Mapping Source

I have been living in Sydney since 2008, hence I would like to investigate the OpenStreetMap dataset of Sydney:

- https://www.openstreetmap.org/relation/1251066
- http://metro.teczno.com/#sydney

Encountered Problems:

Unknown Tags Investigation

With 'print_all_tags' function, it loops through the entire file and counts all the possible tags for Sydney's osm. As a result, there are 8 types of tag shown in below table:

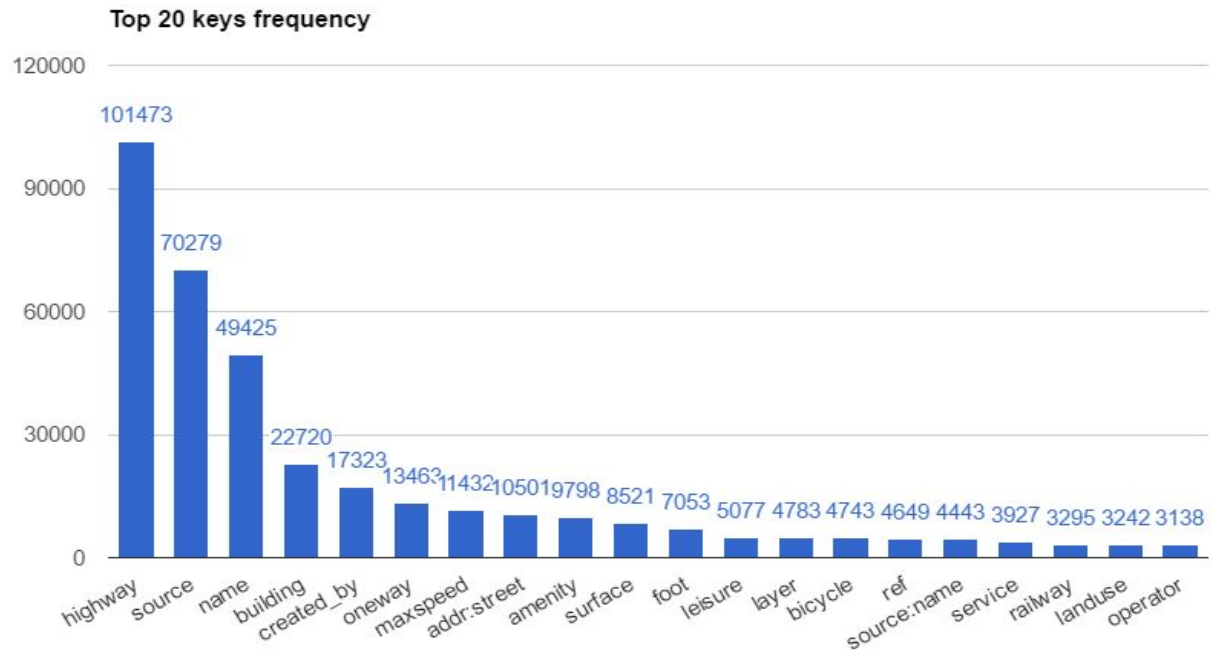
Tag type	'bounds'	'member'	'nd'	'node'	'osm'	'relation'	'tag'	'way'
counts	1	26489	1152606	949441	1	1675	451420	128755

According to OSM Wiki [1], **Elements** are the basic components of OpenStreetMap's conceptual data model of the physical world. They are

- Node
- Way
- Relation

With the definition of ‘Relation’, it is used to define the relationships for different logical or geographic elements. For the purpose of this project, only ‘node’ and ‘way’ elements will be mainly investigated.

In order to investigate the key and values under each type of tags, the ‘print_top_keys’ function was called to list out the top 20 keys with most appearing frequency in the database. As shown in the following chart, ‘highway’ key has dominated the majority of the keys, follow by ‘source’ and ‘name’ keys.



With the top 20 keys shown above, the values of each key were then examined in the dataset. The additional function ‘key_values’ was used to collect the first 20 unique values of each key. Furthermore, all result was saved as an csv file (i.e. ‘key_values.csv’).
The csv file was then import into a SQL database as a table. By fetching and reviewing each key. There are some unexpected values appeared on the list. For example,

Key	**Expected Values**	**Unexpected Values**
‘highway’	‘cycleway’	'crossing; traffic_signals'
‘layer’	‘1’	‘+1’
‘maxspeed’	‘80’ (e.g. integer)	‘8 knots’, ‘40 mph’ , ‘signals’
‘Oneway’	‘No’ (e.g. boolean)	‘-1’
‘addr:street’	‘Road’	‘Rd’

Max Speed

In Sydney, the speed limits range from 10 kilometers per hour to 110 kilometers per hour [2]. Any values out of this range should be reviewed. In the first 20 values of the 'maxspeed' key, there are few unexpected values such as '8 knots', '40 mph. These values need to be converted to standard unit which is kilometers per hour. Hence 8 knots will be 15 km/hr and 40 mph will be 60 km/hr. In addition, few values is out of the speed limits range, such as '8' and '5', it could be again caused by the unit conversion issue. Moreover, there are lots of 'signals' values that was used as speed derestriction signs for prima facie allowances. By default, these values should be equivalent to be 80 km/h in Sydney. Finally, any undefined value or values with special character (e.g. ";") will be removed. The function 'improve_maxspeed' is carried out to update the speed values.

Update Street Name

The street data has inconsistent format for its values. For instance, both "Av." and "Ave." are used for "Avenue". In order to standardise all street names, the Australian street types and abbreviations database[3] is used as the reference for updating all street names. Such that "st." is amended to be "Street" and "Row" is added to be a street type. All the updating process is done by 'improve_street_names' function.

Data Overview:

File sizes:

- sydney.osm : 208 MB
- sydney.db: 142 MB
- nodes.csv: 77.6 MB
- nodes_tags.csv: 3.2MB
- ways.csv: 7.5MB
- ways_nodes.csv: 27.7MB
- ways_tags.csv: 12.6MB

Number of nodes:

```
SELECT COUNT(*)
FROM nodes;
> 949441
```

Number of ways:

```
SELECT COUNT(*)
FROM ways;
> 128755
```

Number of unique users:

```
SELECT COUNT(DISTINCT(e.uid))
FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) e;
> 1093
```

Number of users appearing only once:

```
SELECT COUNT(*)
FROM
(SELECT e.user, COUNT(*) as num
FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e
GROUP BY e.user
HAVING num=1)  u;
> 202
```

Top 10 contributing users:

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

user	**num**
behemoth14	120015
inas	105189
OSMF Redaction Account	59278
Rhubarb	40648
MCH	39526
aharvey	33622
Rub21	33210
Ebenezer	32194
GrantDelandro	24611
bentrails	23661

Note: by statistics, the top one user (behemoth14) has contributed 11.13% of the entire database and the sum of the top ten users contributed nearly 50 % of the database (i.e. 47.48%)

Top 5 Postal Codes:

The data quality of the postal codes are almost perfect as the format and the value is quite consistent. Only 2 of the postal codes contain the state abbreviation in the value (e.g. NSW 2000) .

```
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
LIMIT 5;
>
```

value	count
2213	252
2062	220
2166	215
2150	181
2065	126

Top 5 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
LIMIT 5;
>
```

value	count
Sydney	504
Panania	240
Cabramatta	215
Coogee	79
Padstow	41

Most Popular cuisines:

Australia is a multicultural country[4] and 2011, migrant tended to be most concentrated around a number of key urban centres in Sydney. A large majority of residents in Sydney CBD (78%) were born overseas.
By running the following SQL code, we can find out what is the most popular cuisines in Sydney area and hence reflecting the major groups of oversea migration.

```
SELECT nodes_tags.value, COUNT(*) as num
FROM nodes_tags
    JOIN (SELECT DISTINCT(id) FROM nodes_tags WHERE value='restaurant') i
    ON nodes_tags.id=i.id
WHERE nodes_tags.key='cuisine'
GROUP BY nodes_tags.value
ORDER BY num DESC;
```

>

value	**num**
chinese	33
thai	28
italian	24
pizza	18
indian	14
japanese	14
vietnamese	6
asian	5
greek	5
korean	5

As shown on the table above, Chinese and Thai are the most popular cuisines in Sydney and this table could imply that the major migration groups are from Asia, However, the size of ‘restaurant’ keys is not sufficient to reflect the real population.

Conclusion

The data has been well-cleaned for the purposes of this exercise and there should be more features of the this dataset we could look for, such as reviewing the ‘highway’ components and exploring the lifestyles in Sydney with its ‘leisure’ keys.

Reference

1. OSM Wiki, http://wiki.openstreetmap.org/wiki/Elements

2. Speed limits in Australia, https://en.wikipedia.org/wiki/Speed_limits_in_Australia#Signage

3. Australian Street Types and Abbreviations Database,

http://jaspreetchahal.org/australian-street-types-and-abbreviations-database/

4. Australian Social Trends, 2014,

http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/4102.0main+features102014#SYDNEY