OpenStreetMap Sample Project

Data Wrangling with MongoDB

Ronan Brady

Map Area: Dublin, Ireland

https://mapzen.com/metro-extracts/

https://s3.amazonaws.com/metro-extracts.mapzen.com/dublin_ireland.osm.bz2

1. Reasons for Selection of Dublin Region

I selected Dublin as it is my native city and I have spent most of my life living there. I was interested to see the extent to which the OpenStreetMap data was accurate and up-to-date.

2. Problems Encountered in the Map

2.1 Initial Audit

Initially I ran a script to ensure I had all the address data, mapped and shaped.

Notably several of the fields were in both Irish (Gaeilge) and English.

I counted instances of each address field type:

Key	Count
addr:town	1
addr:city	9,360
addr:city:ga	29
addr:county	2
addr:postcode	32
addr:postal_district	18
addr:street	72,788
addr:street:ga	29
addr:full	8

Notably, there were relatively few entries that contained City or Town or Postcode. This is surprising as a Dublin address is typically of the form:

<House Num> <Street Name>

<Town>

<City> or <County>

Where <City> would be of the form: Dublin <Postcode> or <County> outside of central Dublin specified as County Dublin, typically abbreviated to Co. Dublin

There was only 1 entry for Town and only 9,360 for City and only 2 entries for county.

1.2 Audit of City

I then conducted an audit of the City data noting the following issues:

{'_id': 'Booterstown, Blackrock', 'count': 1} // Comma within City field

{'_id': 'Clontarf Dublin 3', 'count': 1} // Town and City and Postcode in City field

{'_id': 'Corballis, Donabate', 'count': 2} // Two towns listed

{'_id': 'County Wicklow', 'count': 1} // County Wicklow (adjacent County) listed in Dublin extract.

{'_id': 'DUNBOYNE', 'count': 1} All caps

```
{'_id': 'Dublin Co. Dublin', 'count': 1}
{'_id': 'Dun Laoghaire', 'count': 5}  Two different spellings
{'_id': 'Dún Laoghaire', 'count': 12}
{'_id': 'Enniskerry, Co. Wicklow', 'count': 1}  County listed in the City field
{'_id': 'Kinsealey', 'count': 1}  Mispelling
{‘_id’: ‘‘Lucan, County Dublin', 'count': 1}  County in the field
```

1.3 Audit of Street Names

Then Audited the Street names.

Noted commas in streetnames and inconsistency of use of Ave. Avenue etc.

```
'Ave': 2, // non-standard abbreviation
'Avevnue': 1, // mispelling
'Cente': 1, // mispelling
'Center': 1, // non-Irish-British spelling of the word ‘Centre’
'Donghmede': 1, // mispelling
'Nouth': 1, // mispelling
'Roafd': 1, // mispelling
'Sreet': 1, // mispelling
'St': 2, // non-standard abbreviation
'heights': 2, // all lowercase
'lane': 2, // all lowercase
'park': 1, // all lowercase
'place': 1, // all lowercase
'road': 1, // all lowercase
'street': 2 // all lowercase
```

2. Opportunities for Improving Data Quality

2.1 Fix Street Name Data

Based on the audit above there was an opportunity to fix the mispellings in the street data by script. To do this I created a mapping of the incorrect data, mapped to the corrected data.

```
mapping = { "Ave": "Avenue",
      "Avevnue": "Avenue",
      "Cente": "Centre",
      "Center": "Centre",
      "Center": "Centre",
      "Donghmede": "Donaghmede",
      "Nouth": "North",
      "Roafd": "Road",
      "Sreet": "Street",
      "St": "Street",
      "street": "Street",
      "heights": "Heights",
      "lane": "Lane",
      "park": "Park",
      "place": "Place",
      "road": "Road"
      }
```

See fixdata.py for full script.

Running the script resulted in 81 updated entries.

2.1 Fix Postal Region

The majority of addresses listed do not contain a Postal Region. The Postal Region is a basic postcode used in the Dublin area, based on 22 regions within the City. The boundaries of these regions (geo-coordinates) could be determined or sourced from the national Postal operator, and then a script could be written to automatically update addresses with the correct Postcode.

2. Data Overview

2.1 Basic Statistics

This section contains basic statistics about the dataset and the MongoDB queries used to gather them.

File sizes

```
dublin.osm ......... 231.5 MB
dublin.osm.json .... 308.3 MB
```

# Number of documents

```
    db.openmap.find().count()
    1154249
```

# Number of nodes

```
    db.openmap.find({"type":"node"}).count()
    1471349
```

# Number of ways

```
    db.openmap.find({"type":"way"}).count()
    Ways: 170079
```

# Number of unique users

```
    count = len(db.openmap.distinct("created.user"))
    print("Distinct Users: {0}".format(count))

    Distinct Users: 1022
```

# Top 1 contributing user

```
result = db.openmap.aggregate([{"$group":
                                   {"_id" : "$created.user",
"count" : { "$sum" : 1 } } },
                              { "$sort" : {"count": -1} },
                              {"$limit":1 }  ] )
```

```
{'_id': 'Nick Burrett', 'count': 227720}
```

# Number of users appearing only once (having 1 post)

```
result = db.openmap.aggregate([{"$group":
                                   {"_id" : "$created.user",
"count" : { "$sum" : 1 } } },
                              {"$group":
                                   {"_id" : "$count",
"num_users" : { "$sum" : 1 } } },
                              { "$sort" : {"num_users": -1} },
                              {"$limit":1 }  ] )
```

```
{'_id': 1, 'num_users': 246}
# “_id” represents postcount
```

2.2 Other Queries

Top 10 Amenities

Amenity	**Count**
Parking	2,022
Pub	700
Restaurant	614
Fast Food	567

School	543
Cafe	523
Post Box	413
Place of Worship	392
Bench	277
Bicycle Parking	269

Top Contributors For Pubs

Rank	**Name**	**Count**
#1	VictorIE	123
#2	mackerski	103
#3	Nick Burrett	97
#4	ManAboutCouch	73
#5	Meenatuggart	28
#6	IrlJidel	19
#7	rorym	14
#8	plush	12
#8	Eoin OMahony	12
#8	wheelmap_visitor	12
#8	Dafo43	12

Notably, the top 10 contributors account for 70% of the pub entries.

Fast Food Cuisine

Rank	**Name**	**Count**
#1	'fish_and_chips'	77

#2	'burger'	68
#3	chinese	58
#4	pizza	44
#5	kebab	21

Other Analysis Attempted
I attempted to see what the total capacity of the carparks was however there was no carpark with ‘capacity’ attribute. In addition I attempted to see who many pubs had beergardens but there were no pubs listed with the beer_garden attribute.

3. Additional Ideas

3.1 Agree Consistent Format for Irish Addresses

As Ireland is relatively unique in its lack of a Postcode for all addresses, it is off additional importance that to have a consistent address structure for the fields of County, City and Town. The national postal service ‘An Post’ identifies a ‘correct’ address as:

> A correct postal address normally includes:
>
> 1. Name of addresee.
> 2. Number or name of house and the name of the street, road, etc. In rural areas the name of the locality or townland should be shown.
> 3. Post Town. For Dublin addresses the postal district number, where applicable, should always be shown.
> 4. County Name.

Source: http://correctaddress.anpost.ie/pages/Search.aspx

I would suggest that a new standard for Ireland is agreed in the OpenStreetMap community and then the community members work towards consistently applying the new standard. This would avoid the many occurrences of incomplete Dublin addresses, (with No Town or City).

3.2 Impending Implementation of Postcodes

To address lack of postcodes, in 2015 the Irish Government is introducing new postcodes Eircodes.

> The Eircode Address Database (ECAD) is a database containing address information for all 2.2 million addresses in Ireland and includes the following for each address:
>
> - USP postal address
> - The Eircode
> - Geo-coordinates of the centroid of the address
> - Aliases of address elements relating to the address
>
> The ECAD also contains additional data relating to each address including but not limited to boundary data and building information.
>
> Source: https://www.eircode.ie/business/products-and-services

This data is obviously similar to the OpenMap data. This presents an opportunity and a challenge. The OpenMap community may be able to source the data from Eircode to update OpenMap. This would serve adoption of Eircodes; however, reviewing the licensing structure of the Eircode data product, this may not be commercially in the interest of Eircode to do. The challenge then to the OpenMap community is to maintain consistency with what is, or will be, the authorative dataset on Irish addresses. As each residence and business is receiving notification of their Eircode data, it would be possible for individuals to update OpenMap, if a simple mechanism, e.g. online submission form, were provided by the OpenMap community.

3.3 Gaelic Version of Address Data

Gaelic (Irish) is the official first language of the Irish State. However, for historical and cultural reasons English is the de facto first language, spoken by most citizens, with Gaelic a minority language. However, every address in the Status has an Irish (Gaelic) equivalent. This was though, not reflected in the OpenMap DB data for Dublin, with only 29 out of 9,360 cities had a Gaelic entry.

addr:city:ga	29

Though being a minority language, there are ardent supporters in the community, notably Gaelic is one of few global minority language with a course on DuoLingo.

Correspondingly, it may be possible to mobilise the community by highlighting to them to deficit in the OpenMap database, and how accurate listing of the data would further the aims of the minority language movement.

Conclusion

After this review of the data it is clear the the Dublin area data is incomplete and of somewhat poorer quality than expected. The introduction of Eircodes (postcodes) in Ireland presents an opportunity to mass update the OpenMap data and improve accuracy. In addition, engagement with the Irish language community represents another, separate opportunity.