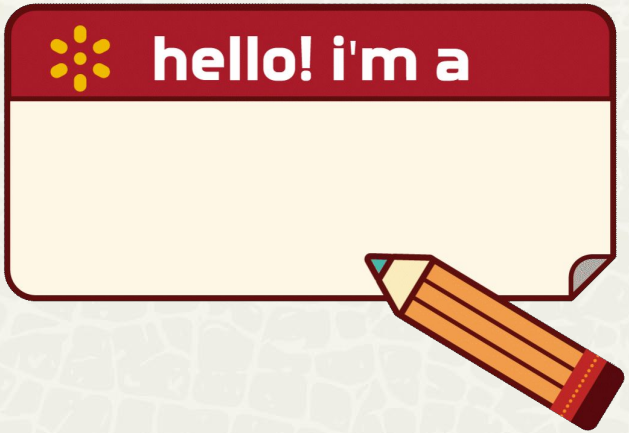
Duplicating Everywhere All At Once

Fixing geographic duplicates in Wikidata, Cebuano Wikipedia and Geonames

WIKIMANIA SINGAPORE

Alex Lum / Canley

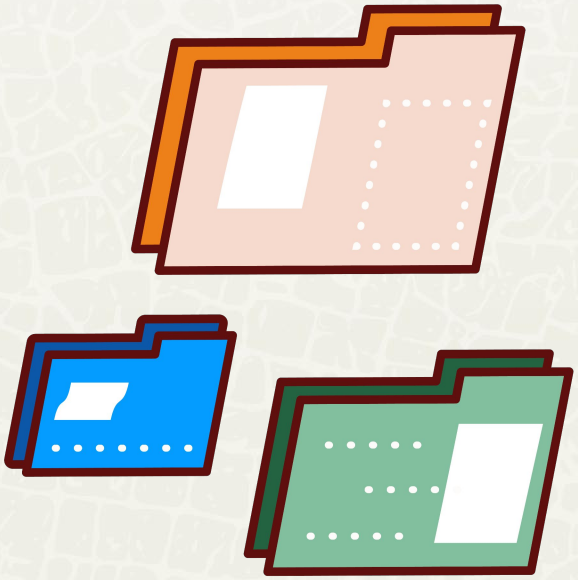


WIKIMANIA
SINGAPORE

Hi, I'm Alex (User:Canley)

metacoretechs
on Twitter X/Mastodon and GitHub

I want everywhere to have comprehensive geographic data on Wikidata!



WIKIMANIA
SINGAPORE

GitHub repository:
https://github.com/metacoretechs/geo-dedup

What's the problem?

WIKIMANIA
SINGAPORE

What's the problem?

From 2014 to 2019, Lsjbot generated 9.5 million articles on the Cebuano and Swedish Wikipedias.

In the years since, other bots created Wikidata items for these articles.

https://en.wikipedia.org/wiki/Lsjbot

What's the problem?

Lsjbot generated articles from various datasets.

For geographic places and objects, it used GeoNames, an openly-licenced, user-contributed global gazetteer.

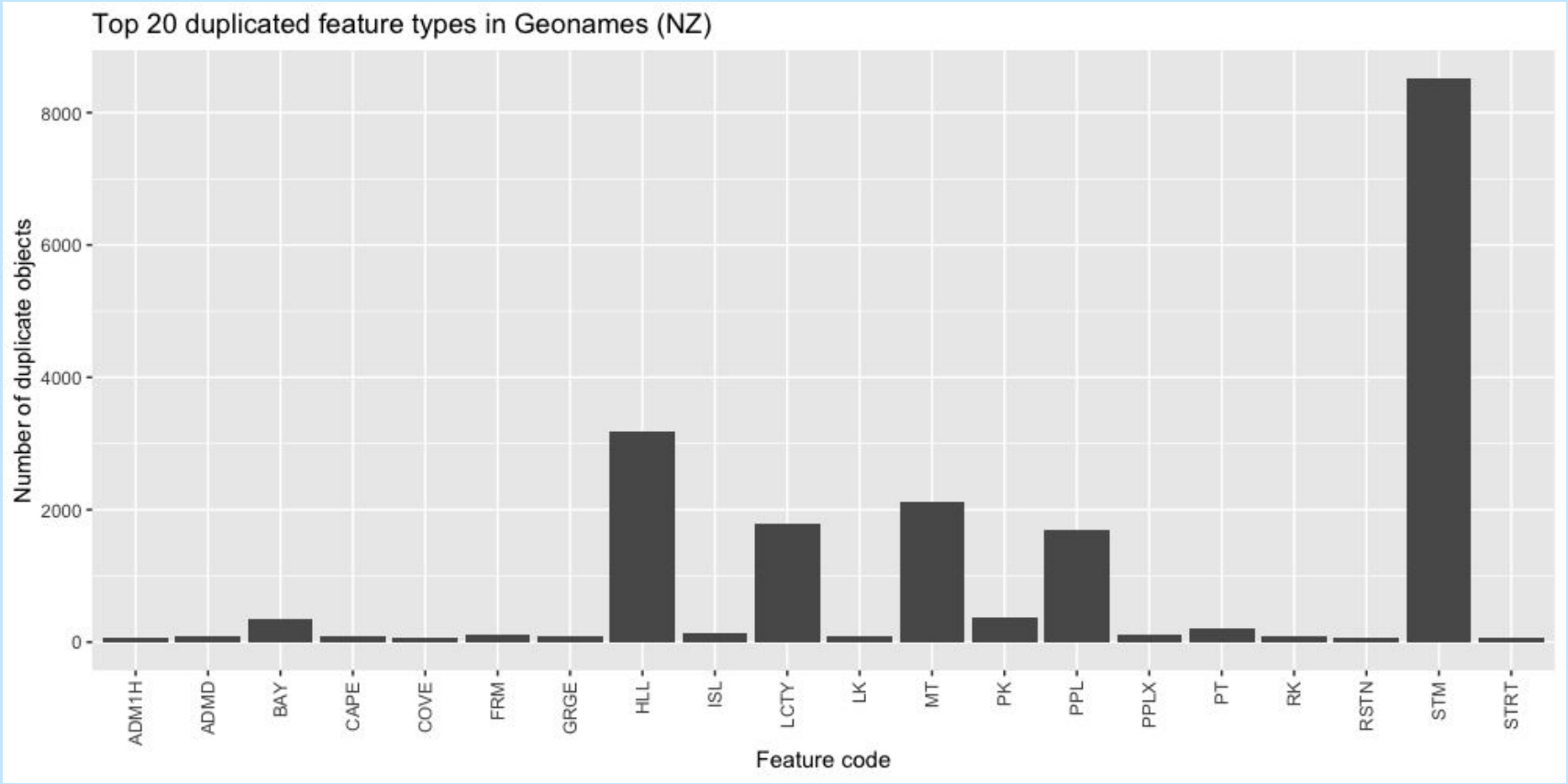
https://www.geonames.org/

What's the problem?

Unfortunately, there are issues with the accuracy of GeoNames data:

- coordinates are often highly inaccurate
- duplicated items, where places or objects have been imported to GeoNames multiple times

D. Ahlers, "Assessment of the Accuracy of GeoNames Gazetteer Data," in Proceedings of the 7th Workshop on Geographic Information Retrieval, Lyon: Association for Computing Machinery, Nov. 2013, pp. 74–81. doi: 10.1145/2533888.2533938.



What's the problem?

The duplicates problem seems to be occurring mainly for certain feature types.

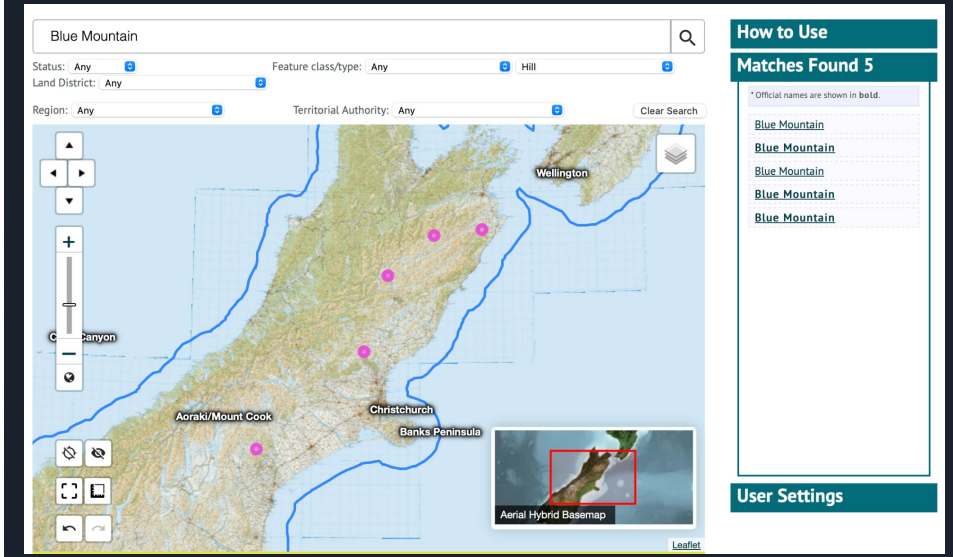
- Mountains and hills
- Rivers and streams

WIKIMANIA SINGAPORE

Let’s look at some examples...

“Blue Mountain” filtered by Hill feature type in the New Zealand Gazetteer:

5 results



WIKIMANIA SINGAPORE

Let’s look at some examples...

“Blue Mountain” filtered by mountain/hill/rock feature type in GeoNames:

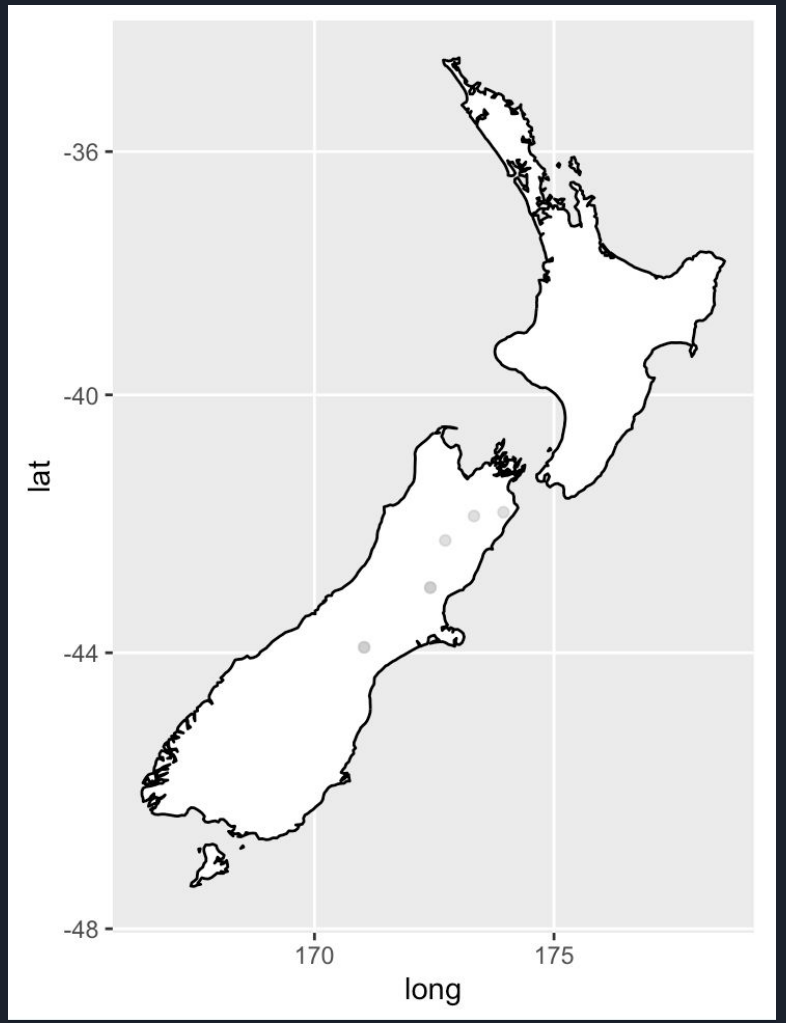
7 results

WIKIMANIA SINGAPORE

	Name	Country	Feature class	Latitude	Longitude
1	Blue Mountain	New Zealand, Marlborough Marlborough District	mountain elevation 2061m	S 41° 54' 38''	E 173° 19' 44''
2	Blue Hill	New Zealand, Nelson Nelson City	mountain	S 41° 12' 0''	E 173° 27' 0''
3	Blue Mountain	New Zealand, Canterbury Mackenzie District	mountain	S 43° 55' 0''	E 171° 2' 0''
4	Blue Mountain	New Zealand, Canterbury Hurunui District	mountain	S 43° 1' 0''	E 172° 25' 0''
5	Blue Mountain	New Zealand, Marlborough Marlborough District	mountain elevation 1243m	S 41° 51' 7''	E 173° 56' 44''
6	Blue Mountain	New Zealand, Canterbury Hurunui District	hill	S 42° 17' 17''	E 172° 43' 48''
7	Blue Mountain	New Zealand, Canterbury Hurunui District	hill	S 43° 0' 29''	E 172° 25' 12''
8	Blue Mountain Pass	New Zealand, Canterbury Mackenzie District	pass	S 43° 56' 54''	E 171° 2' 24''
9	Blue Mountain	New Zealand, Canterbury Mackenzie District	hill	S 43° 55' 6''	E 171° 2' 24''
10	BLUE MOUNTAIN RANGE	New Zealand, Marlborough Marlborough District	mountains	S 41° 51' 26''	E 173° 55' 37''
11	Blue Hill	New Zealand, Canterbury Selwyn District	mountain	S 43° 17' 0''	E 171° 38' 0''

Let’s look at some examples...

“Blue Mountain” items in GeoNames plotted on a map of New Zealand

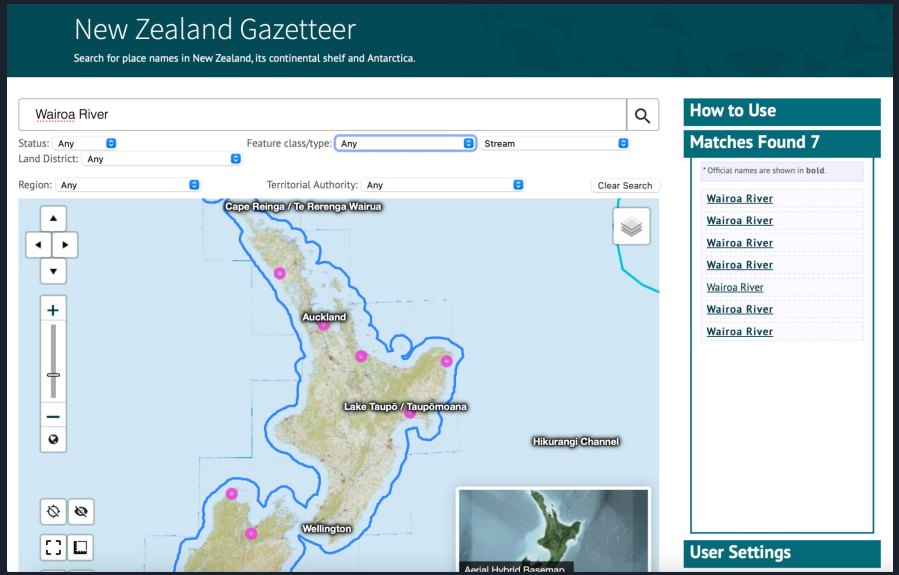


WIKIMANIA SINGAPORE

Let’s look at some examples...

“Wairoa River” filtered by Stream feature type in the New Zealand Gazetteer:

7 results



WIKIMANIA SINGAPORE

Let’s look at some examples...

“Wairoa River” filtered by stream/river feature type in GeoNames:

23 results

WIKIMANIA SINGAPORE

Wairoa River | New Zealand

Feature Class: stream, lake, ...

Continent: all

fuzzy search :

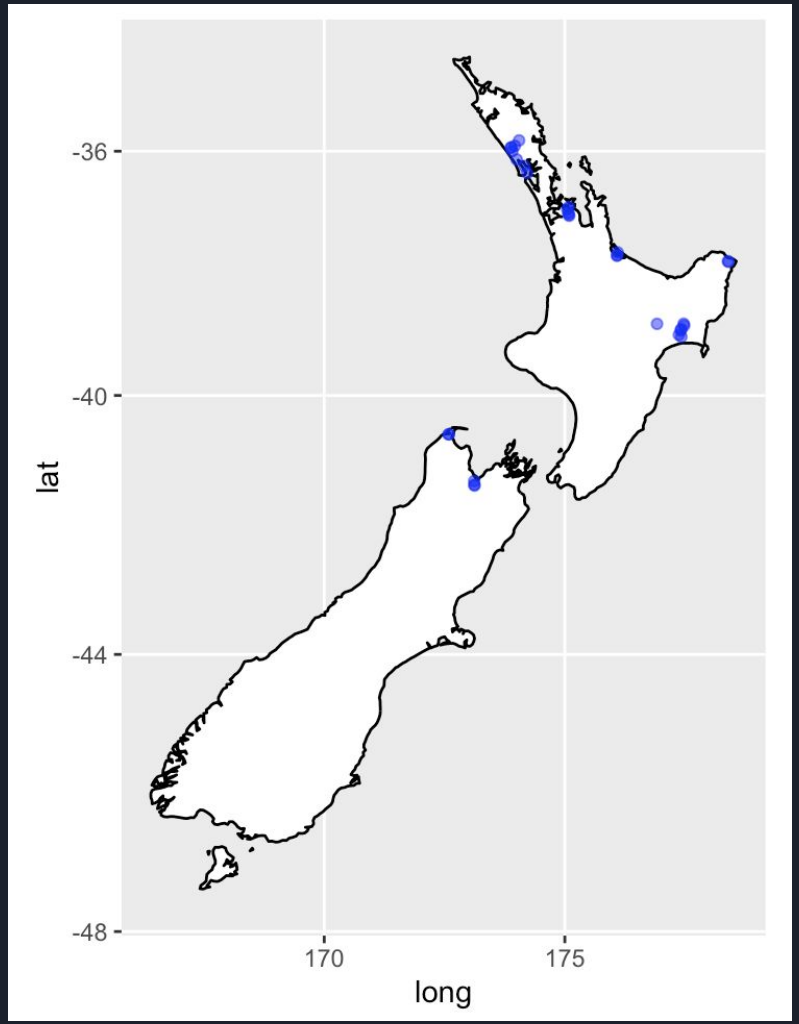
search

32 records found for "Wairoa River"

	Name	Country	Feature class	Latitude	Longitude
1	Wairoa River	New Zealand,	stream	S 37° 41' 0''	E 176° 6' 0''
2	Wairoa River	New Zealand, Nelson	stream	S 41° 24' 29''	E 173° 7' 12''
3	Wairoa Stream Wairoa River	New Zealand,	stream	S 39° 50' 0''	E 174° 37' 0''
4	Wairoa River	New Zealand, Hawke's Bay	stream	S 38° 57' 29''	E 177° 24' 36''
5	Wairoa River	New Zealand, Gisborne	stream	S 38° 50' 53''	E 177° 28' 12''
6	Wairoa River	New Zealand, Auckland	stream	S 36° 14' 53''	E 174° 9' 36''
7	Wairoa River	New Zealand, Auckland	stream	S 36° 58' 5''	E 175° 3' 0''
8	Right Branch Wairoa River Right Branch,Wairoa River Right Branch	New Zealand,	stream	S 41° 29' 0''	E 173° 5' 0''
9	Left Branch Wairoa River Left Branch,Wairoa River Left Branch	New Zealand,	stream	S 41° 29' 0''	E 173° 5' 0''
10	Wairoa River	New Zealand,	stream	S 36° 22' 0''	E 174° 13' 0''
11	Wairoa River	New Zealand,	stream	S 40° 37' 0''	E 172° 35' 0''
12	Wairoa River	New Zealand,	stream	S 39° 3' 35''	E 177° 25' 20''
13	Wairoa River	New Zealand,	stream	S 36° 56' 0''	E 175° 5' 0''
14	Wairoa River	New Zealand,	stream	S 38° 51' 0''	E 176° 55' 0''
15	Wairoa River Waitoa River	New Zealand,	stream	S 41° 21' 0''	E 173° 7' 0''
16	South Branch Wairoa River South Branch	New Zealand,	stream	S 40° 37' 0''	E 172° 36' 0''
17	North Branch Wairoa River North Branch	New Zealand,	stream	S 40° 37' 0''	E 172° 36' 0''
18	Left Branch Wairoa River	New Zealand, Nelson	stream	S 41° 35' 17''	E 173° 6' 0''
19	Right Branch Wairoa River	New Zealand, Nelson	stream	S 41° 30' 29''	E 173° 2' 24''
20	Left Branch Wairoa River	New Zealand, Nelson	stream	S 41° 31' 5''	E 173° 6' 0''
21	Wairoa Stream Wairoa River	New Zealand,	stream	S 35° 27' 0''	E 173° 18' 0''
22	Wairoa River	New Zealand, Gisborne	stream	S 37° 49' 41''	E 178° 23' 24''
23	Wairoa River	New Zealand, Hawke's Bay	stream	S 39° 1' 41''	E 177° 22' 12''
24	Wairoa River	New Zealand, Gisborne	stream	S 38° 52' 41''	E 177° 28' 48''
25	Wairoa River	New Zealand, Auckland	stream	S 36° 8' 17''	E 174° 0' 0''
26	Wairoa River	New Zealand, Auckland	stream	S 37° 43' 57''	E 176° 5' 19''
27	Wairoa River	New Zealand, Auckland	stream	S 35° 59' 53''	E 173° 54' 36''
28	Wairoa River	New Zealand, Auckland	stream	S 35° 49' 5''	E 174° 3' 0''
29	Wairoa River	New Zealand, Auckland	stream	S 35° 54' 29''	E 173° 57' 36''
30	Wairoa River	New Zealand, Auckland	stream	S 35° 56' 17''	E 173° 52' 48''
31	Wairoa River	New Zealand, Auckland	stream	S 37° 1' 41''	E 175° 4' 12''
32	Wairoa River	New Zealand, Auckland	stream	S 37° 4' 41''	E 175° 5' 24''

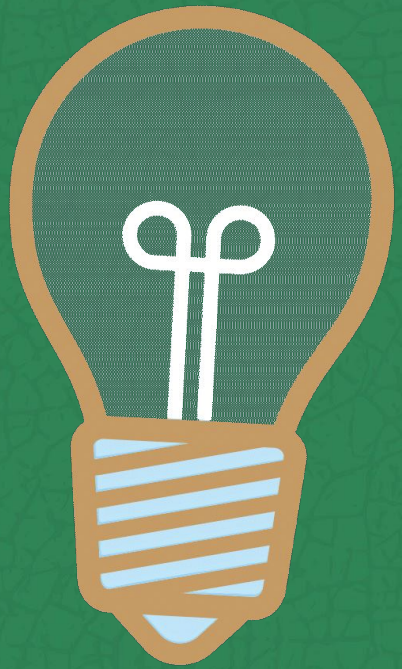
Let’s look at some examples...

“Wairoa River” items in GeoNames plotted on a map of New Zealand



WIKIMANIA SINGAPORE

How can this be fixed?



WIKIMANIA
SINGAPORE

Fixing issues on Wikidata

New Zealand Gazetteer Mix ‘n’ Match catalogue:

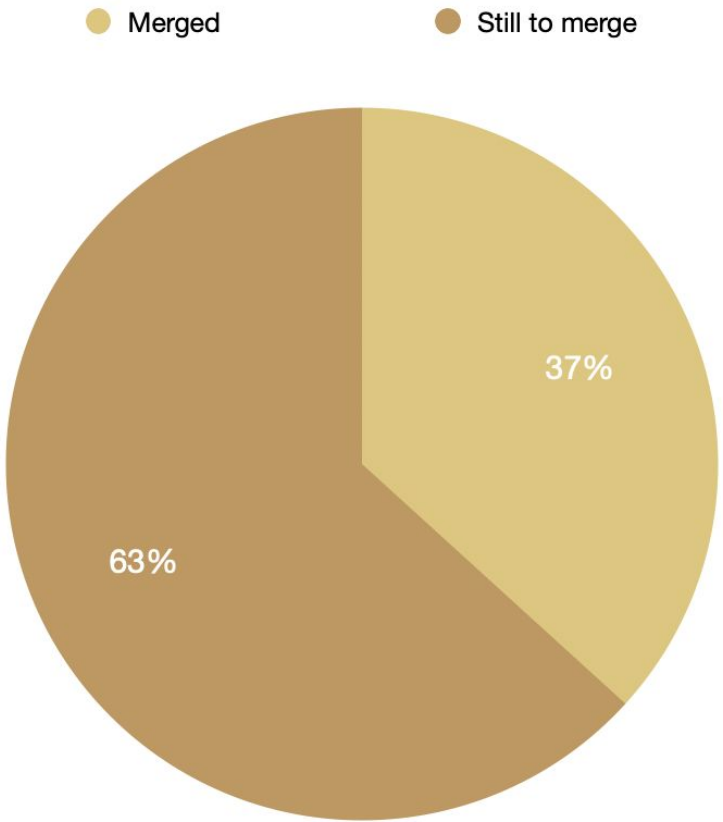
https://mix-n-match.toolforge.org/#/catalog/2857

- Reconcile items on Wikidata to the LINZ ID in Mix ‘n’ Match
- Identify duplicated items using clustering algorithms
- Remove Cebuano sitelinks and merge duplicate items into single item
- Replace coordinates referenced to Cebuano Wikipedia with LINZ gazetteer coordinates

The last two can be quite easily done in the QuickStatements tool.

Fixing issues on Wikidata

- 3,487 items have been manually merged on Wikidata into 1,499 distinct items
- This is about 37% of identified duplicate items



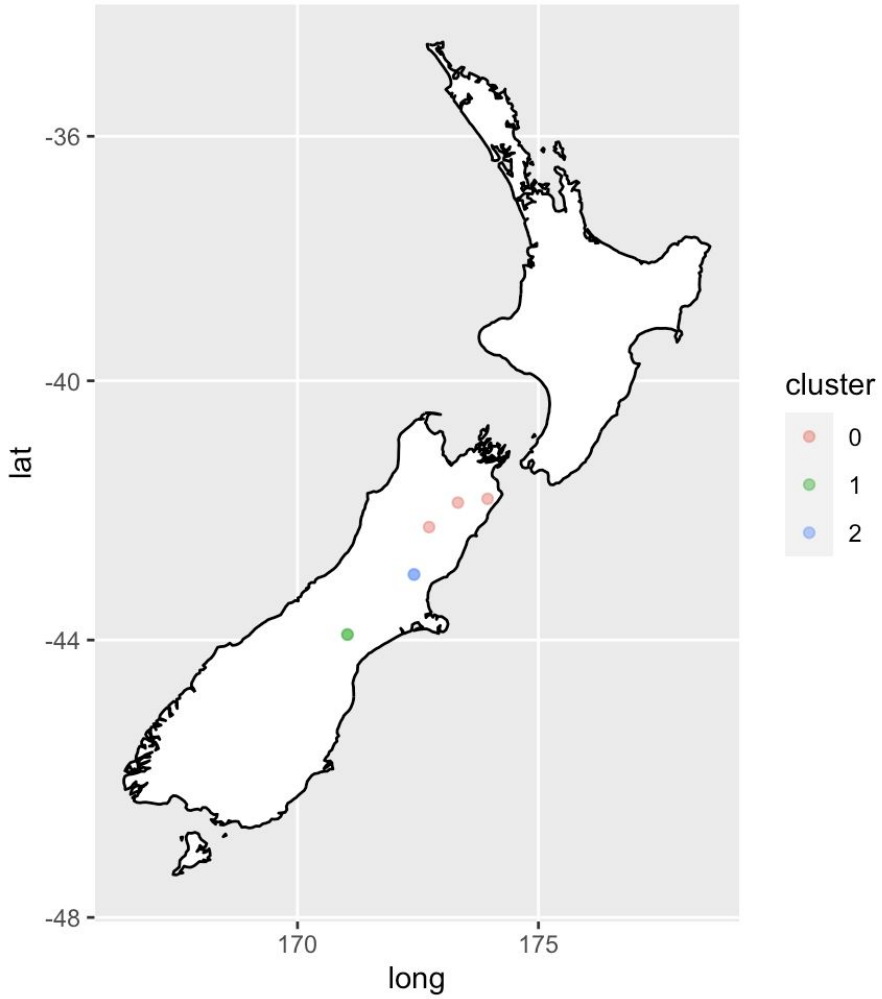
WIKIMANIA SINGAPORE

Identifying duplicates using clustering

“Blue Mountain”

- k-means clustering works well for mountains
- items without a duplicate are not clustered (cluster = 0)

WIKIMANIA SINGAPORE

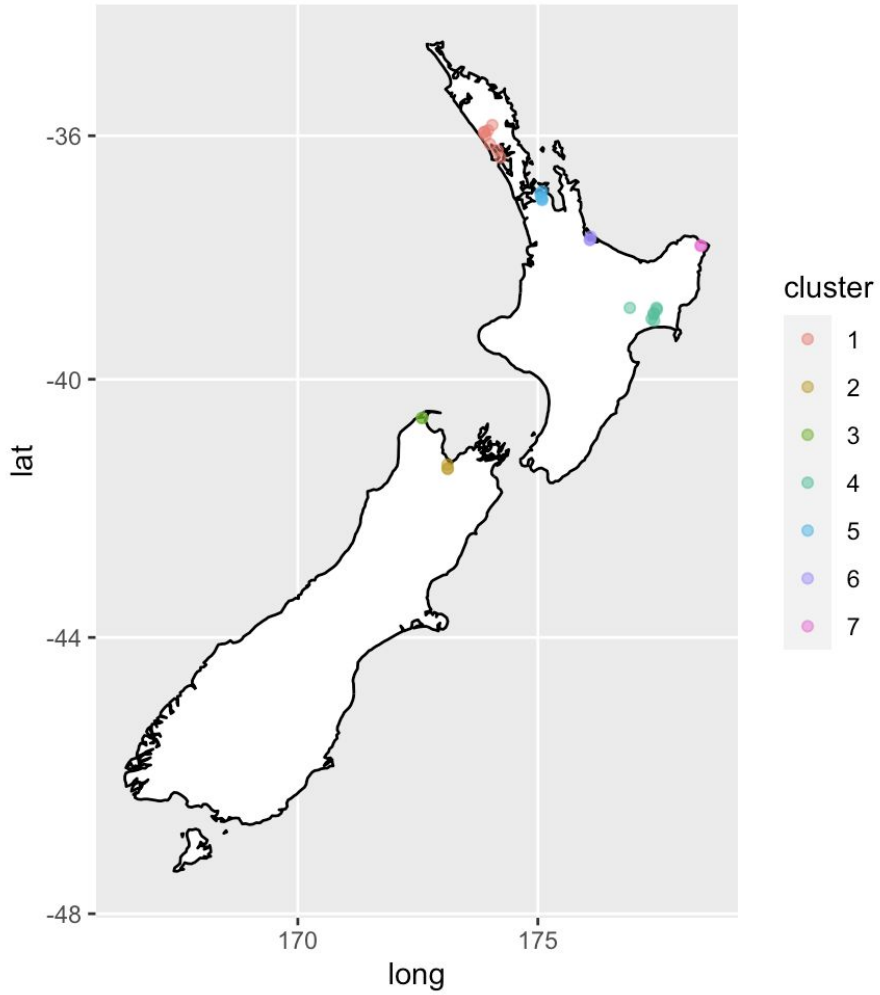


Identifying duplicates using clustering

“Wairoa River”

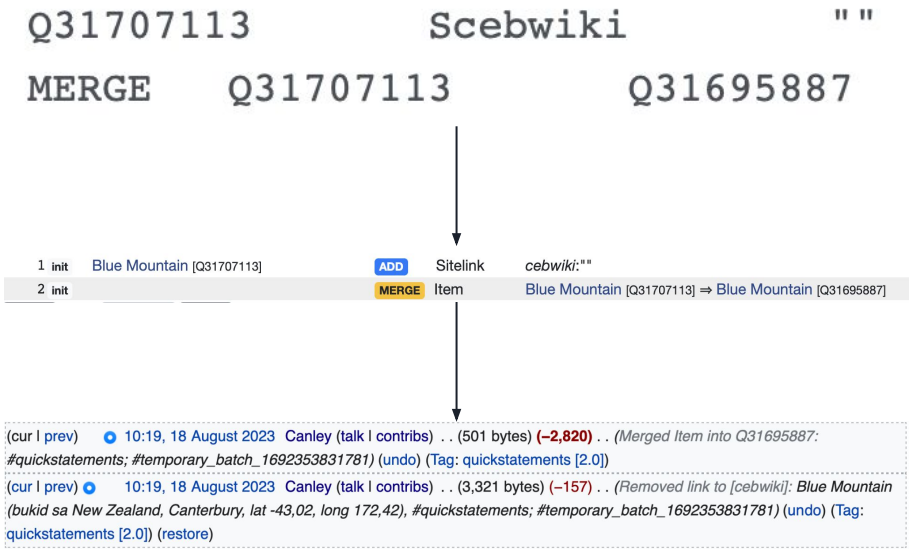
- DBSCAN works well for linear clusters with multiple points, such as rivers, streams and mountain ranges

WIKIMANIA SINGAPORE



Merging duplicates in Wikidata

- Can be done in QuickStatements
- First remove the Cebwiki sitelink by replacing it with an empty string “”
- Then use the MERGE command to merge and redirect to the desired target item

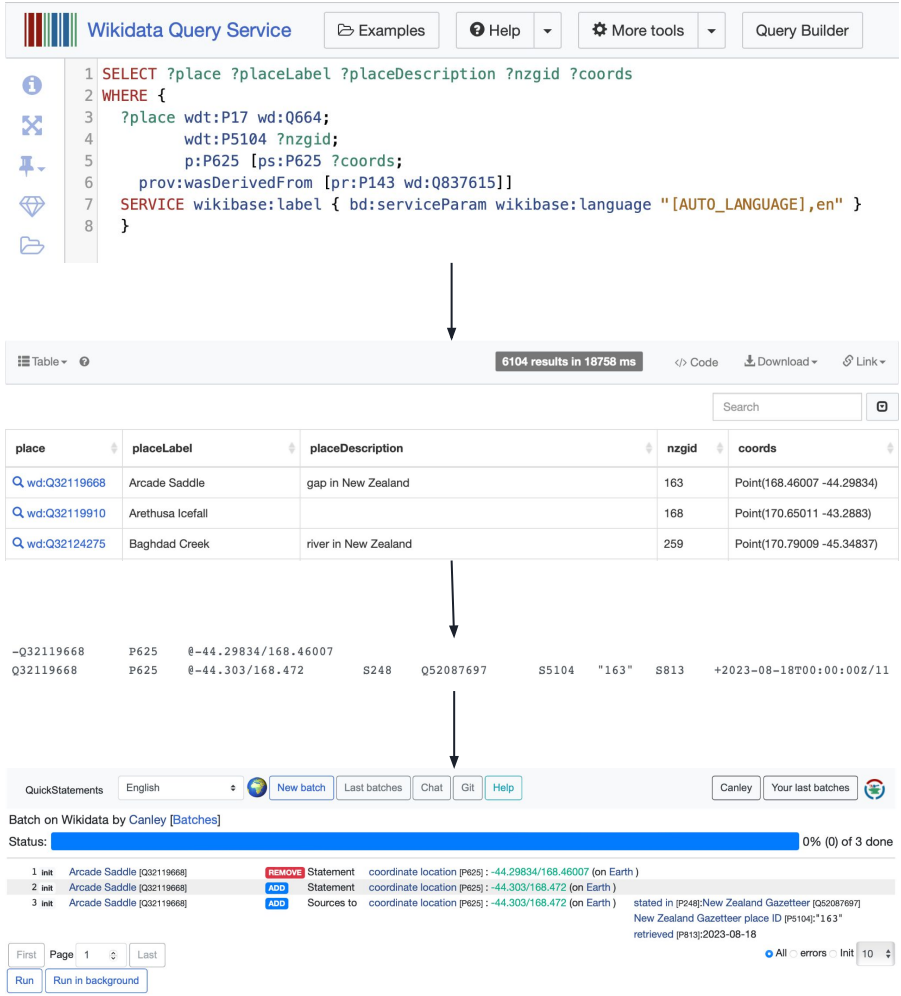


WIKIMANIA SINGAPORE

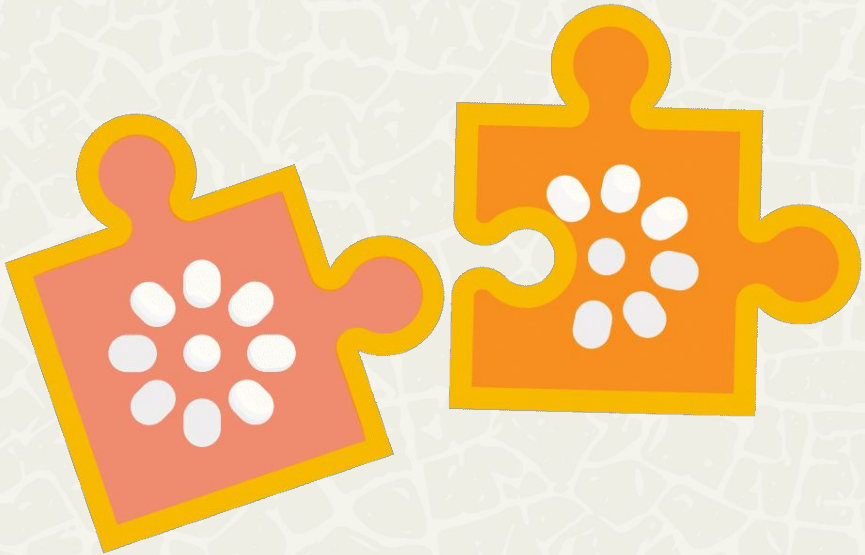
Fixing coordinates

- Run a SPARQL query to list New Zealand geographic features with coordinates cited to the Cebuano Wikipedia: https://w.wiki/7CCb
- Use QuickStatements to replace the coordinates with the more accurate ones from the LINZ Gazetteer

WIKIMANIA SINGAPORE



What about Cebuano Wikipedia and GeoNames?

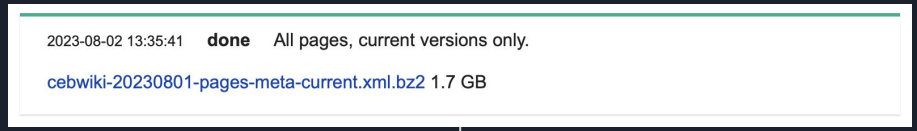


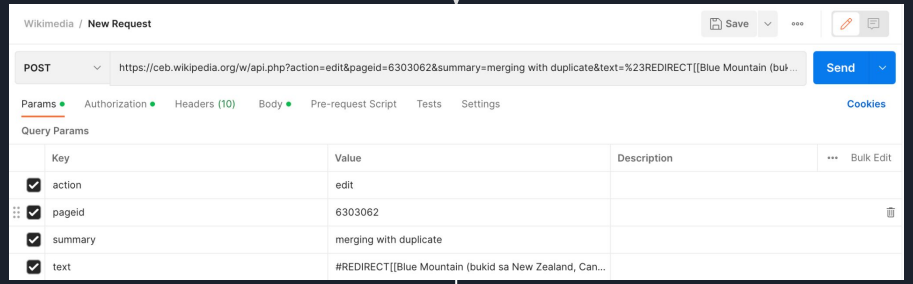
WIKIMANIA
SINGAPORE

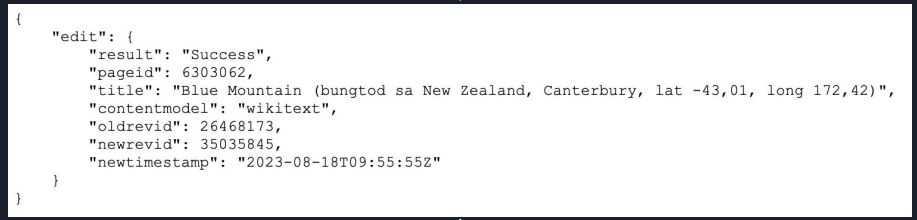
Merging duplicates on Cebuano Wikipedia

- It’s possible using a dump of the Cebuano Wikipedia and the MediaWiki API
- You will need to authenticate your account to get an OAuth token to edit article text

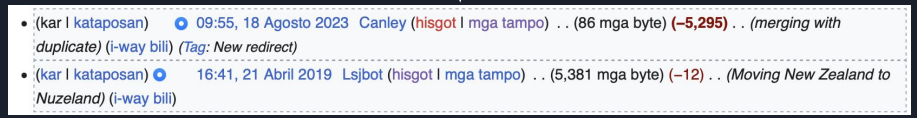
WIKIMANIA SINGAPORE







```
{
    "edit": {
        "result": "Success",
        "pageid": 6303062,
        "title": "Blue Mountain (bungtod sa New Zealand, Canterbury, lat -43,01, long 172,42)",
        "contentmodel": "wikitext",
        "oldrevid": 26468173,
        "newrevid": 35035845,
        "newtimestamp": "2023-08-18T09:55:55Z"
    }
}
```



Merging duplicates on Cebuano Wikipedia

See the GitHub repository for instructions on how to use the API to merge/redirect duplicate articles on a Wikipedia edition

WIKIMANIA SINGAPORE

Merging duplicates on GeoNames

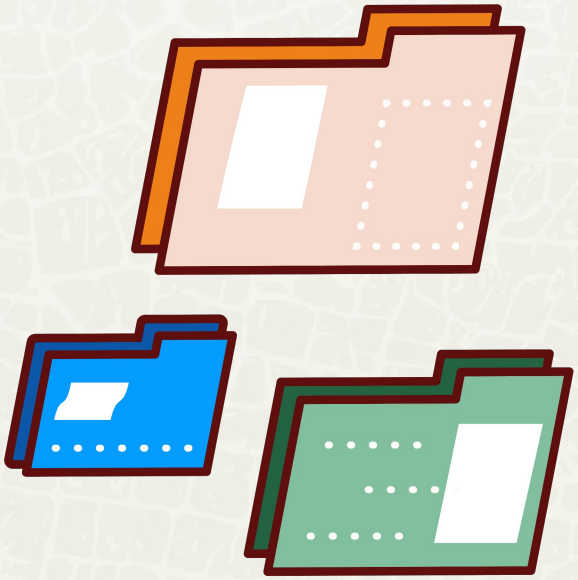
Duplicates and data issues can be reported on the GeoNames discussion board:

http://forum.geonames.org/

An administrator can (hopefully) then update and fix the data...

WIKIMANIA SINGAPORE

Topic	Answers	Author	Views	Last message
Any Sparql End Point for Geonames	1	sashedher	151188	25/05/2023 11:05:42 HoraceKHering
GeoNames on Facebook	6	geotree	248305	16/11/2011 22:51:13 Voyagepedia
Profile not available	1	hotroduhoc001	88149	07/11/2021 15:00:06 marc
feature codes should not use https:	0	valexiev	72189	19/02/2021 17:25:33 valexiev
Collection of geonames SQL queries	5	a110y	228855	19/02/2019 11:05:01 AlexandraHudson
Get city by id	1	randomtrip	7136	01/08/2023 17:16:39 marc
Listed in cities500 yet population 0	1	jimmyff	2371	01/08/2023 17:07:34 marc
places locked in user level 1	6	LibTed	157875	01/08/2023 17:03:05 marc
"The" in "The Netherlands" interfering with search results	0	nfriedly	2376	01/08/2023 16:49:10 nfriedly
GeoNames information correct but JSON is different	0	jwhitney	2376	01/08/2023 16:49:10 jwhitney
Uniformity of Japanese Prefecture Names	0	yuokada	2379	01/08/2023 16:49:10 yuokada
modification date in geoname table	0	tmbdrogba	7686	14/07/2023 11:33:17 tmbdrogba
Japanese Mountains mixed up	1	Perfunct	17982	16/06/2023 10:07:04 marc
Complete list of tags used in get "alternateNames" field	0	mfeltes	10313	20/06/2023 23:22:15 mfeltes
Adress field / IBAN validation	0	HoraceKHering	12837	05/06/2023 09:35:14 HoraceKHering



WIKIMANIA
SINGAPORE

GitHub repository:
https://github.com/metacoretechs/geo-dedup

WIKIMANIA
SINGAPORE