# Data Wrangle OpenStreetData Project

*Loïc Le Tessier*

***Map area: Sucy en Brie, France, where I live***
***(***http://www.openstreetmap.org/#map=13/48.7656/2.5124***)***

***Manual export***

## Problem encountered in the map

I thought that I might encounter some errors, but overall, the data quality was very good.

Running a python script to analyse data within the OSM, I encountered the following:

Tags analysis:

- 0 problematic characters
- 154 788 lower strings
- 12 402 tag with lower colon
- 610 others

Contrary to English, French does not possess that many abbreviations for streets and ways. French using a lot of accents, I had to raise the re.UNICODE flag and use Unicode strings in the code. A script going through all the way and node tags and comparing it to expected street types and close cities gave the following results:

```
Streets
{u'Cit\xe9': set([u'Cit\xe9 des Ch\xeanes']), u'boulevard': set([u'bouleva
rd du g\xe9n\xe9ral Giraud']), 'avenue': set(['avenue Georges Pompidou']),
u'all\xe9e': set([u'all\xe9e du Val de Marne', u'all\xe9e Blaise Cendrars'
])}
```

Everything looks fine and is omission from the expected street type (allée) or is a case problem (bouleva rd vs. Boulevard).

Cities:

```
set(['Bonneuil sur Marne',
      'Sucy en Brie',
      u'limeil-Br\xe9vannes',
      'sucy-en-Brie',
      'valenton'])
```

Those are problematic. Writing convention dictates that Sucy en Brie should be Sucy-en-Brie but most of the time both are accepted. I decided to clean those names during the JSON preparation nonetheless to have a more coherent dataset.

## Overview of the data:

OSM file size: 83.6 Mo

JSON file size: 93.3 Mo

Number of elements in the file:

```
{'bounds': 1,
 'member': 34272,
 'meta': 1,
 'nd': 453319,
 'node': 345356,
 'note': 1,
 'osm': 1,
 'relation': 682,
 'tag': 167800,
 'way': 59091}
```

After importing in MongoDB, I ran the following queries:

**Number of documents in the collection:**

*db.maps.find().count()*

*404447*

**Number of buildings:**

*db.maps.aggregate([{"$match":{"building":{"$exists":True}}},*
*{"$group":{"_id":"Building","count":{"$sum":1}}}])*

*53045*

**Rail stations:**

*db.maps.aggregate([{"$match":{"railway":{"$exists":True},"name":{"$exists":True}}},*
*{"$match":{"railway":"station"}},*
*{"$group":{"_id":"$name"}}])*

```
{u'_id': u'La Varenne - Chennevi\xe8res'},
{u'_id': u'Sucy - Bonneuil'},
{u'_id': u'Boissy-Saint-L\xe9ger'}
```

**Top 10 users:**

```
db.maps.aggregate([{"$match":{"created.user":{"$exists":True}}},
                   {"$project":{"created.user":1,"_id":-1}},
                {"$group":{"_id":"$created.user","count":{"$sum":1}}},
                {"$sort":{"count":-1}},
                {"$limit":10}])
```

```
{u'_id': u'cquest', u'count': 225487},
{u'_id': u'osmmaker', u'count': 65204},
{u'_id': u'Alexandre Pliarchopoulos', u'count': 32550},
{u'_id': u'\xcbdz\xebronK', u'count': 21685},
{u'_id': u'PierenBot', u'count': 17159},
{u'_id': u'Esperanza36', u'count': 10917},
{u'_id': u'didier2020', u'count': 8420},
{u'_id': u'Utilisateur anonyme', u'count': 5980},
{u'_id': u'sevenup', u'count': 3112},
{u'_id': u'Super-Map', u'count': 1831}
```

**Top 10 Amenities:**

```
db.maps.aggregate([{"$match":{"amenity":{"$exists":True}}},
                {"$project":{"amenity":1,"_id":-1}},
                {"$group":{"_id":"$amenity","count":{"$sum":1}}},
                {"$sort":{"count":-1}},
                {"$limit":10}])
```

```
{u'_id': u'parking', u'count': 131},
{u'_id': u'school', u'count': 82},
{u'_id': u'restaurant', u'count': 61},
{u'_id': u'parking_space', u'count': 46},
{u'_id': u'bench', u'count': 40},
{u'_id': u'post_box', u'count': 34},
{u'_id': u'waste_basket', u'count': 31},
{u'_id': u'kindergarten', u'count': 26},
{u'_id': u'bank', u'count': 25},
{u'_id': u'recycling', u'count': 24}
```

**Bottom 10 Amenities:**

```
db.maps.aggregate([{"$match":{"amenity":{"$exists":True}}},
                   {"$project":{"amenity":1,"_id":-1}},
                  {"$group":{"_id":"$amenity","count":{"$sum":1}}},
                   {"$sort":{"count":1}},
                   {"$limit":10}])
```

```
{u'_id': u'clinic', u'count': 1},
{u'_id': u'fire_hydrant', u'count': 1},
{u'_id': u'courthouse', u'count': 1},
{u'_id': u'nightclub', u'count': 1},
{u'_id': u'arts_centre', u'count': 1},
{u'_id': u'clock', u'count': 1},
{u'_id': u'parking_entrance', u'count': 1},
{u'_id': u'dentist', u'count': 1},
{u'_id': u'driving_school', u'count': 1},
{u'_id': u'university', u'count': 1}
```

## Other ideas about the dataset

I wanted to check the rarer cuisine types:

```
db.maps.aggregate([{"$match":{"amenity":"restaurant"}},
                   {"$project":{"cuisine":1,"_id":-1}},
                   {"$group":{"_id":"$cuisine","count":{"$sum":1}}},
                   {"$sort":{"count":1}},
                   {"$limit":10}])
```

```
[{u'_id': u'kebab', u'count': 1},
 {u'_id': u'indian', u'count': 1},
 {u'_id': u'italian;pizza', u'count': 1},
 {u'_id': u'fish', u'count': 1},
 {u'_id': u'regional', u'count': 1},
 {u'_id': u'burger', u'count': 1},
 {u'_id': u'steak_house', u'count': 1},
 {u'_id': u'couscous', u'count': 1},
 {u'_id': u'seafood', u'count': 2},
 {u'_id': u'chinese', u'count': 3}]
```

I know more than 3 Chinese restaurants in the area. Businesses might not be well inputted or be lacking descriptions. Maybe proposing in OpenStreetData to check Yelp or other French restaurant review sites when inputting restaurant can help to gather more information.

Going through the data, I saw a “source” tag and decided to explore it:

**Top 10 Sources:**

```
db.maps.aggregate([{"$match":{"source":{"$exists":True}}},
                   {"$project":{"source":1,"_id":-1}},
                   {"$group":{"_id":"$source","count":{"$sum":1}}},
                   {"$sort":{"count":-1}},
                   {"$limit":10}])
[{u'_id': u'cadastre-dgi-fr source : Direction G\xe9n\xe9rale des Imp\xf4t
s - Cadastre. Mise \xe0 jour : 2011',
  u'count': 22918},
 {u'_id': u'cadastre-dgi-fr source : Direction G\xe9n\xe9rale des Imp\xf4t
s - Cadastre. Mise \xe0 jour : 2010',
  u'count': 19227},
 {u'_id': u'cadastre-dgi-fr source : Direction G\xe9n\xe9rale des Imp\xf4t
s - Cadastre. Mise \xe0 jour : 2012',
  u'count': 13651},
 {u'_id': u'cadastre-dgi-fr source : Direction Generale des Impots - Cadas
tre. Mise a jour : 2014',
  u'count': 5923},
 {u'_id': u'extraction vectorielle v1 cadastre-dgi-fr source : Direction G
\xe9n\xe9rale des Imp\xf4ts - Cadas. Mise \xe0 jour : 2010',
  u'count': 1373},
 {u'_id': u'cadastre-dgi-fr source : Direction G\xe9n\xe9rale des Imp\xf4t
s - Cadastre. Mise \xe0 jour : 2013',
  u'count': 1011},
 {u'_id': u'Bing', u'count': 557},
 {u'_id': u'cadastre-dgi-fr source : Direction G\xe9n\xe9rale des Finances
Publiques - Cadastre. Mise \xe0 jour : 2014',
  u'count': 479},
 {u'_id': u'cadastre-dgi-fr source : Direction G\xe9n\xe9rale des Imp\xf4t
s - Cadastre ; mise \xe0 jour : 2008',
  u'count': 455},
 {u'_id': u'cadastre-dgi-fr source : Direction G\xe9n\xe9rale des Imp\xf4t
s - Cadastre ; mise \xe0 jour : 2009',
  u'count': 440}]
```

A lot of source refers to cadaster data. Someone might have build a bot to parse the French cadaster and get the data. Checking internet, it turns out that as the French cadaster is online, a community was created to import those data in openstreetmap after getting permission by the French government in 2009 (as long as the source is mentioned):

http://wiki.openstreetmap.org/wiki/WikiProject_France/Cadastre

This underlines the possibilities offered by public available data and open source movement. This same logic should be shared, if not already, with other contributors in other countries.

**Bottom 10 sources:**

```
db.maps.aggregate([{"$match":{"source":{"$exists":True}}},
                   {"$project":{"source":1,"_id":-1}},
                   {"$group":{"_id":"$source","count":{"$sum":1}}},
                   {"$sort":{"count":1}},
                   {"$limit":10}])
[{u'_id': u'data.gouv.fr', u'count': 1},
 {u'_id': u'BDCarthage 2012', u'count': 1},
 {u'_id': u'IGN, Service de G\xe9od\xe9sie et Nivellement: 2006', u'count'
: 1},
 {u'_id': u'Service-Public.fr - 06/2013', u'count': 1},
 {u'_id': u"Minist\xe8re de l'\xc9ducation nationale, de l'Enseignement su
p\xe9rieur et de la Recherche nov-2014",
  u'count': 1},
 {u'_id': u'knowledge', u'count': 1},
 {u'_id': u'cadastre-dgi-fr source : Direction G\xe9n\xe9rale des Imp\xf4t
s - Cadastre. Mise \xe0 jour : 2011;data.gouv.fr:LaPoste - 04/2012',
  u'count': 1},
 {u'_id': u'Celtipharm - 10/2014', u'count': 1},
 {u'_id': u"Minist\xe8re de l'Education nationale, de la Jeunesse et de la
Vie associative;cadastre-dgi-fr source : Direction G\xe9n\xe9rale des Imp\
xf4ts - Cadastre. Mise \xe0 jour : 2012",
  u'count': 1},
 {u'_id': u'Yahoo', u'count': 1}]
```

Here we can see that sources are more diverse but are still official data.

## Conclusion

First of all, I loved working with MongoDB. Coming from SQL, the flexibility and syntax are really uplifting.

Regarding the dataset, I am really impressed of the level of detail available for the area. Sourcing from the cadaster is a good idea that should be generalized to other countries whenever possible. But information on businesses and instutions are scarce. Asking user or contributors for information would be great. Looking at wikidata, asking to improve or validate a random element could bear fruit quickly.