

Project 3

Data Wrangling with MongoDB

Derek Gurney

Map Area: Vancouver, BC, Canada

<https://mapzen.com/data/metro-extracts/#vancouver-canada>

[1. Problems encountered in my map](#)

[2. Overview of the Data](#)

[File sizes](#)

[Number of documents](#)

[Number of nodes](#)

[Other ideas about the datasets](#)

1. Problems encountered in my map

An inspection of the data revealed several problems with the map data: abbreviated, inconsistent, or misspelled street types; misspelled city names; inconsistent province name; and misformatted postal codes. All of these were cleaned programmatically; additional issues are discussed below.

Abbreviated, inconsistent, or misspelled street types

Street types in the data set were abbreviated; for example, “Street” was represented as “St.” The abbreviations were also inconsistent, with “Street” sometimes abbreviated as “St” and sometimes as “St.” Lastly, street types were sometimes misspelled; for example “venue” instead of “Avenue.”

Misspelled city names

“Vancouver” was sometimes spelled as “Vancovuer” and was sometimes uncapitalized.

Inconsistent province name

“British Columbia” was alternately represented as “BC”, “B.C.”, and “bc”.

Misformatted postal codes

Postal codes in Canada are formatted as “LetterNumberLetter NumberLetterNumber”, but many of the postal codes were represented either without capitalization or without the middle space.

Unaddressed issues

A number of other issues were left unaddressed: city names that are outside of the map area (e.g. Dresden), provinces outside map area (e.g. Ontario), and incomplete postal codes (e.g. V6C).

2. Overview of the Data

File sizes

Vancouver_canada.osm: 175.8 MB

Vancouver_canada.osm.json: 203.3 MB

Number of documents

```
> db.yvr.find().count()  
3235218
```

Number of nodes

```
> db.yvr.find({"type":"node"}).count()  
2772348
```

Number of ways

```
> db.yvr.find({"type":"way"}).count()  
462801
```

Number of unique users

```
> db.yvr.distinct("created.user").length
```

```
1147
```

<https://workflowy.com/#/6c1576358cfd>

Number of cities

```
> db.yvr.distinct("address.city").length
```

```
24
```

Number of documents with amenity = "cafe"

```
> db.yvr.find({"amenity":"cafe"}).count()
```

```
706
```

Number of cafes

```
> db.yvr.distinct( "id" ,{"amenity" : "cafe"}).length
```

```
353
```

Number of Crematoriums

```
> db.yvr.distinct( "id" ,{"amenity" : "Crematorium"}).length
```

```
1
```

Number of documents with name "Starbucks" or "Starbucks"

```
> db.yvr.distinct({$or:[{"name":"Starbucks"}, {"name":"Starbucks"}]}).count()
```

```
183
```

Number of Starbucks (or Starbucks')

```
> db.yvr.distinct( "id" ,{$or:[{"name":"Starbucks"}, {"name":"Starbucks"}]}).length
```

```
61
```

Number of Tim Hortons (or Tim Hortons'), a competing coffee chain

```
> db.yvr.distinct( "id" ,{$or:[{"name":"Tim Hortons"}, {"name":"Tim Hortons"}]}).length
```

Other ideas about the dataset

The map data can provide additional insights about the geography and economics of the area. Continuing the comparison of Starbucks and Tim Hortons, it would be interesting to compare how the two chains locate themselves:

- Does Starbucks locate in areas with younger people?
- Are Tim Hortons more likely to be located near highways?
- How close are Starbucks located together?

Combined with user data, the map data can also provide insights into how errors are introduced in the map and how they might be prevented. Questions in this vein would include:

- Do most new users create errors when they are new and then improve, or do only some users create most of the errors, both while they are new and later? The answer to this question would inform how to train new users: broadly, or targetted at users who have started with errors. The answer may also suggest editing policy: limit the ability of all new users to make changes from the beginning, or limit new users' ability to make changes only after they've made a certain number of errors.
- What types of geographical entities are most associated with errors? Ways or nodes? Addresses or names? Amenity classification or location? Answers to these questions would help site administrators develop "guard rails" that limited errors while encouraging participation.