OpenStreetMap Data Case Study

Introduction:

This project involves choosing an area of the World in https://www.openstreetmap.org (https://www.openstreetmap.org) and using Data munging, wrangling and cleaning techniques to access data for valitidy, accuracy and completeness and hopefully clean the data.

To complete this project, I will be choosing an area of the World in https://www.openstreetmap.org (https://www.openstreetmap.org) that I care about and using SQL data schema to complete this Project.

Map Area

Lagos, Nigeria http://overpass-api.de/api/map?bbox=2.8482,6.4040,3.9894,6.6755 (http://overpass-api.de/api/map?bbox=2.8482,6.4040,3.9894,6.6755)

This map is an approximate map of Lagos state where I was born and I am quite familiar with the places and quirks. I tried to use the metro extract for Lagos but it seemed to focus on what was known as Lagos City and actually excludes other parts of Lagos state! Additionally it was only 49MB in size.

Problems encountered in Map

I downloaded the Map and running then map through audit.py, check_street_names.py and Investigating the data with SQL, I noticed the following issues.

1. Street names have a number of different types for Rd, Street, and there seems to be a Huge number of 'Thomases'

   Aguda: 1 Avenue: 4 close: 1 Close: 1 Cres: 2 Crescent: 9 Drive: 2 Iyalla: 1 kirikiri: 1 Rd: 4 Rd.: 1 road: 1 Road: 19 Street: 51 Thomas: 183 Way: 8
2. The key type city had three different designations for Lagos:

   Lagos,41 Ikeja,7 Ketu,7 lagos,4 Lagaos,3 50,1 "Ilupeju, Lagos",1 "Lekki lagos",1 "Lekki, Lagos",1 Surulere,1 "Surulere, Lagos",1
3. Incorrect PostCode Values

   I used the audit_post code function to print out any postcodes that didn't match the expected 6 digit format of
   lagos state postcodes. The incorrect post codes are:

   {'+234': ['+234', '+234', '+234', '+234'], '23401': ['23401', '23401'], '71510': ['71510'], 'Yaba': ['Yaba']}
4. I noticed quite a number of tags that had the key "Fixme":

> sqlite> select tags.key, count() *from (select key from nodes_tags union all select key from ways_tags) as tags ...> group by tags.key ...> order by count()* desc ...> limit 10;
>
> highway|38363 source|9775 created_by|6561 name|5328 oneway|4552 building|3627 surface|1931 fixme|1093 country|983 GNS:dsg_code|975

Inconsistent Street Names

Having had a look at the data using my audit.py script and confirming with a quick check with SQL (I had initally imported the data without any cleaning), I decided to use scripts from the case study that utilised regular expressions and the mapping function to correct the most obvuius inconsistent street names.

I used the following update_name function in cleaning.py to fix the inconsistent street names.

However, I was left with Aguda which is a town, Kirikiri is a Road and a town and Iyalla and the still mysterious case of the Thomases. Looking at the individual data for Kirikiri, this value corresponds to the key street, so i can safely assume that this refers to the Road and I will map it accordingly.

Running audit.py to look at the list of unexpected names for the unexpected street types, I can see that one street accounts for all the 'Thomas' street type.

'Thomas': set(['Bode Thomas'])

This is a street in the Surulere LGA commonly called Bode Thomas Street but as far as I can remember, it always just been called Bode Thomas. It is therefore no surprise that those contribuing have only put in "Bode Thomas" without the 'Street' at the end!

To resolve this, I added "Thomas": "Thomas Street" to my mapping dictionary so that all the "Bode Thomas" entries were changed to Bode Thomas Street.

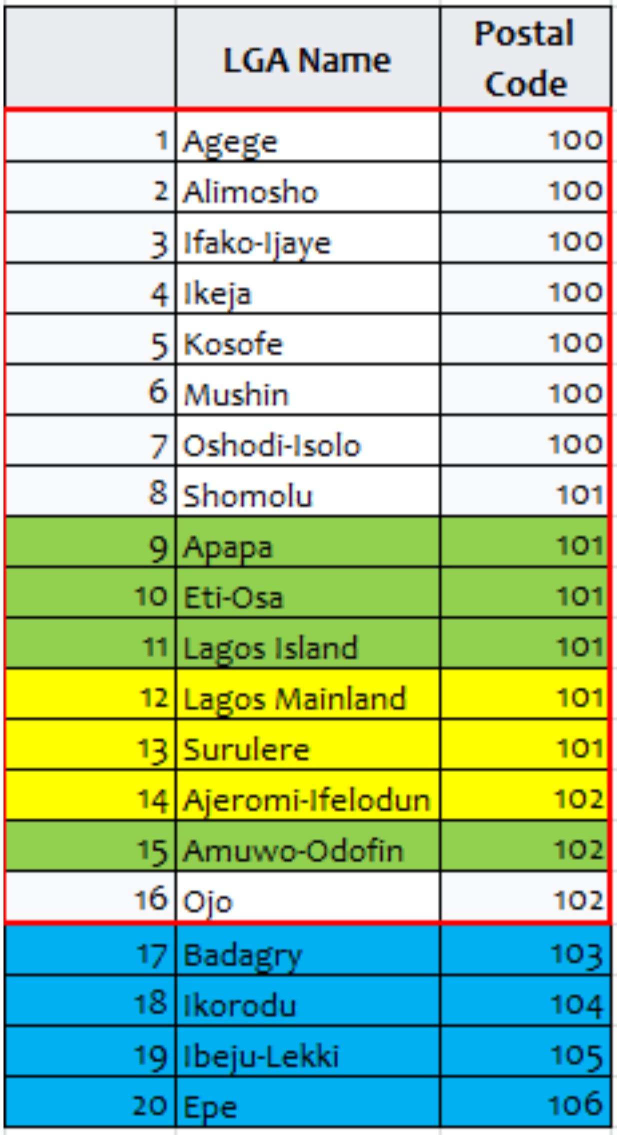
mapping = { "St": "Street", "St.": "Street", "Rd": "Road", "Rd.": "Road", "road.": "Road", "Ave.": "Avenue", "Ave" : "Avenue", "close": "Close", "Cres": "Crescent", "kirikiri": "Kirikiri Road", "Thomas": "Thomas Street"}

City Names

The city names are a strange one; using SQL to look at the city values I found that there were three different designations for value Lagos and some other different values:

Lagos,41 lagos,4 Lagaos,3 50,1 "Ilupeju, Lagos",1 "Lekki lagos",1 "Lekki, Lagos",1 Surulere,1 "Surulere, Lagos",1

The city issue is a tough one due to the many adminstrative changes that Lagos state has gone through over the years. Lagos state consists of 20 Local Goverment Areas (LGA) - these are administrative areas - a bit like counties in the US.



	LGA Name	Postal Code
1	Agege	100
2	Alimosho	100
3	Ifako-Ijaye	100
4	Ikeja	100
5	Kosofe	100
6	Mushin	100
7	Oshodi-Isolo	100
8	Shomolu	101
9	Apapa	101
10	Eti-Osa	101
11	Lagos Island	101
12	Lagos Mainland	101
13	Surulere	101
14	Ajeromi-Ifelodun	102
15	Amuwo-Odofin	102
16	Ojo	102
17	Badagry	103
18	Ikorodu	104
19	Ibeju-Lekki	105
20	Epe	106

- Lagos city initially consisted of 4 LGA's highlighted in Green. The indigenious Lagosians Awori's lived in these areas. Lagos was known as Eko.
- Lagos city quickly expanded to the Yellow areas and this areas would be known as Lagos city.(7 LGA's)
- The first 16 of the above LGAs (boredered by red) now comprise the statistical area of Metropolitan Lagos.
- The Last four LGAs (Badagry, Ikorodu, Ibeju-Lekki and Epe) are within Lagos State but are not part of Metropolitan Lagos.

It's no surprise that the city values are all over the place!

Becuase of the complexity of these values, I will only correct the mispellings and update the lower case values to all become 'Lagos'. I have a suggestion for further fixing these values in the additional ideas section.

Post Codes

The problem Postcodes were grouped quite nicely into the set of issues they represent:

- +234' - This is the Nigerian international dialing code. If I can find any other key describing the location of these entries, I can find the appropriate post codes and update them individually.
- 23401 - This is the Nigerian international dialing code plus the Lagos state regional dialing code; whic is 01
- Yaba is town in the in Ebutte Metta L.G.A
- 71510 - Actually the Post office box number of the Address

I used the element id attribute to match the postcode values to street or city keys to check if this could be used to create a mapping and insert the correct post codes. Some of the elements with Postcode had no street or city values and trying to find the post codes for the street addresses was very tedious so I decided to use an update_postcode function replace these incorrect values with None.

Fixme Tags

I noticed quite a number of tags that had the key "Fixme" . taking a deeper look at these tags, they seemed to be mostly due to lack of popoluation estimate and defaulted to village.

sqlite> select count()*, value from (select key,value from nodes_tags union all select key,value from ways_tags) as tags ...> where tags.key like '%fixme%' ...> group by value ...> order by count()* desc ...> limit 5;

874|no population estimate available, defaulted to village 108|to be surveyed 40|cloud cover present 28|yes 10|unfinished

I joined the nodes tags to the nodes table to pick all the keys and values for nodes with a fixme key to check some of these values. I found a lot of places that certainly can't be classed as villages: For example:

fixme|no population estimate available, defaulted to village country|Nigeria name|Mushin place|village source|GNS GNS:dsg_code|PPL GNS:dsg_string|populated place GNS:id|-2010492

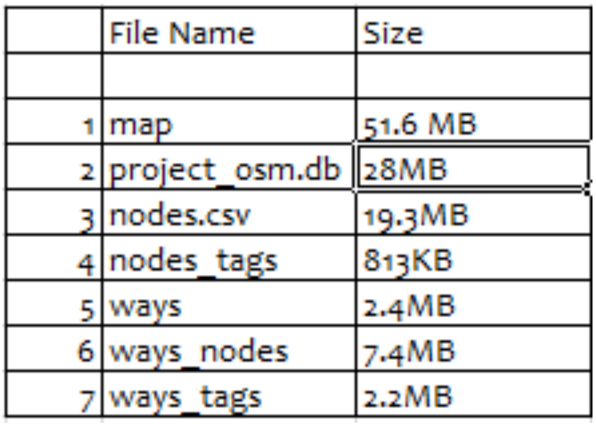
Mushin is an LGA in Lagos state so certianly not a village! However, it appears that most these entries were made by GNS(GEOnet Names Server).The population fields for the data are no longer maintained and has just defaulted the place value to Villages (incorrectly).

Data Overview and Additional Ideas

Data Overview

This section contains statistics about the data and the Database queries are used.

File Sizes



	File Name	Size
1	map	51.6 MB
2	project_osm.db	28MB
3	nodes.csv	19.3MB
4	nodes_tags	813KB
5	ways	2.4MB
6	ways_nodes	7.4MB
7	ways_tags	2.2MB

Number of Nodes|

```
sqlite> select count(*) from nodes;
242119
```

Number of Ways

```
sqlite> select count(*) from ways;
42663
```

Number of Unique users

```
sqlite> select count(distinct n.uid) from (select uid from nodes union all
select uid from ways) n ;
284
```

The Top 10 Contributing Users

```
sqlite> select user, count(*) from (select user, id from nodes union select
user,id from ways)
   ...> group by user
   ...> order by count(*) desc
   ...> limit 10;

BZVPNR|98386
crackers250|53812
JBenoit|17114
Oluwaseun Egbinola|15808
Rub21|12683
ediyes|5698
chdr|5503
aurel_joys|4286
Latze|4233
Coastlines-RJB|4101
```

Addtional Data Exploration

Number of Waterways

```
sqlite> select tags.value, count(*) as count
   ...> from (select * from nodes_tags UNION ALL select * from ways_tags) tags
   ...> where tags.key like '%waterway%'
   ...> group by tags.value
   ...> order by count desc;
stream|67
river|41
drain|36
canal|9
riverbank|6
ditch|5
dock|2
```

Number of Bridges

```
select tags.value, count(*) as count
from (select * from nodes_tags UNION ALL select * from ways_tags) tags
where tags.key like '%bridge%'
group by tags.value
order by count desc;

yes|455
```

```
Lagos state is a consists of many islands joined togther by a sphagetti of
bridges!! For an area the size of lagos, that's a large number of bridges.
```

Top 10 amenities

```
sqlite> select tags.value, count(*) as total
   ...> from (select * from nodes_tags UNION ALL select * from ways_tags) tags
   ...> where tags.key like '%amenity%'
   ...> group by tags.value
   ...> order by total desc
   ...> limit 10;

fuel|39
place_of_worship|30
restaurant|27
bank|26
school|26
parking|25
embassy|15
hospital|12
fast_food|10
atm|8
```

Place of worship and Resturants

Although the Top 10 amenities showed Place of worship and resturant, I couldn't go as deep as I would have liked as there are a lot of the restuarants that didn't have cuisine attributes and the places of worship which didn't have religion attributes. For resturants we had 27 values but only two with cuisines and for religion, we had 30 values, but only 19 with religion as an attribute. There's a church on almost every street corner in Lagos so I would expect way more values.

Additional Ideas

City Validation

As seen earlier, the evolution of Lagos city and Lagos state has led to different designations to different areas over the years. The current administraive divisions of Lagos state are LGA's.

I would suggest updating any Tag with a key type of addr to include "LGA". There is a site of Lagos zipcodes(http://nigeriazipcodes.com/lagos-state-zip-code-complete-towns-villages-list/ (http://nigeriazipcodes.com/lagos-state-zip-code-complete-towns-villages-list/)), which matches every street name and zipcode to an LGA. We could iterate through every "addr:street" value; validate the street name through our Lagos zipcodes file and return the LGA. Clearly for this to be effective, the street names have to have been entered correctly in the first place.

One of the issues I anticipate with this is that it would need to be implemented by someone very familiar with Lagos state and knew about the history. of the State. I knew a little about it because I had heard my parents and Uncles/Aunts talk about it but it would difficult for the average contributor to notice or appreciate the difference(s) and to do this even if they were Lagosians.

GNS Data Source

As mentioned earlier, the GNS(GEONet Names Server) data source is defaulting places to village if it doesn't have population estimates for them. A solution might be to use LGA/Population data on this page(https://en.wikipedia.org/wiki/List_of_Lagos_State_local_government_areas_by_population (https://en.wikipedia.org/wiki/List_of_Lagos_State_local_government_areas_by_population)) to validate population data and change the values of Place keys as needed. Interestingly, looking at the lagos state area using the GNS viewer, I can see the Local Goverment areas designations. However, they haven't been put in the downloaded data as key attributes.

If the LGA attributes were present, it would have been much easier to update the place 'key' with correct values. An issue with this though would be; How is a place designated a city,town, suburb or village (truly)? If a lack of population estimate designated a place as a village, is there some population to area ratio or calculation used to determine the place designation? If so, it would need to be published so anyone doing the updates would be able to make the same calcutions.

Conclusion

In summary, after the audit and review of the Lagos area data I find that the number of values in the data are too small for keys like street, for amenities like places of worship and the size of the file is way too small for an area the size of Lagos. This shows that the data for this area is very much incomplete.

While the data is incomplete, I think updating the LGA data for cities and place data for the fixme tags could go a long way into correcting the data we currently have.

One of the top 10 contributing users was Oluwaseun Egbinola|15808 - He worked on a health project tracking Polio vaccines and that contributed heavily to this area; Data from other projects like this could greatly improve accuracy and data for this area.