

OpenRefine for Data Cleaning

What is OpenRefine?
A data cleaning tool.

Why use OpenRefine?
Tracks changes and is easily reversible.



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



Conventions

✕ Genus invert reset

^[a-z]

case sensitive regular expression

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

^[a-z]



- Column menu

Genus Subgenus Species Subspecies

Facet ▶ Text facet

Text filter ▶ Numeric facet

Edit cells ▶ Timeline facet

Edit column ▶ Scatterplot facet

Transpose ▶ Custom text facet...

Sort... ▶ Custom Numeric Facet...

View ▶ Customized facets ▶

Reconcile ▶

Tegula mariana

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

Show as: rows records Show: 5 10 25 50 100 500

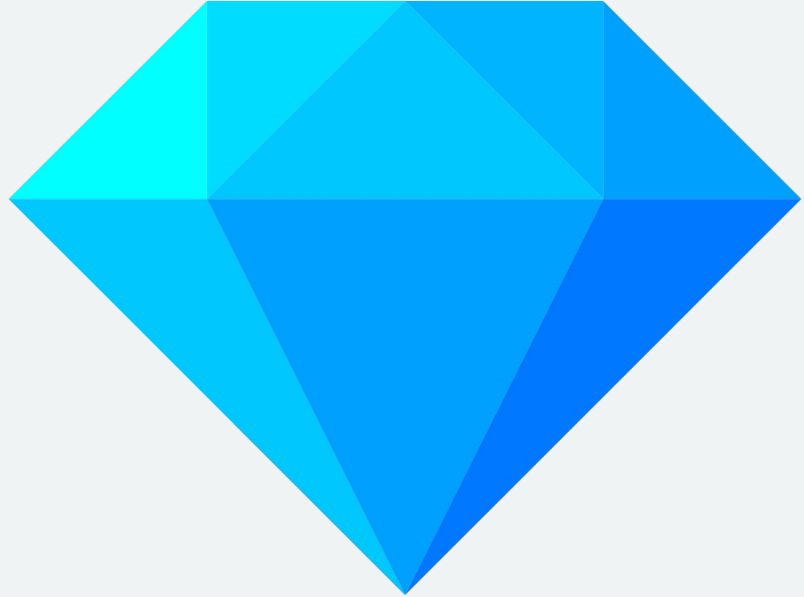
« first < previous 1 of 1

All	ADP number	Cat Numb	Accession year
☆ ↻ 1.	1	ABCD:1	1993 edit
☆ ↻ 2.	2	ABCD:2	1993
☆ ↻ 3.	3	ABCD:3	1993
☆ ↻ 4.	4	ABCD:4	1993
☆ ↻ 5.	5	ABCD:5	1993
☆ ↻ 6.	6	ABCD:6	1993
☆ ↻ 7.	7	ABCD:7	1993
☆ ↻ 8.	8	ABCD:8	1993

- 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	<input checked="" type="checkbox"/>
Charts & Graphs	<input type="checkbox"/>
Auto-formats fields	<input type="checkbox"/>
Allow creation of lookup lists	<input type="checkbox"/>
Allow character sets besides Latin1 (e.g., UTF-8)	<input checked="" type="checkbox"/>
Track changes after closing	<input checked="" type="checkbox"/>
Maintain integrity between rows and columns	<input checked="" type="checkbox"/>
Enforce referential integrity between different sheets	<input type="checkbox"/>
Creates a copy of original data	<input checked="" type="checkbox"/>
Allow you to share steps	<input checked="" type="checkbox"/>

Opening OpenRefine

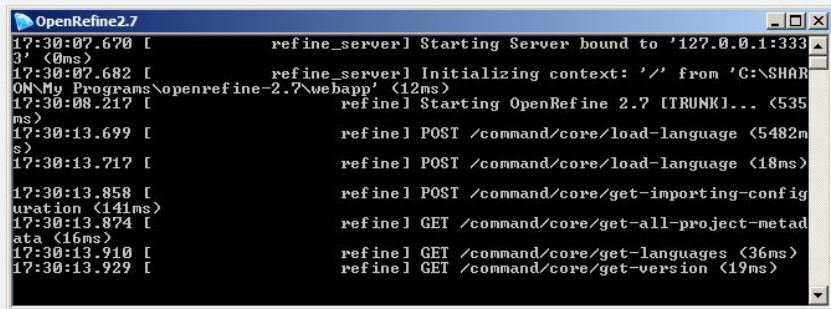
Installation

- 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>



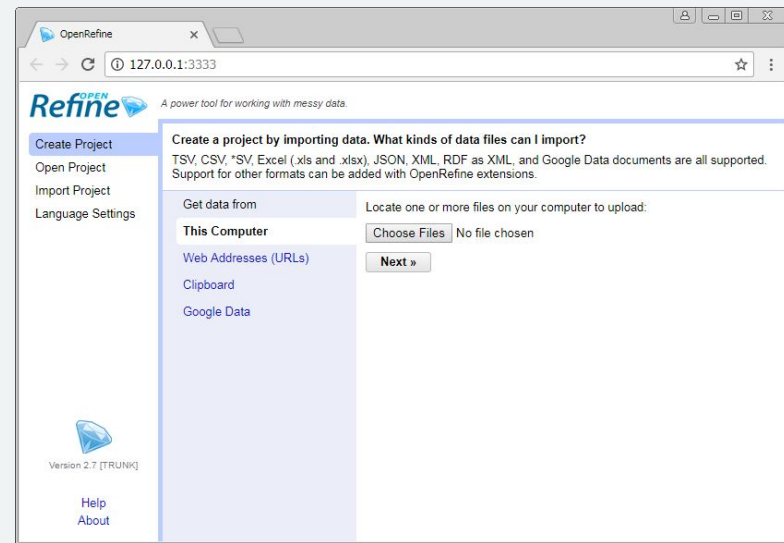
Opening OpenRefine

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



```
OpenRefine2.7
17:30:07.670 [      refine_server] Starting Server bound to '127.0.0.1:3333' (0ms)
17:30:07.682 [      refine_server] Initializing context: '/' from 'C:\SHARON\My Programs\openrefine-2.7\webapp' (12ms)
17:30:08.217 [      refine] Starting OpenRefine 2.7 [TRUNK]... (535ms)
17:30:13.699 [      refine] POST /command/core/load-language (5482ms)
17:30:13.717 [      refine] POST /command/core/load-language (18ms)
17:30:13.858 [      refine] POST /command/core/get-importing-config (141ms)
17:30:13.874 [      refine] GET /command/core/get-all-project-metadata (16ms)
17:30:13.910 [      refine] GET /command/core/get-languages (36ms)
17:30:13.929 [      refine] GET /command/core/get-version (19ms)
```

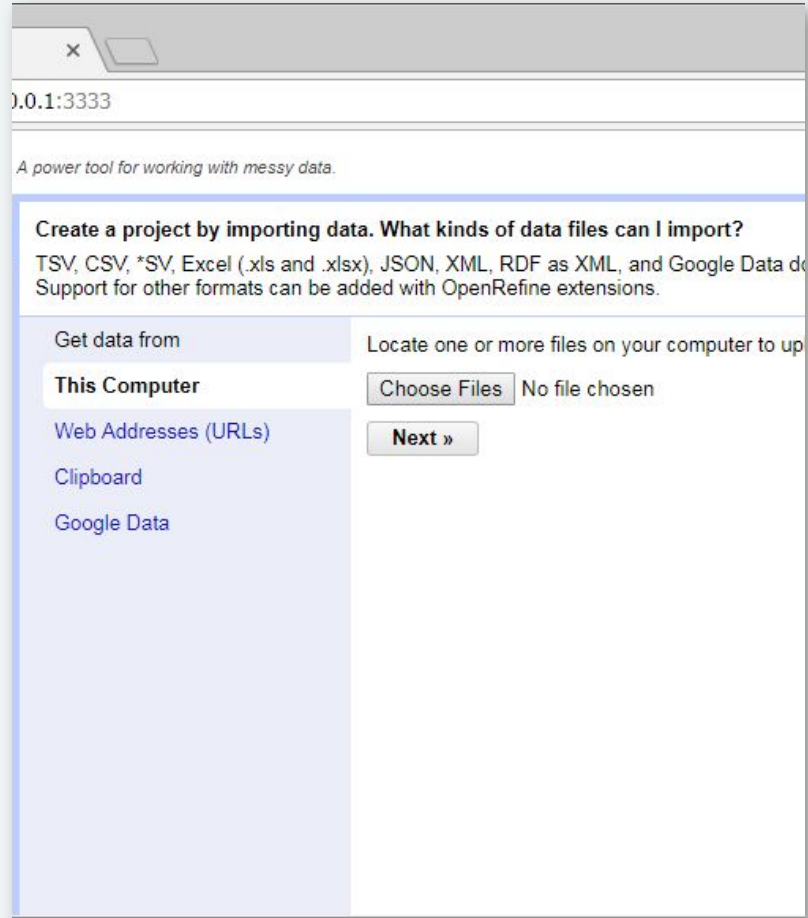
When you open the program, a black screen with code will open and launch OpenRefine in your web browser. Leave the application running to use OpenRefine.



Getting Data into OpenRefine

There are two steps:

1. Loading a file
2. Creating a project

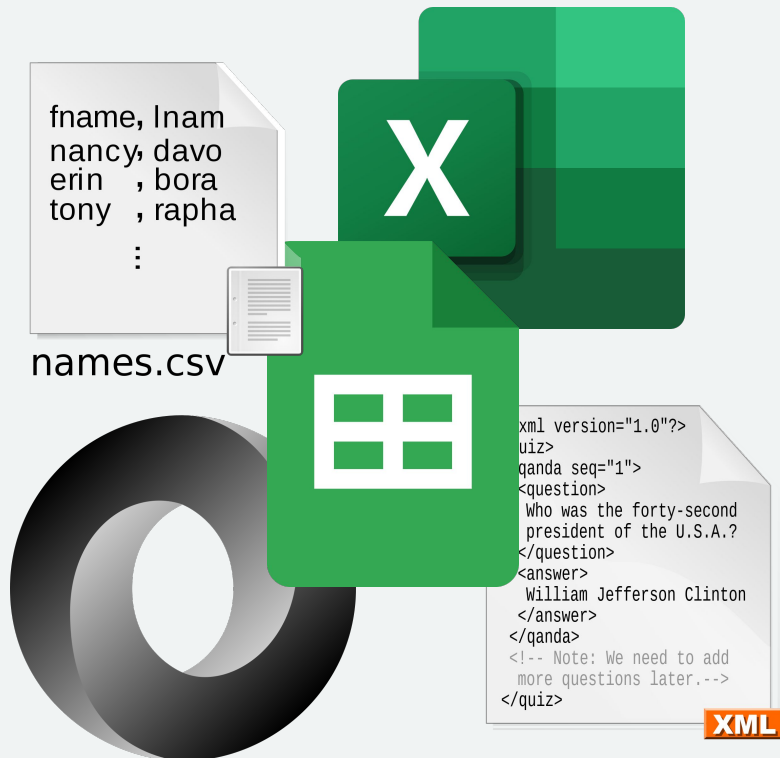


Getting Data into OpenRefine

Load File

OpenRefine can load data in a number of formats:

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

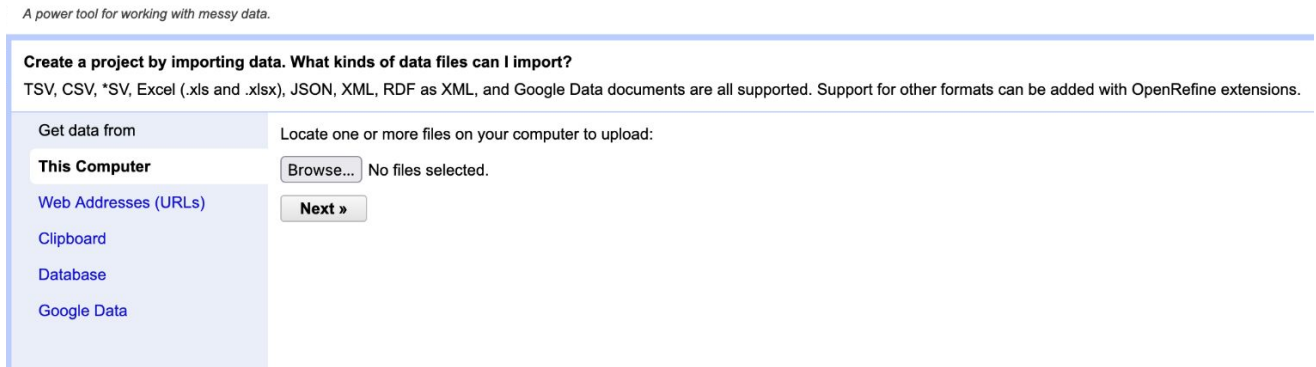


Reference:
User:Dreflymac, CC BY-SA 2.5 <<https://creativecommons.org/licenses/by-sa/2.5>>, via Wikimedia Commons
Google, Public domain, via Wikimedia Commons
Douglas Crockford, Public domain, via Wikimedia Commons
en:User:Dreflymac, CC BY-SA 2.5 <<https://creativecommons.org/licenses/by-sa/2.5>>, via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

Parse data as

Character encoding

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

MARC files

RDF/N3 files

XML files

Open Document Format spreadsheets (.ods)

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

Update Preview

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	Accession 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	Lymnaeidea	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.

Character encoding

Columns are separated by

- commas (CSV)
- tabs (TSV)
- custom , _____

Escape special characters with \

Select Encoding

Common Encodings | All Encodings

Encoding	Aliases
ISO-8859-1	819, ISO8859-1, I1, ISO_8859-1:1987, ISO_8859-1, 8859_1, iso-ir-100, latin1, cp819, ISO8859_1, IBM819, ISO_8859_1, IBM-819, csISOLatin1
US-ASCII	ANSI_X3.4-1968, cp367, csASCII, iso-ir-6, ASCII, iso_646.irv:1983, ANSI_X3.4-1986, ascii7, default, ISO_646.irv:1991, ISO646-US, IBM367, 646, us
UTF-16	UTF_16, unicode, utf16, UnicodeBig
UTF-16BE	X-UTF-16BE, UTF_16BE, ISO-10646-UCS-2, UnicodeBigUnmarked
UTF-16LE	UnicodeLittleUnmarked, UTF_16LE, X-UTF-16LE
UTF-8	unicode-1-1-utf-8, UTF8

Encoding	Aliases
Big5	csBig5
Big5-HKSCS	big5-hkscs, big5hk, Big5_HKSCS, big5hkscs
CESU-8	CESU8, csCESU-8
EUC-JP	csEUCPkdFmtjapanese, x-euc-jp, eucjis, Extended_UNIX_Code_Packed_Format_for_Japanese, euc_jp, eucjp, x-eucjp
EUC-KR	ksc5601-1987, csEUCKR, ksc5601_1987, ksc5601, 5601, euc_kr, ksc_5601, ks_c_5601-1987, euckr
GB18030	gb18030-2000
GB2312	gb2312, euc-cn, x-EUC-CN, euccn, EUC_CN, gb2312-80, gb2312-1980

Cancel

Getting Data into OpenRefine

Create Project

A project is a combination of:

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

Dataset csv		Create Project »	
Family	Subfamily	Genus	Subg ▲
Haliotidae		Haliotis	
Haliotidae		Haliotis	
Trochidae	Trochinae	Clanculus	
Trochidae	Calliostomatinae	Calliostoma	
		Unidentified	
Trochidae	Gibbulinae	Cittarium	
Trochidae	Monodontinae	Monodonta	
Trochidae	Monodontinae	Tegula	
Lymnaeidae		Lymnaea	
Vermetidae		Dendropoma	

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	Accession 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	Lymnaeidea	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.

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](#)

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

The screenshot shows the Refine web interface for 'FMNH Data Cleaning Workshop 2017'. The browser address bar shows '127.0.0.1:3333/project=1516325293115'. The interface includes a 'Facet / Filter' sidebar on the left, a main data table, and a top navigation bar with 'Open...', 'Export', and 'Help' buttons. The table displays 11793 rows with columns for 'Size (mm)', 'condition', 'Data', 'Valves', 'Kingdom', 'Superfamily', 'Family', 'Subfamily', 'Genus', 'Subgenus', 'Species', and 'Subspecies'. The 'Kingdom' column is highlighted with a green box, and a green arrow points to it from above. The 'Facet / Filter' sidebar shows 'Kingdom' as the selected facet with 3 choices: 'Animal' (314), 'Animalia' (11390), and 'Animnalia' (88). The table data is as follows:

Size (mm)	condition	Data	Valves	Kingdom	Superfamily	Family	Subfamily	Genus	Subgenus	Species	Subspecies
92		None			Pleurotomarioidea	Haliotidae		Haliotis		cracherodii	
34		None		Animalia	Pleurotomarioidea	Haliotidae		Haliotis		ancile	
15		None		Animalia	Trochoidea	Trochidae	Trochinae	Clanculus		punicus	
15		None		Animalia	Trochoidea	Trochidae	Calliostomatinae	Calliostoma		ligatum	
20		None		Animalia	Trochoidea			Unidentified			
55		None		Animalia	Trochoidea	Trochidae	Gibbulinae	Cittarium		pica	
17		None		Animalia	Trochoidea	Trochidae	Monodontinae	Monodonta		labis	
19	Good	None		Animalia	Trochoidea	Trochidae	Monodontinae	Tegula		mariana	
44	Good	Basic		Animalia	Lymnaeoidae	Lymnaeidae		Lymnaea		stagnalis	
28		None		Animalia	Cerithioidea	Vermetidae		Dendropoma		irregularis	

Faceting

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

The screenshot shows the Refine web interface for the FMNH Data Cleaning Workshop 2017. The browser address bar shows the URL `127.0.0.1:3333/project?project=1516325293115`. The page title is "Refine OPEN FMNH Data Cleaning Workshop 2017". The interface includes a "Facet / Filter" section with "Undo / Redo" options. A sidebar on the left contains a "Using facets and filters" help box. The main area displays a table with 11793 rows. The table headers are: Size (mm), condition, Data, Valves, Kingdom, Superfamily, Family, Subfamily, Genus, Subgenus, Species, and Subspecies. The "Kingdom" column header has a dropdown menu open, showing options: Facet (selected), Text filter, Edit cells, Edit column, Transpose, Sort..., View, and Reconcile. The "Facet" option is further expanded to show: Text facet (highlighted), Numeric facet, Timeline facet, Scatterplot facet, Custom text facet..., Custom Numeric Facet..., and Customized facets. A red arrow points to the "Facet" option in the dropdown menu.

Size (mm)	condition	Data	Valves	Kingdom	Superfamily	Family	Subfamily	Genus	Subgenus	Species	Subspecies
92		None		Facet				Heliotis		cracherodii	
34		None		Text filter				Heliotis		ancile	
15		None		Edit cells				Clanculus		punicus	
15		None		Edit column			matinae	Calliostoma		ligatum	
20		None		Transpose				Unidentified			
55		None		Sort...				Cittarium		pica	
17		None		View			atinae	Monodonta		labis	
19	Good	None		Reconcile	ea	Trochidae	Monodontinae	Tegula		mariana	
44	Good	Basic			ymnae	Lymnaeidae		Lymnaea		stagnalis	
28		None		Animalia	Cerithioidea	Vermetidae		Dendropoma		irregularis	

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.

The screenshot shows the Refine tool interface for 'FMNH Data Cleaning Workshop 2017'. The main table displays 11793 rows with columns for Size (mm), condition, Data, Valves, Kingdom, Superfamily, Family, Subfamily, Genus, Subgenus, Species, and Subspecies. A sidebar on the left is faceted on the 'Kingdom' column, showing 3 choices: Animal (314), Animalia (11390), and Animmalia (88). The sidebar also shows a '(blank)' category with 1 count. The table data includes rows with various species names like cracherodii, ancile, puniceus, ligatum, Unidentified, pica, labis, mariana, stagnalis, and irregularis.

Size (mm)	condition	Data	Valves	Kingdom	Superfamily	Family	Subfamily	Genus	Subgenus	Species	Subspecies
92		None			Pleurotomarioidea	Haliotidae		Haliotis		cracherodii	
34		None		Animalia	Pleurotomarioidea	Haliotidae		Haliotis		ancile	
15		None		Animalia	Trochoidea	Trochidae	Trochinae	Clanculus		puniceus	
15		None		Animalia	Trochoidea	Trochidae	Calliostomatinae	Calliostoma		ligatum	
20		None		Animalia	Trochoidea			Unidentified			
55		None		Animalia	Trochoidea	Trochidae	Gibbulinae	Cittarium		pica	
17		None		Animalia	Trochoidea	Trochidae	Monodontinae	Monodonta		labis	
19	Good	None		Animalia	Trochoidea	Trochidae	Monodontinae	Tegula		mariana	
44	Good	Basic		Animalia	Lymnaeoidea	Lymnaeidae		Lymnaea		stagnalis	
26		None		Animalia	Cerithioidea	Vermetidae		Dendropoma		irregularis	

Manipulating/ Mangling Data

Filtering

Filtering is a feature that allows users to:

1. Search for values
2. Group records

Full name

case sensitive regular expression

Full name

case sensitive regular expression

Full name

case sensitive regular expression

Filtering

Go to column **Full name** and perform a **Text filter**.

Click the  drop-down icon.

FMNH Data Cleaning Workshop 2017

11793 rows

Using facets and filters

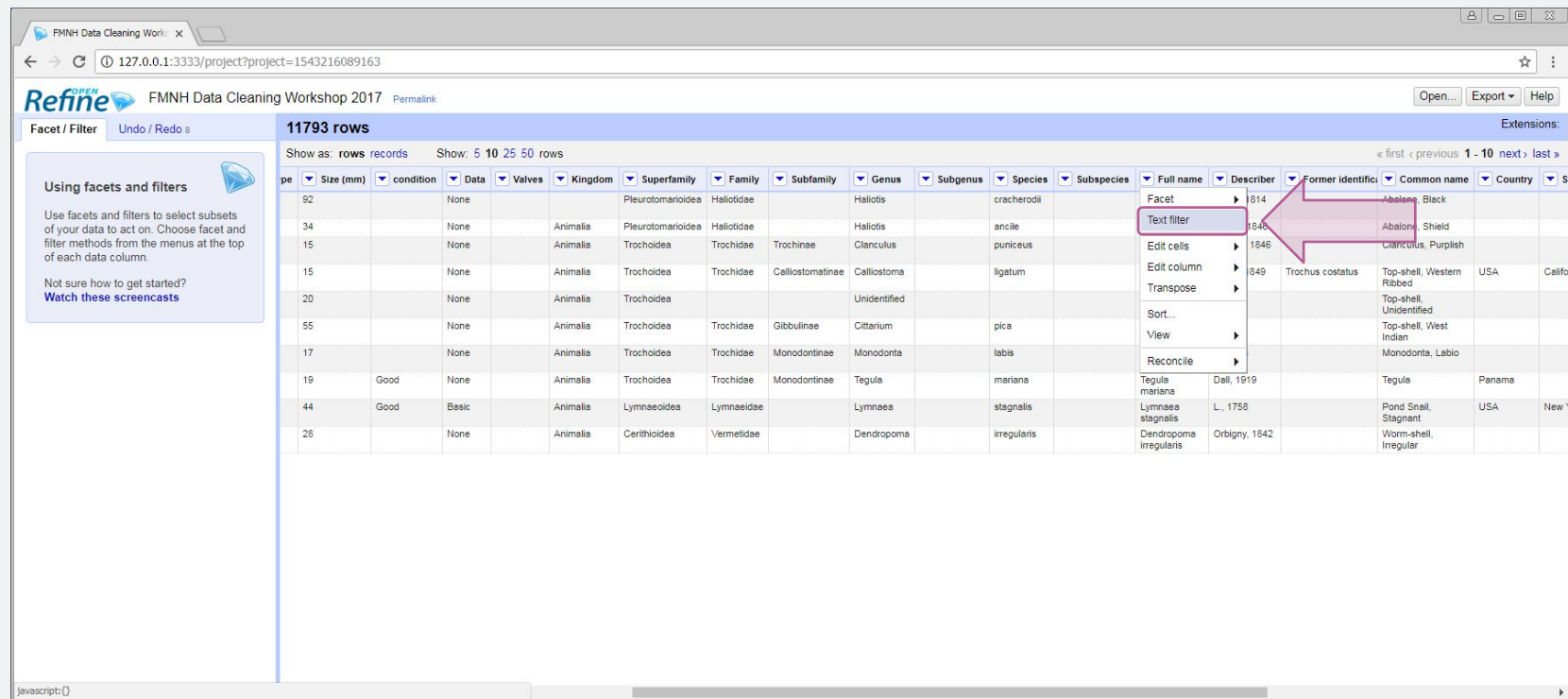
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

pe	Size (mm)	condition	Data	Valves	Kingdom	Superfamily	Family	Subfamily	Genus	Subgenus	Species	Subspecies	Full name	Describer	Former identifier	Common name	Country	S
92			None			Pleurotomoidea	Haliotidae		Haliotis		cracherodii		Haliotis cracherodii	Leach, 1814		Abalone, Black		
34			None		Animalia	Pleurotomoidea	Haliotidae		Haliotis		ancile		Haliotis ancile	Reeve, 1846		Abalone, Shield		
15			None		Animalia	Trochoidea	Trochidae	Trochineae	Clanculus		puniceus		Clanculus puniceus	Philippi, 1846		Clanculus, Purpleish		
15			None		Animalia	Trochoidea	Trochidae	Calliostomatinae	Calliostoma		ligatum		Calliostoma ligatum	Gould, 1849	Trochus costatus	Top-shell, Western Ribbed	USA	Calif
20			None		Animalia	Trochoidea	Trochidae		Unidentified				Unidentified			Top-shell, Unidentified		
55			None		Animalia	Trochoidea	Trochidae	Gibbulinae	Cittarium		pica		Cittarium pica	L., 1758		Top-shell, West Indian		
17			None		Animalia	Trochoidea	Trochidae	Monodontinae	Monodonta		labris		Monodonta labris	L., 1758		Monodonta, Labio		
19	Good		None		Animalia	Trochoidea	Trochidae	Monodontinae	Tegula		mariana		Tegula mariana	Dall, 1919		Tegula	Panama	
44	Good	Basic			Animalia	Lymnaeidea	Lymnaeidae		Lymnaea		stagnalis		Lymnaea stagnalis	L., 1758		Pond Snail, Stagnant	USA	New
28			None		Animalia	Cerithioidea	Vermetidae		Dendropoma		irregularis		Dendropoma irregularis	Orbigny, 1842		Worm-shell, Irregular		

Filtering

Text filter is the second option in the column menu.



The screenshot shows the Refine web application interface. At the top, there's a browser address bar with the URL `127.0.0.1:3333/project?project=1543216089163`. Below it, the Refine logo and 'FMNH Data Cleaning Workshop 2017' are visible. The main area displays a table with 11793 rows. A column menu is open for the 'Former identifier' column, with 'Text filter' highlighted. A pink arrow points to this option. The table columns include 'pe', 'Size (mm)', 'condition', 'Data', 'Valves', 'Kingdom', 'Superfamily', 'Family', 'Subfamily', 'Genus', 'Subgenus', 'Species', 'Subspecies', 'Full name', 'Describer', 'Former identifier', 'Common name', 'Country', and 'State'.

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.

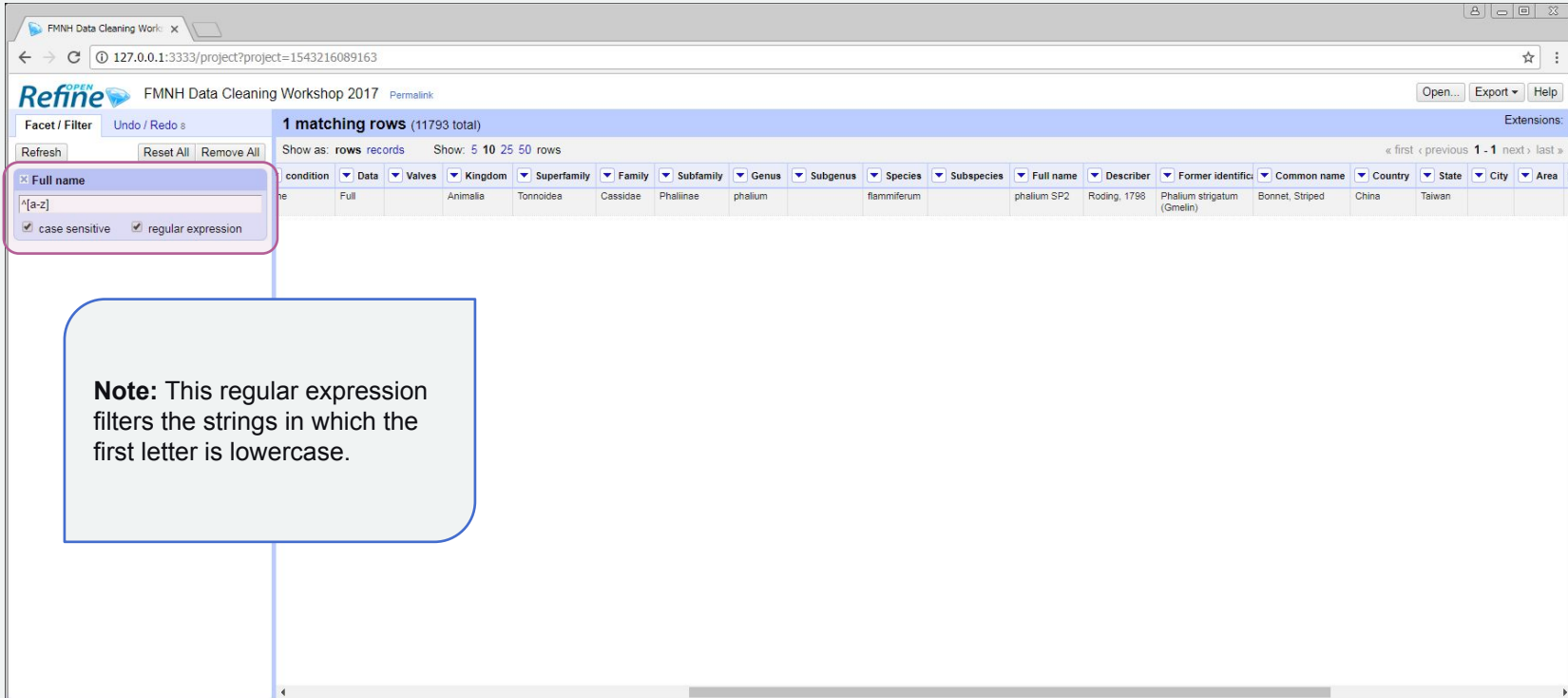
Not sure how to get started?
[Watch these screencasts](#)

pe	Size (mm)	condition	Data	Valves	Kingdom	Superfamily	Family	Subfamily	Genus	Subgenus	Species	Subspecies	Full name	Describer	Former identifier	Common name	Country	State
92			None			Pleurotomoidea	Haliotidae		Haliotis		cracherodii			1814		Abalone, Black		
34		None			Animalia	Pleurotomoidea	Haliotidae		Haliotis		ancile			1846		Abalone, Shield		
15		None			Animalia	Trochoidea	Trochidae	Trochineae	Clanculus		puniceus			1846		Clanculus, Purpleish		
15		None			Animalia	Trochoidea	Trochidae	Calliostomatinae	Calliostoma		ligatum			1849	Trochus costatus	Top-shell, Western Ribbed	USA	Calif
20		None			Animalia	Trochoidea	Trochidae		Unidentified							Top-shell, Unidentified		
55		None			Animalia	Trochoidea	Trochidae	Gibbulinae	Cittarium		pica					Top-shell, West Indian		
17		None			Animalia	Trochoidea	Trochidae	Monodontinae	Monodonta		labis					Monodonta, Labio		
19	Good	None			Animalia	Trochoidea	Trochidae	Monodontinae	Tegula		mariana		Tegula mariana	Dall, 1919		Tegula	Panama	
44	Good	Basic			Animalia	Lymnaeidea	Lymnaeidae		Lymnaea		stagnalis		Lymnaea stagnalis	L., 1758		Pond Snail, Stagnant	USA	New
28		None			Animalia	Certhioidea	Vermetidae		Dendropoma		irregularis		Dendropoma irregularis	Orbigny, 1842		Worm-shell, Irregular		

Filtering

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

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



The screenshot shows the Refine interface for the 'Full name' column. The filter is set to the regular expression `^[a-z]`. The 'case sensitive' and 'regular expression' options are checked. The table below shows one matching row for 'Phallium strigatum'.

condition	Data	Valves	Kingdom	Superfamily	Family	Subfamily	Genus	Subgenus	Species	Subspecies	Full name	Describer	Former identifi	Common name	Country	State	City	Area
	Full		Animalia	Tonnoidea	Cassidae	Phallinae	phallium		flammiferum		phallium SP2	Roding, 1798	Phallium strigatum (Gmelin)	Bonnet, Striped	China	Taiwan		

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

Filtering

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

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

The screenshot shows the Refine web interface for the FMNH Data Cleaning Workshop 2017. A search filter for 'Full name' is active, displaying the regular expression `^[A-Z].*\s[A-Z]`. The 'case sensitive' and 'regular expression' checkboxes are checked. The interface shows 15 matching rows out of 11793 total. A table of results is visible, with columns for condition, data, valves, kingdom, superfamily, family, subfamily, genus, subgenus, species, subspecies, full name, describer, former identifier, common name, country, state, city, and area.

condition	Data	Valves	Kingdom	Superfamily	Family	Subfamily	Genus	Subgenus	Species	Subspecies	Full name	Describer	Former identifier	Common name	Country	State	City	Area
ood	Full		Animalia	Muricoidea	Vasidae	Columbarinae	Columbarium		Unidentified		Columbarium Unidentified		Ancistrosyrinx radiata	Pagoda Shell	USA	Florida		Dry Tortugas
rm	Basic		Animalia	Cypraeoidea	Cypraeidae		Cypraea	Mauritia	maculifera		Mauritia SP2	Schilder, 1932		Cowrie, Reticulated	Cook Islands		Mauke	
ood	Full	1	Animalia	Orthalicoidea	Orthalicoidea	Bulimulinae	Bulimulus		Dealbatus	ragdalei	Bulimulus Dealbatus ragdalei	Pilsbry			USA	Texas	Spofford	
ood	Full	1	Animalia	Helicoidea	Helicellidae	Helicellinae	Helicella	carthusiana	carthusianella	Drap.	Helicella carthusianella Drap.				Switzerland		Geneva	
ia	Helicoidea		Helicellidae	Helicellinae	Helicella		conoidea	St. Simoniana	Helicella conoidea St. Simoniana		Caziot				France	Corsica	Bonifacio	
ia	Helicoidea		Helicidae	Helicinae	Helicina		lueticella	Fer.	Helicina lueticella Fer.						Canary Islands	Grand Canary		
ia	Muricoidea		Muricoidea	Thaidinae	Acanthina		Tuberculata		Acanthina Tuberculata		Sowerby				Mexico			Yzvaros
ia	Cypraeoidea		Ovulidae	Ovulinae	Cyphoma		Gibbosum		Cyphoma Gibbosum		L., 1758			Flamingo Tongue	USA	Florida		Long Ke
ia	Cypraeoidea		Ovulidae	Ovulinae	Cyphoma		Gibbosum		Cyphoma Gibbosum		L., 1758			Flamingo Tongue	USA	Florida		Indian K
ia	Cypraeoidea		Ovulidae	Ovulinae	Cyphoma		Gibbosum		Cyphoma Gibbosum		L., 1758			Flamingo Tongue	USA	Florida		Lower Matecur Key

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

Manipulating/ Mangling Data

Clustering

Clustering is a feature that allows users to:

1. Find different versions
2. Group records

Cluster & Edit column "State"

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 "Godel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method Keying Function 9 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none">Turk's Island (1 rows)[Turk's Island] (1 rows)	<input type="checkbox"/>	<input type="text" value="Turk's Island"/>
2	3	<ul style="list-style-type: none">[Crooked Island] (2 rows)Crooked Island (1 rows)	<input type="checkbox"/>	<input type="text" value="[Crooked Island]"/>
2	3	<ul style="list-style-type: none">Alpes-Mar (2 rows)Alpes-Mar. (1 rows)	<input type="checkbox"/>	<input type="text" value="Alpes-Mar"/>
2	6	<ul style="list-style-type: none">Inagua Island (4 rows)[Inagua Island] (2 rows)	<input type="checkbox"/>	<input type="text" value="Inagua Island"/>
2	2	<ul style="list-style-type: none">Tanega-shima (1 rows)Tanegashima (1 rows)	<input type="checkbox"/>	<input type="text" value="Tanega-shima"/>
2	58	<ul style="list-style-type: none">Illinois (56 rows)ILLINOIS (2 rows)	<input type="checkbox"/>	<input type="text" value="Illinois"/>
2	1279	<ul style="list-style-type: none">Florida (1278 rows)[Florida] (1 rows)	<input type="checkbox"/>	<input type="text" value="Florida"/>
2	25	<ul style="list-style-type: none">Georgia (24 rows)	<input type="checkbox"/>	<input type="text" value="Georgia"/>

Rows in Cluster
0 — 1300

Average Length of Choices
8 — 15

Length Variance of Choices
0 — 1

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Clustering

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

The screenshot shows the Refine web interface for the FMNH Data Cleaning Workshop 2017. The main table displays 11793 rows of specimen records. The sidebar on the left is faceted by the 'State' column, showing 445 choices sorted by name count. A 'Cluster' button is visible in the bottom right corner of the sidebar panel, highlighted with a red circle and a red arrow pointing to the main table. The 'State' column in the table header is highlighted with a green box.

Family	Subfamily	Genus	Subgenus	Species	Subspecies	Full name	Describer	Former identifi	Common name	Country	State	City	Area	Site	Year	Month	Day	Collectors name
Haliotidae		Haliotis		cracherodii		Haliotis cracherodii	Leach, 1814		Abalone, Black									
Haliotidae		Haliotis		ancile		Haliotis ancile	Reeve, 1846		Abalone, Shield									
Trochidae	Trochineae	Clanculus		puniceus		Clanculus puniceus	Philippi, 1846		Clanculus, Purplish									
Trochidae	Calliostomatinae	Calliostoma		ligatum		Calliostoma ligatum	Gould, 1849	Trochus costatus	Top-shell, Western Ribbed	USA	California							
Trochidae						Unidentified			Top-shell, Unidentified									
Trochidae	Gibbulinae	Cittarium		pica		Cittarium pica	L., 1758		Top-shell, West Indian									
Trochidae	Monodontinae	Monodonta		labis		Monodonta labis	L., 1758		Monodonta, Labio									
Trochidae	Monodontinae	Tegula		mariana		Tegula mariana	Dall, 1919		Tegula	Panama								
Lymnaeidae		Lymnaea		stagnalis		Lymnaea stagnalis	L., 1758		Pond Snail, Stagnant	USA	New York		Seneca Lake					
Vermetidae	Dendropoma			irregularis		Dendropoma irregularis	Orbigny, 1842		Worm-shell, Irregular									
Modiolidae		Modiolus		modulus		Modiolus modulus	L., 1758	Modulus floridanus	Modulus, Atlantic	USA	Florida	Sarasota	Sarasota Bay					
Xenophoridae		Xenophora	Onustus	giganteum		Xenophora giganteum	Schepman, 1909		Carrier-shell, Great									
Strombidae		Strombus	Lentigo	granulatus		Strombus granulatus	Swainson, 1822		Conch, Granulated									
Strombidae		Strombus	Tricornis	raninus		Strombus raninus	Gmelin, 1791	Strombus bituberculatus	Conch, Hawk-wing	Haiti								
Strombidae		Strombus	Conomurex	luhuanus		Strombus luhuanus	L., 1758		Conch, Strawberry				South Seas					
Strombidae		Strombus	Lentigo	lentiginosus		Strombus lentiginosus	L., 1758	Pterocera surantia	Conch, Granulated									
Strombidae		Strombus	Plicatus	pulchellus		Strombus pulchellus	Reeve, 1851		Conch, Pretty									
Strombidae		Strombus	Tricornis	tricornis		Strombus tricornis	Lightfoot, 1786		Conch, Three-knobbed									
Strombidae		Strombus	Vomer	iredalei		Strombus iredalei	Abbott, 1960		Conch, Iredale's									

Clustering

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

Cluster & Edit column "State"

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 "Godel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method key collision Keying Function fingerprint 9 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none">Turk's Island (1 rows)[Turk's Island] (1 rows)	<input type="checkbox"/>	<input type="text" value="Turk's Island"/>
2	3	<ul style="list-style-type: none">[Crooked Island] (2 rows)Crooked Island (1 rows)	<input type="checkbox"/>	<input type="text" value="[Crooked Island]"/>
2	3	<ul style="list-style-type: none">Alpes-Mar (2 rows)Alpes-Mar. (1 rows)	<input type="checkbox"/>	<input type="text" value="Alpes-Mar"/>
2	6	<ul style="list-style-type: none">Inagua Island (4 rows)[Inagua Island] (2 rows)	<input type="checkbox"/>	<input type="text" value="Inagua Island"/>
2	2	<ul style="list-style-type: none">Tanega-shima (1 rows)Tanegashima (1 rows)	<input type="checkbox"/>	<input type="text" value="Tanega-shima"/>
2	58	<ul style="list-style-type: none">Illinois (56 rows)ILLINOIS (2 rows)	<input type="checkbox"/>	<input type="text" value="Illinois"/>
2	1279	<ul style="list-style-type: none">Florida (1278 rows)[Florida] (1 rows)	<input type="checkbox"/>	<input type="text" value="Florida"/>
2	25	<ul style="list-style-type: none">Georgia (24 rows)	<input type="checkbox"/>	<input type="text" value="Georgia"/>

Rows in Cluster

Average Length of Choices

Length Variance of Choices

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Clustering

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

Cluster & Edit column "State"

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** 41 clusters found

3	272	<ul style="list-style-type: none">• Alde (2 rows)• ID (1 rows)	<input type="checkbox"/>	California
3	183	<ul style="list-style-type: none">• California (268 rows)• Callifornia (3 rows)• Califorina (1 rows)	<input type="checkbox"/>	California
3	183	<ul style="list-style-type: none">• Hawaii (181 rows)• Hai (1 rows)• Hawii (1 rows)	<input type="checkbox"/>	Hawaii
3	59	<ul style="list-style-type: none">• Illinois (56 rows)• ILLINOIS (2 rows)• Illnois (1 rows)	<input type="checkbox"/>	Illinois
3	71	<ul style="list-style-type: none">• Maine (69 rows)• Main (1 rows)• Mayenne (1 rows)	<input type="checkbox"/>	Maine
3	287	<ul style="list-style-type: none">• District of Columbia (231 rows)• District of Columbia (51 rows)• District of Columbus (5 rows)	<input type="checkbox"/>	District of Columbia
3	43	<ul style="list-style-type: none">• Mindanao (33 rows)• Mindinao (5 rows)	<input type="checkbox"/>	Mindanao

Choices in Cluster
2 — 5

Rows in Cluster
0 — 1300

Average Length of Choices
3 — 20

Length Variance of Choices
0 — 2

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

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

Add column by fetching URLs based on column Full name

New column name Throttle delay
milliseconds

On error set to blank store error Cache responses

Formulate the URLs to fetch:

Expression Language No syntax error.

Preview History Starred Help

row	value	"http://api.gbif.org/v1/species/match?verbose=true&name="+escape(value,'url')
8097.	Hydrobia longinqua	http://api.gbif.org/v1/species/match?verbose=true&name=Hydrobia+longinqua+
8217.	Physa humerosa	http://api.gbif.org/v1/species/match?verbose=true&name=Physa+humerosa+
8400.	Planorbis trivolvis	http://api.gbif.org/v1/species/match?verbose=true&name=Planorbis+trivolvis+

OK Cancel

Reconciliation

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

The screenshot shows the Refine web interface for 'FMNH Data Cleaning Workshop 2017'. The main table displays 19 matching rows with columns for taxonomic classification and metadata. A facet menu on the left allows filtering by 'Collectors name', with 'Baldwin, D.D.' selected. A green arrow points to the table, and a pink arrow points to the selected facet entry.

Collectors name	Family	Subfamily	Genus	Subgenus	Species	Subspecies	Full name	Api_name	RANK 1	RANK 2	RANK 3	RANK 4	RANK 5	Descriptor	Former identifier	Common name	Country	State
Baldwin, D.D. 19	Ilidae	Achatinellinae	Achatinella		dolei	baldwin	Achatinella dolei baldwin										USA	Hawaii
	Ilidae	Achatinellinae	Achatinella		dolei	white variety	Achatinella dolei white variety										USA	Hawaii
	Ilidae	Achatinellinae	Achatinella		violacea		Achatinella violacea							Newc.			USA	Hawaii
	Ilidae	Achatinellinae	Achatinella		hawaiiensis		Achatinella hawaiiensis							Baldwin			USA	Hawaii
	Ilidae	Achatinellinae	Achatinella		ilcozona		Achatinella ilcozona							Gulick			USA	Hawaii
	Ilidae	Achatinellinae	Achatinella		nivea		Achatinella nivea							Baldwin			USA	Hawaii
	Ilidae	Achatinellinae	Achatinella		helvina		Achatinella helvina							Baldwin			USA	Hawaii
	Ilidae	Achatinellinae	Achatinella		zebrina		Achatinella zebrina							Pfeiffer			USA	Hawaii
	Ilidae	Achatinellinae	Achatinella		mehogani		Achatinella mehogani							Gulick			USA	Hawaii
	Ilidae	Achatinellinae	Achatinella		porcellana		Achatinella porcellana							Newc.			USA	Hawaii

Reconciliation

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

The screenshot shows the Refine tool interface for 'FMNH Data Cleaning Workshop 2017'. The main table displays 19 matching rows (out of 11793 total) with columns for taxonomic classification, full name, and rank. A context menu is open over the 'Full name' column, with the option 'Add column by fetching URLs...' highlighted. A pink arrow points to this option.

ity	Subfamily	Genus	Subgenus	Species	Subspecies	Full name	Api_name	RANK 1	RANK 2	RANK 3	RANK 4	RANK 5	Describer	Former identifier	Common name	Country	State
Ilidae	Achatinellinae	Achatinella		dolei	baldwin	Facet							Baldwin			USA	Hawaii
Ilidae	Achatinellinae	Achatinella		dolei	white variety	Text filter							Baldwin			USA	Hawaii
Ilidae	Achatinellinae	Achatinella		violacea		Edit cells							Newc.			USA	Hawaii
Ilidae	Achatinellinae	Achatinella		hawaliensis		Transpose							Baldwin			USA	Hawaii
Ilidae	Achatinellinae	Achatinella		lilcozona		Sort...							Gulick			USA	Hawaii
Ilidae	Achatinellinae	Achatinella		nivea		View							Baldwin			USA	Hawaii
Ilidae	Achatinellinae	Achatinella		helvina		Reconcile							Baldwin			USA	Hawaii
Ilidae	Achatinellinae	Achatinella		zebrina		Move column to beginning							Pfeiffer			USA	Hawaii
Ilidae	Achatinellinae	Achatinella		mahogani		Move column to end							Gulick			USA	Hawaii
Ilidae	Achatinellinae	Achatinella		porcellana		Move column left							Newc.			USA	Hawaii

Reconciliation

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

Add column by fetching URLs based on column Full name

New column name Throttle delay
milliseconds

On error set to blank store error Cache responses

Formulate the URLs to fetch:

Expression Language No syntax error.

Preview History Starred Help

row	value	"http://api.gbif.org/v1/species/match?verbose=true&name="+escape(value,'url')
8097.	Hydrobia longinqua	http://api.gbif.org/v1/species/match?verbose=true&name=Hydrobia+longinqua+
8217.	Physa humerosa	http://api.gbif.org/v1/species/match?verbose=true&name=Physa+humerosa+
8400.	Planorbis trivolvis	http://api.gbif.org/v1/species/match?verbose=true&name=Planorbis+trivolvis+

OK Cancel

Reconciliation

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

The screenshot shows the Refine Data Cleaning Workshop 2017 interface. The main table displays 19 matching rows. The columns are 'Full name' and 'Api_name'. The 'Api_name' column contains JSON-like strings representing taxonomic data. A purple arrow points to the 'Api_name' column header. The interface includes a 'Facet / Filter' sidebar on the left with 'Collectors name' selected, showing 765 choices. The top navigation bar includes 'Open...', 'Export', and 'Help' buttons. The bottom right corner features the 'FILE ID.' logo.



Reconciliation

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

3 matching rows (11793 total) Extensions: Wikidata

Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first < previous 1 of 1 page next > last »

Family	Subfamily	Genus	Subgenus	Species	Subspecies	Full name	Describer	Former identification	Common name	Country	State
Succineidae	Succineinae	Succinea		grosvenorii						USA	Utah
Anodontidae	Anodontinae	Anodonta		nuttalliana						USA	Utah
Anodontidae	Anodontinae	Anodonta		nuttalliana						USA	Utah

- Facet
- Text filter
- Edit cells
- Edit column**
 - Split into several columns...
 - Join columns...
 - Add column based on this column...**
 - Add column by fetching URLs...
 - Add columns from reconciled values...
 - Rename this column
 - Remove this column
 - Move column to beginning
 - Move column to end
 - Move column left
 - Move column right
- Transpose
- Sort...
- View
- Reconcile

Reconciliation

In the **Expression** field, use the equation

`value.parseJson().get(" ")`

to select the categories for the program to separate by.

Add column based on column Api_name

New column name:

set to blank store error copy value from original column

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

Language: **General Refine Expression Language (GREL)** ▼

No syntax error.

Preview History Starred Help

row	value
8097.	{\"usageKey\":2299908,\"scientificName\": \"Hydrobia ulmaria\", \"rank\": \"Gastropoda\", \"matchType\": \"HIGHERRANK\"}

OK Cancel

Reconciliation

The data will be placed in a new column.

The screenshot shows the Refine data cleaning interface. The top navigation bar includes 'Facet / Filter', 'Undo / Redo', and '19 matching rows (11793 total)'. Below this, there are buttons for 'Refresh', 'Reset All', and 'Remove All'. A facet for 'Collectors name' is visible on the left, listing names like Baker, Baldwin, and Bartlett. The main table displays data with columns for 'RANK', 'Descriptor', 'Former identifier', 'Common name', 'Country', 'State', 'City', and 'Area'. A red arrow points to the 'RANK' column, which contains taxonomic classification strings such as 'Animalia, Mollusca, Gastropoda, Stylommatophora, Achatinellidae'. The table also shows metadata like 'y: name=110; authorship=0; classification=-2; rank=5; status=1;'. A red box highlights the 'RANK' column header and its contents, with a red arrow pointing to it from the right.

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.

Split column Rank 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

OK Cancel

Reconciliation

The data will be placed in new separate columns.

The screenshot shows the Refine web interface for 'FMNH Data Cleaning Workshop 2017'. The main content area displays '19 matching rows (11793 total)'. A table is shown with columns for taxonomic ranks and identifiers. A green arrow points to the table header, and a green box highlights the RANK 1-5 columns.

RANK 1	RANK 2	RANK 3	RANK 4	RANK 5	Describer	Former identifier	Common name
Animalia	Mollusca	Gastropoda	Stylommatophora	Achatinellidae	Baldwin		
Animalia	Mollusca	Gastropoda	Stylommatophora	Achatinellidae	Baldwin		
Animalia	Mollusca	Gastropoda	Stylommatophora	Achatinellidae	Newc.		
Animalia	Mollusca	Gastropoda	Stylommatophora	Achatinellidae	Baldwin		
Animalia	Mollusca	Gastropoda	Stylommatophora	Achatinellidae	Gulick		
Animalia	Mollusca	Gastropoda	Stylommatophora	Achatinellidae	Baldwin		
Animalia	Mollusca	Gastropoda	Stylommatophora	Achatinellidae	Baldwin		

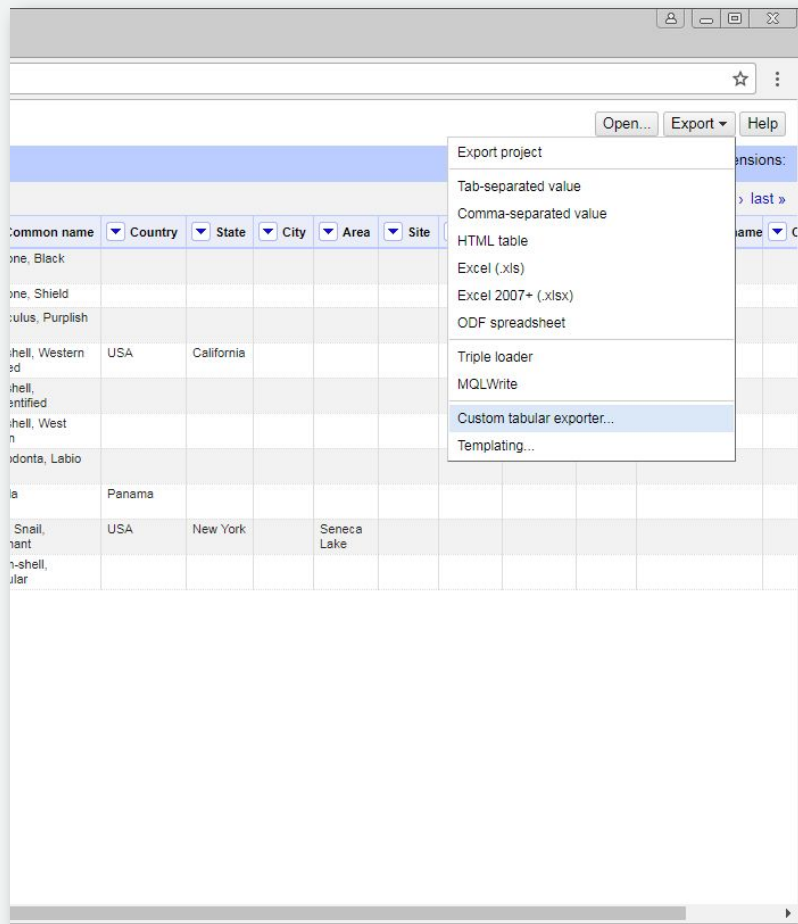
Getting Data out of OpenRefine

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...**

The screenshot shows the OpenRefine web interface. The browser address bar displays '127.0.0.1:3333/project?project=1612458256580'. The page title is 'FMNH Data Cleaning Workshop 2017'. The main content area shows a table with 11793 rows. The table columns include Family, Subfamily, Genus, Subgenus, Species, Subspecies, Full name, Descriptor, Former identifier, Common name, Country, State, City, Area, and Site. A 'Using facets and filters' sidebar is visible on the left. In the top right corner, the 'Export' menu is open, and 'Custom tabular exporter...' is highlighted. A pink circle highlights the 'Export' button, and a pink arrow points to the selected menu item.

Family	Subfamily	Genus	Subgenus	Species	Subspecies	Full name	Descriptor	Former identifier	Common name	Country	State	City	Area	Site
aliotidae		Haliotis		cracherodii		Haliotis cracherodii	Leach, 1814		Abalone, Black					
aliotidae		Haliotis		ancile		Haliotis ancile	Reeve, 1846		Abalone, Shield					
rochidae	Trochineae	Clanculus		puniceus		Clanculus puniceus	Philippi, 1846		Clanculus, Purplish					
rochidae	Calliostomatinae	Calliostoma		ligatum		Calliostoma ligatum	Gould, 1849	Trochus costatus	Top-shell, Western Ribbed	USA	California			
rochidae		Unidentified							Top-shell, Unidentified					
rochidae	Gibbulinae	Cittarium		pica		Cittarium pica	L., 1758		Top-shell, West Indian					
rochidae	Monodontinae	Monodonta		labis		Monodonta labis	L., 1758		Monodonta, Labio					
rochidae	Monodontinae	Tegula		mariana		Tegula mariana	Dall, 1919		Tegula	Panama				
lymnaeidae		Lymnaea		stagnalis		Lymnaea stagnalis	L., 1758		Pond Snail, Stagnant	USA	New York		Seneca Lake	
ermetidae		Dendropoma		irregularis		Dendropoma irregularis	Orbigny, 1842		Worm-shell, Irregular					

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.

Custom Tabular Exporter

Content
Download
Upload
Option Code

Select and Order Columns to Export

ADP number

Cat Numb

Accession year

ACC_N

Former number

count in lot

Specimen identifier's name

Type

Size (mm)

Options for **ADP number**

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

[Help](#)

Use local time zone
 Omit time

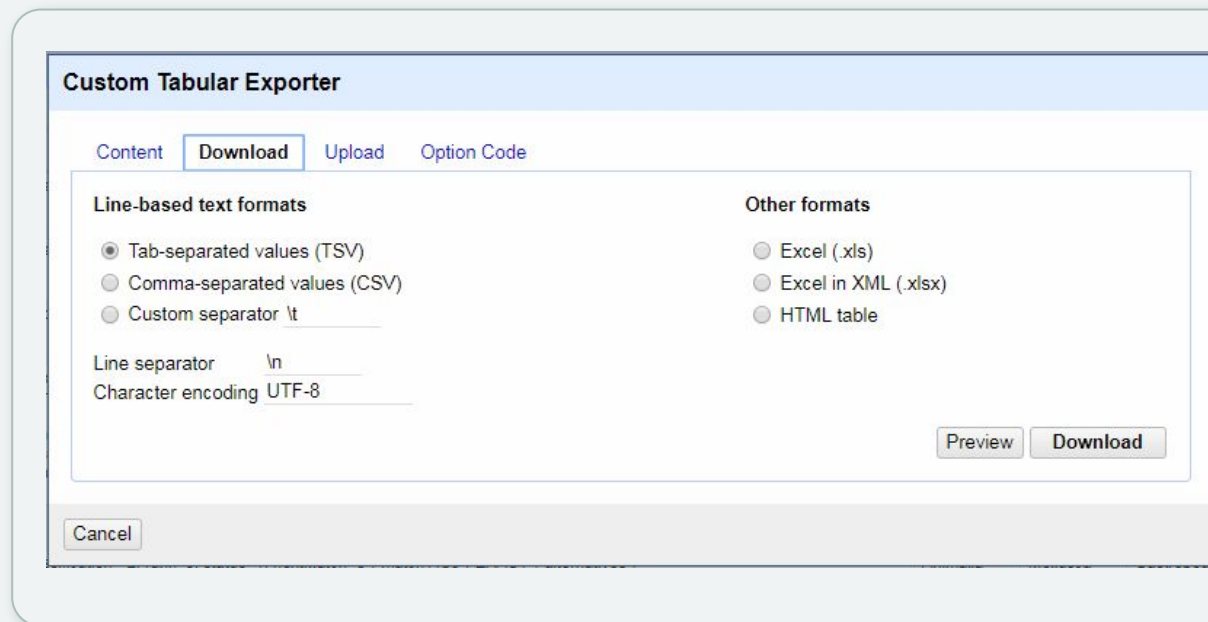
Select All
De-select All

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

Cancel

Getting Data Out of OpenRefine

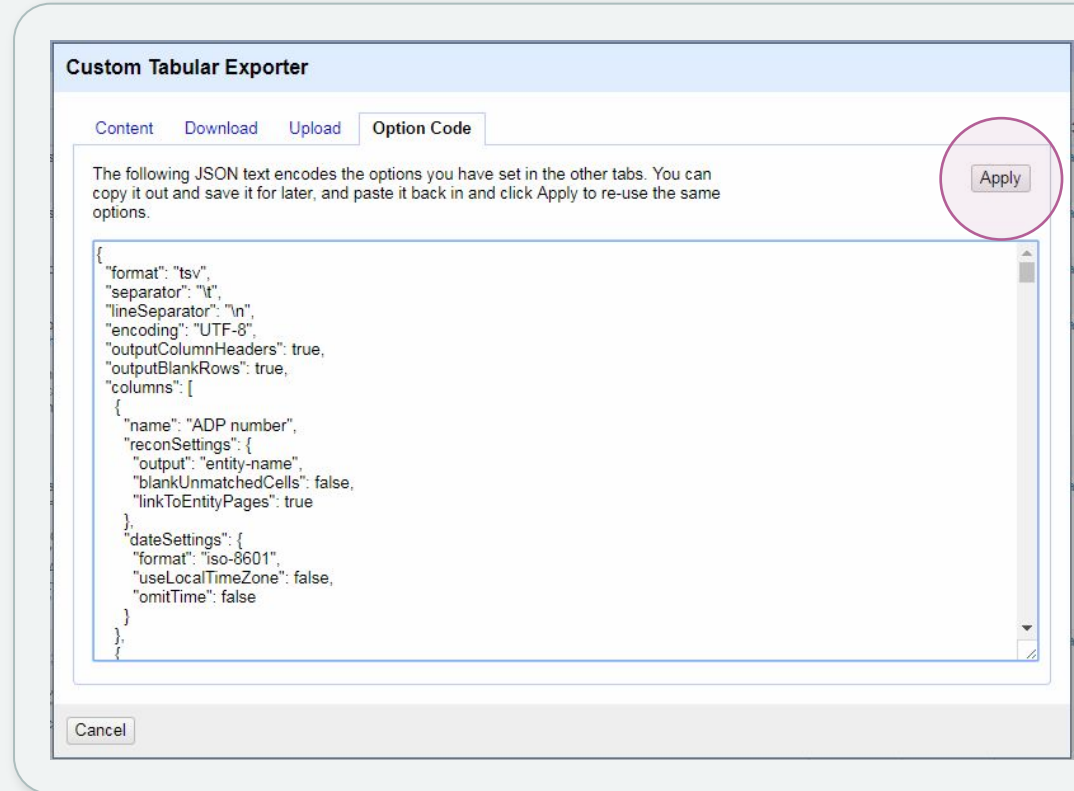
- Go to the **Download** tab and select a separator.
 - Don't modify the other options unless necessary.



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_SAqTb776E



Example Dataset

https://docs.google.com/spreadsheets/d/1Sdi-eT--SyaJFywSyFrLLwtCpQu2x6MOmK_EvXe4F0



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/>