



OpenRefine

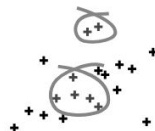
Richard Littauer, Wikicon

Main features



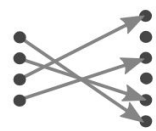
Faceting

Drill through large datasets using facets and apply operations on filtered views of your dataset.



Clustering

Fix inconsistencies by merging similar values thanks to powerful heuristics.



Reconciliation

Match your dataset to external databases via reconciliation services.



Infinite undo/redo

Rewind to any previous state of your dataset and replay your operation history on a new version of it.



Privacy

Your data is cleaned on your machine, not in some dubious data laundering cloud.



Wikibase

Contribute to Wikidata, the free knowledge base anyone can edit, and other Wikibase instances.

How

- Stored locally
- Runs in your browser
- Memory size is changeable
- Has extensions

What formats?

- comma-separated values (CSV) or text-separated values (TSV)
- Text files
- Fixed-width columns
- JSON
- XML
- OpenDocument spreadsheet (ODS)
- Excel spreadsheet (XLS or XLSX)
- PC-Axis (PX)
- MARC
- RDF data (JSON-LD, N3, N-Triples, Turtle, RDF/XML)
- Wikitext

Data types

- String
- Number
- Bool
- Dates - to ISO
- Null
- Error

Also, records:

Work	Actor	Role
The Wizard of Oz	Judy Garland	Dorothy Gale
	Ray Bolger	"Hunk"
		The Scarecrow
	Jack Haley	"Hickory"
		The Tin Man

Facet / Filter

Undo / Redo 0 / 0



10 rows

Extensions

Wikibase ▾

Refresh

Reset all

Remove all

countryLabel

change

7 choices Sort by: name count

Cluster

People's Republic of China 4

Bangladesh 1

Brazil 1

India 1

Japan 1

Nigeria 1

Turkey 1

Facet by choice counts

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

« first

< previous

1



- 10

next >

last »

All

cityLabel

populz

countryLabel



1. Shanghai 23390000 People's Republic of China



2. Beijing 21710000 People's Republic of China



3. Lagos 21324000 Nigeria



4. Dhaka 16800000 Bangladesh



5. Mumbai 15414288 India



6. Istanbul 14657434 Turkey



7. Tokyo 13942856 Japan



8. Tianjin 13245000 People's Republic of China



9. Guangzhou 13080500 People's Republic of China



10. São Paulo 12106920 Brazil

Facets



 **OpenRefine** Tutorial [Permalink](#)

Facet / Filter [Undo / Redo](#) 0 / 0

Refresh

Reset all

Remove all

  **countryLabel** [change](#)

7 choices Sort by: [name](#) [count](#)

Cluster

People's Republic of China 4

Bangladesh 1

Brazil 1

India 1

Japan 1

























Nigeria 1

Turkey 1

Facet by choice counts

10 rows

Show as: **rows** [records](#) Show: [5](#) [10](#) [25](#) [50](#) [100](#)

 All	 cityLabel	 popula	 countryLabel
 	1. Shanghai	23390000	People's Republic of China
 	2. Beijing	21710000	People's Republic of China
 	3. Lagos	21324000	Nigeria
 	4. Dhaka	16800000	Bangladesh
 	5. Mumbai	15414288	India
 	6. Istanbul	14657434	Turkey
 	7. Tokyo	13942856	Japan
 	8. Tianjin	13245000	People's Republic of China
 	9. Guangzhou	13080500	People's Republic of China
 	10. São Paulo	12106920	Brazil

Facets

10 rows

Show as: **rows** records Show: 5 10 25 50 100 500 1000 rows

▼ All	▼ cityLabel	▼ popula	▼ countryLabel
★	🗨	1.	Shanghai
★	🗨	2.	Beijing
★	🗨	3.	Lagos
★	🗨	4.	Dhaka
★	🗨	5.	Mumbai
★	🗨	6.	Istanbul
★	🗨	7.	Tokyo
★	🗨	8.	Tianjin
★	🗨	9.	Guangzhou
★	🗨	10.	São Paulo

Facet

Text filter

Edit cells

Edit column

Transpose

Sort...

View

Reconcile

Text facet

Numeric facet

Timeline facet

Scatterplot facet...

Custom text facet...

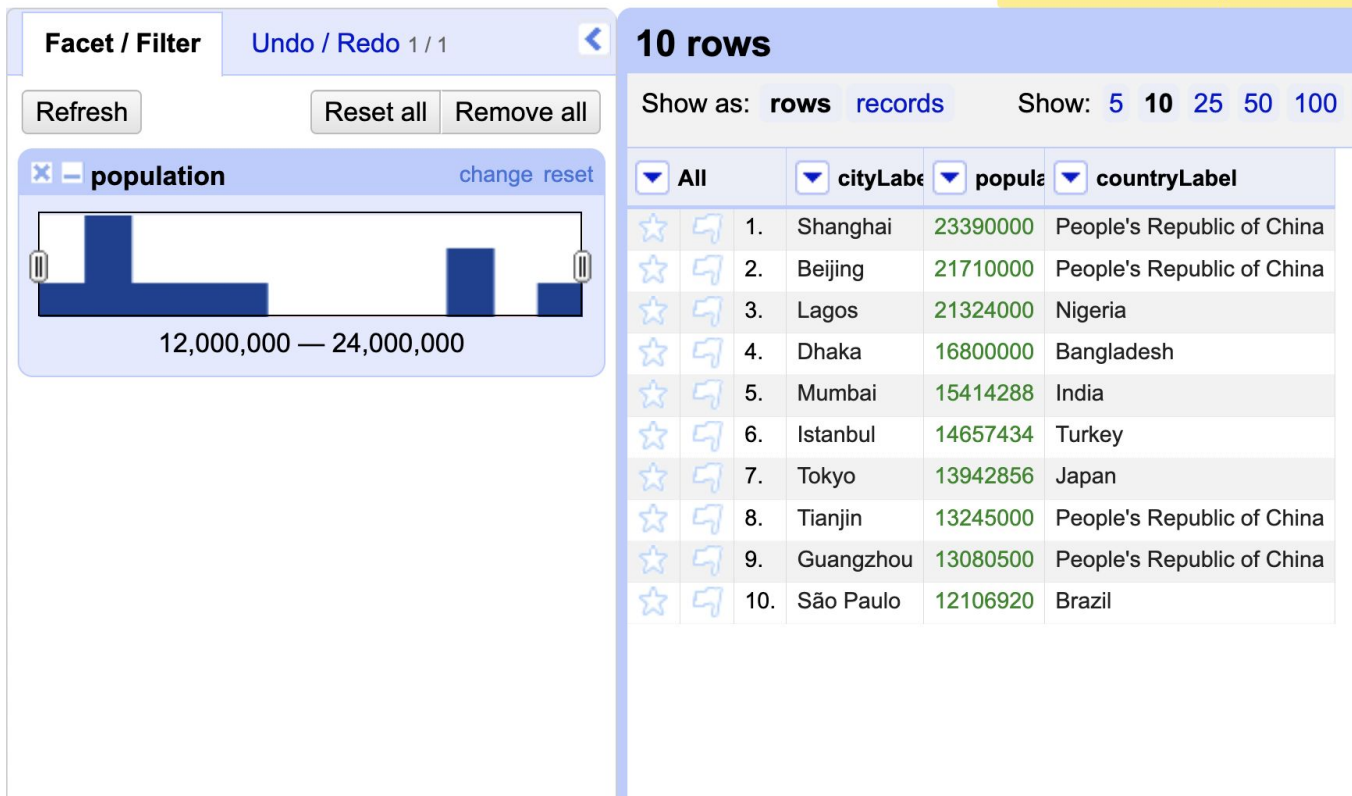
Custom numeric facet...

Customized facets

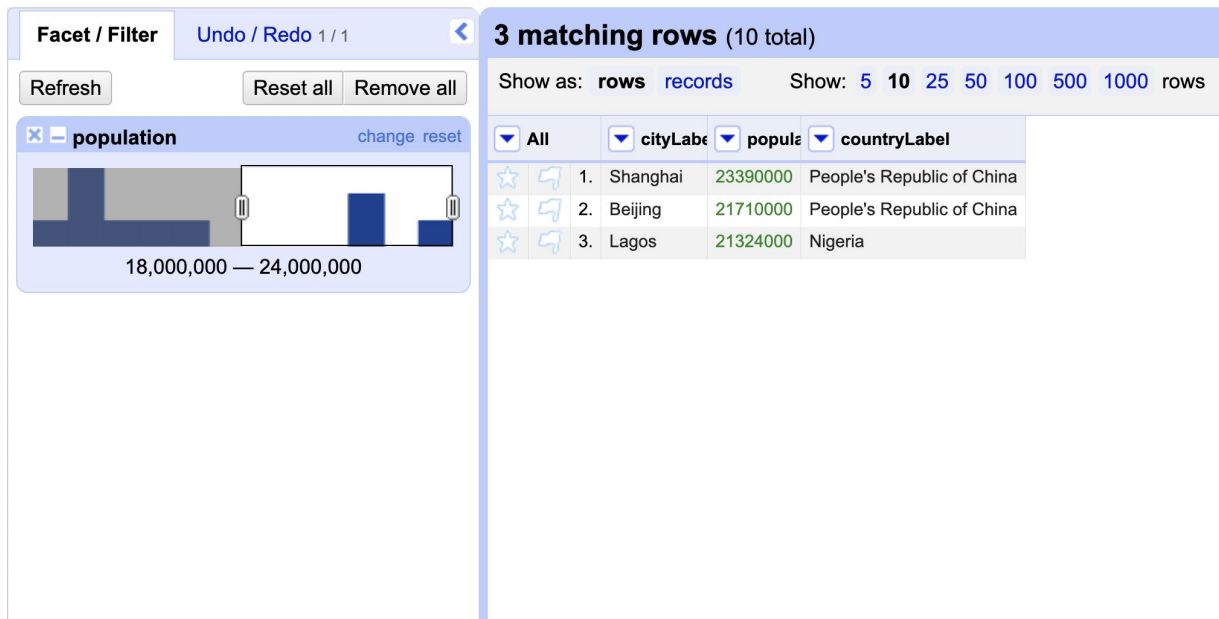
Facets

0 rows				
Show as: rows records		Show: 5 10 25 50 100 500 1000 rows		
		« first		
All	cityLabel	populz	countryLabel	
1.	Shanghai	Facet	public of China	
2.	Beijing	Text filter	public of China	
3.	Lagos			
4.	Dhaka	Edit cells	Transform...	
5.	Mumbai	Edit column	Common transforms	Trim leading and trailing whitespace
6.	Istanbul	Transpose	Fill down	Collapse consecutive whitespace
7.	Tokyo	Sort...	Blank down	Unescape HTML entities
8.	Tianjin	View	Split multi-valued cells...	Replace smart quotes with ASCII
9.	Guangzhou	Reconcile	Join multi-valued cells...	To titlecase
10.	São Paulo		Cluster and edit...	To uppercase
			Replace...	To lowercase
				To number
				To date
				To text
				To null
				To empty string

Facets

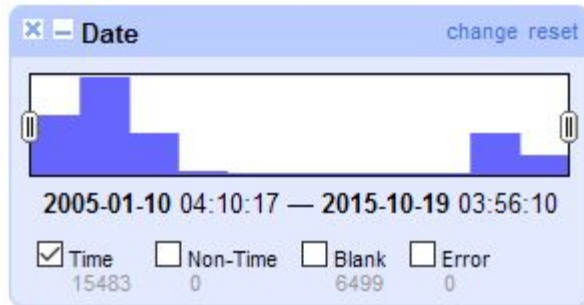


Facets

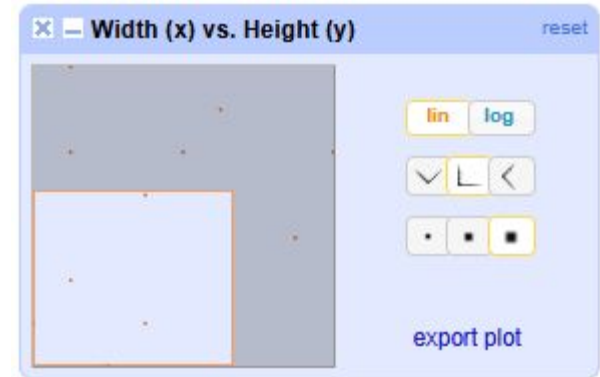


Facets

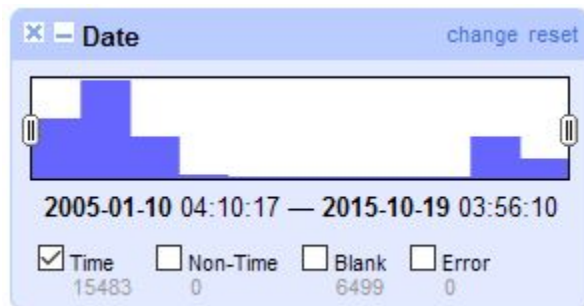
Timeline



Scatterplot



Facets



Facets

Regex

Facet / Filter

Undo / Redo 1 / 1

Refresh

Reset all

Remove all

countryLabel

invert

reset

☐ case sensitive ☒ regular expression

9 matching rows (10 total)

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

	All	cityLabel	popul	countryLabel
1.	Shanghai	23390000	People's Republic of China	
2.	Beijing	21710000	People's Republic of China	
3.	Lagos	21324000	Nigeria	
4.	Dhaka	16800000	Bangladesh	
5.	Mumbai	15414288	India	
7.	Tokyo	13942856	Japan	
8.	Tianjin	13245000	People's Republic of China	
9.	Guangzhou	13080500	People's Republic of China	
10.	São Paulo	12106920	Brazil	

Transforms

- Trim whitespace
- Collapse whitespace
- Unescape HTML
- Case tRaNsFoRmS
- Fill down, blank down
- Replace

Transforms

Clustering

Cluster & edit column "Place"

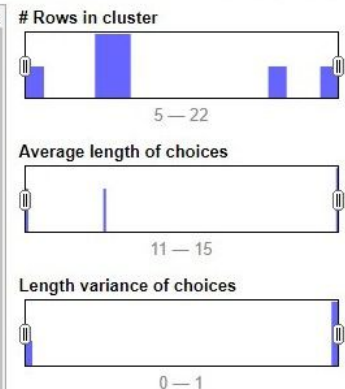
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer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

Method key collision

Keying Function Fingerprint

7 clusters found

Cluster size	Row Count	Values in cluster	Merge?	New cell value
2	10	<ul style="list-style-type: none">PRINCESS FIELD (6 rows)PRINCESS FIELD (4 rows)	<input type="checkbox"/>	<input type="text" value="PRINCESS FIELD"/>
2	22	<ul style="list-style-type: none">MOUNT ZION (18 rows)MOUNT ZION (4 rows)	<input type="checkbox"/>	<input type="text" value="MOUNT ZION"/>
2	10	<ul style="list-style-type: none">SWABYS HOPE (5 rows)SWABYS HOPE (5 rows)	<input type="checkbox"/>	<input type="text" value="SWABYS HOPE"/>
2	9	<ul style="list-style-type: none">BROGUE HILL (5 rows)BROGUE HILL (4 rows)	<input type="checkbox"/>	<input type="text" value="BROGUE HILL"/>
2	9	<ul style="list-style-type: none">BALLARDS RIVER (7 rows)BALLARDS RIVER (2 rows)	<input type="checkbox"/>	<input type="text" value="BALLARDS RIVER"/>
2	5	<ul style="list-style-type: none">SPRING MOUNTAIN (3 rows)MOUNTAIN SPRING (2 rows)	<input type="checkbox"/>	<input type="text" value="SPRING MOUNTAIN"/>
2	19	<ul style="list-style-type: none">GRAVEL HILL (17 rows)GRAVEL HILL (2 rows)	<input type="checkbox"/>	<input type="text" value="GRAVEL HILL"/>



Select all Deselect all

Export clusters

Merge selected & re-cluster

Merge selected & Close

Close

Editing columns

- Split
- Join
- Add column based on this column
- Transposing

Editing columns

Add column by fetching URLs

Add column by fetching URLs based on column Wikidata Entities

New column name

Throttle delay

5000

 milliseconds

On error

☒ set to blank ☐ store error

☒ Cache responses

HTTP headers to be used when fetching URLs: [Hide](#)

Authorization:

User-Agent:

OpenRefine 3.4-beta2 [c67e13b]

Accept:

/

Formulate the URLs to fetch:

Expression

Language

General Refine Expression Language (GREL) ▼

"https://www.wikidata.org/wiki/Special:EntityData/" + value +
".json"

No syntax error.

Preview

History

Starred

Help

row	value	"https://www.wikidata.org/wiki ..."
1.	Q1125633	https://www.wikidata.org/wiki/Special:EntityData/Q1125633.json
2.	Q1376408	https://www.wikidata.org/wiki/Special:EntityData/Q1376408.json
3.	Q1741127	https://www.wikidata.org/wiki/Special:EntityData/Q1741127.json
4.	Q2156146	https://www.wikidata.org/wiki/Special:EntityData/Q2156146.json
5.	Q2405269	https://www.wikidata.org/wiki/Special:EntityData/Q2405269.json
6.	Q2864894	https://www.wikidata.org/wiki/Special:EntityData/Q2864894.json
7.	Q2864894	https://www.wikidata.org/wiki/Special:EntityData/Q2864894.json



OK

Cancel

Reconciling

Use it to:

- fix spelling or variations in proper names
- clean up manually-entered subject headings against authorities such as the Library of Congress Subject Headings (LCSH)
- link your data to an existing dataset
- add to an editable platform such as Wikidata
- or see whether entities in your project appear in some specific list, such as the Panama Papers.

▼ All	▼ Artist	▼ Lifespan	▼ Profession
★ 	1. Anna Jóelsdóttir <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Hübschmannova, Anna (11) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Macková, Anna (11) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Neborová, Anna (11) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Tweeddale, Anna (11) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Anders, Anna (11) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Beeck, Anna (11) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Heindl, Anna (11) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Jermolaewa, Anna (11) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Waser, Anna (11) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Matoušková, Anna (11) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new item	born 1947	contemporary artist
★ 	2. Fridriksdottir, Gabriela Choose new match	born 1971	painter, sculptor
★ 	3. Helgadóttir, Gerður Choose new match	1928–1975	sculptor, stained-glass artist
★ 	4. Gunnfríður Jónsdóttir <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Jónsdóttir, Jóni (19) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new item	edit 1889–1968	sculptor
★ 	5. Sveinsdóttir, Júlíana Choose new match		
★ 	6. Sigurdardóttir, Katrín Choose new match		
★ 	7. Kristín Jónsdóttir <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Jónsdóttir, Jóni (19) <input checked="" type="checkbox"/> <input checked="" type="checkbox"/> Create new item		
★ 	8. Matthíasdóttir, Louisa Choose new match		
★ 	9. Margret the Adroit		

Match this Cell Match All Identical Cells Cancel

Jónsdóttir, Jóni
(Icelandic artist, born 1972)

Reconciling

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		▼ chemical eleme
☆	🗨	1. hydrogen Choose new match
☆	🗨	2. helium Choose new match
☆	🗨	3. lithium Choose new match
☆	🗨	4. beryllium Choose new match
☆	🗨	5. boron Choose new match
☆	🗨	6. carbon Choose new match
☆	🗨	7. nitrogen Choose new match
☆	🗨	8. oxygen Choose new match
☆	🗨	9. fluorine Choose new match
☆	🗨	10. neon Choose new match
☆	🗨	11. sodium Choose new match
☆	🗨	12. magnesium Choose new match
☆	🗨	13. aluminum Choose new match
☆	🗨	14. silicon Choose new match
☆	🗨	15. francium Choose new match
☆	🗨	16. phosphorus Choose new match
☆	🗨	17. iron Choose new match
☆	🗨	18. sulfur Choose new match
☆	🗨	19. chlorine Choose new match
☆	🗨	20. argon Choose new match
☆	🗨	21. potassium Choose new match
☆	🗨	22. calcium Choose new match
☆	🗨	23. lead Choose new match

Expressions

Add column based on column Numbers

New column name

On error

☒ set to blank ☐ store error ☐ copy value from original column

Expression

Language

General Refine Expression Language (GREL) ▾

value + 10

No syntax error.

Preview

History

Starred


Help

row	value	value + 10
1.	1	11
2.	2	12
3.	3	13
4.	4	14
5.	5	15
6.	6	16

OK

Cancel

Expressions

 **OpenRefine** Avilist [Permalink](#)

Facet / Filter

Undo / Redo 6 / 6

Refresh

Reset all

Remove all

scientific_name

invert reset

genia

☐ case sensitive ☒ regular expression

12 matching records (33,844 total)

Show as: rows records Show: 5 10 25 50 100 500 1000 records « first < previous

All	sort	taxon_r	protonym	order	family_na	family	famil	scientific_name		R	
☆	3107.	3107	species	Iotreron Eugeniae	Columbiformes	Columbidae	Pigeons, Doves	21	Pt scientific_name iae		
☆	9632.	9632	subspecies	Galbula melanogenia	Galbuliformes	Galbulidae	Jacamar s	93	Galbula ruficauda melanogenia		
☆	11695.	11695	subspecies	Conurus xanthogenius	Psittaciformes	Psittacidae	African & New World Parrots	106	Eupsittula pertinax xanthogenia		
☆	12098.	12098	species	Psittacula melanogenia	Psittaciformes	Psittaculidae	Old World Parrots	107	Nannopsittacus melanogenia		
☆	12099.	12099	subspecies	Psittacula melanogenia	Psittaciformes	Psittaculidae	Old World Parrots	107	Nannopsittacus melanogenia melanogenia		
☆	12100.	12100	subspecies	Cyclopsitta suavisima	Psittaciformes	Psittaculidae	Old World Parrots	107	Nannopsittacus melanogenia suavisissimus		
☆	12101.	12101	subspecies	Cydopsittacus fuscifrons	Psittaciformes	Psittaculidae	Old World Parrots	107	Nannopsittacus melanogenia fuscifrons		
☆	12235.	12235	species	Eos cyanogenia	Psittaciformes	Psittaculidae	Old World Parrots	107	Trichoglossus cyanogenia		

Expressions

- Python!

Add column based on column scientific_name

New column name

On error ☒ set to blank ☐ store error ☐ copy value from original column

Expression Language

```
import re

# Ensure values are strings and strip whitespace
scientific_name = str(value).strip()
protonym = str(cells["protonym"].value).strip()

# Define regex pattern for detecting the suffix
pattern = r"(pterus|ptera|pterus)$"

# Check if both scientific_name and protonym match the pattern
if re.search(pattern, scientific_name) and re.search(pattern, protonym):
    # Extract the suffix from protonym
    protonym_suffix = re.search(pattern, protonym).group(1)

    # Replace the suffix in scientific_name with the suffix from protonym
    new_name = re.sub(pattern, protonym_suffix, scientific_name)
    return new_name

# Return the original name if no change is needed
return scientific_name
```

No syntax error.

Preview [History](#) [Starred](#) [Help](#)

row	value	import re # Ensure values are ...
333.	Chloephaga melanoptera	Chloephaga melanopterus
364.	Tachyeres brachypterus	Tachyeres brachyptera
1139.	Lagopus lagopus leucoptera	Lagopus lagopus leucopterus
1633.	Rollandia microptera	Rollandia micropterus
2894.	Treron chloropterus	Treron chloroptera
2924.	Treron phoenicopterus	Treron phoenicoptera

29,884 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

« first < previous 1

▼ All	▼ bird	▼ common_name	▼ latin_name	▼ Subspe	▼ Birds	▼ cate	▼ species	▼ family	▼ Obs
☆ 🔊	1.	135	Levantine Shearwater	Yelkouan Shearwater Choose new match	yelkouan	false	Procellariinae	PROCELLARIIFORMES	Banica (C
☆ 🔊	2.	136	Roseate Tern	Roseate Tern Choose new match		false	Sterninae	CHARADRIIFORMES	Govt Mus
☆ 🔊	3.	137	Levantine Shearwater	Yelkouan Shearwater Choose new match	yelkouan	false	Procellariinae	PROCELLARIIFORMES	Banica (C
☆ 🔊	4.	138	Levantine Shearwater	Yelkouan Shearwater Choose new match	yelkouan	false	Procellariinae	PROCELLARIIFORMES	Banica (C
☆ 🔊	5.	139	Swinhoe's Storm-petrel	Swinhoe's Storm Petrel Choose new match	[leucorhoa]	false	Hydrobatinae	PROCELLARIIFORMES	Tomlinso
☆ 🔊	6.	140	Wedge-tailed Shearwater	Puffinus pacificus Choose new match		false	Procellariinae	PROCELLARIIFORMES	Anon (AN
☆ 🔊	7.	141	Little Shearwater	Little Shearwater Choose new match		true	Procellariinae	PROCELLARIIFORMES	Bourne (I
☆ 🔊	8.	142	Gould's Petrel	Gould's Petrel Choose new match		true	Fulmarinae	PROCELLARIIFORMES	Anon (AN
☆ 🔊	9.	143	Relict Gull	Relict Gull Choose new match		false	Laridae	CHARADRIIFORMES	Soderbor
☆ 🔊	10.	144	Albatrosses	albatross Choose new match		false	Diomedidae	PROCELLARIIFORMES	Mullen (J
☆ 🔊	11.	145	Emperor Penguin	Emperor Penguin Choose new match		false	Spheniscidae	SPHENISCIFORMES	Hamilton
☆ 🔊	12.	146	Black-bellied Storm-petrel	Black-bellied Storm Petrel Choose new match		true	Hydrobatinae	PROCELLARIIFORMES	Mayo (AL
☆ 🔊	13.	147	Leach's Storm-petrel	Leach's Storm Petrel Choose new match		false	Hydrobatinae	PROCELLARIIFORMES	Mayo (AL
☆ 🔊	14.	148	Arctic Skua	Parasitic Jaeger Choose new match		true	Stercorariida e	CHARADRIIFORMES	Jeans (B

Add columns from reconciled column latin_name

Add property

Suggested properties

[host](#)
[hymenium type](#)
[iNaturalist taxon ID](#)
[main food source](#)
[maximum viable temperature](#)
[minimum viable temperature](#)
[natural reservoir of](#)
[optimum viable temperature](#)
[parent taxon](#)
[taxon common name](#)
[taxon name](#)
[taxon range map image](#)
[taxon rank](#)
[taxonomic type](#)
[this taxon is source of](#)

Preview

Reset

Western Gull	4345	Larus occidentalis	Larus
Calonectris leucomelas	4189	Calonectris leucomelas	Calonectris
Brown Pelican	4328	Pelecanus occidentalis	Pelican
Heermann's Gull	4353	Larus heermanni	Larus
Phalaropus	3957	Phalaropus	Scolopacidae
Calonectris leucomelas	4189	Calonectris leucomelas	Calonectris
Common Murre	4519	Uria aalge	Uria
Larus glaucescens	4399	Larus glaucescens	Larus
Tufted Puffin	4509	Fratercula cirrhata	Puffin
Western Gull	4345	Larus occidentalis	Larus
Diomedea nigripes		Diomedea nigripes	Diomedea
Bonaparte's Gull	144502	Chroicocephalus philadelphia	Chroicocephalus
Ring-billed Gull	4364	Larus delawarensis	Larus
Larus glaucescens	4399	Larus glaucescens	Larus
Sterna anaethetus		Sterna anaethetus	Sterna

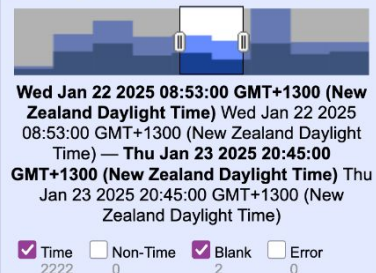
OK

Cancel

Facet / Filter [Undo / Redo](#) 20 / 20

Refresh Reset all Remove all

time_observed_at [change](#) [reset](#)



scientific_name [change](#)

206 choices Sort by: **name count** [Cluster](#)

- [Veronica hectorii](#) 1
- [Veronica macrantha](#) 1
- [Veronica odora](#) 1
- [Veronica pinguifolia](#) 1
- [Veronica salicifolia](#) 3
- [Veronica vernicosa](#) 2
- [Vespula vulgaris](#) 1
- [Wahlenbergia albomarginata](#) 7
- [Zeopocilus](#) 1
- [Zosterops lateralis lateralis](#) 2

Facet by choice counts

quality_grade [change](#) [invert](#) [reset](#)

2 choices Sort by: **name count** [Cluster](#)

- [needs_id](#) 380
- [research](#) 402 [exclude](#)

Facet by choice counts

402 matching rows (4,001 total)

Extensions [Wikibase](#)

Show as: **rows** records Show: 5 10 25 50 100 500 1000 rows

« first < previous 1 - 1258 next > last »

▼	priv	▼	priv	▼	publ	▼	geop	▼	taxo	▼	coor	▼	posit	▼	posit	▼	species_guess	▼	scientific_name	▼	NZTCS cont	▼	NZTCS	▼	common_name	▼	icon	▼	taxon	▼	v
					10				open		false						Kelp Gull		Larus dominicanus						Kelp Gull		Aves		4388		
					10						false						New Zealand tree fern		Dicksonia squarrosa						New Zealand tree fern		Plantae		125927		
					10						false						Māhoe Wao		Melicytus lanceolatus						Māhoe Wao		Plantae		366618		
					10						false						Blechnum membranaceum		Blechnum membranaceum								Plantae		400029		
					2398						false						Slender Path Rush		Juncus tenuis						Slender Path Rush		Plantae		69930		
					58						false						mangrove-leaved daisy-bush		Olearia avicenniifolia						mangrove-leaved daisy-bush		Plantae		336340		
					5						false						Raukaua anomalus		Raukaua anomalus								Plantae		366708		
					3						false						Ozothamnus vauvilliersii		Ozothamnus vauvilliersii								Plantae		1600651		
					4						false						Exocarpos bidwillii		Exocarpos bidwillii								Plantae		401834		
					27754				obscure d		true						Tauhinu		Ozothamnus leptophyllus						Tauhinu		Plantae		349798		

Extensions

- RefineJS
- Google Sheets
- GeoJSON
- Open Street Maps
- FAIR
- AI

Thanks