Quality analysis of the Parisian OSM toponyms evolution

Vyron Antoniou, Guillaume Touya, Ana-Maria Olteanu-Raimond

CHAPTER 8

Quality analysis of the Parisian OSM toponyms evolution

Vyron Antoniou*, Guillaume Touya† and Ana-Maria Raimond†

*Hellenic Military Academy, Greece, v.antoniou@ucl.ac.uk
†Laboratoire COGIT, Institut National de l'Information Géographique et Forestière, 73 Avenue de Paris, 94165 Saint-Mandé, France

Abstract

The paper presents empirical research on the quality of the toponyms that can be retrieved from OpenStreetMap (OSM) under the purpose of enriching authoritative toponymic databases and gazetteers. An analysis on the volatility of places and points-of-interest (POIs) is presented. We examine how named features behave and change in terms of type, name and location. The challenge is to understand the behavior and consequently the fitness-for-purpose of OSM data when it comes to a possible use and integration with authoritative datasets. We show that, depending on the OSM feature type, the volatility can vary considerably and we elucidate which feature types are consistent, and thus could be used in authoritative gazetteers despite their grassroots nature and if there are spatial patterns behind the location changes of features during their lifespan.

Keywords

Toponyms, geographic names, OpenStreetMap, Gazetteer, VGI, data quality.

How to cite this book chapter:
Antoniou, V, Touya, G and Raimond, A-M. 2016. Quality analysis of the Parisian OSM toponyms evolution. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*, Pp. 97–112. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.h.

Introduction

Gazetteers are a vital component of any spatial database irrespectively of the level of detail used (i.e. local, national or international). Gazetteers consist of a list of toponyms, a type and their corresponding geography. This geography can be either a point, a bounding box or the footprint of the place. Gazetteers are usually used as an entrance point to a spatial database. People start exploring geographic data by providing a toponym and search for features, relations, maps or events related to that toponym. In other cases, the outcome of a spatial search is accompanied by toponyms that facilitate the understanding of the result. There are also cases where the result of a search is the toponym itself (e.g. reverse-geocoding). Apart from these practical examples, toponyms and gazetteers play a key role in many aspects of everyday life. Examples can be found in explicit geographic applications like routing, mapping and cartography but also in more general cases such as in government, legislation, security and policing etc. (UN, 2006).

However, National Mapping Agencies (NMAs), which are the de facto agencies responsible for creating and updating gazetteers in a national level, are facing difficulties in keeping toponymic databases up to date due to the lack of resources and due to the extensive field work needed for data collection and verification. On the other hand, Volunteered Geographic Information - VGI (Goodchild 2007) can serve as a promising alternative mechanism for collecting toponyms that could enrich and update official gazetteers (Goodchild & Hill 2008). In this context, the aim of this paper is to examine whether OSM can provide consistent toponymic datasets or the grassroots mechanisms alter constantly the spatial features in such a level that hinder their use in gazetteers. More specifically, the research tries to provide empirical evidence on the following questions: i) What is the population and the types of OSM features that have names and can be used as part of a gazetteer? ii) What kind of changes are taking place for these features? iii) What feature types are affected and how much? iv) Are there any underlying spatial patterns for these changes? This study adopts the definition about toponyms that is provided by the United Nations Group of Experts on Geographical Names (UNGEN), and thus the scope of interest includes populated places, civil divisions, natural features, constructed features and unbounded places or areas that have specific local meaning (UN 2006: 9).

OSM urges its contributors to provide names for spatial objects, if applicable, using the *name* key tag. Contributors can add more than one name for spatial features such as international names or old names by using variations of the *name* key such as *int_name* or *old_name*. Moreover, OSM wiki pages provide detailed guidelines on how to correctly assign a name to spatial objects in order to achieve maximum standardization. In our study only the *name* key tag has been examined of the point-based objects of two broad OSM categories: Places[1]

[1] http://wiki.openstreetmap.org/wiki/Places

and Points of Interest (POIs)[2]. These categories are in accordance with what United Nations (UN) define as a toponym. The remainder of the paper is structured as follows: Section 2 presents briefly a selection of related work on the subject. Section 3 discusses the methodology used to collect and analyze OSM data. Section 4 presents the results followed by discussion and future work in Section 5.

Related Work

The importance of gazetteers and the challenges posed by the nature of VGI data, and especially of toponyms, has drawn the interest of many researchers and there is extensive literature available. Here, we provide few examples of VGI and authoritative data integration efforts so to highlight that VGI quality and stability is an important factor for this task. Such efforts range from creating a gazetteer by harvesting volunteered big geo-data from Web sources (see for example Gao et al. 2014) to combining both administrative and VGI toponyms. For example, Twaroch et al. (2008) use various web sources to create a surface model of the toponyms' footprints. However, the authors highlight the fact that it is difficult to have crisp boundaries when it comes to VGI data and that there is a need to identify outliers. Similarly, Keßler et al. (2009a) proposed the enrichment of authoritative gazetteers with toponyms extracted from geotags of photos. As the authors support, their approach could benefit from quality indicators of the geotags used. The quality of user-contributed data has been also highlighted as a crucial factor in empirical research with geo-tagged photos (see for example Hollenstein & Purves 2010). Regarding OSM, Hahmann and Burghardt (2010) proposed to link OSM with GeoNames gazetteer using semantic web techniques to produce an enriched, multi-lingual gazetteer and Smart et al. (2010) proposed a methodology for the conflation of toponymic data from multiple sources, including both authoritative and VGI datasets, and taking into account the quality differences of each source. However, as Mooney and Corcoran (2012) explain, developers of location-based services should be cautious when it comes to using OSM data as their research on frequently edited features revealed considerable volatility in the naming process. Moreover, Keßler et al. (2009b) underline the importance that gazetteers should cater both for local and small-scale features, as well as timely and user-centric information. In this context, OSM has the potential to become a valuable source of toponyms. Thus the discussion focuses on the nature, the behavior and the evolution of the toponymic datasets that can be retrieved from OSM and how these factors affect quality elements and their use in gazetteers.

[2] http://wiki.openstreetmap.org/wiki/Points_of_interest

OSM Data Extraction

Extending this line of research, this paper focuses on the volatility of OSM features. It goes beyond the naming changes that Mooney and Corcoran (2012) focused and examines also the location changes of OSM features. The area of scope of the research is Paris region (12.012 km^2). The study area is large enough to have a great diversity of named features, and is quite complete due to the large number of OSM contributors. In order to collect the necessary data, the Geofabrick[3] shapefile download service was used. The datasets for the area of scope were downloaded at the first week of December 2014. Shapefiles include as an attribute the unique OSM_ID of every OSM feature. These IDs were used in combination with the OSM API to collect and store in a PostgreSQL/Postgis database all the versions of each feature. This method provided a complete timeline of the OSM edits made in the area of scope.

Analysis

Descriptive statistics

A preliminary analysis on the availability of names for the spatial features (grouped by OSM category) was conducted and the results are shown at Table 1.

It can be seen that, depending on the category, there are considerable variations in the presence of names. For example, in the '*Places*' category, OSM contributors have assigned a name at almost all (i.e. except from 3) features. Arguably, this behavior meets the expectations of an OSM user (including the

Category	Total	With names	%
Land use	36,347	2,201	6.1%
Natural	20,138	2,093	10.4%
Places	4,275	4,272	99.9%
Points	192,228	53,052	27.6%
Railways	16,482	4,471	27.1%
Roads	344,870	152,595	44.2%
Waterways	5,190	2,520	48.6%
Total	**619,530**	**221,204**	**35.7%**

Table 1: Total OSM features and OSM features with names for the study area.

[3] www.geofabrik.de

author of a gazetteer) as it is generally expected that all point features classified as '*Places*' should have a name. In contrast, this cannot be observed at the '*Roads*' category. Although, in reality, roads have a name (especially in urban areas like Paris) or a reference name (e.g. *link to Motorway X*) the percentage of named features is barely 44.2%. Another category that is of interest for a gazetteer is the one of 'Points'. This category includes a variety of local features that OSM contributors deem as Points of Interest (POIs). Here the percentage of named features is just 27.6% but, as it will be explained later, this factor is not indicative of the completeness of the dataset in terms of names as for many POIs' subcategories a name tag is not applicable (such as for 'crossing' or 'bench'). Given these results the research focused into two categories that were deemed as the most interesting when it comes to examining the potential to create or enrich a gazetteer: OSM Places and OSM POIs.

OSM Places

In terms of changes in type, location and name, a Place point can either remain stable in its entire life-cycle or undergo a change in one or any combination of these three factors. In an effort to understand whether the OSM data can serve as a source of consistent toponyms, the percentage of the features that have been changed or remained stable has been recorded (Figure 1).

It can be seen that two thirds of the features have never been changed while the most common change that features undergo is in their geographic location. In this context, the next issue of interest was to examine the types of places and their corresponding population versus the location movements that took place for each '*Place*' type. This classification was used so to examine which types of features, have been moved by OSM contributors. Again, this can give an overview of the consistency of OSM Places. The findings are shown in Table 2.

The findings show that, depending on the type of place, there is considerable variation in terms of location change. For example, while only 8% of the features belonging to the '*locality*' type has been moved, for the features that belong to the '*town*' type this reaches 80%. Following this observation, the next step was to examine the magnitude of location change (calculated in meters) for each type of place. The magnitude of location change is considered as the distance between two points: i) the centroid calculated taking into account all positions of the feature during its life and ii) the last position of the features. The results are shown in Figure 2.

This type of analysis can visualize the volatility in location change of various place types. It can be observed that entities with large spatial extends (either crisp of fuzzy) suffer from large changes in their location in contrast with smaller entities. For example, almost 14% of all '*towns*' have moved over 1,000 m whereas for '*suburbs*' 21% of the features have moved less than 100 m and 65% remained stable (see also Table 2).

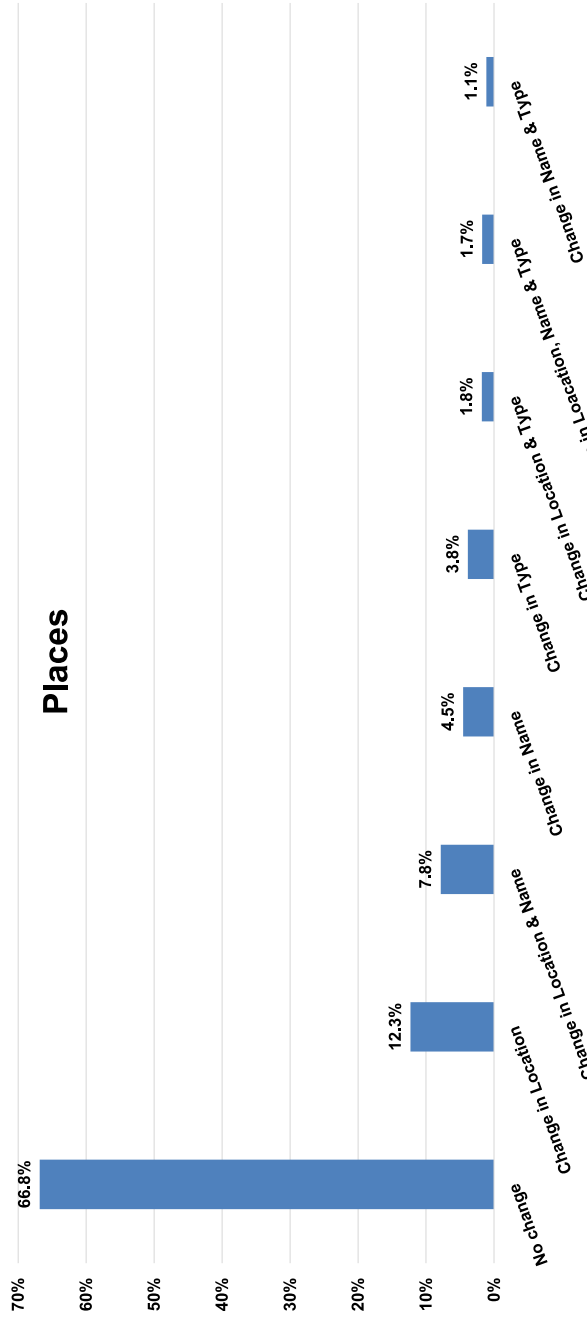


Figure 1: Changes in Places taking into count three factors (location, name and type).

Type	All features	Features moved	%
Allotments	2	2	100%
City	1	1	100%
Hamlet	496	86	17%
Island	9	1	11%
Islet	1	0	0%
Isolated_dwelling	37	3	8%
Locality	2,096	167	8%
Neighbourhood	214	20	9%
State	1	1	100%
Suburb	130	45	35%
Town	248	198	80%
Village	1,039	487	47%
Yes	1	0	0%
Total	**4,275**	**1,011**	**24%**

Table 2: Types of OSM places and number of OSM places that have been geographically moved.

OSM POIs

As noted above, from almost 200K of POIs only 27.6% of them (i.e. 53,052) had a *name* attribute. This is an expected observation as there are types of POIs where the *name* is not an applicable attribute. For example, the POI types of *crossing*, *bench*, *traffic_signals* and *survey_points* have in total 75,819 spatial features that account approximately to the 40% of the total population, and less than 0.04% of them (i.e. only 27 features) have names.

Similar to the Places' analysis, the changes of the same elements (i.e. of location, type and name) have been examined also for the POIs. The findings show that about 60% of features have not been changed since their creation while the most common change this time is the change in their name.

In order to examine which POI types are the most volatile in terms of name and location change, a scatter-plot (Figure 4) has been created. The x-axis in Figure 4, shows the percentage of features that had a change in name for the 30 most populous OSM types. Name changes range from minor changes (e.g. alterations in capital letters or blank spaces) up to changes in the entire name. Although it is not clear which OSM feature types should be included in a gazetteer (see also discussion in Section 5), Figure 4 shows that there are types of POIs that have a large rate of name changes and others that remain relatively

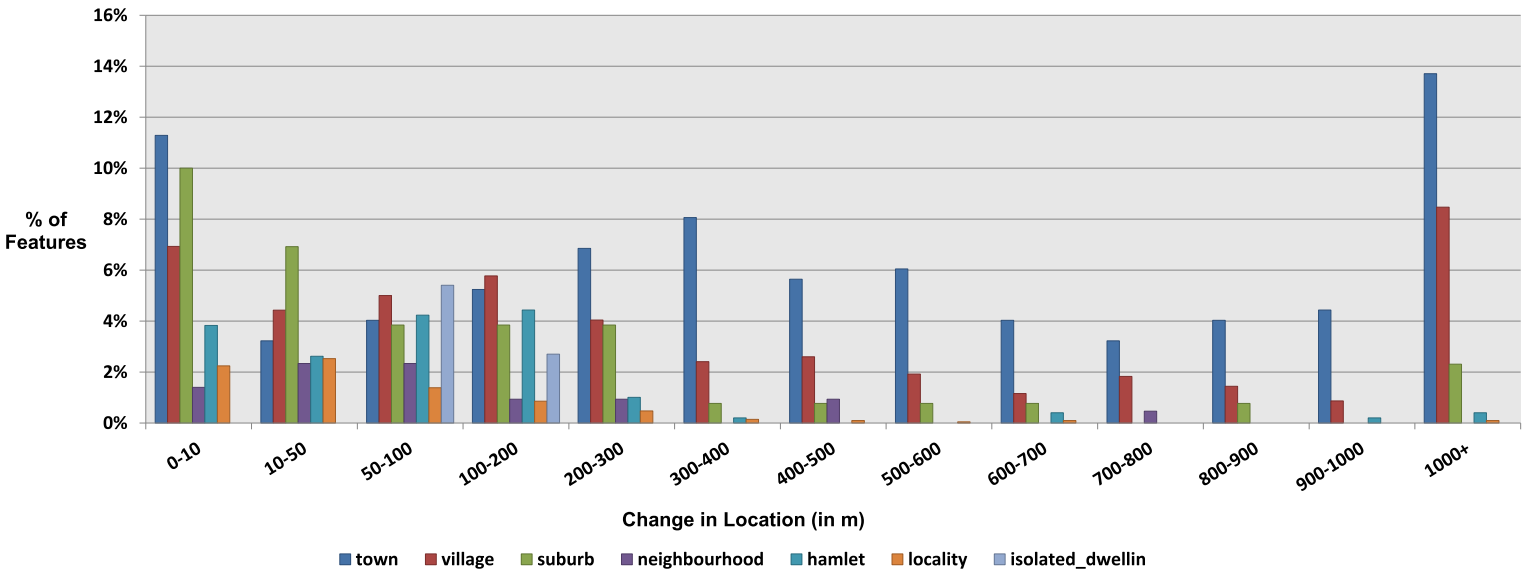


Figure 2: The magnitude of location change (x-axis, in m) per type of OSM 'Place'.

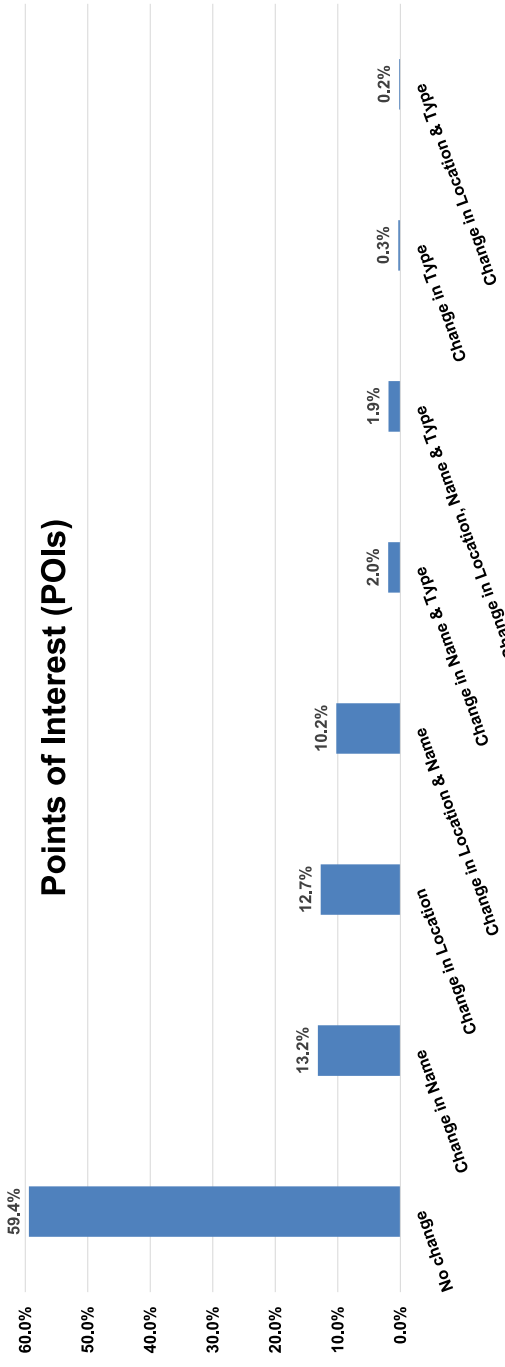


Figure 3: Changes in POIs taking into count three factors (location, name and type).

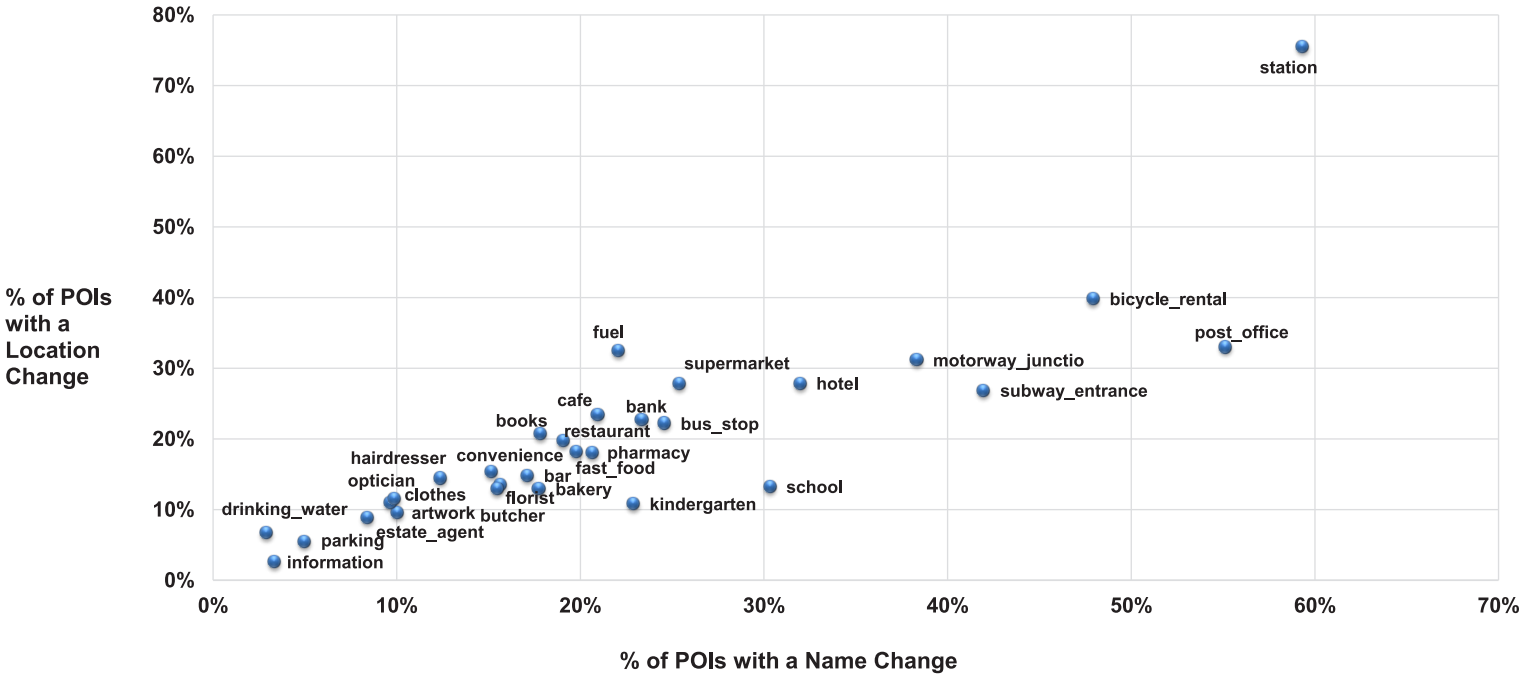
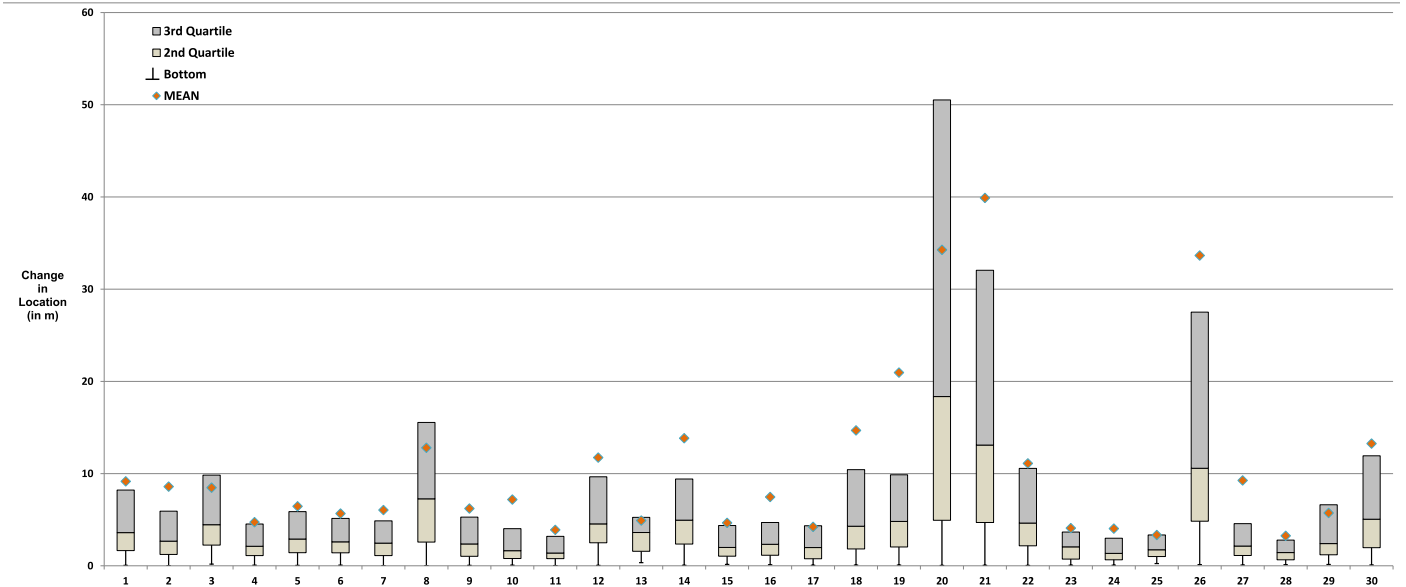


Figure 4: Percentage of POIs that had a name (x-axis) and a location (y-axis) change.



1. bus_stop	7. café	13. information	19. drinking_water	25. butcher
2.restaurant	8. school	14. bicycle_rental	**20. motorway_junction**	**26. fuel**
3. parking	9. fast_food	15. convenience	**21.station**	27. florist
4. bakery	10. clothes	16. bar	22. post_office	28. optician
5.bank	11. hairdresser	17. subway_entrance	23. artwork	29. books
6.Pharmacy	12. supermarket	18. hotel	24. estate_agent	30. kindergarten

Figure 5: Box-Whisker graph of the location change for each POI type.

stable (e.g. lower left corner). The y-axis in Figure 4, shows the % of features that had a location change for each OSM type. Here, again, it can be observed that not all POI types behave the same; certain POI types suffer more than others.

The combined view of changes over these two factors indicates which of these types (if they are to be included in a gazetteer) might be considered as too unstable to populate a gazetteer. However the counter argument can be that the seemingly stable behavior of certain POI types can be explained by poor contributors' attention and thus this stability might indicate obsolete or out-of-date features. In any case, Figure 4 raises awareness of the futures' behavior and gives a better insight on what kind of volatility should be expected per POI type.

After gaining a better understanding on which POI types are volatile in terms of name and location change, the next step was to quantify the latter. In order to better visualize the position change, a Box-Whisker plot has been created in Figure 5 (note that the upper quartile is not marked as outliers in many types make it draw out of scale – for ease of understanding the mean value has been added).

First, this type of analysis can help to understand which spatial features should not be modeled as POIs since the simple geometry of a point appears not to be the best way to model this physical entity. For example, motorway junctions seem not to gather consensus among OSM contributors regarding the position of the POI as the average location change is more than 30 m. On the contrary, there are POI types that despite their location change, the distance between various locations remains well under 10 m (i.e. an arbitrary positional accuracy threshold of hand-held GPS devices). Second, this type of analysis can highlight gross errors and outliers in OSM datasets that might downgrade the overall spatial quality of a dataset. The largest the distance between the mean and the upper level of the second Quartile box (i.e. the 50% of the features), the more outliers and gross positional errors exist in each category. For example, 8% of the features for the *fuel* category have been moved more than 100 m (with a recorded maximum movement of 659 m). Finally, it is made clear that for many POI types a clearer feature extraction guide is needed. For example, when capturing schools or station it needs to be clear for contributors where the point should be positioned: at the entrance of the building, at the centroid of the main building or somewhere else. Let us mention that although there are instructions in the OSM wiki pages how to map each feature, apparently these instructions are not explicit enough and thus inconsistencies occur.

Spatial patterns

Finally, for the entire dataset of POIs a hot-spot analysis was calculated based on the location change for each feature. A visualization based on Z-score is shown in Figure 6. Hot-spot analysis (using the *Getis-Ord Gi* statistic provided in ESRI ArcGIS 10.2.2) can reveal whether a phenomenon is random or not.

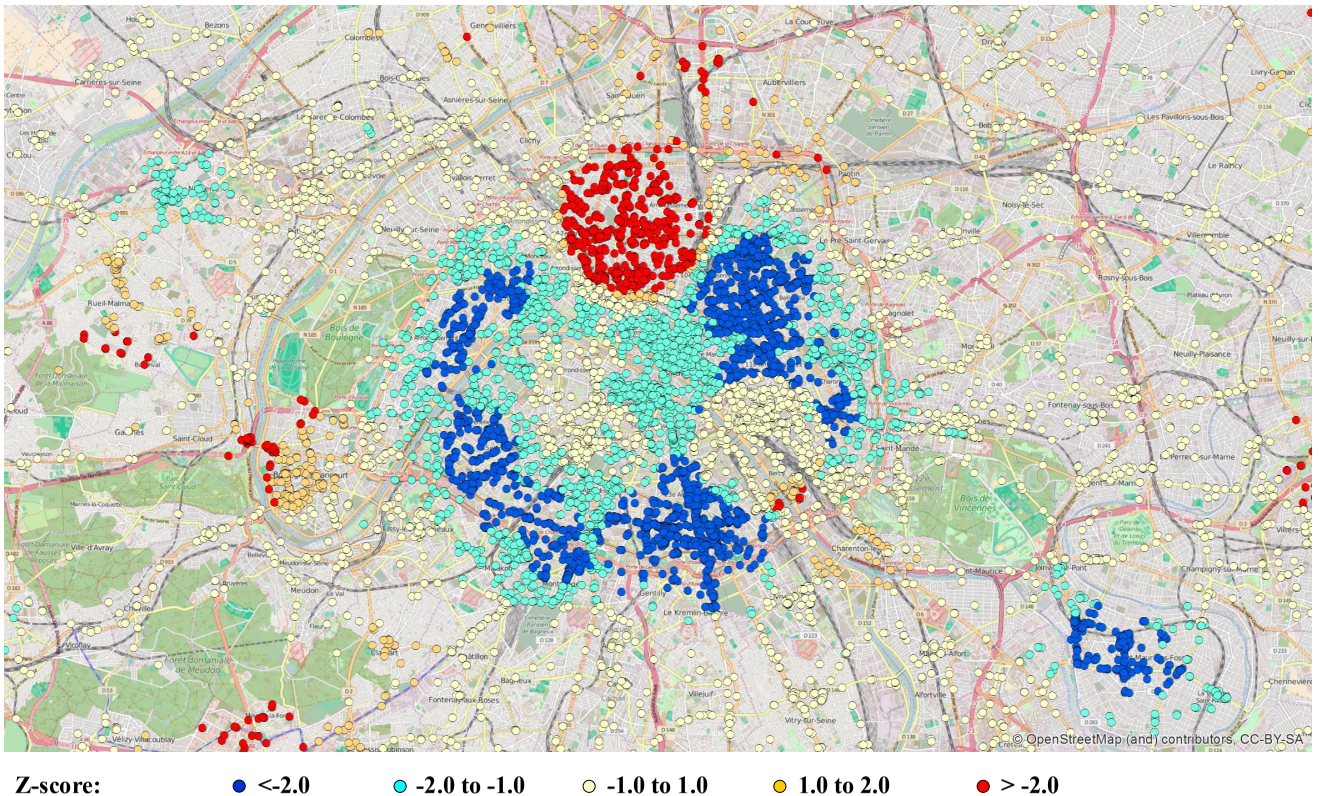


Figure 6: Hot Spot Analysis (using the Getis-Ord Gi* statistic) on the location change of POIs.

Here it can be observed that there are concentrated hot colors areas (i.e. areas with not random large movements) and cold colors areas (i.e. areas with not random small movements) for POIs. While a first observation can be made that hot colors areas appear in the popular and touristic area of Montmartre (north of map) and the cold colors areas appear at the periphery of Paris (more residential than touristic), further analysis is needed to fully understand the causes of the phenomenon. For instance, Montmartre is a hill, and the sources, like the ortho-rectified satellite imagery, used for the positioning of the POIs may be less accurate there.

Discussion and future work

VGI datasets are a dynamic source of spatial information. In particular, OSM datasets, which usually function as a proxy in the research on VGI data, have drawn the interest of researchers regarding their use in helping NMAs to complete or update existing geospatial products or even to create new ones (Antoniou 2011). Improved and enriched authoritative products can be toponymic databases and gazetteers. The importance of gazetteers in acquiring accurate results in spatial searches is paramount and thus the update of official gazetteers with local knowledge should be made with caution and meticulous examination of the VGI quality. Unnecessary changes in the names, types or the geographic position (no matter how subtle or small) can introduce problems to authoritative products or location based services. However, once successful, the presence of local and community-level named features and landmarks can considerably enrich and improve gazetteers and geospatial services. A first point of consideration is the decision on which types of user-contributed features should be used in a gazetteer. For example, certain types of POIs are possible to serve as landmarks that can help to provide eloquent and easily understandable routing directions. Although this paper does not delve into the subject of feature type importance, it provides evidence that the selection of OSM types and features should be examined from a quality point of view as well.

What this paper has examined is the behavior and thus the fitness-for-purpose of OSM data as a source of toponymic data. The aim was to examine whether the OSM datasets are a consistent datasets or the grassroots mechanisms alter constantly the datasets in such a level that in practice hinder the use of OSM data. The findings show that VGI and authoritative data conflation is not a straightforward process as they differ considerably in nature. While authoritative toponyms are largely static and hard to change spatial entities, a considerable percentage of VGI toponyms undergo changes. Not all OSM types are fit to support the enhancement of administrative gazetteers as the OSM specification and contribution practices might generate an unwanted volatility in the data. This observation generates a number of questions that could be the aim of future work. First, it is important to understand the nature of these changes. For example, do

changes in location serve a better mapping outcome, refer to previous mistakes and thus are a spatial quality improvements or are they simply real-life movements that OSM contributors capture? Relating movement to the geographic extent of the named feature, or to some contributor pattern would be useful to understand how and why the changes occur. Using what Goodchild and Li (2008) call the geographic approach to assess named features movement would also be useful: e.g. check whether a Place feature that refers to a town has been move to the centroid of the town hall. Similarly, it could be examined if there are any time patterns in the changes. For example are these changes concentrated at the early period of the creation of a feature and thus it is an indicator of quality improvement (as discussed in Haklay et at. 2010) or are they happening during the entire life-cycle of each feature and indicate an endogenous volatility of the spatial feature? Nevertheless, contributors might alter OSM features (no matter what the reason) and this change can either be very small and thus authoritative products and services that have integrated OSM data will not be affected or might be large enough to introduce unwanted volatility. Finally, it is of interest to compare, in terms of completeness, the OSM toponyms with authoritative data so to understand at what extend VGI data can help NMAs to improve their gazetteers. Thus, future work will include the comparison between OSM and authoritative toponyms (provided by IGN France, the French NMA).

Acknowledgments

This research took place during a two week Short Term Scientific Mission (STSM), funded by COST TD1202 ENERTGIC Action, in IGN with the help of Dr Bénédicte Bucher. Finally, we are grateful to the reviewers for their helpful comments on the original paper.

References

Antoniou, V. 2011. *User generated spatial content: an analysis of the phenomenon and its challenges for mapping agencies*. Thesis (PhD), University College London.

Gao, S., Li, L., Li, W., Janowicz, K., & Zhang, Y. 2014 (In Press). Constructing gazetteers from volunteered big geo-data based on Hadoop. *Computers, Environment and Urban Systems*. DOI: http://dx.doi.org/10.1016/j.compenvurbsys.2014.02.004

Goodchild, F. M. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4): 211–221. DOI: http://dx.doi.org/10.1007/s10708-007-9111-y

Goodchild, F. M., & Hill, L. 2008. Introduction to digital gazetteer research. *International Journal of Geographical Information Science, 22*(10): 1039–1044. DOI: http://dx.doi.org/10.1080/13658810701850497

Goodchild, M. F., & Li, L. 2012. Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*: 110–120. DOI: http://dx.doi.org/10.1016/j.spasta.2012.03.002

Hahmann, S., & Burghardt, D. 2010. Connecting Linked GeoData and geonames in the spatial semantic web. In *Proceedings of extended abstracts, 6th International GIScience Conference,* Zurich, Switzerland.

Haklay, M., Basiouka, S., Antoniou, V., & Ather, A. 2010. How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. *The Cartographic Journal*, *47*(4): 315–322. DOI: http://dx.doi.org/10.1179/000870410X12911304958827

Hollenstein, L., & Purves, R. 2015. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, (1): 21–48.

Keßler, C., Maué, P., Heuer, J. T., & Bartoschek, T. 2009a. Bottom-up gazetteers: Learning from the implicit semantics of geotags. *GeoSpatial semantics*, *5892*: 83–102. DOI: http://dx.doi.org/10.1007/978-3-642-10436-7_6

Keßler, C., Janowicz, K., & Bishr, M. 2009b. An agenda for the next generation gazetteer: Geographic information contribution and retrieval. In: *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems*, pp. 91–100.

Mooney, P., & Corcoran, P. 2012. Characteristics of heavily edited objects in OpenStreetMap. *Future Internet*, *4*(1): 285–305. DOI: http://dx.doi.org/10.3390/fi4010285

Smart, P. D., Jones, C. B., & Twaroch, A. F. 2010. Multi-source toponym data integration and mediation for a meta-gazetteer service. In *Geographic Information Science, 6292*: 234–248. DOI: http://dx.doi.org/10.1007/978-3-642-15300-6_17

Twaroch, A., F., Jones, B., C., &. Abdelmoty A. 2008. Acquisition of a vernacular gazetteer from web sources. In: *Proceedings of the first international workshop on Location and the web.* ACM: pp. 61–64.

United Nations. 2006. *Manual for the National Standardization of Geographical Names: United Nations Group of Experts on Geographical Names*. Department of Economic and Social Affairs, New York, USA: UN.