OpenStreetMap Project

Map Area

Donostia San Sebastian, Spain. Downloaded from https://mapzen.com/data/metro-extracts/

It is one of the most important places for tourism in the north of Spain. The city gets lot of tourists mainly because of the good food and beaches. The place is pretty near from the small town where I live, so I am going to check its data.

Problems encountered in my map

- **Street names don't follow any rules**.

Capital letters. All first letters of every word should be capital**:**

Ex.
Zapatari Kalea
Zapategi kalea
AVENIDA SAN PEDRO

Two languages. The main language in the data is in Basque. Spanish names get their own key addr:street:es (more info about it later on). So, we need to store only Basque name of the Street.

Ex.
Zabaleta Anaien Kalea / Calle Hermanos Zabaleta

Abbreviations. We are going to store the longer version. We also store the Basque word.

Ex.
AVDA. -> Etorbidea (Basque version of Avenida)
CL -> Kalea (Basque version of Calle)
CR CRTA -> Karretera
pz -> Plaza

Wrong order. We need to switch the order. This should be the last step, so we reverse the final correct name.

Ex.
aso,pz De -> Plaza De Aso

- **There are many key and value pairs in nodes that seems to be auto-generated by some software and not by human.**

```
select key, count(*) from node_tags where key="created_by" group by key;
created_by,5729

select key, value, count(*) from node_tags where key="created_by" group
by value order by count(*) DESC;
created_by,JOSM,5668
created_by,almien_coastlines,61
```

To remove those useless tags, I added a conditional:

```
if tag.get("k") != "created_by":
[...]
```

So we can remove 5729 tags. That will make our .csv and sqlite database way smaller. It is not actually a problem by itself, but adds extra work and storage need.

- **There is one amenity with name as "unclassified" that need to be fixed.**

select * from node_tags where nodeid=(SELECT nodeid from node_tags where value="unclassified");
538358775,name,Agifes
538358775,fixme,specify amenity
538358775,amenity,unclassified

Agifes is an association of people and family members to help others with mental illness. Following this link http://wiki.openstreetmap.org/wiki/Key:amenity it should be "social_facility".

- **There are postal codes that are not part of the city but from towns near it.**

Donostia San Sebastian postal codes are from 20001 to 20018 and 20100 and 20160. Checking the data we have these extra postal codes:

```
SELECT value,count(*) from node_tags where key="addr:postcode" and value
not in(
"20001","20002","20003",
"20004","20005","20006",
"20007","20008","20009",
"20010","20011","20012",
"20013","20014","20015",
"20016","20017","20018",
"20100","20160")
group by value order by count(*) DESC;
```

20120,1666
20110,896
20130,813
20170,789
20115,585
20140,513
20180,30
20800,3
20159,2

Those are the postal codes from towns next to Donostia San Sebastian, so there is actually nothing to fix here. The data downloaded from mapzen includes a wider area than expected.

- **Places like restaurants usually have a website as a tag. One of those is not actually a website but an email address:**

SELECT * from node_tags where nodeid="3484549022";

3484549022,name,Morgan
3484549022,phone,+34 943 424 461
3484549022,amenity,restaurant
3484549022,website,info@morgandonostia.com
3484549022,delivery,no
3484549022,addr:city,Donostia - San Sebastián
3484549022,addr:street,Narrika kalea
3484549022,addr:postcode,20003
3484549022,addr:housenumber,7

The real website should be **morgandonostia.com**

- **Not all street names are in Spanish**

```
select key, count(*) from node_tags where key like "addr:street%" group
by key order by count(*) DESC;
```
addr:street|14735
addr:street:es|71

Only 71 include the Spanish translation of the street name.

Data Overview

- File sizes

6.1M node_tags.csv
9.9M nodes.csv
54M san-sebastian_spain.osm.xml
1.2M way_tags.csv
514K ways.csv

- **Number of Unique Users**

```
SELECT count(distinct(n.user))
from (SELECT user from nodes UNION ALL SELECT user from ways) n;
377
```

- **Number of nodes**

```
select count(distinct(nodeid)) from nodes;
242048
```

- **Number of ways**

```
select count(distinct(wayid)) from ways;
26556
```

- **In Donostia San Sebastian going to bars to drink and eat is the most important tourist activity. With a population of 186.000, that makes a ratio of a bar/pub/cafe per 293 inhabitants.**

```
select value, count(*) from node_tags where key="amenity" and
(value="pub" or value="bar" or value="cafe") group by value order by
count(*) DESC;

bar,433
pub,85
cafe,45
```

- **The user with more nodes and ways contributed is equiserre:**

```
sqlite> SELECT user, count(*) from ways group by user ORDER BY count(*)
DESC LIMIT 1;

equiserre,9454

sqlite> SELECT user, count(*) from nodes group by user ORDER BY COUNT(*)
DESC limit 1;

equiserre,81192
```

- **And the Top Ten contributing users:**

```
SELECT n.user, COUNT(*)
FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) n
```

```
GROUP BY n.user ORDER BY COUNT(*) DESC LIMIT 10;
```

```
equiserre,90646
Uranzu,41335
mikeltxo,23452
Mandragora121,14350
karrikas,12803
dr_grijando,8077
Perutzio,5395
Jon_e,4780
Oskarbi,4624
ketari,4382
```

Additional Ideas

In the map area checked there are two official languages, so some streets are in Spanish, some in Basque. To clean the street names someone with knowledge in both languages should be needed. There are already tags to specify the language and the name of the street for that particular language, so there is lot of work to do in that area.

Also, taking in account that restaurants and bars are the most important tourist attractions would be good to get some database that would allow us to update those entries with the correct and updated telephones and websites. Users need services like Google Maps or OpenStreetMaps to all the information of a particular place. For restaurants and bars is really important to have correct phone numbers so possible customers can call and ask for information of reserve a table. We could:

1- Find a database that we can use as a reference to fill/fix information that is not correct. In the case of Spain, http://www.paginasamarillas.es/ is one of the most important search engines for these kind of services. It is important to take in account that it is also a human-maintained database so there could be still some problems to fix, the same way we did in this project. There is a big problem with that service, it doesn't provide an API so we would need to use curl or other methods (like beautifulsoap) to download the results of a particular restaurant and extract the data from it. This could be a slow process and error prone.

We could also use Google Maps, that has an API that we can use to gather information about each restaurant in our database and then compare the most relevant data (street, phone number and web site). Why Google Maps has more up-to-date and reliable information? Because it provides an easy way to business owners to add their business and update the contact information. OpenStreetMap is not user friendly on that particular topic, so business information will be more accurate in Google Maps, so getting the info from it would be good.

2- Once we have found the differences, we need to take note of those places where the information was incorrect and investigate why. The problem could be in OpenStreetMaps and we fixed it, but it could happen that it is correct and we added wrong information from Paginas Amarillas or Google Maps. There is no easy way to correlate to values like a phone programatically and decide witch one is the correct one. So, a manual process would be needed to investigate each of those restaurants with inconsistencies between databases and find which is the correct one (usually going to the business database and calling the number there).

Another problem would be more philosophical, about freedom. Why we want a open licensed map created by volunteers, if we actually use closed databases to correlate and fix the information? Do we want a 100% free and volunteer based map with some mistakes here and there, or we want a not that free map with accurate info? Maybe the final project would be to provide an easy to use interface for business owners, so they do the job of having their data updated.

Additional Data Exploration

- **Top Ten sports with more facilities to practice.**

```
select value, count(*) from way_tags where key="sport" group by value
order by count(*) DESC limit 10;
```
soccer,70
pelota,48
multi,33
basketball,23
swimming,22
tennis,13
athletics,5
skateboard,5
equestrian,4
skating,4

- **Most important religion in the area.**

```
select value, count(*) from way_tags where key="religion" group by value
order by count(*) DESC limit 10;
```
christian,62

- Top 10 Most popular cuisines

```
SELECT node_tags.value, COUNT(*) as num
FROM node_tags
    JOIN (SELECT DISTINCT(nodeid) FROM node_tags WHERE
value='restaurant') i
    ON node_tags.nodeid=i.nodeid
WHERE node_tags.key='cuisine'
GROUP BY node_tags.value
ORDER BY num DESC;
```
basque,31
regional,14
chinese,7
seafood,7
italian,6
spanish,6
pizza,4
kebab,3
Sociedad_Gastronómica,1
burger,1
japanese,1
vegetarian,1