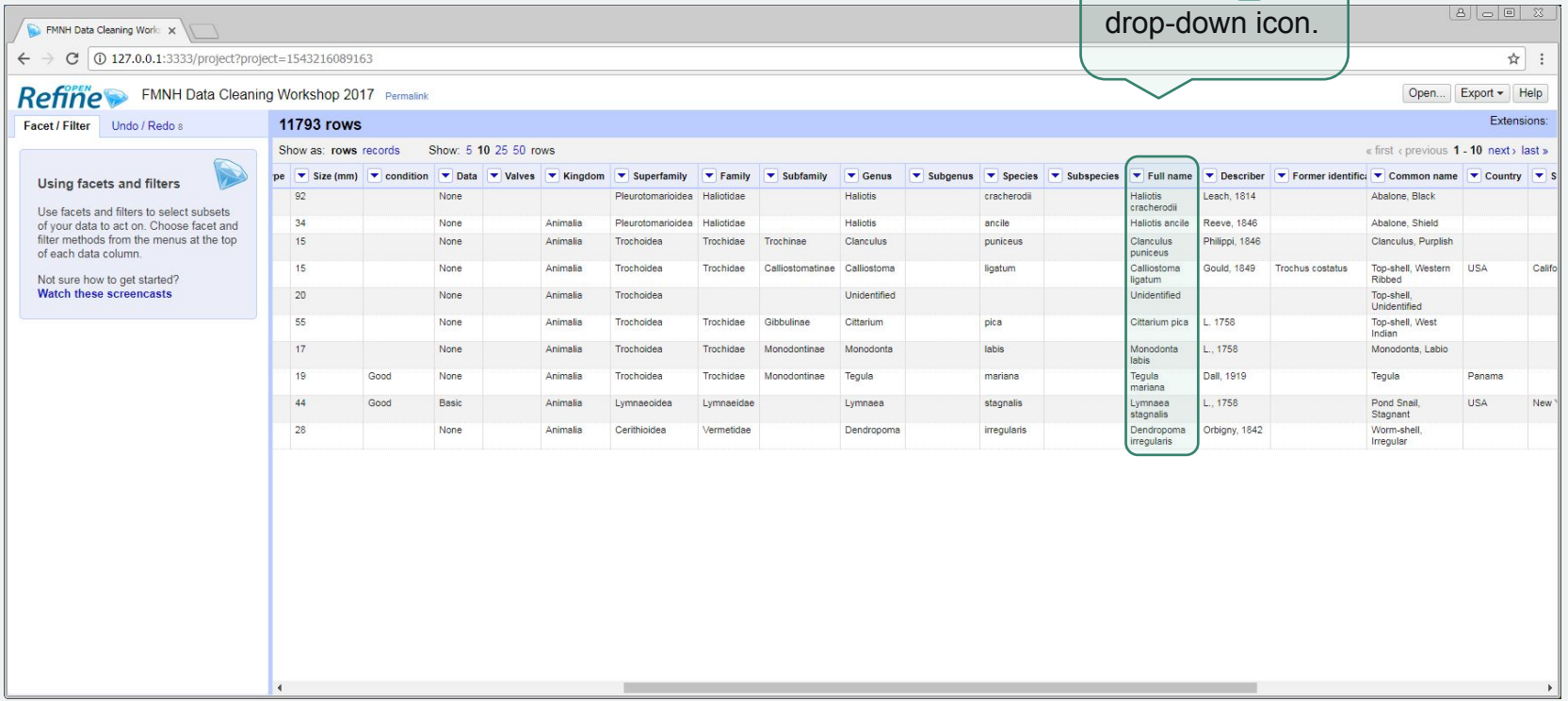
Filtering is a feature that allows users to:

1. Search for values
2. Group records



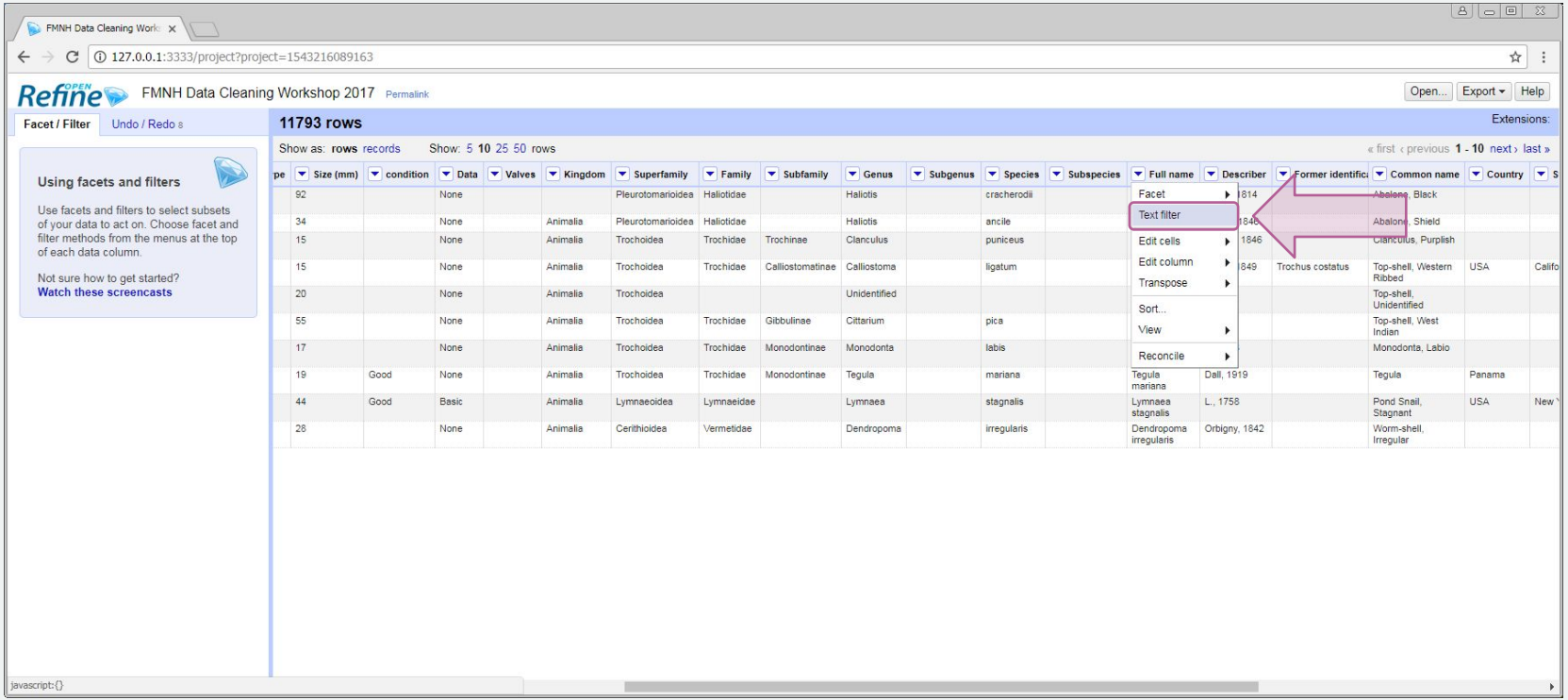
Filtering

Go to column Full name and perform a Text filter.



Filtering

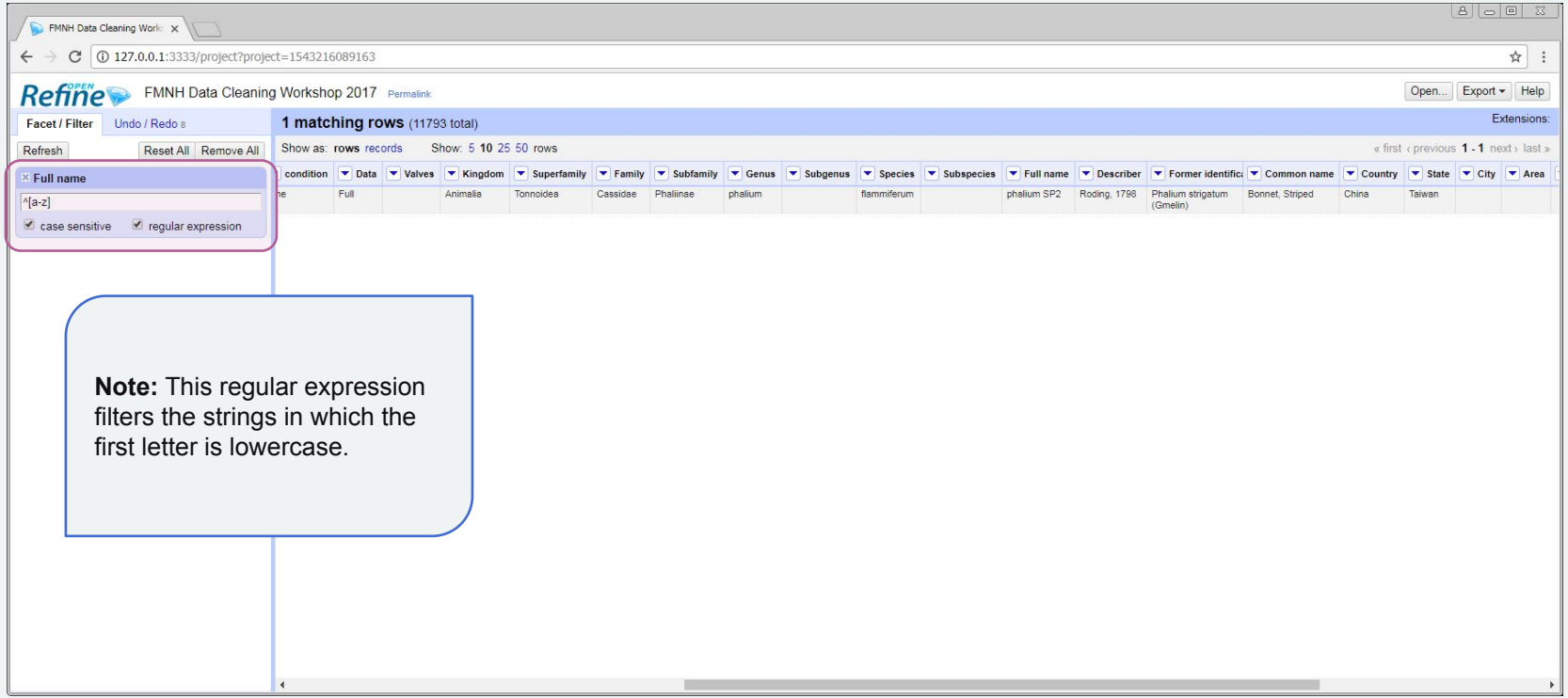
Text filter is the second option in the column menu.



Filtering

Check regular expression and case sensitive, then paste the expression:

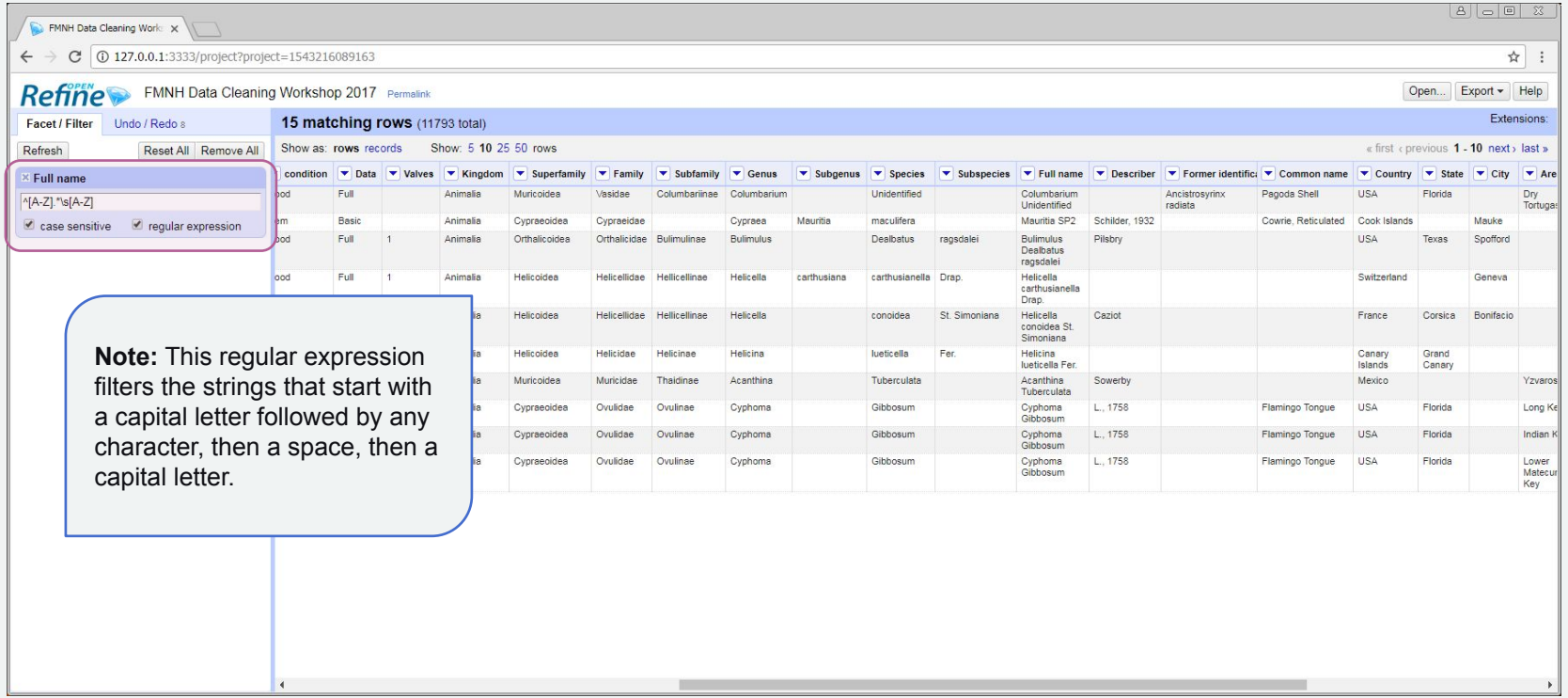
```
^[a-z]
```



Filtering

Check regular expression and case sensitive, then paste the expression:

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

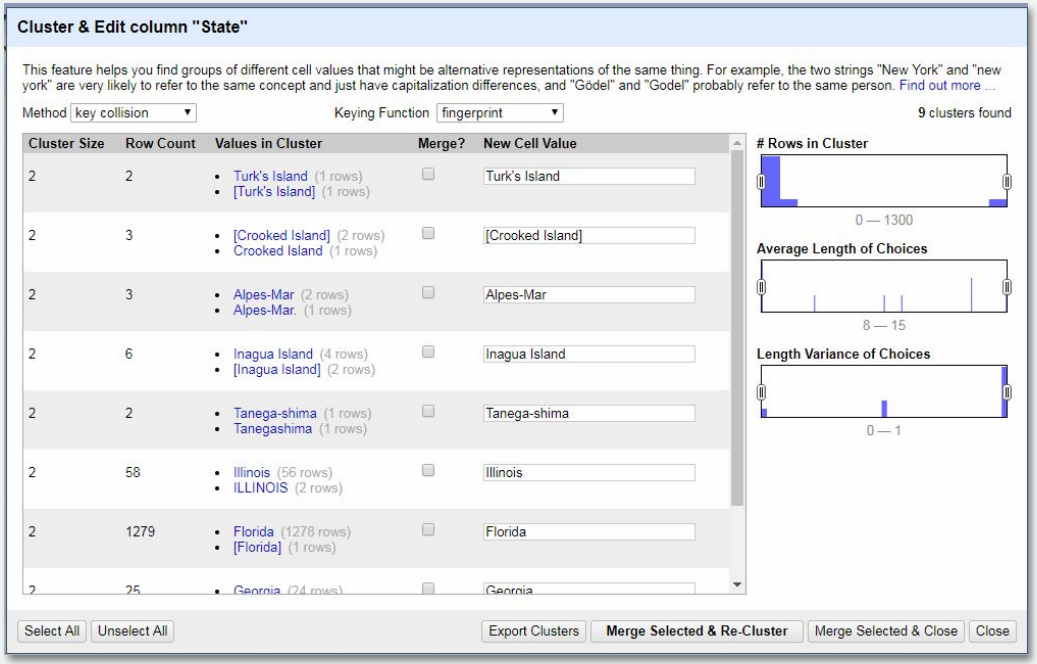


Manipulating/ Mangling Data

Clustering

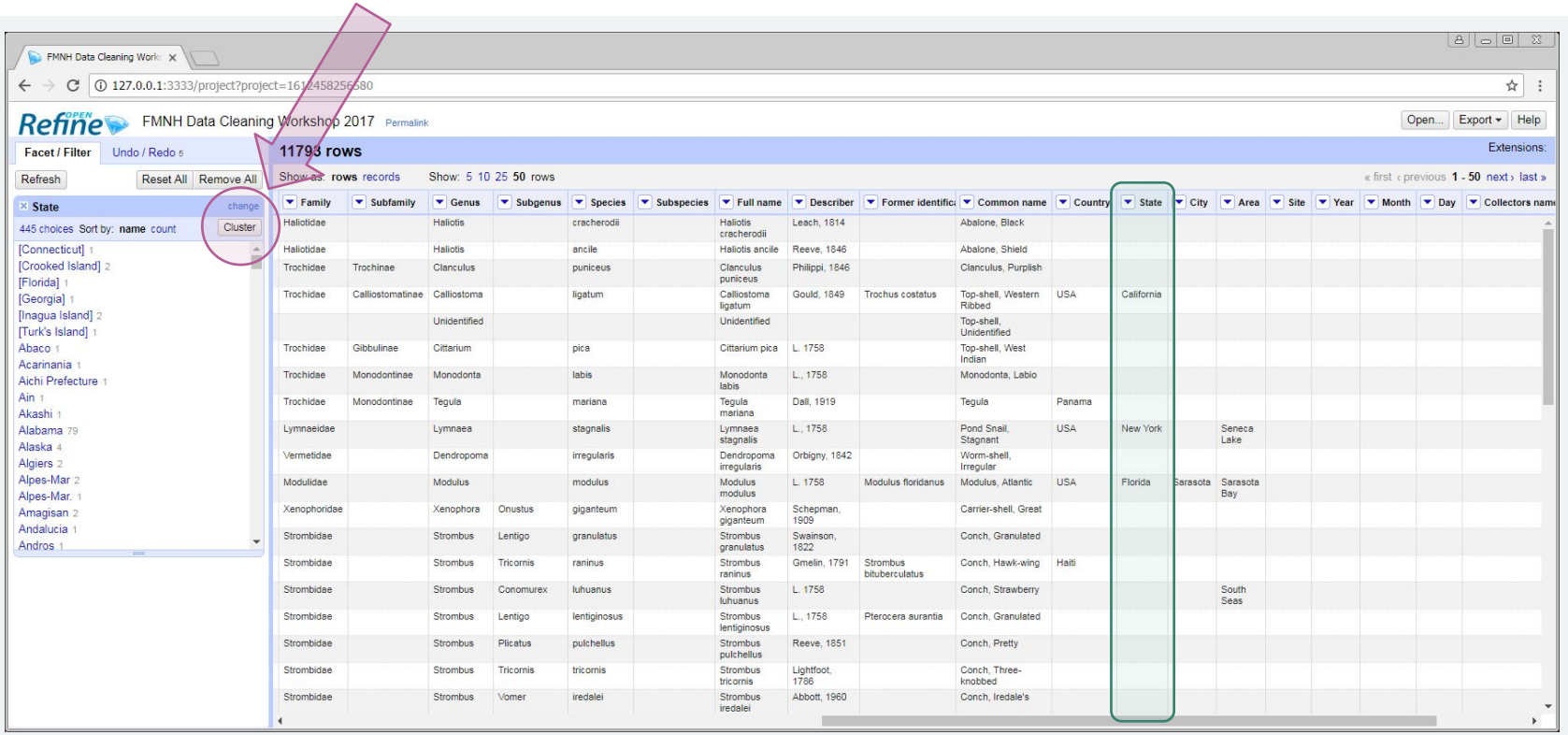
Clustering is a feature that allows users to:

1. Find different versions
2. Group records



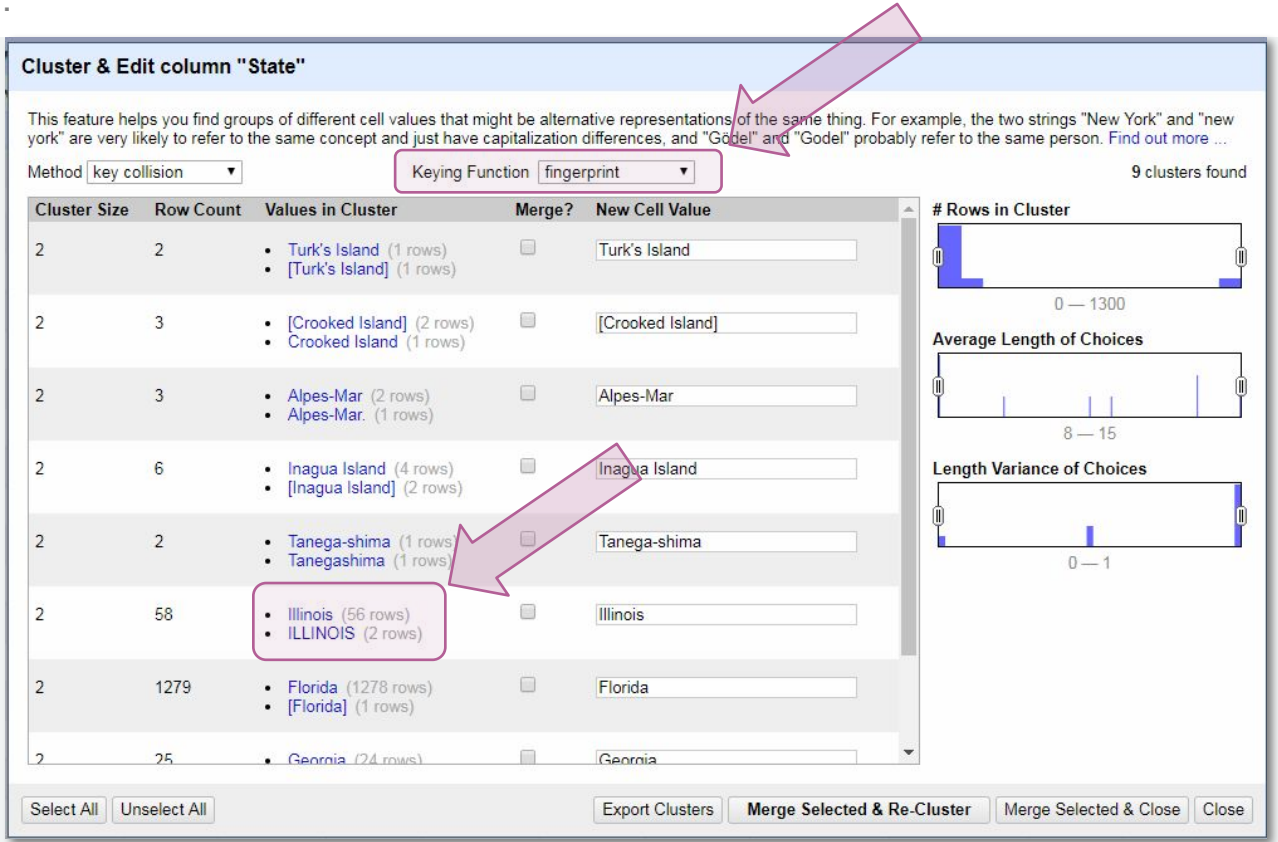
Clustering

When a column is faceted, click the Cluster button in the right-hand corner of the sidebar panel.



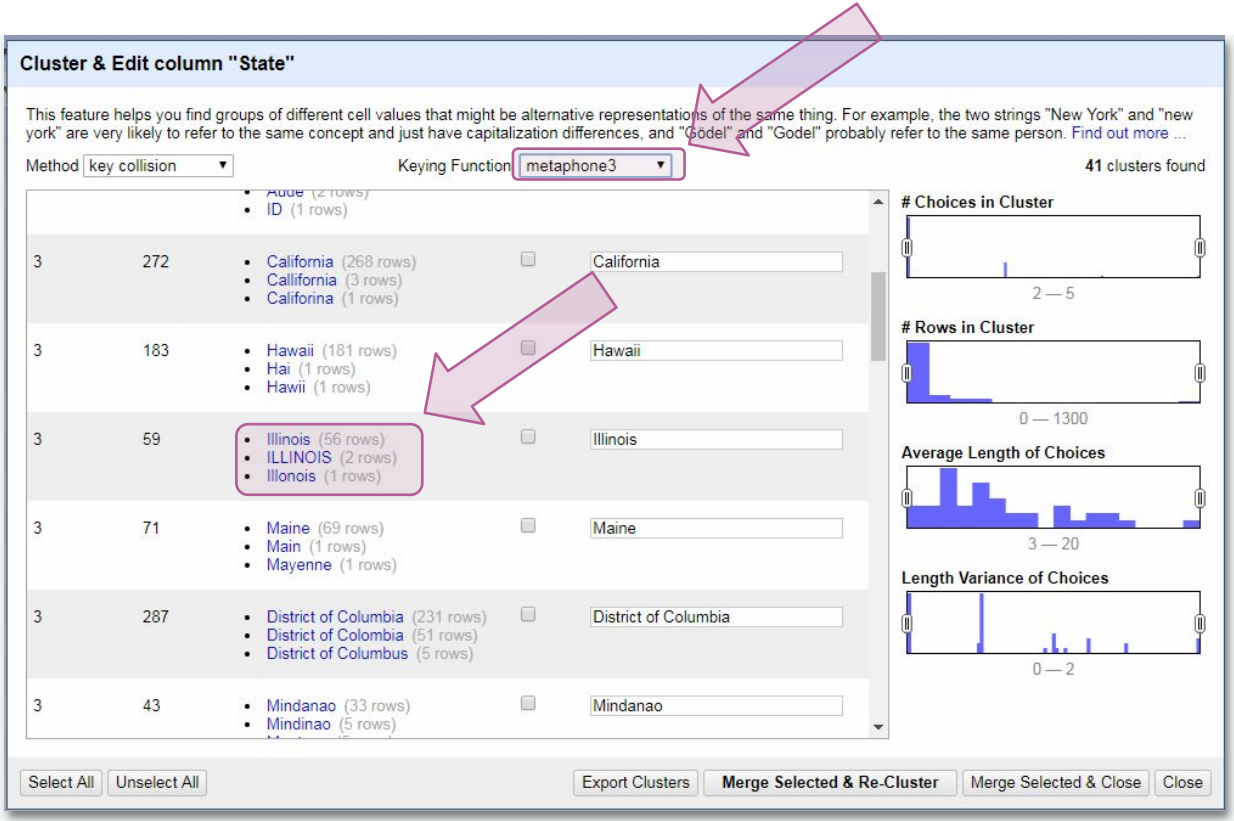
Clustering

The pop-up window displays values that may refer to the same thing. Users can use this screen to standardize them.



Clustering

The Keying Function determines how the values are grouped. Use “metaphone3” to broaden the clusters.

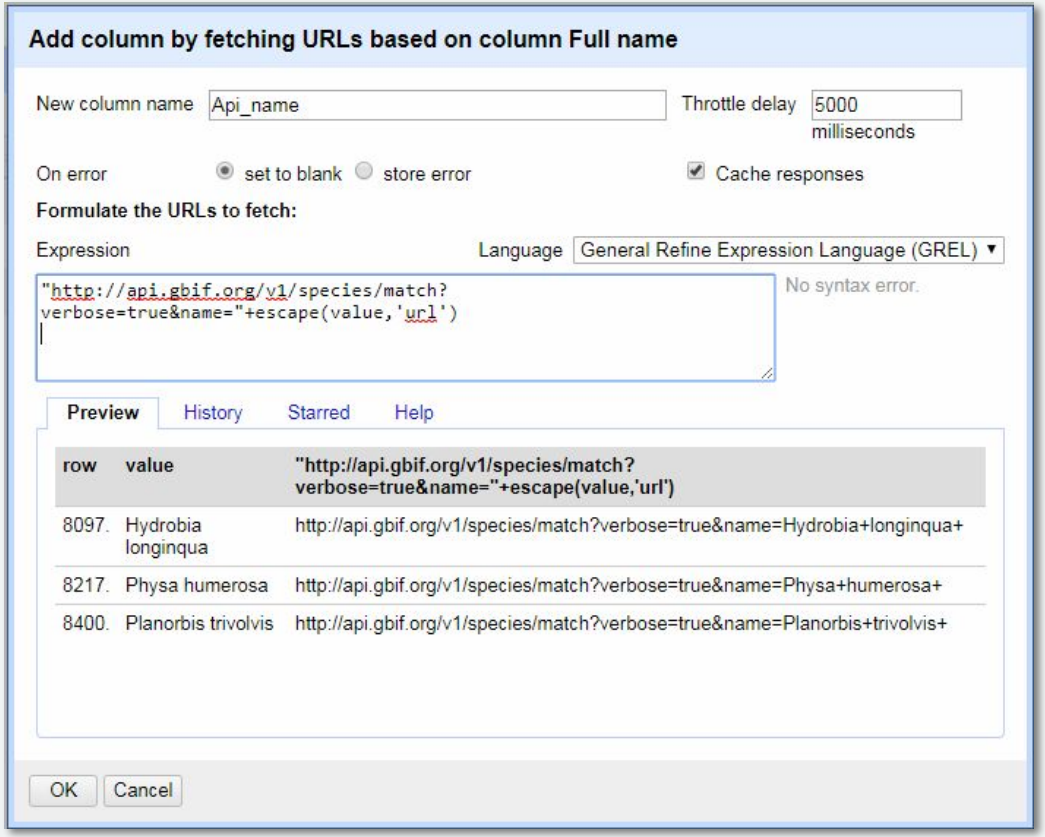


Manipulating/ Mangling Data

Reconciliation

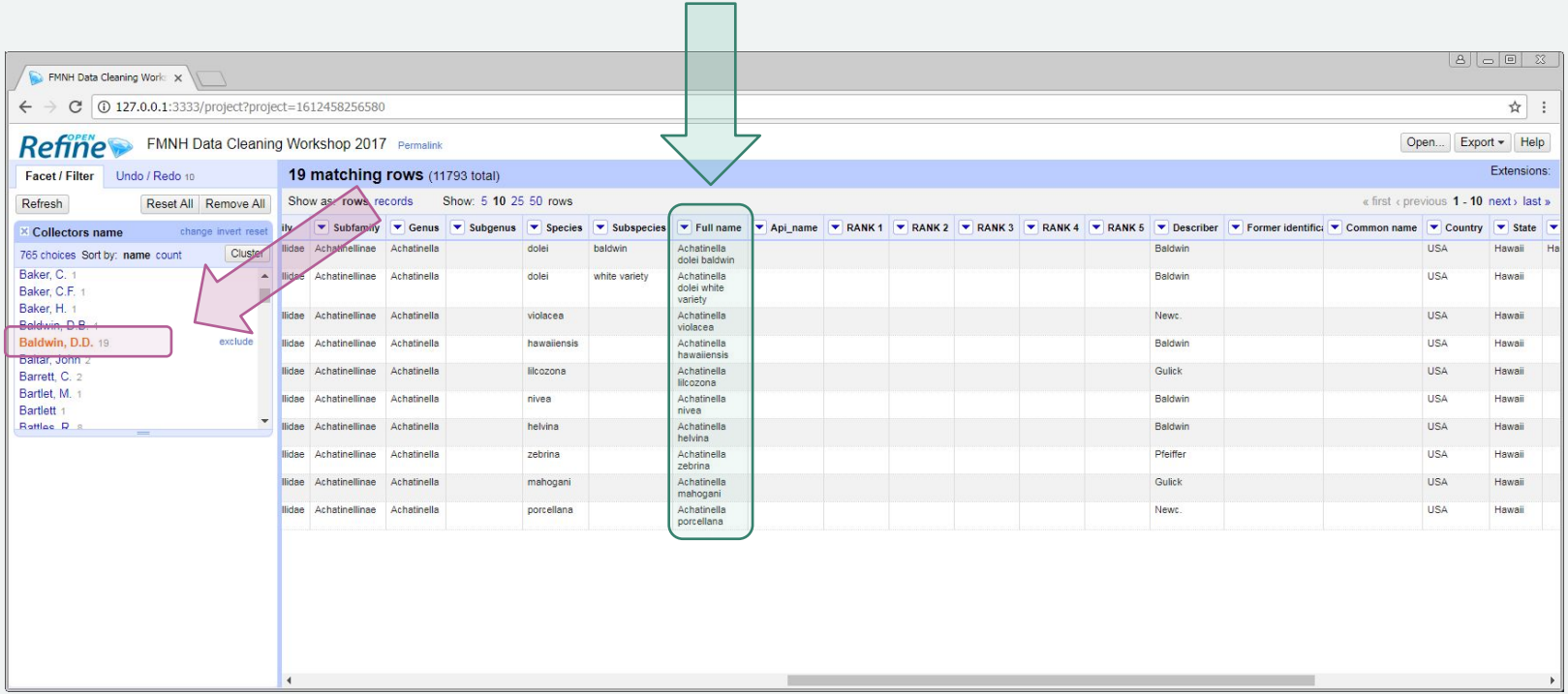
Reconciliation is a feature that allows users to:

1. Check data against an external source
2. Import data from an external source



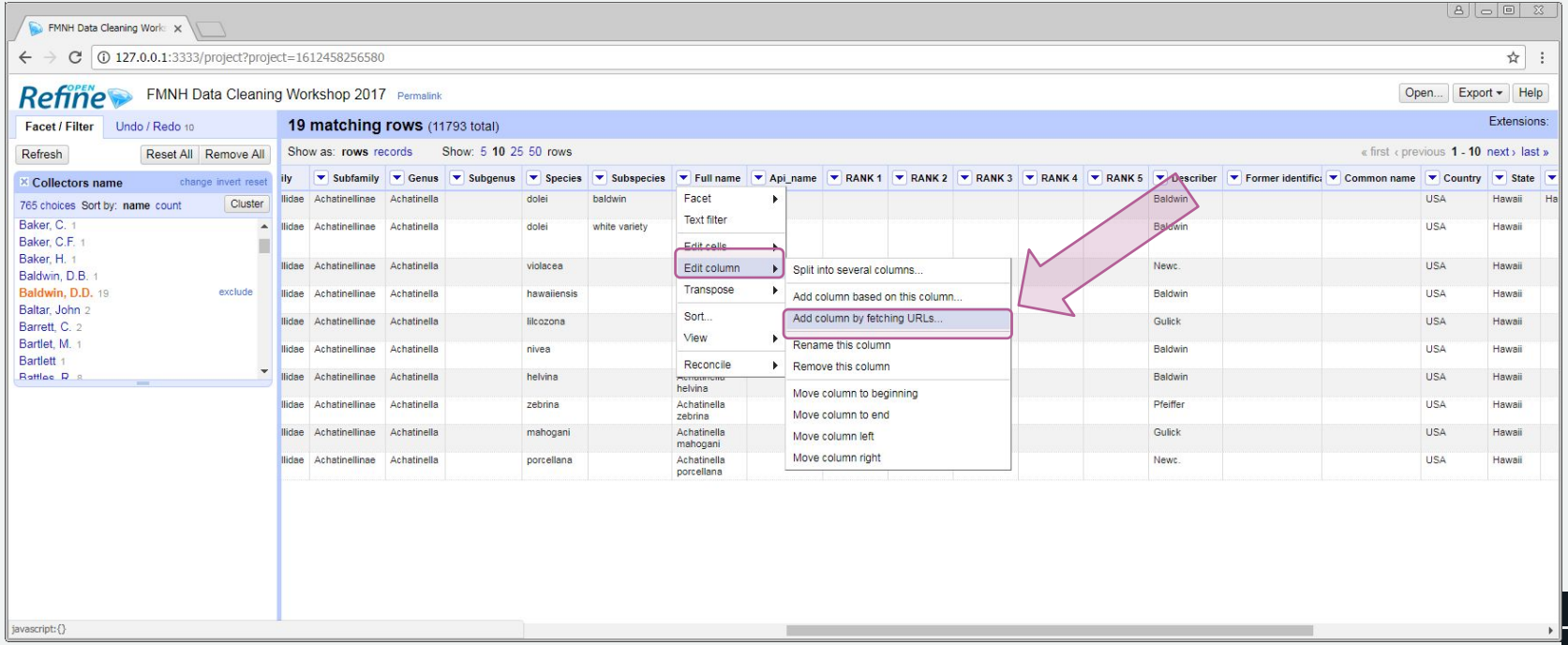
Reconciliation

Select a column and use the facet feature to choose a set of entries.



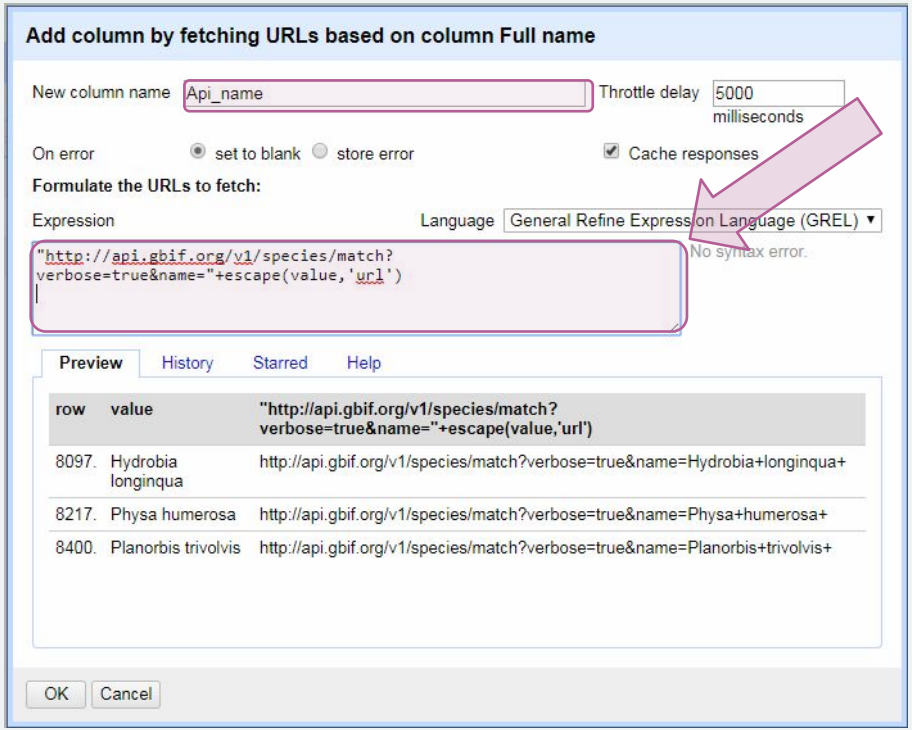
Reconciliation

Click the drop-down icon. Under Edit column select Add column by fetching URLs.



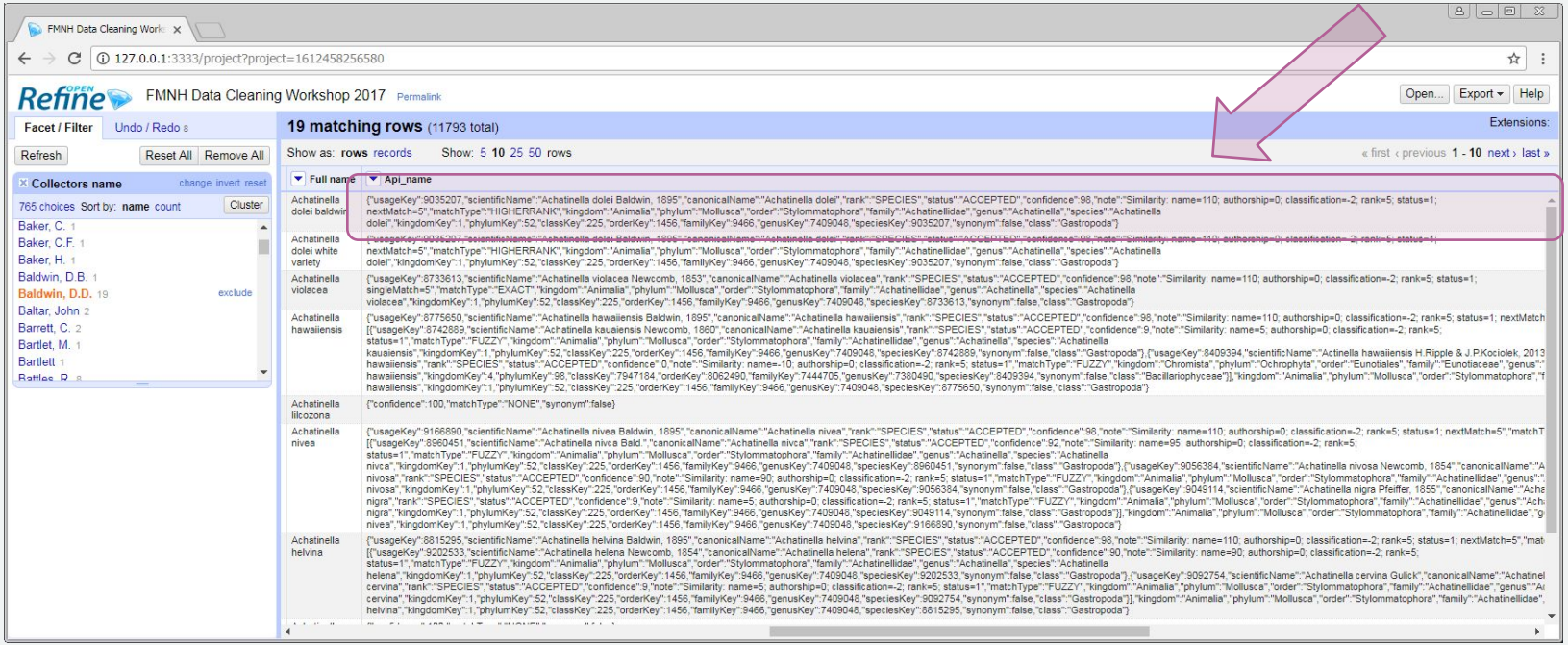
Reconciliation

A new box will open. Enter the URL to draw information from, as shown below.



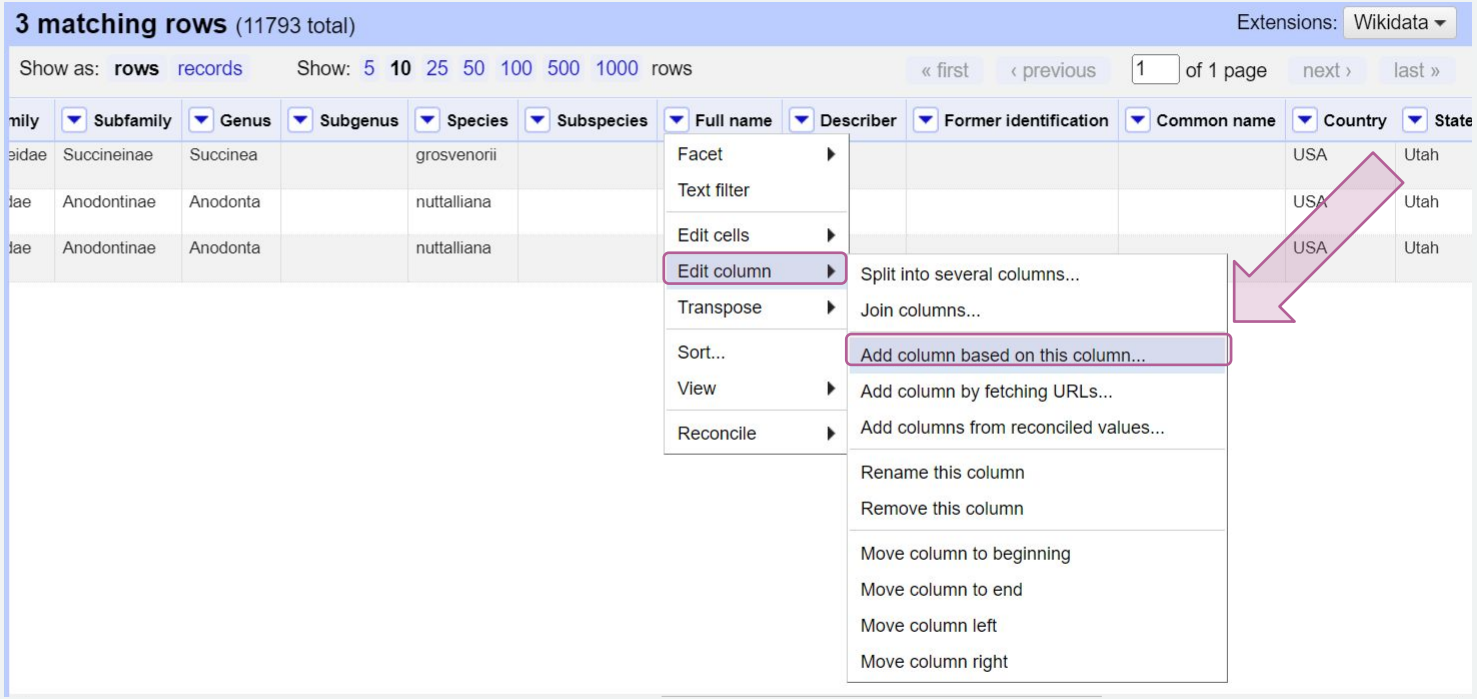
Reconciliation

The data from API will be placed in a new column.



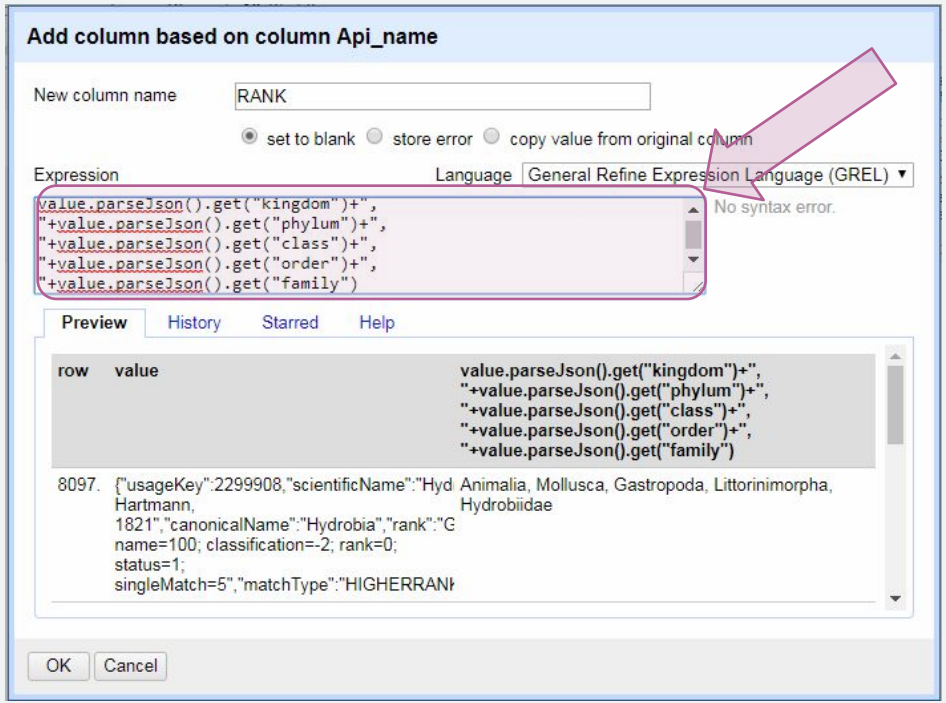
Reconciliation

To extract only the necessary columns, select Add column based on this column from the Edit column options.



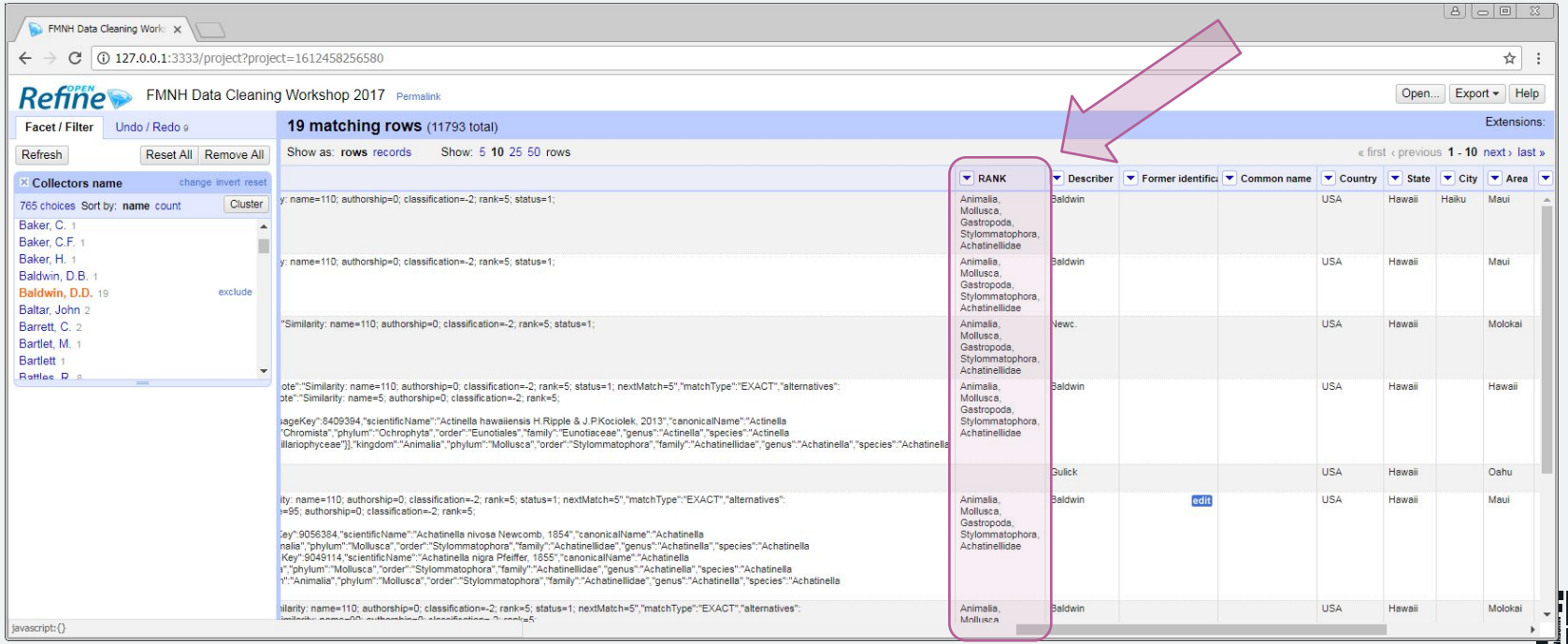
Reconciliation

In the Expression field, use the equation `value.parseJson().get(“ ”)` to select the categories for the program to separate by.



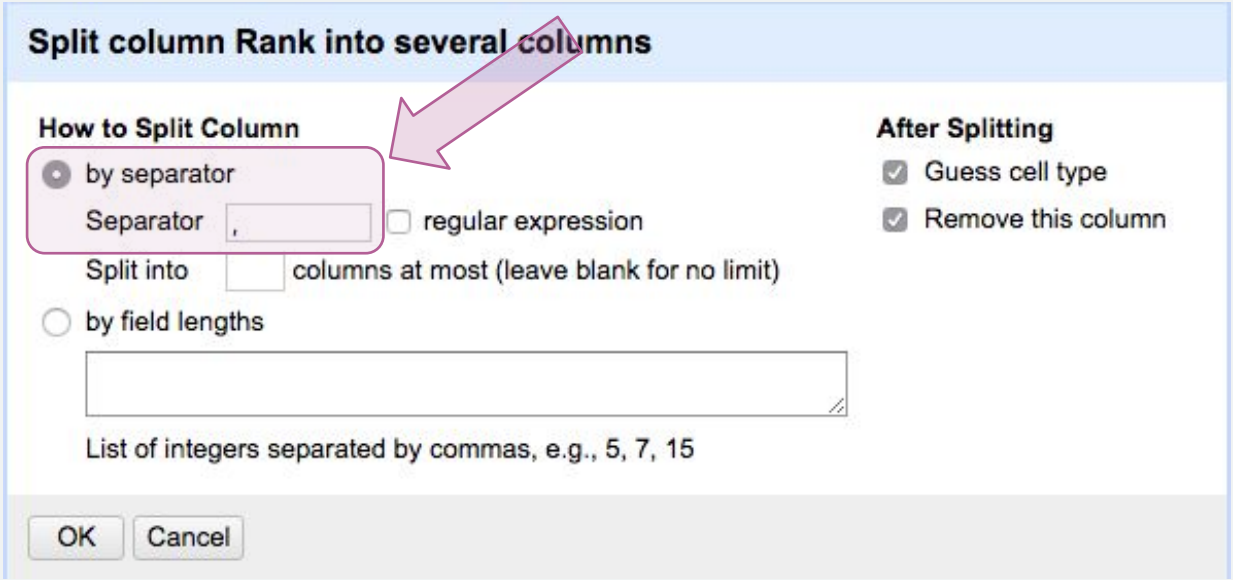
Reconciliation

The data will be placed in a new column.



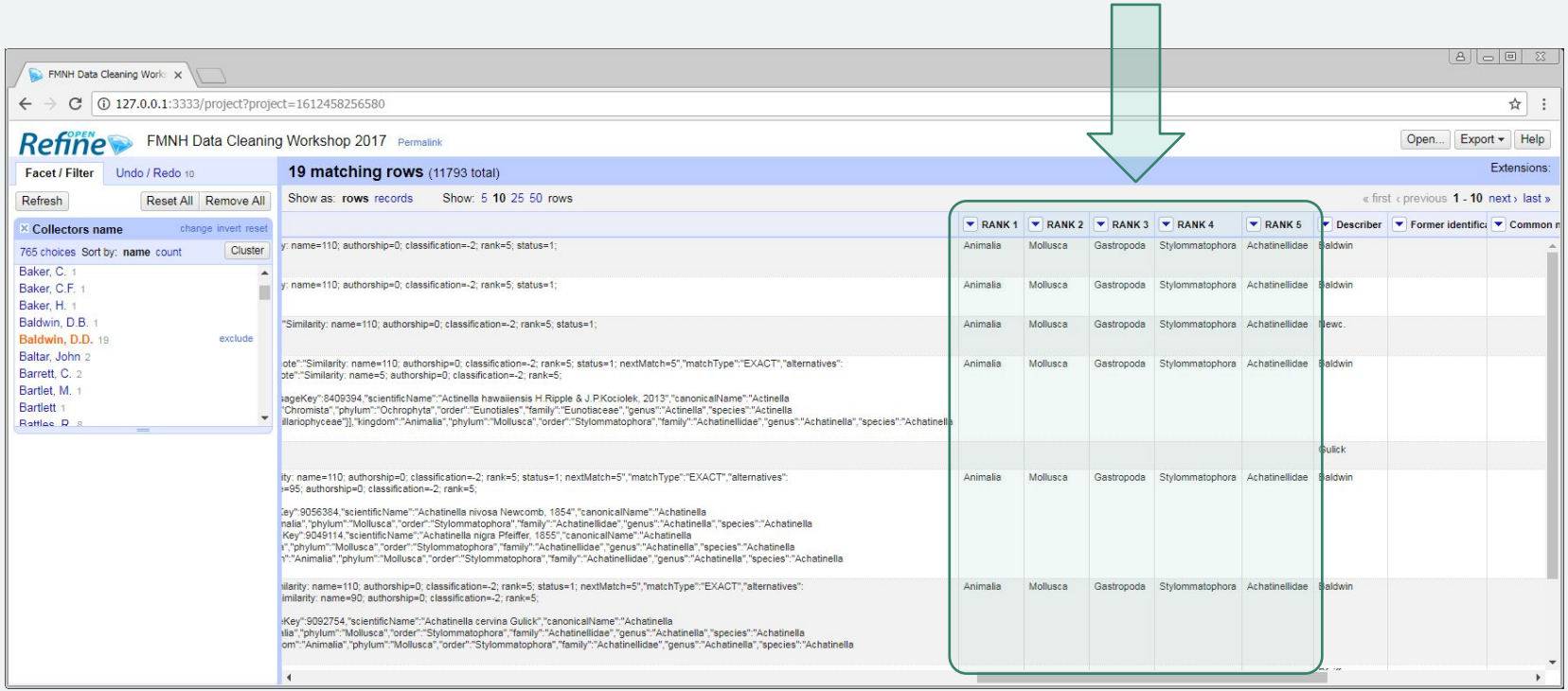
Reconciliation

Under Edit column, you can also select Split into several columns…. Choose a character to separate on. A common separator, or delimiter, is a comma.



Reconciliation

The data will be placed in new separate columns.



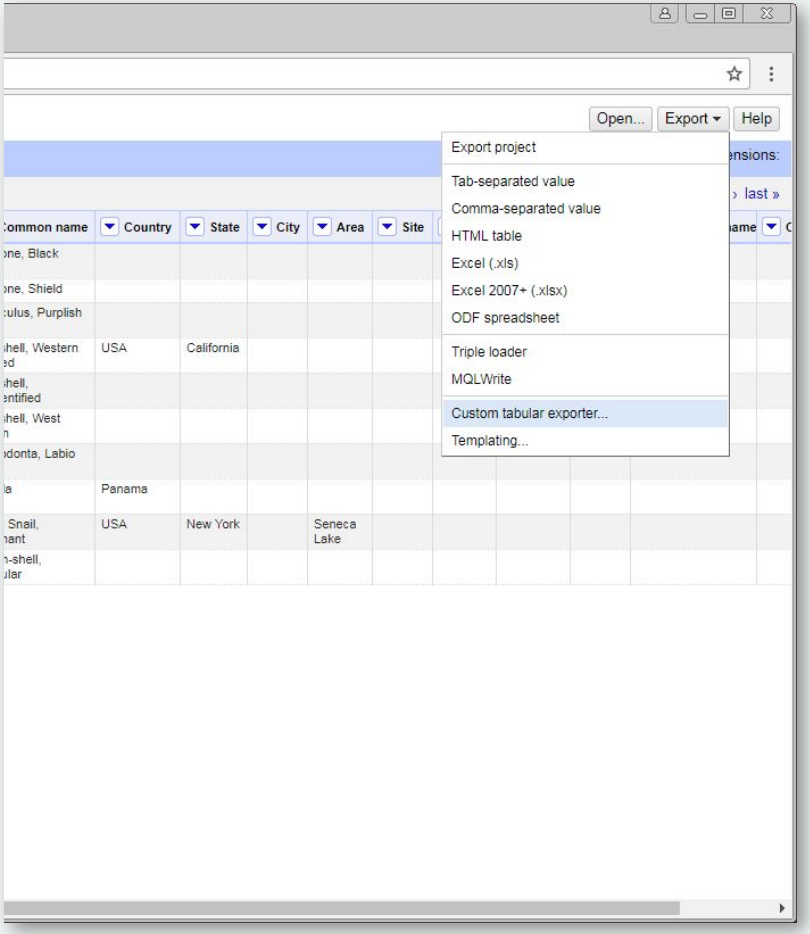
Getting Data out of OpenRefine

Export

Exporting allows you to:

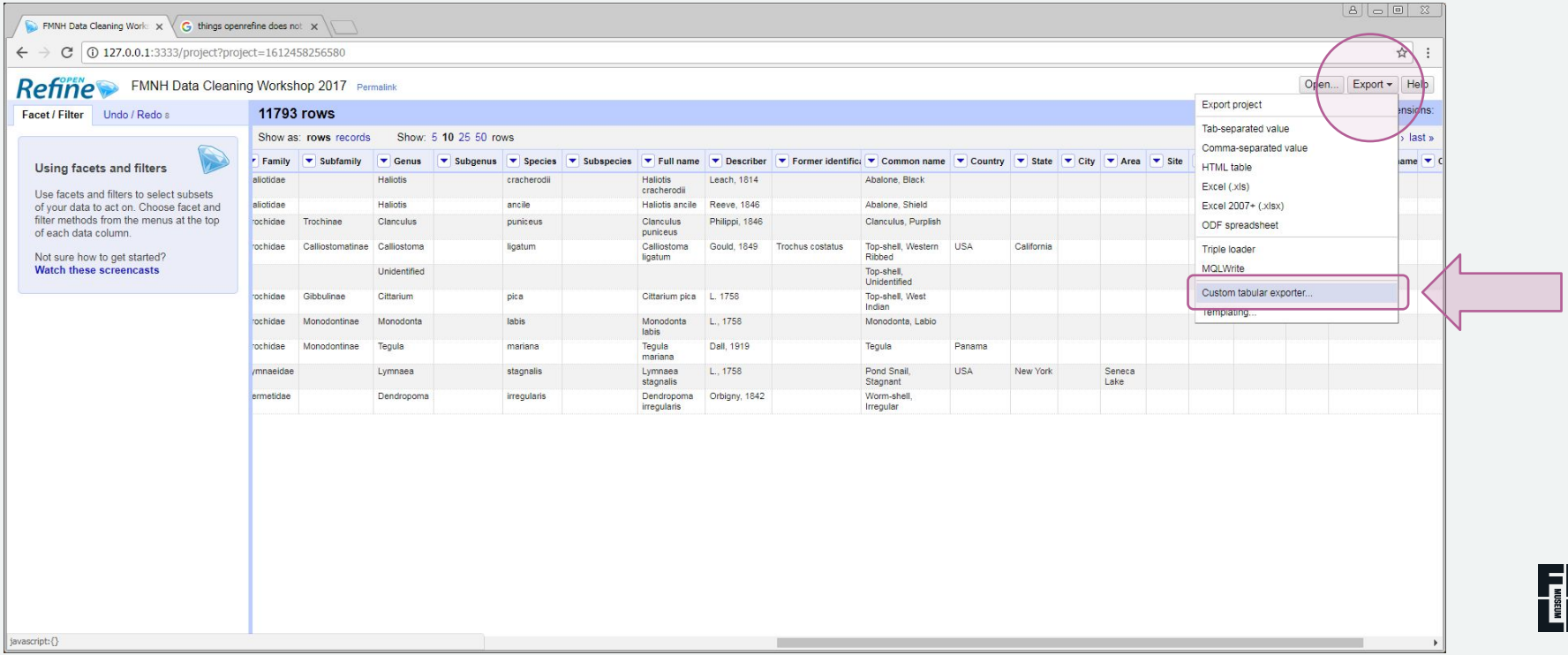
1. Get only the fields you want in the file
2. Format the export file

There are several options for exporting cleaned data, but the following option is useful in most cases.

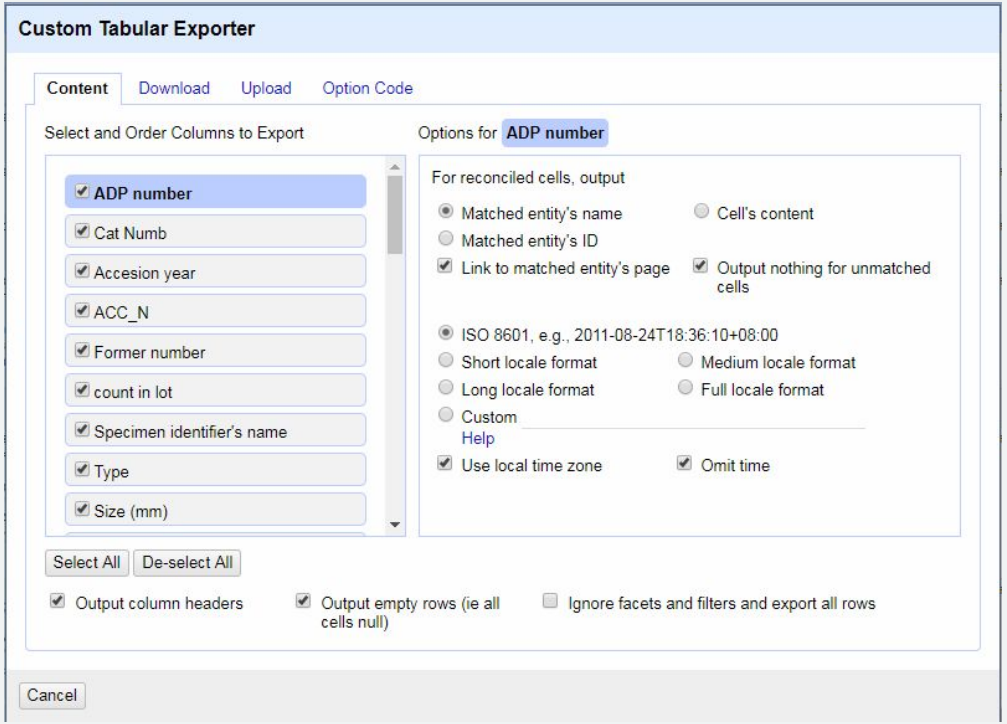


Exporting

In the upper-right corner of the window click Export and select Custom tabular exporter….



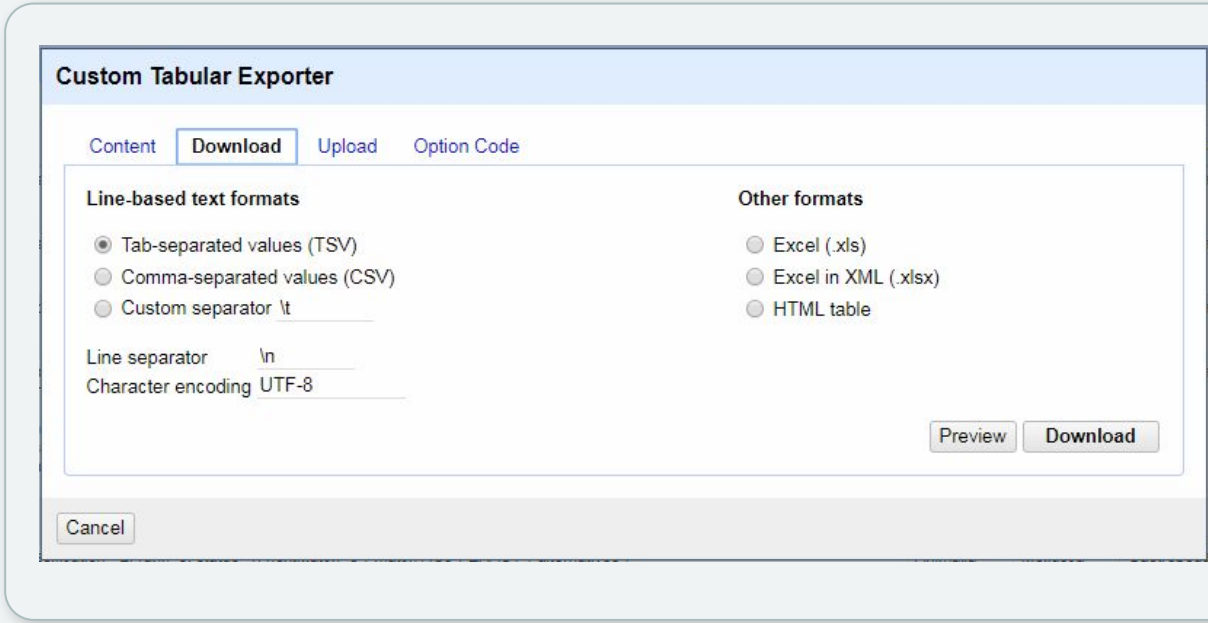
Getting Data Out of OpenRefine



- In the Content tab, users can choose specific columns to export.
  - If you select Ignore facets and filters and export all rows all facets and filterings will be ignored.
  - This is useful if users forget to clear them before exporting.

Getting Data Out of OpenRefine

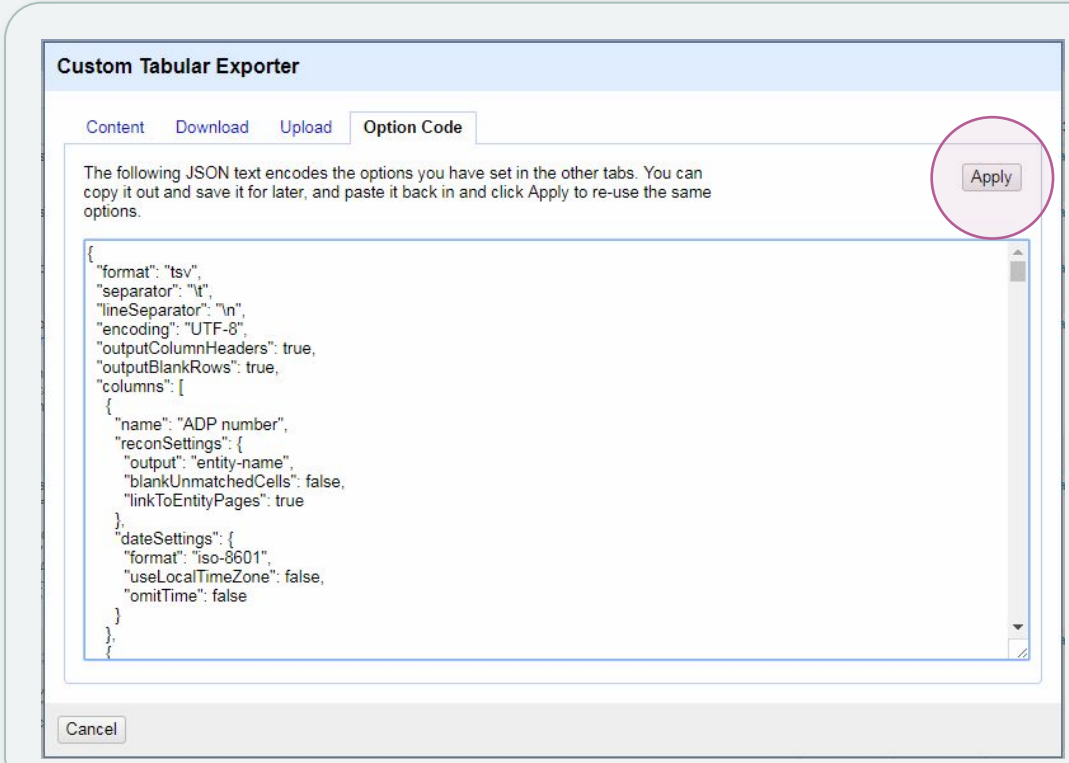
- Go to the Download tab and select a separator.
  - Don’t modify the other options unless necessary.



Getting Data Out of OpenRefine

Code

- Go to the Option Code tab, select and copy all JSON text in the textbox, and save it.
  - The text encodes the options that have been set in the other tabs.
  - To use the code in the future, paste the JSON text into the textbox and click the Apply button.



Practice the skills:

Practice Exercises

https://docs.google.com/document/d/1-l2Uh0zyNLD4GDokUedxApatdz4uFi76e_SAgTb776E

Example Dataset

https://docs.google.com/spreadsheets/d/1Sdi-eT--SyaJFywSyFrLLwtCpQu2x6MOmK_EvXej4F0

Questions?

Ask now; we might have answers!

Presenter name

Title or credentials

- Bio
- Contact Information

Insert Presenter Image

Thank you!

Sharon Grant
sgrant@fieldmuseum.org
https://www.fieldmuseum.org/

Janeen Jones
jjones@fieldmuseum.org
https://www.fieldmuseum.org/

Kate Webbink
kwebbink@fieldmuseum.org
https://www.fieldmuseum.org/

This project was made possible in part by the
Institute of Museum and Library Services
Grant ME-249136-OMS-21 | **IMLS.gov**