OpenStreetMap Data Case

Completed By: Trenton J. McKinney

Date: 2017/08/10

Table of Contents

Back to Top

OSM Map Area

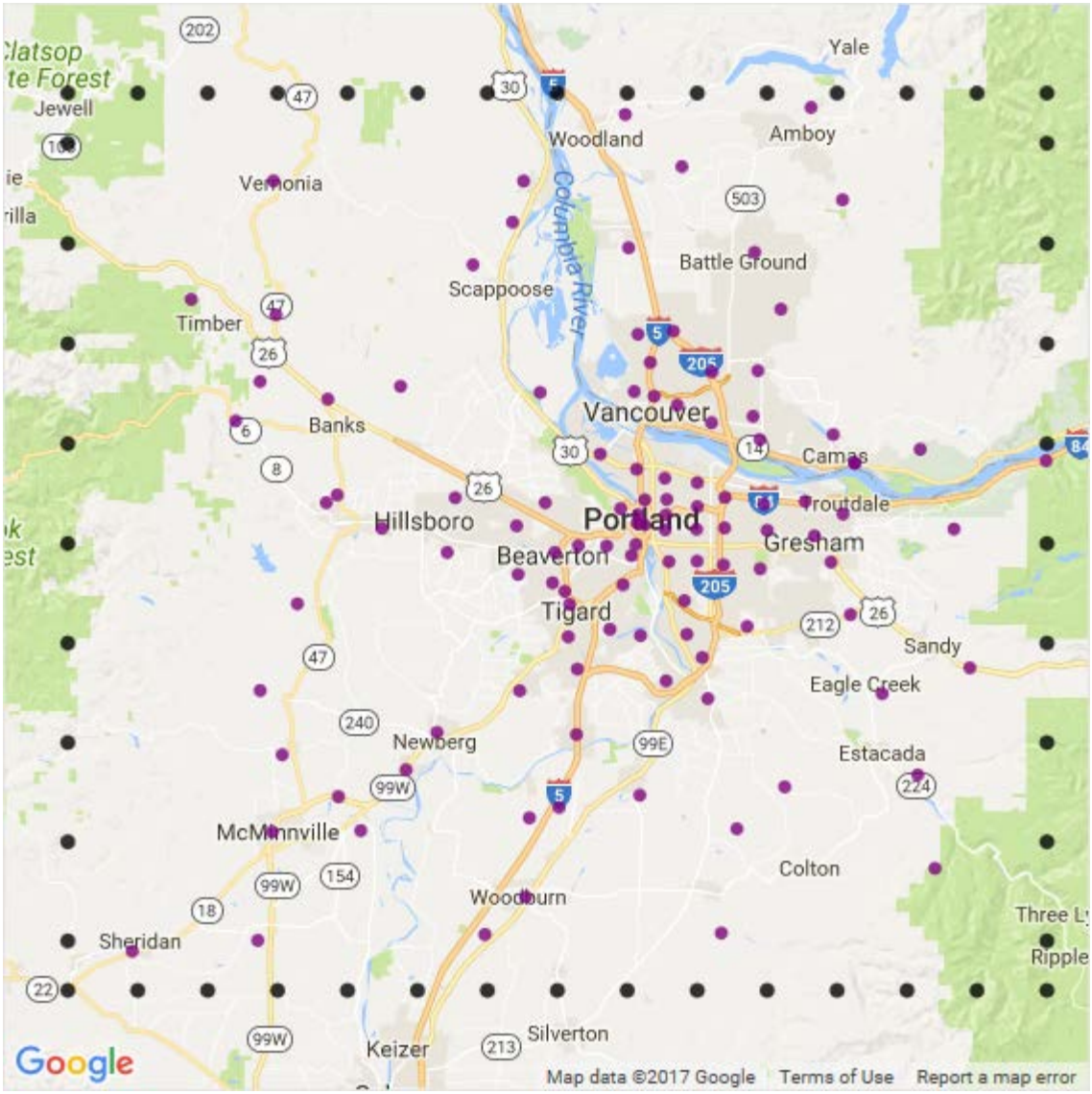
Portland OR, United States (Portland Metro Area)

- Portland at Mapzen

I live within and am interested in determining what type(s) of interesting information can be gleaned from the Portland Metropolitan OSM file. The map below depicts the area encompassed by the OSM file (black dots) and each purple dot represents the unique zip codes discovered within the ways_tags and nodes_tags.

Black dots outline the area of the OSM data & Purple dots are postcodes from ways_tags and nodes_tags

Notebook to generate map



Back To Top

Corrected OSM File Issues

- Before / After Comparison of Corrected City Names
- Before / After Comparison of Corrected Zip Codes
- Before / After Comparison of Street Names
- Additional Cleaning

Before / After Comparison of Corrected City Names

- project_fix_city_name.py
- The table shows the types of errors associated with the city names and the result of correction.

```
def fix_city_name(name, mapping=MAPPING):
    """Splits tag.attrib['v'] and checks each string against MAPPING
.
    If there's a value match, the string is changed to the new value
."""

    if name in mapping:
        name = name.replace(name, mapping[name])
    return name
```

Not Fixed		Fixed		Difference
Portland	223151	Portland	223169	18
Portland, Oregon	10			
Portland, OR	7			
portland	1			
Beaverton	29145	Beaverton	29158	13
Beaverton, OR	13			
Happy Valley	8419	Happy Valley	8420	1
97086	1			
Vancouver	603	Vancouver	604	1
vancouver	1			
Molalla	30	Molalla	31	1
molalla	1			
Vernonia	5	Vernonia	11	6
vernonia	6			

Back to Corrected OSM File Issues

Before / After Comparison of Corrected Zip Codes

- project_fix_zip_code.py
- The table shows the types of errors associated with the zip codes and the result of correction.

```
def fix_zip_codes(zip_codes):
    """Expects a string.  Will search the string for a consecutive 5
 digits and
    return the string as a zip code or leave blank if there's no mat
ch."""

    zip_code = re.compile('\d{5}')
    zip_code = zip_code.findall(zip_codes)

    if zip_code:
        return zip_code[0]
    else:
        return ''
```

Not Fixed		Fixed		Difference
97206	17875	97206	17878	3
97206-2633	1			
97206-2635	2			
97229	17242	97229	17243	1
97229-5688	1			
97123	12080	97123	12081	1
97123-4201	1			
97124	11782	97124	11785	3
97124-5961	2			
97124-9433	1			
97211	11710	97211	11711	1
97211-1798	1			
97213	11013	97213	11014	1
97213-1422	1			
97236	8380	97236	8381	1
97236-4913	1			
97035	7234	97035	7235	1
97035-2557	1			
97225	7176	97225	7178	2
97225-3010	1			
97225-6345	1			
97140	7076	97140	7079	3
97140-9205	3			
98662	72	98662	73	1
98662-6413	1			
97005	5511	97005	5513	2
97005-2823	1			
97005-2992	1			

Not Fixed		Fixed		Difference
97116	6310	97116	6334	24
97116-1675	1			
97116-2426	1			
97116-2427	1			
97116-2428	4			
97116-2429	2			
97116-2430	2			
97116-2431	4			
97116-2446	1			
97116-2463	1			
97116-2464	1			
97116-2497	1			
97116-2800	1			
97116-2863	1			
97116-2872	1			
97116-2896	1			
97116-2897	1			
98683	16	98683	17	1
98683-5227	1			
97113	3402	97113	3403	1
97113-8912	1			
97209	1347	97209	1349	2
Portland, OR 97209	2			
97106	895	97106	896	1
97106-9019	1			
97119	637	97119	638	1
97119-8514	1			

Back to Corrected OSM File Issues

Before / After Comparison of Street Names

- audit_street_names.py
- project_fix_street_name.py
- The table shows a non-exhaustive sample of street name corrections and a link to the full list of corrections is included below.

```
def fix_street_name(name, mapping=MAPPING):
    """Splits tag.attrib['v'] and checks each string against MAPPING
.
    If there's a value match, the string is changed to the new value
."""
    name = name.strip()
    x = name.split()
    for y in x:
        if y in mapping:
            name = name.replace(y, mapping[y])
    return name
```

- List of Street Types - Excluding Expected
- Full list of corrected street names

Sample of Corrected Street Names

Not Fixed => Fixed
SW Tonquin Rd. => Southwest Tonquin Road
SE Enterprise Cir => Southeast Enterprise Circle
SW 125th Ave => Southwest 125th Avenue
NE Cumulus Ave => Northeast Cumulus Avenue
NW 12th Ave => Northwest 12th Avenue
Mollala Ave => Mollala Avenue
Southeast 172nd Ave => Southeast 172nd Avenue
SE 60th Ave => Southeast 60th Avenue
NE 94th Ave => Northeast 94th Avenue
SW 11th Ave => Southwest 11th Avenue
SW Martinazzi Ave => Southwest Martinazzi Avenue
SE 96th Ave => Southeast 96th Avenue
NE 10th Ave => Northeast 10th Avenue
Northwest 185th Ave => Northwest 185th Avenue
SW 78th Ave => Southwest 78th Avenue
NE 22nd Ave => Northeast 22nd Avenue
4th Ave => 4th Avenue
Pacific Ave => Pacific Avenue

Back to Corrected OSM File Issues

Additional Cleaning

```
SELECT value
FROM (SELECT * FROM nodes_tags UNION ALL
    SELECT * FROM ways_tags) tags
WHERE key='phone'
GROUP BY value
```

The table below shows the various formats phone numbers come in. They should be corrected to a standard format for consistency.

Phone Number Formats
+1 (503) 282-9603
+01-503-639-1712
+01 503 352 9306
+1-971-500-9181
(360) 253-5117
+1 360 696 5232
+1 503-864-4592
1+503-692-3773
+15038443400
360 834 6100
360 8342682
360-253-6019
503.236.2970
5032083083

Back to Top

File & Database Overview

- This section contains basic statistics about the Portland Metro OSM dataset and the SQLite queries used.
- Sample OSM
- Full Fixed DB - link will expire 2017/10/08

File Stats

Filename	Filesize
portland_oregon.osm	1,551,530kB
portland_full_fixed.sqlite	934,734kB
nodes.csv	607,859kB
nodes_tags.csv	10,414kB
ways.csv	58,545kB
ways_nodes.csv	177,929kB
ways_tags.csv	152,534kB

Number of Node

```
SELECT COUNT(*) FROM nodes;
```

6,627,751

Number of Ways

```
SELECT COUNT(*) FROM ways;
```

865,354

Number of Distinct Contributers

```
SELECT COUNT(DISTINCT(users.uid))
FROM (SELECT uid FROM nodes UNION ALL
    SELECT uid FROM ways) users;
```

1,392

Back To Top

Database Exploration

- This section highlights the basic topics of exploration from the dataset and the associated SQLite queries.

City Name Count

- The OSM encompasses 74 cities.

```
SELECT tags.value, COUNT(*) as count
FROM (SELECT * FROM nodes_tags UNION ALL
    SELECT * FROM ways_tags) tags
WHERE tags.key LIKE 'city'
GROUP BY tags.value
ORDER BY count DESC;
```

City	Count
Portland	223169
Beaverton	29158
Hillsboro	24511
Tigard	23220
Gresham	19365
Oregon City	16693
Aloha	15314
Lake Oswego	14159
Milwaukie	9978
West Linn	9569

Zip Code Count

- The OSM encompasses 116 zip codes.

```
SELECT tags.value, COUNT(*) as count
FROM (SELECT * FROM nodes_tags
    UNION ALL
        SELECT * FROM ways_tags) tags
WHERE tags.key='postcode'
GROUP BY tags.value
ORDER BY count DESC;
```

Zip Code	Count
97206	17878
97229	17243
97045	16694
97219	14675
97223	13776
97202	13528
97007	12121
97123	12081
97124	11785
97211	11711

Top 10 Contributers

- Total user contributions 7,493,105 by 1,392 users.
- The top 2 contributers constitute %51.5 of the entries and the top 11, %88.7.

```
SELECT contrib.user, COUNT(*) as count
FROM (SELECT user FROM nodes
    UNION ALL SELECT user FROM ways) contrib
GROUP BY contrib.user
ORDER BY count DESC
LIMIT 10;
```

USERS	COUNTS
Peter Dobratz_pdxbuildings	1955428
lyzidiamond_imports	1900707
Mele Sax-Barnett	569515
baradam	545764
Darrell_pdxbuildings	431676
cowdog	334597
Peter Dobratz	304914
Grant Humphries	298520
justin_pdxbuildings	116568
amillar-osm-import	107968

Back To Top

Interesting Explorations

- Delving into the data shows how much Portland appreciates parking, biking and coffee. Apparently we like swimming too, eventhough it's only sunny for 3 months of the year.

Top Amenities

```
SELECT tags.value, COUNT(*) as count
FROM (SELECT * FROM nodes_tags UNION ALL
    SELECT * FROM ways_tags) tags
WHERE tags.key='amenity'
GROUP BY tags.value
ORDER BY count DESC;
```

Amenity	Count
parking	5749
bicycle_parking	3071
restaurant	1696
fast_food	1209
place_of_worship	1199
school	932
bench	903
waste_basket	827
cafe	732
bank	507

Top Cuisine

```
SELECT value, COUNT(*) as count
FROM (SELECT * FROM nodes_tags UNION ALL
    SELECT * FROM ways_tags) tags
```

```
    WHERE key='cuisine'
    GROUP BY value
    ORDER BY count DESC;
```

Cuisine	Count
coffee_shop	470
mexican	305
pizza	296
burger	287
sandwich	197
american	166
chinese	120
thai	120
japanese	72
sushi	65

Sports Facilities

```
    SELECT value, COUNT(*) as count
    FROM (SELECT * FROM nodes_tags UNION ALL
        SELECT * FROM ways_tags) tags
    WHERE key='sport'
    GROUP BY value
    ORDER BY count DESC;
```

Sport	Count
swimming	1424
baseball	876
tennis	675
basketball	511
soccer	426
golf	115
american_football	101
athletics	85
multi	52
yoga	44
skateboard	36
martial_arts	29

Back To Top

Other Ideas About the Dataset

Improving the Dataset

- Increase the number of contributors, partiularly in rural or less frequented locations. We can see, based upon Top 10 Contributers, most of the data comes from the top 11 users and from City Name Count we can see that of the 74 citys in the dataset, the vast majority of the data is for Portland and that some of the smaller cities only have 1 count. The primary idea behind OSM "... is a map of the world, created by people like you and free to use under an open license." I had never heard of OSM prior to this project requirement, so some type of local outreach like Meetup: OpenStreetMap Portland, but in other communities might increase the user base.
- Another idea for improving OSM is to import large datasets from other applications with a large number of users and geospatial data such as Google or Apple Maps or Pokemon Go to name a few.

Benefits:

- The single most obvious benefit is more users equates to more data.

Potential Issues

- The main issue with attracting more users is probably the process of reaching people that may be interested.
  - Meetups are mostly free, but the volume is low.
  - People have a tendency to ignore website ads
  - Commercials cost money
- Once a potential new user is found, there are addition roadblocks
  - Monetary constraints with GPS equipment acquisition
  - Personal time contraints
  - Technical hurdles:
    - How to Contribute
    - Contribute Map Data
- Large data imports from outside sources:
  - Goes against the idea of a community based map
  - "We are only interested in 'free' data. We must be able to release the data with our OpenStreetMap License"
  - There are additional technical hurdles related to importing data
    - OSM Import Guidelines
    - The Tiger Import had to be spread over several months to prevent overloading the OSM servers

If You're Interested in Contributing to OpenStreetMaps

- Beginner's Guide

Back To Top

Conclusion

Based upon the collected data, as shown in Corrected OSM File Issues, there are a relatively small number of issues. Specifically, only 40 city names and 50 zip codes required standardization. Additionally, fewer that 240 street names were transformed from short form to long form.

As mentioned in Other Ideas About the Dataset, the Portland data is very thorough, but the more rural communities surrounding Portland would benefit from more users and data. Bringing awareness of the OSM project and its benefits in terms of data availability to potential new users seems to be an intergral component to the continued success of OSM.