Introduction

For this project I chose to download the OSM sample file for Dublin city (ROI). Dublin is the only city in my country large enough to meet the project requirements, and in choosing this map I hoped to encounter (and resolve) problems unique to data sets dealing with Ireland.

In this report I will address four problems; invalid nodes, inconsitent roof values, inconsistent address keys and invalid mongoDB characters. After that I will present an overview of the data, and finally some suggestions on where the data could be improved.

Problems with the Data

Checking for Valid Nodes

One of the first tasks I did while cleaning the data (norcleaning.py), was to check if all the 'node' elements in the OSM file were valid. A valid node must have 'id', 'lat' and 'lon'. It turned out that all nodes in the sample file and the full file were valid, but this was done mainly as an example of a data point that can be audited for validity.

Incosistent Roof Values

While examining the keys for 'way' elements I found the following entries (and counts) for roof keys :

"building-roof-shape": 327,

"building:roof": 1319,

"building:roof:shape": 20569,

"building:roof:type": 13,

"building:roof_shape": 1031,

These are all pretty similar so I standardised these to : "building:roof:shape" or "building:roof:type".

I also looked at the values for these keys and found the following entries (and counts) :

"flat": 57,

"pitched": 23004,

"slate": 95,

"sloped": 39,

"tile": 56,

"tiled": 8

"pitched" and "sloped" are the same thing, so I standardised these to "pitched". One thing I noted at this point was the length of time it took for my cleaning program to run on the full map file doubled from one minute to two after implementing this step.

Inconsistent Address Keys

The sample file doesn't contain any, but the full map file contained “address” keys that caused my cleaning program to crash. It turned out there was only two of these in the full map file, so I created a set of rules to standardise these to “housename”, “housenumber”, etc. I think this captures the data best for the map file I chose, but it may require better cleaning rules on other map files.

Problem mongoDB Characters

If a value contained characters that might cause problems in mongoDB, the tryFixValue() function was called. In examining these, most were found to be house numbers and telephone numbers beginning with “#” or “+”. These simply had the first character removed. A large number of what remained were “.” or “?”, and were replaced with “empty”. I did no further processing on what remained (mostly "@LS@") and just replaced these with “invalid”.

Data Overview

File Sizes

dublin_full.osm 126MB

dublin_full.json 164MB

Number of Documents

```
dublinDB.dublin.count()
```

637268

Number of Nodes

```
dublinDB.dublin.find({'type':'node'}).count()
```

525263

Number of Ways

```
dublinDB.dublin.find({'type':'way'}).count()
```

109311

Number of Unique Users

```
len(dublinDB.dublin.distinct('user'))
```

947

Top Three Contributors

```
pipeline = [ { "$group" : { "_id" : "$user","count": {"$sum": 1 }}},
        { "$sort" : { "count" : -1 }},
        { "$limit" : 3 }
        ]
result = dublinDB.dublin.aggregate(pipeline)
```

brianh, VictorIE and mackerski

Number of Pubs

```
dublinDB.dublin.find({'amenity':'pub'}).count()
```

493

Additional Ideas

I expected to have more problems with irish language entries than I encountered, but I still think these could be improved. The convention at the moment seems to be to use the dominant language for the area the map represents. This leaves most of the country being mapped in english, and only gaeltacht areas using irish. This causes problems when people have irish housenames for example. Many people also have an english language version of their housename as many online forms won't accept the irish language version. This leads to some entries having both “addr:housename” and “addr:housename:ga” tags.

I would rather standardise this to having the key in the same language as the value. An example would be instead of “addr:streetname:ga” would be simply “seoladh:ainmsráide”.

I also think it would be useful to develop a tool to allow less tech savvy users to add to maps. It wouldn't be without it's challenges, but even a city the size of dublin has only 947 contributors. My own town has no entry for Well Lane. This has a couple of small car parks, an amenity walkway, a recycling drop off point, a children's playground, an outdoor gym and a FIFA sized all weather pitch. I don't have the skill to draw these things in myself, but it could be useful to have a tool to click on the map and at least add the tags; perhaps with a limited amount of tags, keys and values to choose from.

Apart from the overheads for developing and maintaining the tool, this may open the map up to more instances of incorrect, innaccurate or openly malicious use, though the potential for these already exists. I would hope the benefits of opening the map up to more contributors would more than outweight the disadvantages.