

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. Problems Encountered in the Map

Initial Audit

Audit of City

Audit of Street Names

2. Data Overview

3. Additional Ideas

Conclusion

1. Problems Encountered in the Map

1.1 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.

Firstly I created a set of all the address keys used in the data set.

```
'addr:city',  
'addr:city:ga',  
'addr:country',  
'addr:county',  
'addr:full',
```

'addr:housename',
'addr:housename:en',
'addr:housename:ga',
'addr:housename:it',
'addr:housenumber',
'addr:housenumber:source',
'addr:interpolation',
'addr:place',
'addr:postal_district',
'addr:postcode',
'addr:street',
'addr:street:ga',
'addr:street:name',
'addr:terracedname',
'addr:town',
'addr:unit',

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': 'Bray', 'count': 16}`

`{'_id': 'Cabinteely', 'count': 4}`

`{'_id': 'Castleknock', 'count': 1}`

`{'_id': 'Celbridge', 'count': 45}`

`{'_id': 'Churchtown', 'count': 6}`

`{'_id': 'Citywest', 'count': 1}`

`{'_id': 'Cloghran', 'count': 7}`

`{'_id': 'Clondalkin', 'count': 1}`

`{'_id': 'Clongriffin', 'count': 1}`

`{'_id': 'Clontarf', 'count': 43}`

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

`{'_id': 'Co. Dublin', 'count': 1}`

`{'_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': 'Dalkey', 'count': 3}`

`{'_id': 'Donabate', 'count': 2}`

`{'_id': 'Donnybrook', 'count': 1}`

{'_id': 'Dublin', 'count': 4022}
{'_id': 'Dublin 1', 'count': 245}
{'_id': 'Dublin 10', 'count': 3}
{'_id': 'Dublin 11', 'count': 146}
{'_id': 'Dublin 12', 'count': 153}
{'_id': 'Dublin 14', 'count': 1}
{'_id': 'Dublin 15', 'count': 10}
{'_id': 'Dublin 16', 'count': 19}
{'_id': 'Dublin 17', 'count': 7}
{'_id': 'Dublin 18', 'count': 5}
{'_id': 'Dublin 2', 'count': 476}
{'_id': 'Dublin 20', 'count': 1}
{'_id': 'Dublin 24', 'count': 2}
{'_id': 'Dublin 3', 'count': 85}
{'_id': 'Dublin 4', 'count': 37}
{'_id': 'Dublin 5', 'count': 39}
{'_id': 'Dublin 6', 'count': 677}
{'_id': 'Dublin 6W', 'count': 220}
{'_id': 'Dublin 7', 'count': 152}
{'_id': 'Dublin 8', 'count': 328}
{'_id': 'Dublin 9', 'count': 99}
{'_id': 'Dublin Airport', 'count': 12}
{'_id': 'Dublin Airport Logistics Park', 'count': 1}
{'_id': 'Dublin Co. Dublin', 'count': 1}
{'_id': 'Dun Laoghaire', 'count': 5} Two different spellings
{'_id': 'Dún Laoghaire', 'count': 12}
{'_id': 'Dunboyne', 'count': 3}
{'_id': 'Dundrum', 'count': 2}
{'_id': 'Dunshaughlin', 'count': 6}
{'_id': 'Enniskerry, Co. Wicklow', 'count': 1} County listed in the City field
{'_id': 'Finglas', 'count': 2}
{'_id': 'Glenageary', 'count': 1}

{'_id': 'Harristown', 'count': 4}
{'_id': 'Howth', 'count': 10}
{'_id': 'Kilbarrack', 'count': 1}
{'_id': 'Killiney', 'count': 1}
{'_id': 'Kilternan', 'count': 1}
{'_id': 'Kinsealey', 'count': 1} Misspelling
{'_id': 'Leixlip', 'count': 16}
{'_id': 'Leopardstown', 'count': 1}
{'_id': 'Loughlinstown', 'count': 1}
{'_id': 'Lucan', 'count': 1314}
'Lucan, County Dublin', 'count': 1} County in the field
{'_id': 'Malahide', 'count': 19}
{'_id': 'Monkstown', 'count': 1}
{'_id': 'Mount Merrion', 'count': 1}
{'_id': 'Ongar', 'count': 28}
{'_id': 'Palmerstown', 'count': 7}
{'_id': 'Portmarnock', 'count': 3}
{'_id': 'Raheny', 'count': 1}
{'_id': 'Rathfarnham', 'count': 16}
{'_id': 'Rathmines', 'count': 2}
{'_id': 'Sandyford Industrial Estate', 'count': 5}
{'_id': 'Santry', 'count': 4}
{'_id': 'Shankill', 'count': 1}
{'_id': 'Stillorgan', 'count': 26}
{'_id': 'Stillorgan Industrial Park', 'count': 1}
{'_id': 'Sutton', 'count': 1}
{'_id': 'Swords', 'count': 33}
{'_id': 'Tyrrelstown', 'count': 1}

1.3 Audit of Street Names

Then Audited the Street names.

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

{'1': 1,
'4': 1,
'Abbey': 2,
'Airport': 1,
'Albany': 1,
'Alders': 1,
'Alley': 4,
'Apartments': 1,
'Ave': 2,
'Avenue': 246,
'Avevnu': 1,
'Ballycoolin': 1,
'Bank': 2,
'Bar': 1,
'Bay': 1,
'Beeches': 1,
'Blanchardstown': 1,
'Boulevard': 1,
'Brae': 1,
'Brambles': 1,
'Bridge': 1,
'Brook': 2,
'Brugh': 1,
'Buildings': 1,
'Bypass': 3,
'Cairn': 1,
'Campus': 2,
'Castle': 2,
'Cedars': 1,
'Cente': 1,

'Center': 1,
'Centre': 13,
'Chase': 4,
'Cherries': 1,
'Citywest': 1,
'Cliffs': 1,
'Close': 70,
'Coille': 1,
'College': 1,
'Coombe': 1,
'Copse': 7,
'Cornelscourt': 1,
'Corner': 1,
'Cottages': 22,
'Court': 102,
'Courtyard': 1,
'Cove': 2,
'Crescent': 57,
'Cross': 3,
'Crossing': 1,
'Dale': 10,
'Dales': 1,
'Demesne': 1,
'Dene': 1,
'District': 1,
'Dock': 1,
'Donghmede': 1,
'Downs': 10,
'Drive': 117,
'Dublin': 1,
'Dundrum': 1,
'East': 25,

'Elms': 2,
'End': 1,
'Estate': 7,
'Extension': 1,
'Fairways': 1,
'Farm': 1,
'Fitzpatrick': 1,
'Flats': 1,
'Fosters': 1,
'Gardens': 30,
'Garth': 4,
'Gate': 2,
'Gates': 1,
'Glade': 8,
'Glasnevin': 1,
'Glen': 5,
'Grange': 2,
'Great': 1,
'Green': 60,
'Grove': 91,
'Hall': 11,
'Harbour': 3,
'Haven': 1,
'Heath': 7,
'Heights': 19,
'Hill': 25,
'House': 4,
'Island': 1,
'Kilbride': 1,
'Kilmainham': 1,
'Lane': 88,
'Laurels': 1,

'Lawn': 58,
'Lawns': 16,
'Leopardstown': 1,
'Little': 1,
'Lodge': 7,
'Lorcan': 1,
'Lotts': 2,
'Lower': 41,
'Lucan': 1,
'Mall': 4,
'Manor': 10,
'Maples': 1,
'Mara': 1,
'Market': 3,
'Mart': 1,
'Mead': 1,
'Meadow': 4,
'Meadows': 5,
'Meath': 1,
'Mews': 13,
'Middle': 1,
'Moorings': 1,
'Mount': 1,
'North': 27,
'Nouth': 1,
'Oaks': 4,
'Orchard': 3,
'Paddock': 1,
'Parade': 12,
'Park': 233,
'Pinar': 1,
'Pines': 1,

'Place': 75,
'Plaza': 3,
'Pleasant': 1,
'Point': 1,
'Poplars': 1,
'Quay': 21,
'Rathcoole': 1,
'Rathgar': 1,
'Rectory': 1,
'Rest': 1,
'Rise': 24,
'Road': 513,
'Roafd': 1,
'Row': 24,
'Rubrics': 1,
'Shée': 1,
'Slopes': 1,
'Smol': 1,
'South': 28,
'Square': 34,
'Sreet': 1,
'St': 2,
'Stables': 1,
'Strand': 2,
'Street': 205,
'Sycamores': 1,
'Terrace': 109,
'Thicket': 1,
'Thorpe': 1,
'Upper': 39,
'Vale': 10,
'Valley': 2,

'View': 24,
'Villa': 2,
'Village': 5,
'Villas': 30,
'Walk': 36,
'Ward': 1,
'Way': 42,
'West': 25,
'Willows': 1,
'Wood': 20,
'Woods': 1,
'Yard': 1,
'castletown': 1,
'heights': 2,
'lane': 2,
'park': 1,
'place': 1,
'road': 1,
'street': 2}

After initially downloading I queried town = Malahide, however, there were no results and I discovered that Malahide was listed as a city. Further I noticed that addresses in Terenure were listing the city as 'Dublin 6W' but did not have any postcode attribute.

I then queried the definition of City and Town in openstreetmap

2. Data Overview

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: {}".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
```

3. Additional Ideas

Agree Consistent Format for Irish Addresses

As Ireland is relatively unique in it's lack of a Postcode for all addresses, it is of additional importance that to have a consistent address structure for the fields of County, City and Town. This is because these fields contain essential information to accurately specify an address. I would suggest that this is agreed in the OpenMap 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).

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 authoritative 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.

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.