An analysis of tagging practices and patterns in urban areas in OpenStreetMap

Nikola Davidovic
Faculty of Electronic Engineering
University of Nis
Serbia
nikola.davidovic@elfak.ni.ac.rs

Peter Mooney
Computer Science Department
Maynooth University
Ireland
Peter.mooney@nuim.ie

Leonid Stoimenov
Faculty of Electronic Engineering
University of Nis,
Serbia
Leonid.stoimenov@elfak.ni.ac.rs

Abstract

The OpenStreetMap Wiki website provides guidance and advice on how to map in OpenStreetMap (OSM). In particular the 'Map Features' page in the Wiki has become the de-facto guidebook and ontology for applying attributes to geographical objects in OSM. In this paper we outline research carried out in investigating if OSM contributors in 30 selected urban areas are using these guidelines in their mapping and tagging practices. We find that while there is broadly good compliance with the 'Map Features' guidelines there is under-utilisation of tags and this leads to an inhomogeneous set of tags being applied to similar geographical objects in different urban areas. The results of this research could be useful in providing better assistance to contributors in selecting a set of tags to apply to specific geographical objects when they are contributing to OSM.

Introduction and Motivation

One of the most popular contemporary implementations of the Voluntary Geographic Information (VGI) paradigm is OpenStreetMap (OSM). OSM is a very popular project worldwide and is based on a large mapping community. The number of mapped objects and the level of details with which these objects are represented depends on a number of factors including: the number of mappers, types of objects, use of bulk imports of data, etc. How well objects are mapped is also strongly linked to best practices that community adopts.

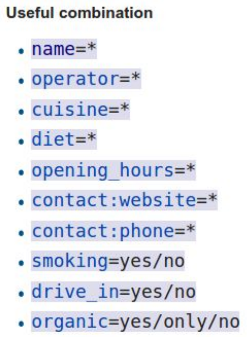
In OSM the attributes of features are represented using tags. Tags are attached to a map feature's basic data structures (nodes, ways, and relations). While the OSM Wiki (the Map Features page predominantly) offers guidance on best practice for tagging OSM's free tagging system does not impose any rules or limitations on the number or the content of the used tags. However it is expected that community, in time, agrees on the set of tags that should be used for a particular type of objects. But achieving such an agreement on worldwide scale or even a national scale is difficult and potentially unworkable.

Tagging in OpenStreetMap has been subject to a good deal of research outlined in the literature. In [X -our paper][8] we analysed the number of tags and changes in tags through editing and found that there is no easily observable distinct pattern to the application or tags to objects. Zielstra et al [7] find that local knowledge of contributors allows the collection and editing of detailed features such as trails, street furniture and attribute (tagging) information that can only be accessed locally. Gröchenig et al [3] analyse OpenStreetMap tagging over several years and find that tagging heavily depends on a number of distinct influences such as geographical or legal borders, data imports, unexpected events or diverse community developments. Barron and Zipf [2] and Keßler and de Groot [5] indicate that an important indicator of trust of OSM data and data completeness would be ensuring a consistent set of tags are applied to different sets of objects. However as of yet there has been no systematic evaluation in the literature of how well used the guidance in the OSM Wiki actually is in practice.

The OSM Wiki can be seen as one of the available sources of tagging guidance. The wiki's 'Map Feature' pages provide guidelines on how particular tags should be applied and which types of objects they are best suited to. These wiki pages also suggest which tags should be used in conjunction with described tag in the 'Useful combination' section of the page. The benefits of applying these suggested useful combination of tags to all objects are numerous. Benefits include: having homogenous descriptions of the objects in OSM providing large-scale analyses possibilities, consistent usage in location-based service applications such as navigation or tourist applications, and more widespread use by the general audience etc.

Unfortunately empirical evidence suggests that the guidance on tagging in these 'Map Features' pages is not always followed by mappers. Moreover the degree to which the information in 'Map Features' pages is followed differs in different parts of the world. This brings us to the key research question in this paper. How well are the 'Useful combinations' suggestions for tagging adopted by communities in different urban areas around the world? We have selected 30 urban areas of the world with different population and area sizes as well as different socio-economic characteristics. From the TagInfo website, we have selected 30 tags (key-value pairs) and analysed how well suggested tags, found on their respective 'Map Feature' wiki pages in the 'Useful combination' section, are followed by the communities. The results of our research are communicated through 9 selected tags which are representative of OSM in urban areas and for which we find that are the best examples of inhomogeneous sets of used tags.

Figure 1: This screenshot from the Map Features Wiki shows the 'Useful Combinations' suggestions for tags and keys to be used with the selected tag amenity=restaurant



Useful combination

- name=*
- operator=*
- cuisine=*
- diet=*
- opening_hours=*
- contact:website=*
- contact:phone=*
- smoking=yes/no
- drive_in=yes/no
- organic=yes/only/no

Source:http://wiki.openstreetmap.org/wiki/Tag:amenity%3Drestaurant

The paper is organized in the following way: After introduction, we are stating main implementation characteristics of our data mining approach for this research, our data sources and generated reports. Section Experimental results is providing rules for tags selection, rating approach we have used and main results of this paper. In section Future Work we are stating direction we will follow in the future OSM tag research that is greatly motivated by these results.

Implementation Details

Research question that we have imposed for this research requires developing a way to extract objects with given tag and then to count how many times their suggested tags appear. Having the number of these appearances, it is possible to give an estimation how well suggested tags are represented in a dataset. In order to solve this problem, we have developed an algorithm for processing these data. The algorithm uses OSM data imported into PostGIS database and then executes proper queries against it. For this research we were using osm2pgsql tool for importing OSM data. In order to import all possible tags, we used default.style file for importing main tags and hstore support for other tags (not listed in default.style file). When importing with hstore support, all tags that do not have their dedicated column are imported into hstore column as a comma separated list of keys and their values. Obtained data are then programmatically processed to extract this information. In order to be able to analyse set of selected urban areas our algorithm is processing them in a batch. First, one urban area OSM dataset is imported using osm2pgsql in slim mode. The OSM data (which is a bounding rectangle around the city extent in OSM) for urban areas that were used in this research, was downloaded from the Mapzen service (https://mapzen.com/data/metro-extracts/) which provides OSM data in several GIS formats for most urban areas around the world. Mapzen updates the OSM data downloads every few days. After data import has finished, we run our script that makes reports (Table 1) for each analysed tag and their respective suggested tags.

Table 1: Report for highway=footway in Frankfurt.

Report for tag: highway = footway		
Total number of records: 53325		
Tag	No. of occurrences	Percentage
name	3257	6.11%
access	2284	4.28%
footway	3185	5.97%
lit	2856	5.36%
surface	10882	20.41%
wheelchair	217	0.41%

After all selected urban areas have been processed and for each of them per-tag reports have been generated, we proceed with scripts that analyse and merge this data further. These reports analyses were going into two main directions. One was estimating how much suggested tag is really used in practice and the other was statistical analysis on the number of different tags that could be found along with the analysed tag.

Experimental Analysis

In order to guide our analysis we consulted the TagInfo application which lists the frequency with which all tags (and their key-value) pairs are used in OSM on a global scale. In order to have good representation of global data, we choose 30 urban areas from different parts of the world with different population, area sizes, socio-economic characteristics and OSM communities. The urban areas selected are as follows: Bangkok, Beijing, Boston, Bucharest, Buenos Aires, Dublin, Düsseldorf, Frankfurt, Helsinki, Johannesburg, Kyoto, London, Lyon, Madrid, Manchester, Mexico City, Milan, Nairobi, New Delhi, Nis, Oslo, Ottawa, Prague, Saint Petersburg, San Francisco, Singapore, Sydney, Vienna, Vilnius and Warsaw. Cities were selected in order to properly cover different parts of the world. However, it can be noted that European cities are somewhat better represented in this cities selection since we expected them to be mapped with greater detail [6].

We choose the 30 most frequently occurring tags as listed by TagInfo for OSM in January 2016. Obviously, not all listed tags were suitable for this research. Some of the most occurring tags do not have 'Useful combination' section on their pages, or are just keys with value 'yes'. The rules we used in order to select these 30 tags are as follows:

1. Tag has dedicated Map Feature Wiki page (URL is ending with Tag:key=value)
2. Tag has at least two suggested tags in the "Useful combination" section of the Wiki page.
3. Tag value are not "yes" since such tags do not have dedicated pages and their suggested tags correspond to the tag key and not the key-value combination.
4. Tag is not listed as suggested tag for any of the previously selected tags for the research (i.e. "service" tags for "highway" tags).
5. Only suggested tags from wiki pages without suggested values are selected for the research.

We run our analysis on all 30 selected tags and then examined results. Based on initial findings, we decided to focus on a smaller subset of popular tags from TagInfo as some of the 30 most frequently occurring will not appear frequently in urban areas such as landuse=meadow or waterway=ditch. Tags that we selected are given in Table 2. For each of these 9 selected tags we created a simple lookup table of the "Useful Combinations" of other tags suggested in the OSM wiki page for each of these 9 tags. The extracted suggested tags are also shown in the Table 2.

Table 2: Suggested tags for the 9 selected tags.

Tag key-value	Suggested Tags
highway=residential	name, oneway
natural=tree	height, circumference, start_date, leaf_type, genus, species, taxon, denotation
highway=footway	name, access, footway, lit, surface, wheelchair
highway=path	access, surface, sac_scale, mtb:scale, width, smoothness, trail_visibility
amenity=parking	access, capacity, fee, name, maxstay, operator
highway=primary	name, ref, lanes

highway=bus_stop	public_transport, name, operator
railway=rail	name, gauge, electrified, frequency, voltage, usage, service, bridge, tunnel
leisure=pitch	sport, surface

Our analysis tool calculated the number of times each of the suggested tags appear on an object along with the particular tag (one of the 9 selected tags), for all 30 cities. We calculated the relative percentage of times that each of the suggested tags were used for all suggested tag objects. So for example if there were 1,000 objects with highway=residential and we found that the oneway tag (see Table 2) was also present with highway=residential on 600 of these objects the relative percentage was 60%. To analyse the results we used a Likert scale ranking [1,4] to measure the compliance of object tagging to the suggested tags in the OSM Wiki as shown in Table 2. A suggested tag shows POOR compliance if 0 – 20% of objects use it with the selected tag, FAIR compliance if 21-40% of objects use it, AVERAGE if 41-60% of objects use it, GOOD if 61-80% of objects use it and EXCELLENT if > 80% of objects use it. As means of an example consider the tag highway=bus_stop (Table 3) in two cities namely London and Warsaw. There are three suggested tags: public_transport, name and operator. In London there are 45,611 OSM objects with the highway=bus_stop tag. There are 3.6% of these objects with public_transport (POOR), 96% of these objects with name (EXCELLENT) and < 1% with operator (POOR). In Warsaw there are 5,868 OSM objects with the highway=bus_stop tag. 47% (AVERAGE) have the public_transport tag, 95% (EXCELLENT) have the name tag while 29% (FAIR) have the operator tag. From previous example we can see that in both cities name tag is used on majority of bus stop objects. But besides having the name tag, bus stops are generally not well described in London. If we take public_transport tag for example, we can see that (using tagInfo) it is used to additionally describe features of the bus stop. Most common values are platform, stop_position and stop_area but can also be info_board, dispatcher etc. Therefore, this tag can add valuable information to users and therefore it would be beneficial if it was used more frequently. On the other hand, operator tag, as per tagInfo application, is used to denote one operator that is using that bus stop. But what happens when multiple operators are using the same bus stop and is operator tag omitted by users on purpose in those cases?

Table 3: Comparison of Likert scale ranked tags compliance for highway= bus_stop in London and Warsaw

highway= bus_stop

Key	London	Warsaw
public_transport	Poor	Average
name	Excellent	Excellent
operator	Poor	Fair

Based on per city and per tag reports we create cumulative reports per tag but for all 30 cities we were using in this research. Results are shown in tables 4 through 12.

Table 4: Results of applied Likert scale ranking for highway=residential in all 30 cities

highway=primary

Key	Poor	Fair	Average	Good	Excellent
lanes	8	7	3	5	7
ref	4	7	5	1	13
name	0	0	4	10	16

Table 5: Results of applied Likert scale ranking for amenity=parking in all 30 cities

amenity=parking

Key	Poor	Fair	Average	Good	Excellent
fee	27	3	0	0	0
capacity	30	0	0	0	0
name	27	2	1	0	0
access	17	9	4	0	0
maxstay	30	0	0	0	0
operator	30	0	0	0	0

Table 6: Results of applied Likert scale ranking for highway=bus_stop in all 30 cities

highway=bus_stop

Key	Poor	Fair	Average	Good	Excellent
operator	21	3	1	3	2
public_transport	15	5	3	3	4
name	0	3	3	5	19

Table 7: Results of applied Likert scale ranking for highway=residential in all 30 cities

highway=residential

Key	Poor	Fair	Average	Good	Excellent
name	5	1	4	5	15
oneway	25	4	1	0	0

Table 8: Results of applied Likert scale ranking for highway=path in all 30 cities

highway=path

Key	Poor	Fair	Average	Good	Excellent
sac_scale	30	0	0	0	0
mtb:scale	30	0	0	0	0
access	30	0	0	0	0
width	30	0	0	0	0
surface	12	12	6	0	0
trail_visibility	29	1	0	0	0
smoothness	30	0	0	0	0

Table 9: Results of applied Likert scale ranking for railway=rail in all 30 cities

railway=rail

Key	Poor	Fair	Average	Good	Excellent
bridge	19	11	0	0	0
name	16	6	5	3	0
service	5	14	9	2	0
tunnel	30	0	0	0	0
electrified	3	6	6	4	11
frequency	12	4	5	8	1
gauge	4	3	3	4	16
voltage	12	4	5	8	1
usage	11	7	6	6	0

Table 10: Results of applied Likert scale ranking for leisure=pitch in all 30 cities

leisure=pitch

Key	Poor	Fair	Average	Good	Excellent
sport	0	1	6	10	13
surface	30	0	0	0	0

Table 11: Results of applied Likert scale ranking for natural=tree in all 30 cities

natural=tree

Key	Poor	Fair	Average	Good	Excellent
circumference	28	0	1	0	1
taxon	28	0	0	0	2
leaf_type	24	2	2	1	1
start_date	29	0	0	1	0
height	27	0	1	0	2
denotation	26	1	2	0	1
genus	28	1	1	0	0
species	25	1	2	0	2

Table 12: Results of applied Likert scale ranking for highway=footway in all 30 cities

highway=footway

Key	Poor	Fair	Average	Good	Excellent
lit	29	1	0	0	0
name	30	0	0	0	0
footway	28	2	0	0	0
wheelchair	30	0	0	0	0
surface	24	5	1	0	0
access	30	0	0	0	0

From these tables we can draw following conclusions:

1. Suggested tags from wiki pages are generally not very well applied by the mapping community. That does not generally mean that tagging is poor since we were able to detect large number of different tags that are used across datasets. It means that there is no common approach to tagging specific objects.
2. Out of all 9 selected tags, the ones that are tagged in accordance with tag suggestions are related to vehicle navigation and are either motorways or streets (highway=primary or highway=residential). These types of objects are very popular amongst contributors and often subject to automated data imports. This indicates that third party data source imports do influence tagging behaviour.
3. Name tags are usually very well applied except for amenity=parking. We feel that this is due to the fact that there may be confusion around what name (if any) should be applied to a car park.
4. Surprisingly, we find that railway=rail is mapped quite well given that it requires detailed domain knowledge from mappers. While some of this data may have come from the imports there are active contributors in the rail transportation theme.
5. Amenity=parking is a tag which can be very useful for car drivers and we were expecting that suggested tags are better applied. Tags like capacity and maxstay are very important when planning navigation but these tags are poorly applied across all cities.
6. Almost all suggested tags of the tag highway=footway have poor or fair compliance. Mapping these objects most likely requires mappers to have on-the-ground knowledge.

From this work there are a number of general conclusions which can be summarised as follows:

1. Objects which can be mapped easily using popular OSM editing software with access to aerial imagery have generally good compliance. These are high level objects like streets, roads and motorways.
2. Compliance (based on our scale) could be influenced when there are bulk imports of 3rd party data to OSM.
3. Tags whose values can be deduced from areal images are generally mapped better.
4. Tags that require verifiability on the field usually have poor compliance.

Future Work

There are a number of very interesting directions for future work related to the research outlined in the paper. As we have indicated in the results section above there is divergence in tagging practices observed in our case-study datasets from the suggested tags in the Map Features documentation. This needs to be addressed urgently because we feel that it is contributing to both the confusion around tagging for new contributors to OSM and also adding to the inhomogeneous picture we observe in tagging patterns for similar objects in different urban areas. Our immediate future work will develop a data mining technique to extract patterns of tagging in OSM where groups or co-occurrences of tags are used frequently by contributors. These patterns or collections of tags may very well diverge from the advice and direction outlined in Map Features but yet are being used extensively by OSM contributors. We shall also investigate the types of influence OSM editing software has on tagging patterns and practice. This will involve analyzing the tagging suggestions which the editor software produce when a contributor creates or edits a specific type of geographical objects in OSM. Integrating a more standardized approach in these software to the selection of a base set of tags for specific types of geographical objects may see more homogeneous usage of tagging between different urban areas. The influence of the import of 3rd party data into OSM must also be investigated in terms of how tagging is affected.

References

[1] Allen, E and Seaman, C . Likert Scales and Data Analyses. Quality Progress 2007, 64-65. (2007)

[2] Barron, C., Neis, P. & Zipf, A. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. Transactions in GIS n/a–n/a (2014).

[3] Gröchenig, S., Brunauer, R. & Rehrl, K. Digging into the history of VGI data-sets: results from a worldwide study on OpenStreetMap mapping activity. Journal of Location Based Services 0, 1–13 (2014).

[4] Harpe, S. E. How to analyze Likert and other rating scale data. Currents in Pharmacy Teaching and Learning 7, 836–850 (2015).

[5] Keßler C., de Groot R T A. Trust as a proxy measure for the quality of Volunteered Geographic Information in the case of OpenStreetMap. In Groot R T A de, Bucher B, and Crompvoets J (eds) Proceedings of the Sixteenth AGILE Conference on Geographic Information Science. Berlin, Springer Lecture Notes inGeoinformation and Cartography: 21–37 (2013)

[6] Neis, P. & Zielstra, D. Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. Future Internet 6, 76–106 (2014).

[7] Zielstra, D., Hochmair, H. H., Neis, P. & Tonini, F. Areal Delineation of Home Regions from Contribution and Editing Patterns in OpenStreetMap. ISPRS International Journal of Geo-Information 3, 1211–1233 (2014).

[8] Mooney, P. & Corcoran, P. Characteristics of Heavily Edited Objects in OpenStreetMap. Future Internet 4, 285–305 (2012)