Area : New Delhi,India
Downloaded from : http://osm-extracted-metros.s3.amazonaws.com/new-delhi.osm.bz2

Familiarizing with the data set:

The New Delhi data set has about .5 million nodes and tags and has main focus on position, the city, the amenities available in that area and routes (highways, streets).

Top level tags :

At the first look, here are the top level tags in the xml :

{"way" : 20767 ,"node" : 502659 , "broad_palmate_leaves" : 1, "boundary" : 1 , "Public" : 1 }

Problems encountered in the dataset:

1. **Auditing of Street names :**

After running the audit_street.py, I realised that a lot of places donot have the standard types such as "Street", "Avenue", "Boulevard", "Drive", "Court", "Place", "Square", "Lane", "Road", "Trail", "Parkway", "Commons" but are named with the local slang for the streets such as marg and chowk (in regional languages).
It was also very surprising that they did not need a lot of cleanup from my end since they are fairly well cleaned up already.

There were a few place where
Rd should be cleaned to Road,
city was mentioned along with the street names and they had to be separated (
eg.

```
'Delhi': set(['Block 35, Trilokpuri, New Delhi',
              'Ganesh Nagar I South, Pandav Nagar, New Delhi',
              'HPR School Main Road, Hira Colony, Siraspur, Delhi',
              'I-64, Laxmi Nagar, New Delhi',
              'Mayur Vihar Phase III New Delhi',
              'Opposite Plaza Market,Sector 9, Rohini, New Delhi, North West Delhi',
              'Phase 1 Ashok Vihar Rd, Delhi',
              'Razapur, Sector 9, New Delhi',
              'Thakkar Bapa Nagar, Karol Bagh, New Delhi',
              'Vinod Nagar West, New Delhi, East Delhi',
              'Vir Banda Bairagi Marg, New Delhi']),
```

Here is the full output from the audit:

```
{
 'Accher': set(['Accher']),
 'Area': set(['Sarita Vihar Institutional Area']),
 'Bagh': set(['Ahuja Sons Shalwale, Karol Bagh',
              'Ajmal Khan Road, Karol Bagh',
              'Karol Bagh',
              'Meera Bagh',
              'Mini Market Mini MarketNanakpura Road Moti GaonMoti Bagh',
              'Mint Market Nankpura, Nr Moti Bagh']),
 'Bazaar': set(['Main Bazaar']),
 'Bhangel': set(['Bhangel', 'Purani Tanki, Bhangel']),
 'Camp': set(['Hudson Lines, Kingsway Camp', 'Tibetan New Camp']),
 'Centre': set(['Plot 9, Jasola District Centre',
                'Sector 10, Rohini Twin District Centre']),
 'Chowk': set(['Chandni Chowk',
               'Jalebi Chowk',
               'Sector 16 Dynamic House, Rajnigandha Chowk']),
 'College': set(['DCE College', 'Sardhanand College']),
 'Colony': set(['Panchsheel Colony']),
 'Complex': set(['Amatash Shopping Complex',
                 'Ganga Shopping Complex',
                 'Shree Balaji Complex']),
 'Delhi': set(['Block 35, Trilokpuri, New Delhi',
               'Ganesh Nagar I South, Pandav Nagar, New Delhi',
               'HPR School Main Road, Hira Colony, Siraspur, Delhi',
               'I-64, Laxmi Nagar, New Delhi',
               'Mayur Vihar Phase III New Delhi',
               'Opposite Plaza Market,Sector 9, Rohini, New Delhi, North West Delhi',
               'Phase 1 Ashok Vihar Rd, Delhi',
               'Razapur, Sector 9, New Delhi',
               'Thakkar Bapa Nagar, Karol Bagh, New Delhi',
               'Vinod Nagar West, New Delhi, East Delhi',
               'Vir Banda Bairagi Marg, New Delhi']),
 'Delhi.': set(['Anarwali Masjid, Block J, New Delhi.']),
 'Dwarka': set(['Dwarka', 'Sec - 19, Poket 3, Dwarka', 'Sector 11 Dwarka']),
 'Enclave': set(['AP Market, shop no.41, Maurya Enclave',
                 'Bhera Enclave',
                 'Safdarjung Enclave',
                 'Vasundhara Enclave']),
 'Estate': set(['Lodhi Estate']),
 'Extension': set(['Block B, Ashok Nagar Extension',
                   'I. P. Extension',
                   'South Extension']),
 'Faridabad': set(['Sector 21-D, Faridabad', 'sector 37, Faridabad']),
 'Firni': set(['Siraspur Firni']),
 'Flats': set(['Street C, Munirka DDA Flats', 'Street E, Munirka DDA Flats']),
```

```
'Gali': set(['Pratap Gali']),
'Gaziabad': set(['Hayat Nagar, Khoda Colony, Gaziabad']),
'Ghar': set(['G.B.Nagar, Bharat Ghar']),
'Govindpuram': set(['Govindpuram']),
'Harola': set(['Hanuman Maket, Harola',
               'Hanuman Market Harola',
               'Hanuman Markrt Harola',
               'Sector 5, Harola']),
'Haryana': set(['udhyog vihar, phase - 2, Gurgaon, Haryana']),
'I.P.Extension': set(['I.P.Extension']),
'II': set(['Delta II', 'South City II']),
'III': set(['Gharoli Dairy Farm, Mayur Vihar III',
            'Knowledge Park III',
            'Neeti Khand III']),
'Indirapuram': set(['Vaibhav Khand, Indirapuram']),
'Janakpuri': set(['A-3 Commercial Complex, Janakpuri', 'C Block, Janakpuri']),
'Janpath': set(['Janpath']),
'K': set(['Pocket K']),
'Kaushambu': set(['Kaushambu']),
'Kendra': set(['Sector-29, Near Vyapar Kendra']),
'Khanpur': set(['Khanpur']),
'Khas': set(['Hauz Khas']),
'Kunj': set(['Bhavani Kunj, Vasant Kunj']),
'Lines': set(['Under Hill Lane, Civil Lines']),
'Lok': set(['Sushant Lok']),
'Marg': set(['Abdul Gaffar Khan Marg',
             'Africa Avenue Marg',
             'Amrita Shergil Marg',
             'Aruna Asif Ali Marg',
             'Aurobindo Marg',
             'Bahadur Shah Zafar Marg',
             'Bahadur shah Zafar Marg',
             'Bakshi Marg',
             'Bhagwan Mahavir Marg',
             'Indraprastha Marg',
             'Kasturba Gandhi Marg',
             'Mahatma Gandhi Marg',
             'Nelson Mandela Marg',
             'Nyaya Marg',
             'Press Enclave Marg',
             'Rafi Marg',
             'Ramakrishna Ashram Marg',
             'Sadhbhawna Marg',
             'Sahakarita Marg',
             'Sansanwal Marg',
             'Sham Nath Marg',
```

```
          'Tees January Marg',
          'Vikas Marg',
          'Vinay Marg']),
'Market': set(['Defence Colony Market',
           'Khan Market',
           'Naoroji Nagar Market',
           'Sujan Singh Park, Subramania Bharti Marg,Behind Khan Market',
           'Wazirpur, Shiv Market']),
'Munirka': set(['DDA Flats, Munirka']),
'NH-IV': set(['Faridabad NH-IV']),
'NIT': set(['NIT']),
'NOIDA': set(['Sector 16 NOIDA']),
'Nagar': set(['Adhyatmik Nagar',
          'Ansari Nagar',
          'Block B, Ashok Nagar Extension, New Ashok Nagar',
          'Block JVishwakarma ParkLaxmi Nagar',
          'Chander Nagar',
          'F Block, Sanjay Nagar',
          'First Floor, C Block, Sector 10, Noida, Gautam Buddh Nagar',
          'GB Nagar',
          'H33, Bali Nagar',
          'Hoshiarpur Village,Gautam Buddh Nagar',
          'Kalesh Complex, Pandav Nagar',
          'Lane E, Sarojini Nagar',
          'Lane K, RBI Staff Quarters, Sarojini Nagar',
          'Lohia Nagar',
          'New Ashok Nagar Rd, Block B, New Ashok Nagar',
          'Rajinder Nagar',
          'Sanjay Nagar',
          'Sector 12, Noida, Gautam Buddh Nagar',
          'Shastri Nagar',
          'Shiv Nagar',
          'Tilak Nagar Round About, Ashok Nagar',
          'Vishwakarma Park Lakshmi Nagar']),
'Nasirpur': set(['Nasirpur Road, Pocket 10, Pink Apartments, Nasirpur']),
'Noida': set(['Block A, Sector 12, Noida',
          'Greater Noida',
          'Nawada Yadav Market,  Noida',
          'Noida',
          'Sector 10, Noida',
          'Sector 100, Noida']),
'Noida,': set(['Village Chhalera & Sadarpur, Sadarpur, Sector 45, Noida,']),
'Paharganj': set(['Chuna Mandi, Paharganj']),
'Park': set(['Subroto Park', 'Swaroop Park']),
'Path': set(['Shanti Path']),
'Pitampura': set(['Complex Pitampura', 'Pitampura']),
```

'Pradesh': set(['Opp. Barqueles Building Khora Colony Sector 62A Noida Uttar Pradesh']),
'Pritampura': set(['Pritampura']),
'Puri': set(['Pocket 10 Durga Park Colony Dashrath Puri']),
'Rajokri': set(['Air Force Station Rajokri']),
'Rani': set(['Hauz Rani']),
'Rd': set(['Arya Samaj Rd']),
'Road,': set(['MVL Coral, Alwar Bypass Road,']),
'Rohini': set(['Jain Mandir Marg, Pocket 6, Sector 9, Rohini',
'Pocket 6, Sector 9, Rohini',
'Sector-18, Rohini']),
'Rohini,Delhi': set(['G-7 Sector-16, Rohini,Delhi']),
'S1': set(['S1']),
'SECTOR-11': set(['SECTOR-11']),
'Sec-8': set(['Sec-8']),
'Sector-10': set(['Sector-18, D D Complex, Opposite Savitri Market, Sector-10']),
'Sector-58': set(['Sector-58']),

'Shahdara': set(['East Gorakh Park, Shahdara']),
'Shakarpur': set(['Main Rd, Block B, Nanakpura, Shakarpur']),
'Society': set(['Green Street Society', 'Vaishali Express Green Society']),
'Station': set(['Eicher service Station',
'Gulab Bhavan, Harola, Sector 5, Opposite Fire Station']),
'Station)': set(['Najafgarh Road (Next to the Tilak Nagar Police Station)']),
'Town': set(['Mayur Vihar, Ganesh Temple, 92, Pocket D, Phase 2, Model Town']),
'U.P)': set(['Kherli Hafizpur, Noida (U.P)']),
'UDSC': set(['UDSC']),
'UP)': set(['Kherli Hafizpur, Noida (UP)']),
'University': set(['Jawaharlal Nehru University']),
'Vasundhara': set(['Vasundhara']),
'Vihar': set(['170, Phase-1 Udyog Vihar',
'Building Materials Market, Ecotech-II, Udyog Vihar',
'Ecotech-II, Udyog Vihar',
'Gandhi Vihar',
'Kendriya Vihar',
'Kumar House, Central Market, Prashant Vihar',
'Mayur Vihar',
'Noida Rd Sector 25 JalVayu Vihar',
'Palam Vihar',
'Preet Vihar Road Block BPreet Vihar',
'Saraswati Vihar',
'Sarita Vihar',
'Sector 25, JalVayu Vihar',
'Sunder Vihar']),
'Vikaspuri': set(['Ansal Majestic Tower, Vikaspuri',
'Ansal majestic tower, Vikaspuri']),

'Village': set(['Carterpuri Village']),
'Wali': set(['Gali Chandi Wali']),
'West': set(['Dundas St. West']),
'colony': set(['Khora colony', 'new friends colony']),
'course': set(['Golf course']),
'daas': set(['sant kabir daas']),
'delhi': set(['delhi']),
'garden': set(['dilshad garden']),
'gurgaon': set(['gurgaon']),
'janakpuri': set(['janakpuri']),
'lane': set(['Basant lane']),
'marg': set(['Dr. Bishamber das marg', 'chaudhary fateh singh marg']),
'mohalla': set(['prakash mohalla']),
'moti': set(['moti']),
'nagar': set(['shastri nagar']),
'noida': set(['Sadarpur, Main Market, noida', 'noida']),
'p': set(['sector 11, block- p']),
'road': set(['Ring road', 'ansari road']),
'vihar': set(['rail vihar',
'shankar vihar',
'swami Narayan Marg, Ashok vihar']),
Arya Samaj Rd => Arya Samaj Road
}

2. Auditing the tag types:

After running audit_tags.py, here are the problems that were captured.

{'lower': 186001, 'lower_colon': 7987, 'other': 168, 'problemchars': 4}
[{'k': 'Gaurav General Store', 'v': 'convenience'},
{'k': 'Aggarwal Sweets', 'v': 'fast_food'},
{'k': 'Ladies Readymade Garments', 'v': ''},
{'k': 'M.S. FLAT', 'v': ''}]

This also shows that there were a lot of issues with problematic chars or missing values . This shows that the data file has been fairly revisited many times and corrected for issues.

3. Auditing the users who edited this file:

After running audit_users.py, the result shows that the file has been edited many times by a lot of users . From this set, it does look like a few people have major contributions .
Total number of unique users: 686
[('56597:Oberaffe', 269365),

("600918:n'garh", 83641),
('451671:Edolis', 52068),
('1306:PlaneMad', 16186),
('1292377:Nepolean', 14967),
('46622:roemcke', 14848),
('338611:marek kleciak', 14635),
('1292385:BalaK', 12166),
('1292408:Naveena', 12086),
('17429:thevikas', 9968),
]

This makes me think that it is either a company (working with maps) , that has taken a special interest in updating the dataset or automated processes that has entered the data.

Preparing the data for MongoDB

The data was cleaned up manually.Then p3_xmltojson.py was run on the data set and relevant json file was created.
Here is how the json object was created:

1. The top level objects were either node (city names, places,amenities) or ways (highways,crossroads).
2. The information on longitute and latitude for every node/way was populated into an array [lon,lat] and made a part of the node/way.
3. The information on who edited the node was populated into a created hash and made a part of the node/way.
4. Amenities such as schools/hospitals/markets/shops were made a part of amenities array and injected into ways.

Here is the basic example of how a node/way looked as a jsin object:

{"changeset": "6914006", "node_refs": ["3512134280", "3512134282", "3512134283", "3512134284", "3512134285"], "created": {"timestamp": "2011-01-09T11:46:40Z", "version": "8", "uid": "17429", "user": "thevikas"}, "address": {"street": "Block A1", "housenumber": "38"}, "position": ["28.4195123", "77.0441854"], "type": "node", "id": "266598403"}

{"changeset": "508147", "amenity": ["post_office"], "node_refs": ["3512134280", "3512134282", "3512134283", "3512134284", "3512134285"], "created": {"timestamp": "2008-05-24T09:47:26Z", "version": "1", "uid": "17429", "user": "thevikas"}, "created_by": "Potlatch 0.9a", "position": ["28.4620809", "77.0315003"], "type": "node", "id": "266599672"}

{"building": "yes", "changeset": "18193716", "amenity": ["school"], "node_refs": ["58043990", "58043991", "58043992", "58043993", "58043994", "58043995", "58043996", "58043997", "58043990"], "name": "School of Computional and Integrative Sciences", "created": {"timestamp": "2013-10-05T11:24:44Z", "version": "2", "uid": "81656", "user": "satyaakam"}, "type": "way", "id": "7891819"}

Once the data was converted to json, p3_write_todb.ipynb was run and data was imported into mongo db.

Data Overview:

Files in xml and json:

166M /Users/indur/Downloads/new-delhi_india.json
107M /Users/indur/Downloads/new-delhi_india.osm

Number of documents:
```
> use p3_osm;
switched to db p3_osm
> db.maps.count()
523429
```

Number of nodes:
```
> db.maps.find({type: "node"}).count()
502659
```

Number of ways:
```
> db.maps.find({type: "way"}).count()
20767
```

Number of unique users:
```
> db.maps.distinct("created.user").length
659
```

Top contributing users:
```
> db.maps.aggregate([{$group:{_id:"$created.user", count:{$sum:1}}},
{$sort:{count:-1}}, {$limit:3}])

{ "_id" : "Oberaffe", "count" : 269265 }
{ "_id" : "n'garh", "count" : 81574 }
{ "_id" : "Edolis", "count" : 52010 }
```

Least contributing users:

```
> db.maps.aggregate([{$group:{_id:"$created.user", count:{$sum:1}}}, {$group:
{_id:"$count", total_users:{$sum: 1}}}, {$sort: {_id: 1}}, {$limit: 3}])

{ "_id" : 1, "total_users" : 132 }
{ "_id" : 2, "total_users" : 62 }
{ "_id" : 3, "total_users" : 30 }

{ "_id" : "secondary", "count" : 1794 }
```

Addresses with postal codes:

```
> db.maps.find({"address.postcode": {$exists: true}}).count()
378
```

Additional ideas:

1. Top amenitites:

The data showed that people captured more schools,fuel,place_of_workship,parking and atms more than other amenities.
One interesting observation here is that there are way more schools/universities and colleges than public libraries , ratio being 1:0.04 .

```
> db.maps.aggregate( [{$group: {_id:"$amenity", count:{$sum: 1}}}, {$sort: {count: -
1}}])
{ "_id" : [ "school" ], "count" : 262 }
{ "_id" : [ "fuel" ], "count" : 177 }
{ "_id" : [ "place_of_worship" ], "count" : 156 }
{ "_id" : [ "parking" ], "count" : 131 }
{ "_id" : [ "atm" ], "count" : 128 }
{ "_id" : [ "restaurant" ], "count" : 120 }
{ "_id" : [ "bank" ], "count" : 99 }
{ "_id" : [ "hospital" ], "count" : 83 }
{ "_id" : [ "fast_food" ], "count" : 70 }
{ "_id" : [ "pharmacy" ], "count" : 61 }
{ "_id" : [ "college" ], "count" : 54 }
{ "_id" : [ "police" ], "count" : 48 }
{ "_id" : [ "cafe" ], "count" : 42 }
{ "_id" : [ "cinema" ], "count" : 39 }
{ "_id" : [ "embassy" ], "count" : 36 }
{ "_id" : [ "bus_station" ], "count" : 36 }
{ "_id" : [ "marketplace" ], "count" : 28 }
{ "_id" : [ "toilets" ], "count" : 26 }
{ "_id" : [ "bar" ], "count" : 21 }
```

{ "_id" : ["post_office"], "count" : 20 }
{ "_id" : ["fountain"], "count" : 17 }
{ "_id" : ["post_box"], "count" : 17 }
{ "_id" : ["public_building"], "count" : 17 }
{ "_id" : ["library"], "count" : 14 }
{ "_id" : ["taxi"], "count" : 12 }
{ "_id" : ["grave_yard"], "count" : 11 }
{ "_id" : ["university"], "count" : 11 }
{ "_id" : ["fire_station"], "count" : 10 }
{ "_id" : ["doctors"], "count" : 9 }
{ "_id" : ["kindergarten"], "count" : 9 }
{ "_id" : ["swimming_pool"], "count" : 8 }
{ "_id" : ["telephone"], "count" : 7 }
{ "_id" : ["parking_space"], "count" : 5 }
{ "_id" : ["waste_basket"], "count" : 5 }
{ "_id" : ["theatre"], "count" : 5 }
{ "_id" : ["bench"], "count" : 5 }
{ "_id" : ["car_rental"], "count" : 5 }
{ "_id" : ["drinking_water"], "count" : 4 }
{ "_id" : ["courthouse"], "count" : 4 }
{ "_id" : ["pub"], "count" : 4 }
{ "_id" : ["clinic"], "count" : 4 }
{ "_id" : ["community_centre"], "count" : 3 }
{ "_id" : ["arts_centre"], "count" : 3 }
{ "_id" : ["veterinary"], "count" : 3 }
{ "_id" : ["prison"], "count" : 2 }
{ "_id" : ["dentist"], "count" : 2 }
{ "_id" : ["Delhi university women's college"], "count" : 1 }
{ "_id" : ["Bawana Bus depot"], "count" : 1 }
{ "_id" : ["auditorium"], "count" : 1 }
{ "_id" : ["orphanage"], "count" : 1 }
{ "_id" : ["nursing_home"], "count" : 1 }
{ "_id" : ["parking_entrance"], "count" : 1 }
{ "_id" : ["car_wash"], "count" : 1 }
{ "_id" : ["recycling"], "count" : 1 }
{ "_id" : ["vending_machine"], "count" : 1 }
{ "_id" : ["Garbage Collection Units"], "count" : 1 }
{ "_id" : ["traffic education"], "count" : 1 }
{ "_id" : ["driving_school"], "count" : 1 }
{ "_id" : ["House"], "count" : 1 }
{ "_id" : ["residential"], "count" : 1 }
{ "_id" : ["Ksan Ghat"], "count" : 1 }
{ "_id" : ["fairgrounds"], "count" : 1 }
{ "_id" : ["biergarten"], "count" : 1 }
{ "_id" : ["club"], "count" : 1 }
{ "_id" : ["bureau_de_change"], "count" : 1 }

{ "_id" : ["bicycle_parking"], "count" : 1 }
{ "_id" : ["baby_hatch"], "count" : 1 }
{ "_id" : ["townhall"], "count" : 1 }
{ "_id" : ["community_hall"], "count" : 1 }
{ "_id" : ["Ayurvedic Hospital"], "count" : 1 }
{ "_id" : ["shelter"], "count" : 1 }
{ "_id" : ["Suvidha Market, Netaji Nagar"], "count" : 1 }
{ "_id" : ["Netaji Nagar Market"], "count" : 1 }
{ "_id" : ["Electricity office"], "count" : 1 }
{ "_id" : ["architect"], "count" : 1 }

Types of highways:
> db.maps.aggregate([{$match: {type: "way"}}, {$group: {_id:"$highway", count:{$sum: 1}}}, {$sort: {count: -1}}])
{ "_id" : "residential", "count" : 8883 }
{ "_id" : null, "count" : 4107 }
{ "_id" : "service", "count" : 3398 }
{ "_id" : "tertiary", "count" : 1536 }
{ "_id" : "secondary", "count" : 867 }
{ "_id" : "living_street", "count" : 672 }
{ "_id" : "unclassified", "count" : 224 }
{ "_id" : "primary", "count" : 206 }
{ "_id" : "footway", "count" : 201 }
{ "_id" : "trunk", "count" : 176 }
{ "_id" : "primary_link", "count" : 128 }
{ "_id" : "trunk_link", "count" : 104 }
{ "_id" : "pedestrian", "count" : 48 }
{ "_id" : "steps", "count" : 44 }
{ "_id" : "motorway_link", "count" : 43 }
{ "_id" : "motorway", "count" : 36 }
{ "_id" : "track", "count" : 30 }
{ "_id" : "path", "count" : 27 }
{ "_id" : "secondary_link", "count" : 25 }
{ "_id" : "tertiary_link", "count" : 6 }
{ "_id" : "cycleway", "count" : 4 }
{ "_id" : "construction", "count" : 2 }

Popular cuisines:
> db.maps.aggregate([{$match: {amenity: {$exists: 1}, amenity:"restaurant"}}, {$group: {_id:"$cuisine", count: {$sum:1}}}, {$match: {_id: {$ne: null}}}, {$sort: {count: -1}}, {$limit: 5}])

{ "_id" : "indian", "count" : 7 }

{ "_id" : "chinese", "count" : 3 }
{ "_id" : "pizza", "count" : 3 }
{ "_id" : "vegetarian", "count" : 3 }
{ "_id" : "asian", "count" : 2 }

Conclusion:

Looking at the dataset, it looks like there is a great wealth of information about the city though the set is far from being complete.
If there is more effort put on the data cleanup with robust scripts and if there is an app that would auto update the data as the users move around the different places, the data would become extremely useful in :
1. planning where more facitilities/infrastructure is needed.
2. attracting tourism.
3. real estate
and so on.