Duplicating Everywhere All At Once

Fixing geographic duplicates in Wikidata, Cebuano Wikipedia and Geonames

Alex Lum / Canley
Hello! I'm a

metacoretechs
on Twitter X/Mastodon and GitHub

I want everywhere to have comprehensive geographic data on Wikidata!
GitHub repository:
https://github.com/metacoretechs/geo-dedup
What’s the problem?
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
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/
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

What's the problem?

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

- Mountains and hills
- Rivers and streams
Let’s look at some examples…

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

5 results
Let’s look at some examples...

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

7 results
Let’s look at some examples…

“Blue Mountain” items in GeoNames plotted on a map of New Zealand
Let’s look at some examples...

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

7 results
Let’s look at some examples...

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

23 results
Let’s look at some examples...

“Wairoa River” items in GeoNames plotted on a map of New Zealand
How can this be fixed?
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
Identifying duplicates using clustering

“Blue Mountain”

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

“Wairoa River”

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

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

**Wikidata Query Service**

```
SELECT ?place ?placeLabel ?placeDescription ?nzgid ?coords
WHERE {
  ?place wdt:P17 wd:0664;
  wdt:P5104 ?nzgid;
  p:P625 [ps:P625 ?coords];
  prov:wasDerivedFrom [pr:P143 wd:0837615]]
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en" }
```
What about Cebuano Wikipedia and GeoNames?
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
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.
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…
GitHub repository: https://github.com/metacoretechs/geo-dedup
GeoNames website and data is licensed under a Creative Commons Attribution 4.0 License

New Zealand Gazetteer by the NZGB is licensed under a Creative Commons Attribution 4.0 International License