OpenStreetMap Project - Data Wrangling with MongoDB

Jorge Saldivar

Map Area: Gran Asunción, Paraguay
Extract URL:
https://s3.amazonaws.com/mapzen.odes/ex_sxYbT8Ujp3e8HGSbSEMzRUN2rqBkA.osm.bz2

Justification of the extract selection

I am originally from Paraguay, specifically from Asunción. I found this project as an exciting opportunity to explore the quality of my hometown map in OpenStreetMap. I obtained the data from the Mapzen website after generating a custom extract of the interested area.

Problems found on the map

After exploring a sample of data, I discovered issues in the name of the streets. Also, I found that some tag elements have keys named “type,” which forced me to use a different key name to identify the type of element in my MongoDB.

Inconsistent street name prefixes

Firstly, the prefix of the street names (e.g., Ave.) was inconsistent. Some of them, use the full name *Avenida* (Avenue in Spanish) while others employ abbreviations (Avda.)

Example before the update
Avda. Bonifacio Obando Esq. Villa Florida

Example after the update
Avenida Bonifacio Obando Esq. Villa Florida

Inconsistent person’s name titles

In Asuncion, streets are primarily named using names of personalities from politics, science, literature, militia, etc. In all cases, the street names of these historical people have prepended the title of the person, e.g., dr., ing. (engineer in Spanish). I found inconsistencies in the use of these titles, sometimes the full word is used (e.g., Doctor) other times an abbreviation of the title is employed (e.g., Dr.).

Example before the update
Dr. G. Rodríguez de Francia

Example after the update

Doctor G. Rodríguez de Francia

Spanish names were wrongly written

I found street names improperly written; specifically, they missed the accent.

Example before the update
Cerro Cora

Example after the update
Cerro Corá

Before inserting into the MongoDB, I fixed these inconsistencies using the function *update_street_name* of the module *audit_street_name*

Overview of the data

Here, I present some high-level information about the data

File size: granasuncion_paraguay.osm ... 88.5 MB
Collection size: asuncion ... 34.5 MB

Number of records

```
> db.asuncion.count()
451918
```

Number of nodes

```
> db.asuncion.find({'type': 'node'}).count()
391867
```

Number of ways

```
> db.asuncion.find({'type': way}).count()
60051
```

Number of unique contributors

```
> db.asuncion.aggregate([{'$group': {'_id': '$created.uid'}}]).toArray().length
728
```

Top-ten most productive contributors

```
> db.asuncion.aggregate([{'$group': {'_id': '$created.uid', 'user': {'$first': '$created.uid'},
'contributions': {'$sum': 1}}}, {'$sort': {'contributions': -1}}, {'$limit': 10}])
{ "_id" : "16881", "user" : "16881", "contributions" : 67948 }
{ "_id" : "6691543", "user" : "6691543", "contributions" : 35845 }
```

{ "_id" : "5634607", "user" : "5634607", "contributions" : 27417 }
{ "_id" : "6691431", "user" : "6691431", "contributions" : 27403 }
{ "_id" : "3093724", "user" : "3093724", "contributions" : 26526 }
{ "_id" : "6691964", "user" : "6691964", "contributions" : 25993 }
{ "_id" : "494122", "user" : "494122", "contributions" : 24762 }
{ "_id" : "6735634", "user" : "6735634", "contributions" : 21920 }
{ "_id" : "6691451", "user" : "6691451", "contributions" : 17543 }
{ "_id" : "5236803", "user" : "5236803", "contributions" : 14620 }

Most common tags in nodes

Among the nodes that are tagged, the top-ten most common tags are:

Tag	Nodes in which the tag appear	Usage
Name	4691	Used to report the name of the node
Source	3932	Used to report the source of the information contained in the node
Address	2506	Used to provide information about addresses
Shop	2230	Used to provide the name of nodes associated with shops
Amenity	2032	Used to describe the service associated with the node (e.g., school, pharmacy, restaurant, fast food, place of worship)
Highway	792	Used in case the node is related to traffic (e.g., traffic_signals, crossing, bus stop, turning circle, stop, rest area)
Natural	284	Used when the node represents an element of nature (e.g., tree, peak, spring, land)
Building	277	Used in case the node represents a building to indicate its type (e.g., church, chapel, school, residential, warehouse)
Place	255	Used when the node is associated with a territory (e.g., neighbourhood, city, village, region)
Phone	220	Used to indicate the phone number associated with the node

Most common amenity

Schools are the most common amenity with 627 occurrences. Pharmacies and restaurants occupied the second and third place with 257 and 204 occurrences, respectively. These results are not surprising for someone living in Asuncion since going around one can see that the city full of pharmacies. It seems that here having a pharmacy is a very profitable business.

Other ideas about the dataset

Most of the nodes and ways do not have tags but only the minimum attributes, namely: id, creation date time, and position.

Number of nodes without tags

In particular, 383916 nodes have not been tagged, which represent about 98% of the total number of nodes (391867). As they are now, nodes are not particularly useful since they do not provide details about their purpose on the map. More work is needed if we want to have a more helpful map of Asuncion. On the other hand, I found that all ways are equipped with tags that describe their type and location, among additional information.

An idea to solve the problem of “anonymous” nodes could be to expand the use of OpenStreetMap in Paraguay. Since the quality of OpenStreetMap maps depends exclusively on contributions of the open-source community and considering that very few people know about the service and use it, it is not strange that the map of Asuncion is incomplete and misses a lot of information. People in Paraguay used more alternative services, like Google Maps or Waze. So, one way to address this problem can be to enlarge the community of users, who would eventually help in making the map more informative.