Patterns of tagging in OpenStreetMap data in urban areas

Nikola Davidovic*[1] and Peter Mooney†[2]

[1]Faculty of Electronic Engineering,University of Nis,18000 Nis(RS), Serbia
[2]Department of Computer Science, Eolas, North Campus, Maynooth University, Ireland

Summary

OpenStreetMap (OSM) contributors are free to apply any tags (key-value pairs) they wish to geographic objects in the OSM database. Guidance for tagging is provided by many sources (Wikis, OSM editors, mailing lists, etc). In general contributors do not always follow these guidelines. In this paper we develop an approach to extract and analyse patterns of tagging of OSM data in urban areas. We find that there are local variations and other factors which see an in-homogeneous approach to tagging in different cities and regions. Our results can contribute to developing 'best practice' approaches for tagging in OSM.

KEYWORDS: OpenStreetMap, Tagging, Pattern Matching, Volunteered Geographic Information.

1 Introduction

As stated by many authors OpenStreetMap (OSM) has a flexible and open approach to the tagging of objects in its database of geographical features (Arsanjani et al., 2015; Ballatore and Mooney, 2015; Mooney and Corcoran, 2012; Jilani et al., 2014). Contributors to OSM who wish to create new geographical features or editing existing ones can obtain advice and guidance on which *tags* (key-value pairs) to use by consulting a number of sources which include: the OSM Map Features pages on the OSM Wiki[1]; automated tagging suggestions and searchable listing of popular tags from OSM editor software such as JOSM or iD; aggregator engines such as TagInfo[2] which provide statistics on the global usage of tags and their key-value pairs in OSM and through online discussion forums such as mailing lists (Ballatore and Mooney, 2015). However contributors are not strictly enforced to apply a specific set of tags to a particular object depending on its thematic area or geographical context. Contributors are still free to apply tags as they see fit due to local and regional variations and their geographical understanding (Estima and Painho, 2013; Schmidt et al., 2013).

What does this mean in reality? As a means of a very simple example suppose three different OSM contributors are creating an object in the OSM database to represent a restaurant building. Suppose

*nikola.davidovic@elfak.ni.ac.rs
†peter.mooney@nuim.ie

[1]`http://wiki.openstreetmap.org/wiki/Map_Features`
[2]`https://taginfo.openstreetmap.org/`

the three contributors are in Kyoto Japan, San Francisco USA and Düsseldorf Germany. Given the flexibility of the OSM ontology and its open approach to tagging all three of our contributors can apply whichever tags they feel are most appropriate to this object. So in the case of San Francisco the contributor might use `cuisine,name,addr:housenumber,addr:street` as four keys for tags. In Düsseldorf the contributor might use `addr:housenumber,addr:street,addr:city,name` as four keys for tags while in Kyoto the contributor might use `name:ja_rm,name:ja,name:en,name`. It is immediately evident from this simple example (where the tagging is extracted from actual OSM data from each city) that there is no broad agreement amongst our three contributors. Certainly in the case of Kyoto there is a strong tendency towards the multilingual nature of the name of the restaurant. Are there similar patterns of tagging being used by contributors to OpenStreetMap in urban areas? Are there local variations or other factors which see a fragmented and in-homogeneous approach to tagging in these areas? Such a situation makes comparing data between OSM areas difficult. It can also be problematic when comparing, combining or fusing OSM data with other sources of geospatial data such as National Mapping Agency (NMA) data.

In this paper we describe an approach which we have developed to extract and analyse patterns of tagging in OSM data with particular focus on urban areas. Are there similar patterns of tagging behaviors carried out in major urban areas in OSM? For objects with more than one tag are there repeated patterns or structures used in other co-occurring tags? The remainder of the abstract is organised as follows. Section 2 outlines our experimental analysis of patterns of tags extract from OSM for a selection of urban areas. In Section 3 we close the paper with some conclusions on this work and some directions for immediate and longer term future work.

2 Experimental Analysis

We selected 22 different city datasets and downloaded their OSM data from the Mapzen Metro Extracts website[3]. Cities used were: Bangkok, Beijing, Boston, Bucharest, Buenos Aires, Dublin, Düsseldorf, Frankfurt, Helsinki, Johannesburg, Kyoto, Madrid, Mexico City, Nairobi, New Delhi, Nis, Oslo, Ottawa, Saint Petersburg, San Francisco, Singapore and Sydney. Cities of all sizes from all continents were chosen. Some cities were especially being included because of their active OSM communities. The OSM data was imported into a PostgreSQL PostGIS database for processing.

Our approach to extracting patterns of tagging is centered around the concept of co-occurring tags. For our analysis we chose a number of *target tags* which are globally very frequently used according to the TagInfo application[4]. These are: `amenity=restaurant`, `highway=primary`, `highway=residential`, `natural=water`, `highway=track` and `building=yes`. In each of our city datasets we extracted nodes or polygons which had one of the target tags. For each object containing a target tag we build a hash-table data structure to store all of the other tags which appear with the target tag (co-occurring). We counted the number of co-occurrences of pairs of tags (the target tag and another) ($N = 1$), triples (the target tag and two other tags) ($N = 2$), and so on. When processing is completed

[3] `https://mapzen.com/data/metro-extracts`
[4] `https://taginfo.openstreetmap.org/tags`

we have aggregated derived datasets containing statistics corresponding to each target-tag and the other tags which co-occur with them. We extend our analysis to $N = 5$. For the co-occurring tags we just store the key for the tag. Table 1 and Table 2 show two examples of the information extracted on co-occurring patterns.

In Table 1 the $N = 5$ co-occurrences for the `building=yes` target tag are shown. Due to space constraints we only show the cities where the percentage of objects with the $N = 5$ co-occurrence is greater than 1% of all objects with the target tag. Table 2 describes in a similar way the $N = 5$ co-occurrences for the `highway=residential` target tag.

There are a number of interesting observations. With the exception of Dublin, Helsinki and San Francisco Table 1 shows broad agreement on the $N = 5$ co-occurrences with `building=yes`. San Francisco appears to be influenced by a bulk import of external geographic data for the Redwood City region. The OSM Wiki for this tag[5] indicates that the following tag values should be considered as suitable combinations with `building=yes`: `addr`, `name`,`height`,`building:levels`, `entrance`,`shop`,`amenity`,`office`,`craft` and `building:architecture`. Addressing via the `addr:` prefix is dominant. Interestingly the `name` or `amenity` tag value do not appear in the most frequently occurring combinations of co-occurrences in Table 1.

For Table 2 there is much less agreement amongst the most frequently occurring patterns of $N = 5$ co-occurrences. Boston, Kyoto, Ottawa and San Francisco are subject to bulk import of external geographic data. This bulk import does not seem to have any direct relationship to the suggested or recommended tags in the OSM Wiki pages. Among the remaining cities there is apparent agreement on tag keys such as `name` and `surface`. Overall there is no structured pattern of tagging for `highway=residential`. The OSM Wiki [6] indicates that `name` and `oneway` are keys which should be used. This only appears for Bucharest and Nis. There are some considerable differences between the co-occurrences of tagging between cities. This will require further investigation potentially involving socio-economic (Haklay, 2013) indicators and other spatial variables. We have many more examples which shall be presented in the conference presentation.

3 Conclusions and Future Work

The problem of extracting co-occurring tags and determining the most frequently patterns of co-occurring tags in different OSM datasets is an interesting and challenging problem. To our current knowledge this type of investigation of tagging structures in crowdsourced geographic data has not been performed yet. We strongly believe that identifying these patterns of tagging and their local variations can greatly contribute to an overall assessment of the temporal and semantic data quality of objects in databases such as OpenStreetMap. A more structured and homogeneous approach to tagging in urban areas can greatly benefit those developing applications and methodologies based on OSM data. These include: Scioscia et al. (2014) who develop a general technique for semantic annotation of OSM data using Location-based Services (LBS) applications; Kunze and Hecht (2015)

[5] `http://wiki.openstreetmap.org/wiki/Key:building`

[6] `http://wiki.openstreetmap.org/wiki/Tag:highway%3Dresidential`

Table 1: ($N = 5$) co-occurrences for the `building=yes` target tag.

City	N=5 Co-occurrences	#Objects	%Obj
Dublin	addr:housenumber; building:cladding; building:use; house; levels	1191	3.57
Düsseldorf	addr:city; addr:country; addr:housenumber; addr:postcode; addr:street	124716	44.98
Frankfurt	addr:city; addr:country; addr:housenumber; addr:postcode; addr:street	176564	19.33
Helsinki	addr:housenumber; addr:street; building:fi:id; building:levels; start_date	11594	9.62
Madrid	addr:city; addr:housenumber; addr:postcode; addr:street; building:levels	2238	5.33
Ottawa	addr:city; addr:housenumber; addr:postcode; addr:street; source:addr	1765	10.62
St Petersburg	addr:city; addr:country; addr:housenumber; addr:street; building:levels	2680	1.45
San Francisco	addr:city; addr:housenumber; addr:street; redwood_city_ca:addr_id; redwood_city_ca:bld_gid	12868	6.02
Singapore	addr:city; addr:country; addr:housenumber; addr:postcode; addr:street	766	1.72
Sydney	addr:city; addr:country; addr:housenumber; addr:postcode; addr:street	360	2.84

who has developed an approach to process semantic information from OSM data to specify non-residential usage in residential buildings or; Bakillah et al. (2014) who developed an approach to accurately estimate population distribution from OSM data but concluded that that further research is needed to understand how tags and their locations can reveal population distribution patterns. In Behrens et al. (2015) the authors assess the usability of the iD OSM editor tool and conclude that there are improvements to be made regarding how users learn the tool. Presenting users, and in particular novice users, with suggested co-occurring tags extracted from the global corpus of OSM co-occurrences may deliver a more structured approach to OSM tagging. We believe that by extending our analysis to other cities and tag combinations we can develop a best practice tagging guide for a large number of features in OSM.

Acknowledgements

This work was supported by the COST Action Short Term Scientific Mission (STSM) COST-STSM-TD1202-30048 and the support of the COST Action TD1202 'Mapping and the Citizen Sensor' is greatly appreciated.

Table 2: ($N = 5$) co-occurrences for the `highway=residential` target tag.

City	N=5 Co-occurrences	#Objects	%Obj
Boston	attribution; condition; lanes; massgis:way_id; name	11621	84.77
Bucharest	is_in:city; maxspeed; name; oneway; surface	1238	6.85
Buenos Aires	lanes; lit; maxspeed; name; surface	937	1.29
Düsseldorf	lit; lit_by_gaslight; maxspeed; name; surface	2550	19.73
Helsinki	maxspeed; name; name:fi; name:sv; surface	8574	39.35
Johannesburg	access; bicycle; foot; motor_vehicle; surface	1096	2.09
Kyoto	yh:STRUCTURE; yh:TOTYUMONO; yh:TYPE; yh:WIDTH; yh:WIDTH_RANK	1808	8.33
Madrid	lanes; lit; maxspeed; name; surface	1400	1.82
Nis	name; name:sr; name:sr-Latn; oneway; surface	41	3.2
Ottawa	attribution; geobase:acquisitionTechnique; is_in; lanes; name	2555	10.45
St Petersburg	lanes; lit; maxspeed; name; surface	1641	7.81
San Francisco	name; tiger:cfcc; tiger:county; tiger:name_base; tiger:name_type	26407	84.6
Sydney	is_in:suburb; maxspeed; name; source:name; surface	1491	3.23

Biography

Mr. Nikola Davidovic is a research assistant and a PhD candidate at the Faculty of Electronic Engineering, University of Nis, Serbia. His PhD is examining methodologies and approaches to extracting patterns of behavior from Volunteered Geographic Information (VGI) datasets. Dr. Peter Mooney is a lecturer and researcher in the Computer Science Department at Maynooth University, Ireland. His interests include Volunteered Geographic Information, User-generated content and spatial database management.

References

Arsanjani, J. J., Zipf, A., Mooney, P., and Helbich, M. (2015). An Introduction to OpenStreetMap in Geographic Information Science: Experiences, Research, and Applications. In Arsanjani, J. J., Zipf, A., Mooney, P., and Helbich, M., editors, *OpenStreetMap in GIScience*, Lecture Notes in Geoinformation and Cartography, pages 1–15. Springer International Publishing. DOI: 10.1007/978-3-319-14280-7_1.

Bakillah, M., Liang, S., Mobasheri, A., Arsanjani, J. J., and Zipf, A. (2014). Fine-resolution population mapping using openstreetmap points-of-interest. *International Journal of Geographical Information Science*, 28(9):1940–1963.

Ballatore, A. and Mooney, P. (2015). Conceptualising the geographic world: the dimensions of nego-

tiation in crowdsourced cartography. *International Journal of Geographical Information Science*, 29(12):2310–2327.

Behrens, J., van Elzakker, C. P. J. M., and Schmidt, M. (2015). Testing the usability of openstreetmap's id tool. *The Cartographic Journal*, 52(2):177–184.

Estima, J. and Painho, M. (2013). Exploratory analysis of openstreetmap for land use classification. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, GEOCROWD '13, pages 39–46, New York, NY, USA. ACM.

Haklay, M. (2013). Neogeography and the delusion of democratisation. *Environment and Planning A*, 45(1):55–69.

Jilani, M., Corcoran, P., and Bertolotto, M. (2014). Automated highway tag assessment of openstreetmap road networks. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '14, pages 449–452, New York, NY, USA. ACM.

Kunze, C. and Hecht, R. (2015). Semantic enrichment of building data with volunteered geographic information to improve mappings of dwelling units and population. *Computers, Environment and Urban Systems*, 53:4 – 18. Special Issue on Volunteered Geographic Information.

Mooney, P. and Corcoran, P. (2012). Characteristics of Heavily Edited Objects in OpenStreetMap. *Future Internet*, 4(1):285–305.

Schmidt, S., Manschitz, S., Rensing, C., and Steinmetz, R. (2013). Extraction of address data from unstructured text using free knowledge resources. In *Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies*, i-Know '13, pages 7:1–7:8, New York, NY, USA. ACM.

Scioscia, F., Binetti, M., Ruta, M., Ieva, S., and Sciascio, E. D. (2014). A framework and a tool for semantic annotation of {POIs} in openstreetmap. *Procedia - Social and Behavioral Sciences*, 111:1092 – 1101. Transportation: Can we do more with less resources? 16th Meeting of the Euro Working Group on Transportation Porto 2013.