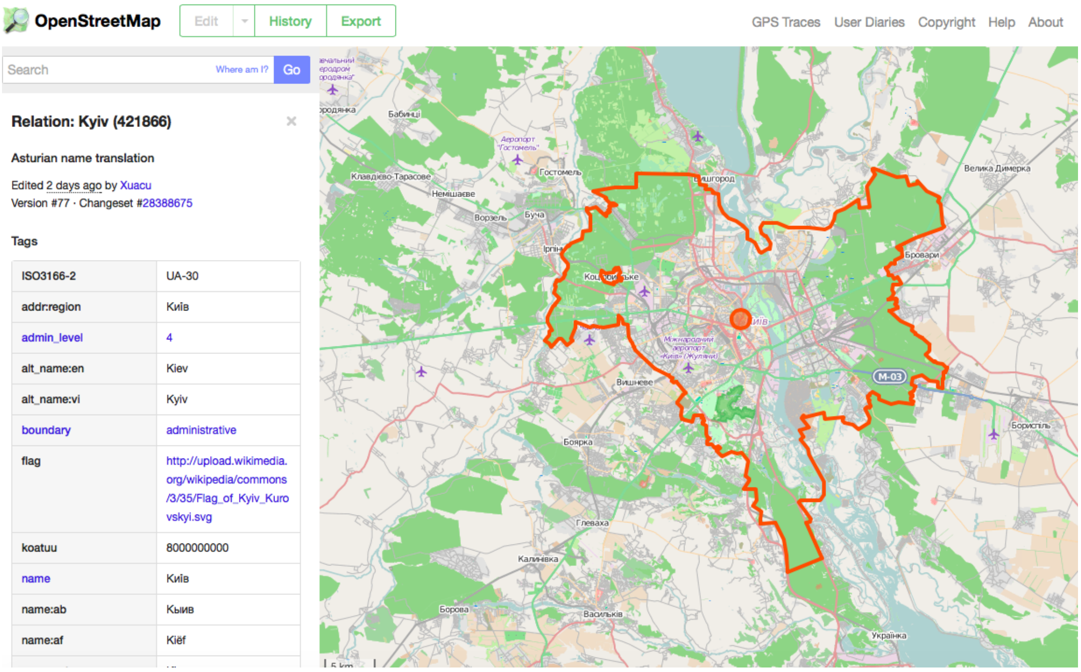
Data Wrangle OpenStreetMaps Data

Oleksii Renov

February 3, 2015

Choosing map area

Choosing map area is very simple for me. I live in Kiev, Ukraine. So i hadn't got another choice. As the very first step I need minimum and maximum longitude and latitude. For this task I'd gone to `http://openstreetmap.org` in the search field I'd typed Kiev and choose the second link. I've the following web page in my browser.



Relation: Kyiv (421866)

Asturian name translation

Edited 2 days ago by Xuacu
Version #77 · Changeset #28388675

Tags

ISO3166-2	UA-30
addr:region	Київ
admin_level	4
alt_name:en	Kiev
alt_name:vi	Kyiv
boundary	administrative
flag	http://upload.wikimedia.org/wikipedia/commons/3/35/Flag_of_Kyiv_Kurovskyi.svg
koatuu	8000000000
name	Київ
name:ab	Кыив
name:af	Kiëf

I've done some zomming and choose next latitude and longitude parameters in my case:

- minimum latitude = 29.9048;
- maximum latitude = 50.0986;
- minimum longitude = 31.1600;
- maximum longitude = 50.7052.

Next step is to download my dataset. I've gone in `http://overpass-api.de/query_form.html`, where i can past some queries to download data what i need.

My query is next:

```
(node(29.9048, 50.0986, 31.1600, 50.7052);<;);out;
```

Anyway there is exists numbers of variants loading data. One of them is simply go to this url `http://overpass-api.de/api/map?bbox=29.9048,50.0986,31.1600,50.7052`. The xml file which I've already downloaded is very big, near 250mb. In my usual work i've chance working with xml, and files near 1mb is very hard to analyze. Let's move to processing this dataset.

Process Dataset

As very first step, let's check what tags are in our dataset. Also would be helpful to check for attributes. Let's start from tags.

```
{'bounds': 1,
 'member': 75124,
 'meta': 1,
 'nd': 1316743,
 'node': 1079619,
 'note': 1,
 'osm': 1,
 'relation': 5016,
 'tag': 417122,
 'way': 153102}
```

Next json is corresponded to tag attributes.

```
{'changeset': 1237737,
 'generator': 1,
 'id': 1237737,
 'k': 417122,
 'lat': 1079619,
 'lon': 1079619,
 'maxlat': 1,
 'maxlon': 1,
 'minlat': 1,
 'minlon': 1,
 'osm_base': 1,
 'ref': 1391867,
 'role': 75124,
 'timestamp': 1237737,
 'type': 75124,
 'uid': 1237737,
 'user': 1237737,
 'v': 417122,
 'version': 1237738}
```

First of all i would like to notice, that ’k’ is key and ’v’ is value. There are equal number of keys and values what is obvious. But it isn’t clear what this keys are holding. I would like to explore more details and understand what keys and values have the most frequency in my dataset. In my opinion top 10 would be perfect for start.

```
{('surface', 9780),
 ('amenity', 9919),
 ('name:en', 10396),
 ('name:uk', 10646),
 ('name:ru', 12823),
 ('addr:street', 14084),
 ('name', 23862),
 ('addr:housenumber', 30646),
 ('highway', 67770),
 ('building', 73142)}
```