Proceedings of the Academic Track

Editors:

Marco Minghini
Christina Ludwig
Jennings Anderson
Peter Mooney
A. Yair Grinberger

DOI: 10.5281/zenodo.5116434

Contents

OpenStreetMap as a multi-faceted research subject: the Academic Track at State of the Map 2021

A. Yair Grinberger[1,*], Jennings Anderson[2], Peter Mooney[3], Christina Ludwig[4] and Marco Minghini[5,†]

[1] Department of Geography, The Hebrew University of Jerusalem, Israel; yair.grinberger@mail.huji.ac.il
[2] YetiGeoLabs, Montana, USA; jennings.anderson@gmail.com
[3] Department of Computer Science, Maynooth University, Ireland; peter.mooney@mu.ie
[4] GIScience Research Group, Institute of Geography, Heidelberg University, Germany; christina.ludwig@uni-heidelberg.de
[5] European Commission, Joint Research Centre (JRC), Ispra, Italy; marco.minghini@ec.europa.eu

* Author to whom correspondence should be addressed.
† The views expressed are purely those of the author and may not in any circumstances be regarded as stating an official position of the European Commission.
This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

The OpenStreetMap (OSM) project has come a long way since its establishment in 2004, celebrating the uploading of its 100 millionth changeset on 25 February 2021 [1]. From a project led by a small group of mapping enthusiasts, OSM has grown into a comprehensive global geographic database produced by a global community of contributors, attracting attention from governments, NGOs, and recently, also global tech giants [2–4]. Along with this increasing interest from organizations, the academic community had also kindled an interest in OSM, with the first scientific paper in English focusing on OSM (i.e. having OpenStreetMap in its title) published as early as 2007, according to Google Scholar [5]. Since then, the study of OSM has grown into a distinct research stream and community, with 119 publications focusing on OSM in 2020 alone (again, per Google Scholar) [6].

Yet, OSM is a unique study object for the fields that frequently interact with it (mainly geo-information, computer science, geography and engineering [7]). It is not an abstract notion, distant from the researcher, but rather a constantly evolving multifaceted entity, defined not only by the data it contains but also by its users and the communities they form, in which researchers can take an active role. Accordingly, there is no one way to engage with OSM in research—one may use its data for other applications, study the quality of OSM data, the dynamics of data production or the behavior of contributors, and even produce tools and approaches for enriching the data and supporting data production [7]. Furthermore, OSM research may include direct and indirect interactions with the ever-growing OSM community, starting from engagement with data and ending with becoming an active member of the OSM Foundation and affiliated organizations, such as the Humanitarian OpenStreetMap Team (HOT). The Academic Track at the State of the Map conference is perhaps the most

Grinberger, A.Y., Anderson, J., Mooney, P., Ludwig, C. & Minghini, M. (2021). OpenStreetMap as a multi-faceted research subject: the academic track at State of the Map 2021
In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 1-5. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5111623

explicit example of such interactions. In the first three iterations of the Track (2018-2020), authors gave 40 talks, presented 18 posters, and enjoyed the opportunity to both receive direct feedback on their research from the larger OSM community, and interact with mappers. The nine studies included in the 4th edition of the Track, published in these proceedings, not only represent the variety of topics studied through OSM research, but also the full plethora of ways in which research and researchers choose to engage with OSM.

Two studies discuss use cases of OSM data. Will [8] provides a specific example, considering the development of a web-based interface for the purposes of allowing the specification of custom machine-readable language (MRL) queries which are executed on the Overpass service. These queries can be presented to the system as natural language questions such as "Which restaurants in Vienna are wheelchair-accessible?", hence increasing the accessibility of OSM data. Graux and Michel [9] adopt a broader perspective, investigating the usage and involvement of OSM in Horizon 2020 (H2020) projects funded by the European Union during the course of the H2020 programme. They analyse the presence of OSM and other geographic services in over 92,000 deliverables produced by over 8,000 projects and report that OSM has around 18,600 mentions across all deliverables. This suggests that the impact of OSM, which is traditionally difficult to measure, is substantial as the authors indicate that the projects involving OSM were funded with an overall budget of almost 4 billion euros of public money. In addition, such findings can inform the FAIR (Findable, Accessible, Interoperable and Reusable) use of results, which is a requirement in H2020, for projects that have to do with OSM.

Other studies engage with OSM by investigating aspects of data quality. Juhász [10] proposes a new framework for evaluating the temporal accuracy of OSM with regards to the time it takes for mappers to incorporate real-world changes into the map. Through a case study using publicly available data from the Florida Department of Transportation (FDOT), Juhász is able to validate the approach by limiting the comparison to only highway features and successfully identifies the elapsed time between the known construction date and the update in OSM. The success of applying the framework to highway features gives hope to future development involving more authoritative data sources. Li & Anderson [11] propose a new method to detect vandalism in OSM through the concept of embeddings, e.g. an embedding contains users who have edited the same OSM feature. These embeddings are fed into a Gradient Boosting Decision Tree (GBDT) model which is trained using manually labeled OSM vandalism corpus for the name attribute of OSM features. Finally, Elias et al. [12] describe a plugin for the open source QGIS software offering spatio-temporal analysis of OSM data quality. The developed QGIS plugin allows for visualisation of the results of positional and thematic accuracy of OSM. In particular, the authors draw attention to the relevance of identifying aspects of quality and heterogeneity in OSM contributions for national mapping purposes.

These latter two studies show engagement that goes beyond inspecting OSM data: Elias et al. [12] focus is on providing direct utility to OSM users, editors, and researchers through the development of a new tool; Li & Anderson's model [11] is also used for an analysis distinguishing between different communities of editors by calculating the cosine similarity between the user embeddings, detecting new communities of non-corporate editors alongside the known communities from different corporations such as Apple or Facebook. Li and Anderson's work is joined by other user-centric studies, i.e. studies that focus on the dynamics of data production and editor behaviors. Veselovsky et al. [13] also

relate to the issue of corporate editors (CEs) employed by corporations such as Facebook, Apple, or Amazon, proposing a method to distinguish volunteer mappers from corporate mappers based on their editing behaviour in order to better analyse how the corporate editing activity which has emerged in the past years has influenced OSM. They use different machine learning algorithms and a new method for identifying the local time zone of each editor so that it can be compared to a “corporate editing signature” (marked by 8 hour workday and no activity on the weekend) to identify between 700 and 2000 additional corporate mappers. Sarkar & Anderson [14] offer an additional study of CEs, describing how they have influenced the editing interaction patterns between mappers in OSM from 2015 to 2019 using network analyses to represent editors as nodes and editing interactions between them as edges. The authors find that, despite CEs most often editing other CEs' work, there is still ample interaction between CEs and non-CEs with both of them editing each other’s work. Finally, Owusu et al. [15] develop a classification schema to assess local data in OSM and its “fitness-for-purpose” for local OSM communities. Broken into two larger categories “core” and “specific”, the schema contains four levels that increase in specificity from mapping buildings, roads, or other easy-to-trace objects from satellite imagery objects to specific, localized knowledge. Utilizing the Ohsome API, the authors categorize OSM editing into this schema over time and look for distinct patterns of evolution between the levels as time progresses. Quantitatively investigating these patterns across three different regions validates the schema and the framework’s ability to characterize the evolution of OSM in each region with respect to when highly localized information was added to the map.

The issue of research providing utility for OSM, evident in [12] as discussed above, is also considered by Mooney and Galván [16], who address the question of what machine learning has ever done for OSM through a comprehensive literature review of approximately 50 research papers involving machine learning and OSM. The authors acknowledge that as a massive open geographic dataset, OSM is a tempting data source for machine learning researchers to investigate. By categorizing the types of machine learning papers involving OSM from quality-improvements to simple training datasets, the authors make clear the wide-range of OSM-related machine learning research. The authors warn researchers not to use machine learning simply for the sake of it as a popular new research technology, but to critically assess cases where machine learning provides a clear advantage to answering the research question.

Together, these 9 abstracts highlight three major themes in OSM research: data quality [10–12], the identity of contributors and the nature of their contributions [11, 13–15], and the use of new approaches and techniques such as machine learning within OSM data use and production [8, 16]. These themes, especially the latter two, are also the subject of intense and sometimes heated discussions within the OSM mapping community where questions and doubts regarding the utility and effects of CEs and machine-produced data are raised [4]. This stresses the importance of the interactions between the mapping and the research community, where the latter takes inspiration from the former while the former potentially enjoys the insights provided by the latter. Looking into the future and based on the experience from previous editions, we, the Academic Track Scientific Committee (see Figure 1) are certain that the Academic Track will continue to provide a welcomed stage supporting the development of these symbiotic relations. We wish to end this editorial by expressing our thanks to the authors who submitted their work to the Track and participated in it, to the State of the Map Working Group in charge of the organization, to all the

volunteers supporting the conference, and to the members of the OSM mapping and research communities who continuously work to improve OSM.

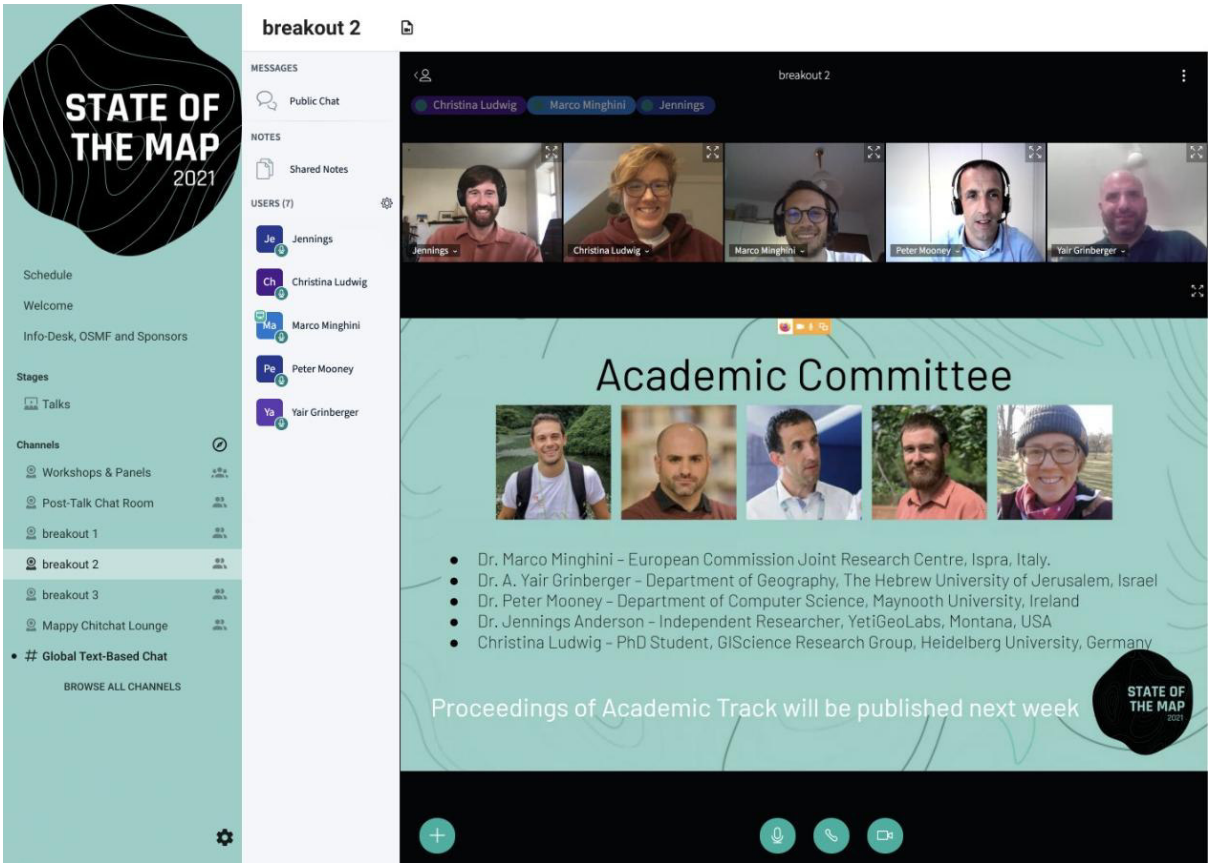

Figure 1. The Academic Track Scientific Committee during the online State of the Map 2021 conference.

References

[1] OpenStreetMap Contributors (2021). History of OpenStreetMap. Retrieved from https://wiki.openstreetmap.org/wiki/History_of_OpenStreetMap

[2] Arsanjani, J. J., Zipf, A., Mooney, P., & Helbich, M. (2015). An introduction to OpenStreetMap in geographic information science: Experiences, research and applications. In: Arsanjani, J. J., Zipf, A., Mooney, P., & Helbich, M. (Eds.) *OpenStreetMap in GIScience*, 1-15. Springer, Cham.

[3] Mooney, P., & Minghini, M. (2017). A review of OpenStreetMap data. In: Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C. C., & Antoniou, V. (Eds.) *Mapping and the Citizen Sensor*, 37-59. Ubiquity Press, London.

[4] Anderson, J., Sarkar, D., & Palen, L. (2019). Corporate editors in the evolving landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information,* 8(5), 232.

[5] Google Scholar search. Retrieved from https://scholar.google.com/scholar?q=allintitle%3A+openstreetmap&hl=en&as_sdt=0%2C5&as_ylo=2008&as_yhi=2004

[6] Google Scholar search. Retrieved from https://scholar.google.com/scholar?q=allintitle%3A+openstreetmap&hl=en&as_sdt=0%2C5&as_ylo=2020&as_yhi=2020

[7] Grinberger, A. Y., Minghini, M., Juhász, L., Mooney, P., & Yeboah, G. (2019). Bridging the map? Exploring interactions between the academic and mapping communities in OpenStreetMap. In: Minghini, M., Grinberger, A. Y., Juhász, L., Yeboah, G., & Mooney, P. (Eds.) *Proceedings of the Academic Track at the State of the Map 2019*, 1-2.

[8] Will, S. (2021). NLMaps Web: A Natural Language Interface to OpenStreetMap. In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., & Grinberger, A.Y. (Eds.) *Proceedings of the Academic Track at the State of the Map 2021 Online Conference*, 13-15.
[9] Graux, D., & Michel, T. (2021). Involvement of OpenStreetMap in European H2020 projects. In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., & Grinberger, A.Y. (Eds.) *Proceedings of the Academic Track at the State of the Map 2021 Online Conference*, 34-36.
[10] Juhász, L. (2021). Towards understanding the temporal accuracy of OpenStreetMap: A quantitative experiment. In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., & Grinberger, A.Y. (Eds.) *Proceedings of the Academic Track at the State of the Map 2021 Online Conference*, 19-22.
[11] Li, Y., & Anderson, J. (2021). Introducing OpenStreetMap user embeddings: Promising steps toward automated vandalism and community detection. In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., & Grinberger, A.Y. (Eds.) *Proceedings of the Academic Track at the State of the Map 2021 Online Conference*, 23-26.
[12] Elias, E., Amorim, F., Silva, L., Schmidt, M., Camboim, S., & Fernandes, V. (2021). A proposal for a QGIS plug-in for spatio-temporal analysis of OSM data quality: The case for the city of Salvador, Brazil. In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., & Grinberger, A.Y. (Eds.) *Proceedings of the Academic Track at the State of the Map 2021 Online Conference*, 27-30.
[13] Veselovsky, V., Sarkar, D., Anderson, J., & Soden, R. (2021). An automated approach to identifying corporate editing activity in OpenStreetMap. In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., & Grinberger, A.Y. (Eds.) *Proceedings of the Academic Track at the State of the Map 2021 Online Conference*, 31-33.
[14] Sarkar, D., & Anderson, J. (2021). Community interactions in OSM editing. In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., & Grinberger, A.Y. (Eds.) *Proceedings of the Academic Track at the State of the Map 2021 Online Conference*, 6-8.
[15] Owusu, M., Herfort, B., & Lautenbach, S. (2021). Towards a framework for measuring local data contribution in OpenStreetMap. In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., & Grinberger, A.Y. (Eds.) *Proceedings of the Academic Track at the State of the Map 2021 Online Conference*, 16-18.
[16] Mooney, P., & Galván, E. (2021). What has machine learning ever done for us? In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., & Grinberger, A.Y. (Eds.) *Proceedings of the Academic Track at the State of the Map 2021 Online Conference*, 9-12.

Community interactions in OSM editing

Dipto Sarkar[1,*] and Jennings Anderson[2]

[1] Department of Geography and Environmental Studies, Carleton University, Ottawa, Canada; diptosarkar@cunet.carleton.ca
[2] YetiGeoLabs, Montana, USA; jennings.anderson@gmail.com

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

OpenStreetMap (OSM) data is produced by a vibrant community of mappers. These geospatial data produsers (a portmanteau of "producer" and "user" commonly used when describing members of peer-production communities) represent a plethora of individuals with different motivations, methods of contribution, and usage [1, 2]. Thus, the 1.6M OSM contributors have been aptly described as a community of communities [3]. In recent years, corporate editing teams have introduced a new dynamic to the discussion of communities in OSM; editing teams hired by corporations, such as, Apple, Facebook, Microsoft, Uber, are capable of contributing thousands of changesets each day, outpacing the average volunteer contributor [4, 5]. Additionally, corporate editors (CEs) tend to focus their editing on particular types of map features. These two attributes of corporate editing can lead to CEs breaking off into a siloed group of their own with little or no interaction with the rest of the editors on the map.

Previous research on the OSM community using network analysis methods showed there was limited collaboration between editors with most objects being edited only a few times [6]. Senior editors in particular perform a majority of the mapping work on their own, but do indeed interact with others through co-editing—where subsequent mappers edit the same objects [7]. Since these studies were performed, the OSM community has grown significantly and the community dynamics have significantly evolved with more individual and organized participation, such as corporate editing.

Here, we use a data driven approach to characterize the interactions between the CEs and the rest of the OSM community. We define interactions through editing patterns, constructing a network of interactions where each node represents an editor, and two nodes are connected by an edge if the two mappers have edited the same map object. If the mapper represented by node A edits an object last edited by the mapper represented by node B, then the edge connecting these nodes is directed from A to B. Thus, the node's degree is the number of co-editors of the node with the in-degree representing the mappers who edited objects after the user and the out-degree representing the number of mappers editing objects before the user. The network can have multiple disconnected components as not everyone co-edits with everyone else in the network. We utilized the OSM-Interactions tilesets to construct these networks [8]. OSM-Interactions vector tiles contain the editing history of all highway and building objects at zoom level 14. They include minor changes to

Sarkar, D. & Anderson, J. (2021). Community interactions in OSM editing
In: Minghini, M., Ludwing, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 6-8. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112211

the geometry of objects in which only nodes are moved, but the parent way is left untouched. In this way, we are capturing the complete history of map objects in OSM, as opposed to just changes to the basic OSM elements (primarily nodes or ways).

In keeping with the objects which are primarily edited by CEs, our analysis focuses only on highway and building objects for construction of the network. The nodes in the network are further annotated with a binary category representing whether they are a CE or not. We classify a mapper as being a CE or not by comparing usernames in the network to the disclosed lists of usernames on a corporation's OSM wiki or Github page.

We focus on 4 locations: Egypt, Jamaica, Thailand, and Singapore. We create networks for each of these locations at 3 timepoints, 2015, 2017, and 2019 to characterize the changes between over time. Thus, we constructed and analyzed 12 networks. The locations were chosen as they all have different groups of CEs active.

Across all networks, the Largest Connected Component (LC) accounted for 94% of all nodes highlighting significant interactions amongst all mappers. The LC is the largest group of connected nodes, meaning there exists a path between any two users via other co-editors. The more prolifit a mapper edits, the more likely they are to be a part of the LC. Within the LC, the rate of growth of CE nodes exceeds the rate of growth of non-CE nodes at a rate of 11:1 between 2015 and 2019. However, both types of editors (CE and non-CE) have a comparable number of in and out degrees in most networks, indicating that they edit other people's work and have their work edited at a similar rate. Figure 1 shows specific characteristics of each network:

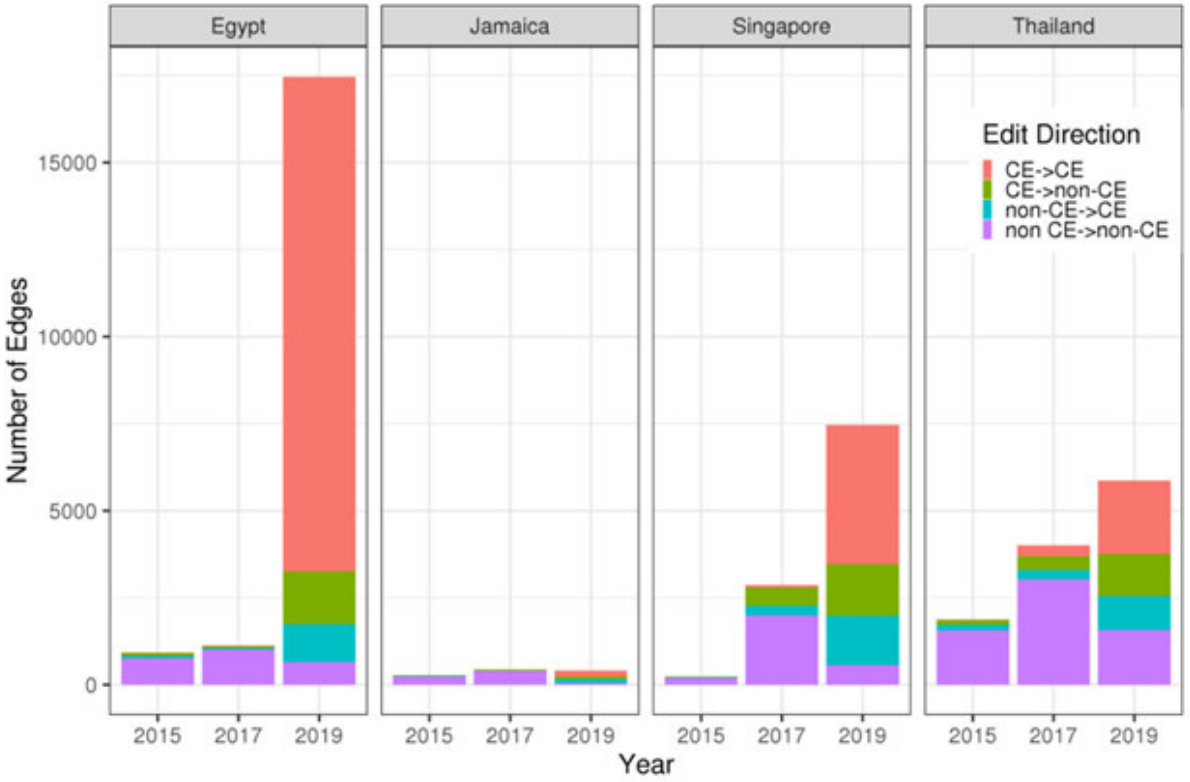

Figure 1. Number of edges in each network each year. Edges are colored by the type of source and target nodes. Orange represents corporate editors co-editing each other, green represents corporate editors editing the work of non-corporate editors, blue represents the work of non-corporate editors editing corporate editors, and purple represents volunteer co-editing activity.

In terms of who edits whose work, CE's edit other CE's work most often, but interaction between CEs and non-CEs have also grown through time, keeping the network connected (Figure 1). The massive growth in the orange edges in Figure 1 in 2019 represents many CEs editing each other. The increase in the green bars shows that corporate editors are also editing (at a much lesser rate) the existing work on the map—however, the blue area shows that volunteers are also editing corporate-sponsored work at a similarly increasing rate.

With regards to age of the mappers (calculated in terms of their enrollment date in OSM) and the volume of edits they perform, younger mappers in both groups tend to edit others' work at a higher rate than senior mappers, but there is more variation in these statistics for non-CE mappers. This is a finding contrary to previous research on editing interaction patterns mentioned above. Additionally, characterizing the time between edits shows that edits made by CE's persist for a slightly shorter duration than edits made by non-CE, primarily due to other CEs editing the same object soon after.

In conclusion, the editing networks highlight the vibrancy of data co-production. The volunteer editor and CEs are interacting with each other's edits to produce the map. The per-group interaction is nuanced and shows unique editing patterns which warrant further investigation. During the timespan of this study, the rate of growth of the CE community was faster than the non-CE community, but whether the pattern will hold over time and whether other locations exhibit the same pattern will require further research.

References

[1] Budhathoki, N. R., & Haythornthwaite, C. (2013). Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap. *American Behavioral Scientist,* 57(5), 548-575.

[2] Coleman, D. D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered geographic information: The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research,* 4(1), 332-358.

[3] Solís, P., Anderson, J., and Rajagopalan, S. (2020). Open Geospatial Tools for Humanitarian Data Creation, Analysis, and Learning through the Global Lens of YouthMappers. *Journal of Geographical Systems*.

[4] Anderson, J., Sarkar, D., & Palen, L. (2019). Corporate editors in the evolving landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 8(5), 232.

[5] Anderson, J., & Sarkar, D. (2020). Curious Cases of Corporations in OpenStreetMap. In: Minghini, M., Coetzee, S., Juhász, L., Yeboah, G., Mooney, P., & Grinberger, A. Y. (Eds.). *Proceedings of the Academic Track at the State of the Map 2020 Online Conference*, 13-14.

[6] Mooney, P., & Corcoran, P. (2012). How social is OpenStreetMap? In: *Proceedings of the 15th AGILE International Conference on Geographic Information Science*, 282-287.

[7] Mooney, P., & Corcoran, P. (2014). Analysis of interaction and co-editing patterns amongst openstreetmap contributors. *Transactions in GIS*, 18(5), 633-659.

[8] Anderson, J (2020). Analyzing OpenStreetMap Contributions at Scale: Introducing osm-interactions Tilesets. In: Abramowicz, W., & Klein, G. (Eds.) *Business Information Systems Workshops, BIS 2020 International Workshops Lecture Notes in Business Information Processing*, 394, 267-271.

What has machine learning ever done for us?

Peter Mooney[1,*] and Edgar Galvan[1]

[1] Naturally Inspired Computation Research Group, Department of Computer Science, Maynooth University, Maynooth, Ireland; peter.mooney@mu.ie, edgar.galvan@mu.ie

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

Recently, machine learning (ML) based approaches have been applied frequently to many different types of problems in OpenStreetMap (OSM). Indeed, ML approaches have been used extensively by the research community for a plethora of applications and problems both related and unrelated to OSM. Wagstaff suggests ML offers "a cornucopia of useful ways to approach problems which defy manual solutions" [1]. In specific relation to the geospatial domain, ML approaches have been reported for at least the last two decades with the remote sensing community being particularly active in ML usage. The number of works appearing around ML in the geospatial domain began to noticeably increase around a decade ago with work by authors such as Werder et al. [2] on interpretation of buildings in settlements and detecting road intersections from GPS traces by Fathi and Krumm [3]. Around this time interest in the combination of ML and OSM began to emerge. Funke et al. argued that many aspects of OSM data might be suitable for "extrapolation or classification using ML" [4]. Many examples have emerged with ML approaches being used to consider problems such as: predicting or recommending tagging for objects, object classification based on contextual or proximity information, tag usage checking, automated mapping approaches etc. Anderson et al. showed that Facebook's recent mapping campaign in OSM used ML to detect road networks from satellite imagery which are then validated by OSM editors and the local OSM communities [5]. Examples also exist where OSM is used in ML approaches for other geospatial classification problems while authors such as Feldmeyer et al. used machine and deep learning algorithms with OSM for developing socio-economic indicators [6]. Audebert et al. [7] argued that OSM's richness means it can be used in difficult problems such as semantic labeling of aerial and satellite images.

In addition to the observations by Vargas-Munoz et al. [8] in their recent review of ML approaches in OSM we can usually observe ML and OSM interaction in one of two ways: (1) ML approaches are used to improve the quality and coverage of OSM layers by using GIS and Remote Sensing and (2) instances where OSM layers are used as a means of training ML models for some specific task such as building segmentation, population estimation, navigation or land use classification. In this research, we ask the following question: With all of the many applications and integration of ML with OSM over the past number of years, how many of these applications and approaches have been adopted or used by the OSM community? Furthermore, what are the benefits or impact of these efforts from the research community with ML approaches to the OSM project and OSM community?

Mooney, P. & Galván, E. (2021). What has machine learning ever done for us?
In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 9-12. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112219

We performed a systematic review of approximately 50 peer-reviewed academic journal and conference papers. These papers are selected on the following basis: the paper(s): (1) clearly outlines an ML approach using OSM data and (2) tackle a problem known in the OSM community such as tag prediction, contribution patterns, or geometry correction. We used the Google Scholar search engine to retrieve the papers for review. Paper metadata such as title, keywords, and abstract contents were used to select the papers. We processed the results in linear fashion, as returned by the Google Scholar search, and selected the first 50 papers. This included a manual check of all the papers to ensure that the content of each paper related to our selection criteria. From the initial search results, five of the papers were either literature review papers or used OSM as just a background map layer in visualisations. These were replaced by the next five qualifying papers in the search results. The GitHub repository at [9] contains the links to all papers and our classifications. We acknowledge that this sampling is far from representative of the whole field. With greater resources (search and analysis time, research availability), a much larger sample of papers could be assembled. These papers were analysed using the following set of questions for guidance:

- What are the most common ML approaches used by researchers for the three instances outlined above?
- What are the most common types of problems in OSM tackled by ML approaches?
- Are the approaches reproducible and replicable by others?
- What, if any, is the awareness and/or understanding of the OSM project or community outlined in the paper?

Using these questions, we now report a narrative on our findings on the benefits and impacts of these efforts to the OSM project and the OSM community. We do not evaluate the results of the papers such as analysis of the accuracy of ML approaches nor do we advocate a specific ML approach. Both tasks are outside the scope of this work.

- In Vargas-Munoz et al. [8] one interaction of ML and OSM is to improve the quality and coverage of OSM. We found that 23 papers (46%) could be classified as displaying this type of approach. Furthermore, here we attempted to group papers into different types of coverage and quality tasks for OSM. We used our own classification and allowed multiple selections. From these 23 papers we found that: 3 papers dealt with contribution patterns in OSM, 9 papers considered data quality issues in general, 14 papers were connected to annotation and tagging in OSM while the remaining 10 papers dealt with topological issues and geometric analysis (correction, alignment, creation, and other tasks)
- Also in Vargas-Munoz et al. [8] the other interaction of ML and OSM is to use OSM data to train models and serve applications. We found that 31 papers (62%) used this approach. We believe that 2 papers shared both classifications. Here, as before, we attempted to group papers into different types of application domains. We created our own classification as follows: Navigation and transportation (7 papers), use of OSM to generate ML training datasets (6 papers), Socio-economic analysis (6 papers), image analysis (3 papers) and miscellaneous (5 papers). We found great diversity in these applications with examples taken from areas such as multilane road extraction, electric vehicle routing, training datasets for urban areas, population estimation, air quality forecasting and image labelling using OSM.

- OSM community understanding. Grinberger et al. [10] explored interactions between the academic and mapping communities in OSM. We have based our classifications on this work with 5 classes representing the understanding of OSM as outlined in the academic paper. Multiple selections were allowed for each paper. The five classes used are: OSM as a data source (25 papers, 50%), OSM as a data source produced by contributors (25 papers, 50%), OSM as a social data product (14 papers, 28%), or no understanding or perception of OSM specified (25 papers, 50%). We acknowledge that this particular aspect of the review is very subjective and reflects our personal interpretation. We also understand that authors may discuss their understanding of the OSM community in different ways with the constraints of the paper.
- There was great variation in the types of machine learning approaches used in these works. Random forests are by far the most popular but many other approaches are also implemented: ensemble ML, convolutional neural networks, logistic regression, supervised and unsupervised clustering, support vector machines, AdaBoost, Boosting, Latent Dirichlet Allocation and others.
- Finally, we considered the reproducibility and replicability of the studies in each of the papers. Would it be possible for other scientists in the academic community or members of the OSM community to reproduce or replicate the work in a given paper? We decided to classify reproducibility and replicability in three ways: red (appears very difficult to reproduce or replicate for reasons related to proprietary software usage, lack of description, etc); amber (appears that reproduction and replication is possible but access to certain APIs or datasets is required) and green (reproduction and replication appears to be a priority in the work). After our review, we found: red (8 papers, 16%), amber (29 papers, 58%) and green (13 papers, 26%). Grippa et al [11] and Zurbaran et al [12] are two excellent examples of papers with reproducibility and replicability as priority issues.

As stated previously, we acknowledge the subjective nature of these results and our Github repository will allow others to consider our classifications. There is incredible diversity, even within this small sample, of applications of ML with OSM. We believe that the ability to integrate OSM data with other data sources greatly adds to these opportunities. At this stage in this work, it is difficult to ascertain the level of adoption or indeed opportunities for adoption of published ML approaches by the OSM community. We were disappointed to see that in our evaluation 50% of the papers surveyed did not clearly indicate any understanding or connection to the OSM community. OSM provides incredible value to ML research. Indeed, we believe it is impossible to estimate the value of OSM as a data source to ML research and studies. In a similar way, and as shown above, many ML approaches do lend themselves very well as potential candidates for implementation by the OSM community. We conclude that it still remains difficult to understand how all of this ML could contribute effectively to the OSM database and OSM community without improved interactions between both communities.

We must not get carried away with the combination of ML and OSM purely for the sake of it. OSM, as a massive open geospatial database, is a very attractive source of (geo-)data for researchers and practitioners looking to train, benchmark and test ML approaches. Consequently, we can confidently state that, after well over a decade of reported results in this domain, researchers have produced many excellent research and knowledge outputs using the ML and OSM combination. Now is a good time to ask, as we

have done in this paper, how could all of this ML knowledge contribute effectively to the OSM database and OSM community. Grinberger et al. [10] argue that efforts to establish and strengthen interaction between the research community interested in working with or in OSM and the OSM community itself have generally been positive. However, opportunities exist to enhance interactions between these two communities and perhaps ML could be the catalyst for a new interaction. Based on this the scientific contribution of this work is multi-faceted. Firstly, this paper will stimulate debate about the contribution of these ML approaches to the improvement of OSM data and enhancement of the OSM community. Secondly, this work will highlight situations where these ML approaches have delivered genuinely new and novel outputs of interest to OSM in general. Finally, this work will issue the challenge to the academic community to apply ML to several interesting and open problems which are of mutual interest to both the academic and OSM community, returning in kind the considerable impact OSM has had on this academic field.

References

[1] Wagstaff, K. (2012). Machine learning that matters. *arXiv preprint*, arXiv:1206.4656.
[2] Werder, S., Kieler, B., & Sester, M. (2010). Semi-automatic interpretation of buildings and settlement areas in user-generated spatial data. In: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 330-339.
[3] Fathi, A., & Krumm, J. (2010). Detecting road intersections from GPS traces. In: *International conference on geographic information science*, 56-69. Berlin, Heidelberg: Springer.
[4] Funke, S., Schirrmeister, R., & Storandt, S. (2015). Automatic extrapolation of missing road network data in OpenStreetMap. In: *Proceedings of the 2nd International Conference on Mining Urban Data*, 1392, 27-35.
[5] Anderson, J., Sarkar, D., & Palen, L. (2019). Corporate editors in the evolving landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 8(5), 232.
[6] Feldmeyer, D., Meisch, C., Sauter, H., & Birkmann, J. (2020). Using OpenStreetMap Data and Machine Learning to Generate Socio-Economic Indicators. *ISPRS International Journal of Geo-Information*, 9(9), 498.
[7] Audebert, N., Le Saux, B., & Lefèvre, S. (2017). Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 67-75.
[8] Vargas-Munoz, J. E., Srivastava, S., Tuia, D., & Falcao, A. X. (2020). OpenStreetMap: Challenges and opportunities in machine learning and remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 9(1), 184-199.
[9] Mooney, P., & Galvan, E. (2021). What has machine learning ever done for us?. Retrieved from https://github.com/petermooney/sotm2021
[10] Grinberger, A. Y., Minghini, M., Juhász, L., Mooney, P., & Yeboah, G. (2019). Bridging the map? Exploring interactions between the academic and mapping communities in OpenStreetMap. In: Minghini, M., Grinberger, A. Y., Juhász, L., Yeboah, G., & Mooney, P. (Eds.) *Proceedings of the Academic Track at the State of the Map 2019*, 1-2.
[11] Grippa, T., Georganos, S., Zarougui, S., Bognounou, P., Diboulo, E., Forget, Y., Lennert, M., Vanhuysse, S., Mboga, N., & Wolff, E. (2018). Mapping Urban Land Use at Street Block Level Using OpenStreetMap, Remote Sensing Data, and Spatial Metrics. *ISPRS International Journal of Geo-Information*, 7(7), 246.
[12] Zurbaran, M. A., Wightman, P., & Brovelli, M. A. (2019). A machine learning pipeline articulating satellite imagery and OpenStreetMap for road detection. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-4/W14, 255-260.

NLMaps Web: A natural language interface to OpenStreetMap

Simon Will[1,*]

[1] Department of Computational Linguistics, Heidelberg University, Heidelberg, Germany;
simon.will@gorgor.de

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

Nominatim [1] and Overpass [2] are powerful ways of querying OSM, but the Overpass Query Language is somewhat impractical for quick queries for unfamiliar users. In order to query OSM using natural language (NL) queries such as “Show me where I can find drinking water within 500m of the Louvre in Paris”, Lawrence and Riezler [3] created the first NLMaps dataset mapping NL queries to a custom machine-readable language (MRL), which can then be used to retrieve the answer from OSM via a combination of queries to Nominatim and Overpass. They extended their dataset in a subsequent work by auto-generating synthetic queries from a table mapping NL terms to OSM tags – calling the combined dataset NLMaps v2. [4] The proposed purpose of these datasets is training a parser that can parse NL queries into their MRL representation, as done in [4–7].

The main aim of this work was to build a web-based NLMaps interface that can be used to issue queries and to view the result. In addition, the web interface should enable the user to give feedback on the returned, either by simply marking the parser-produced MRL query as correct or incorrect, or by explicitly correcting it with the help of a web form. This feedback should be directly used to improve the parser by training it in an asynchronous online learning procedure.

After observing that parsers trained on NLMaps v2 perform poorly on new queries, an investigation into the causes for this revealed several shortcomings in NLMaps v2, mainly: (1) Train and test split are extremely similar limiting the informativeness of evaluating on the test split. (2) Various inconsistencies exist mapping from NL terms to OSM tags (e.g. “forest” sometimes mapping to natural=wood, sometimes to landuse=forest). (3) The NL queries’ linguistic diversity is limited since most of them were generated with a very simple templating procedure, which leads to parsers trained on the data not being very robust to new wordings of a query. (4) In a similar vein, there is only a small amount of different area names in NLMaps v2 with the names “Paris”, “Heidelberg” and “Edinburgh” being so dominant that parsers are biased towards producing them. (5) Some generated NL queries are not a good representation of natural language, which makes them counter-productive learning examples. (6) Usage of OSM tags is sometimes incorrect, which affects the usefulness of produced parses.

The detailed analysis is used to eliminate some of the shortcomings – such as incorrect tag usage – from NLMaps v2. Additionally, a new approach of auto-generating

Will, S. (2021). NLMaps web: A natural language interface to OpenStreetMap
In: Minghini, M., Ludwing, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 13-15. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112227

NL-MRL pairs with probabilistic templates is used to create a dataset of synthetic queries that features a significantly higher linguistic diversity and a large set of different area names. The combination of the improved NLMaps v2 and the new synthetic queries is called NLMaps v3.

A character-based GRU encoder-decoder model with attention [8] is used for parsing NL queries into MRL queries using the configuration that performed best in previous work [7]. This model is trained on NLMaps v3 and used as the parser in the newly developed web interface. Mainly through advertising on the OSM talk list and the OSM subreddit, 12 annotators were hired from all over the world to use the web interface to issue new NL queries and to correct the parser-produced MRL query if it is incorrect. They were required to complete a tutorial before the annotation job and received help compiled from taginfo [9], TagFinder [10] and custom suggestions for difficult tag combinations. The collected dataset contains 3773 NL-MRL pairs and is called NLMaps v4.

With the help of NLMaps v4, an informative evaluation can be performed revealing that a parser trained on NLMaps v2 data achieves an exact match accuracy of 5.2% on the MRL queries of the test split of NLMaps v4 while a parser trained on NLMaps v3 performs significantly better with 28.9%. Pre-training on NLMaps v3 and fine-tuning on NLMaps v4 achieves an accuracy of 58.8%.

Since the goal of this work is to deliver an online learning system – i.e. a system that updates the parser directly after receiving feedback in the form of an NL-MRL pair –, various online learning simulations are conducted in order to find the best setup. In all cases, the parser is pre-trained on NLMaps v3 and then receives the NL-MRL pairs in NLMaps v4 one by one, updating the model after each step. The most simple variant of the experiment uses only the one NL-MRL pair for the update, another variant adds NL-MRL pairs from NLMaps v3 to the minibatch and a third variant additionally adds further “memorized” NL-MRL pairs from previously given feedback to the minibatch. The main findings of the simulation are that all variants improve performance on NLMaps v4 with respect to the pre-trained parser, but with some of them the performance on NLMaps v3 degrades. The simple variant that updates only on the one NL-MRL pair is particularly unstable, while adding NLMaps v3 instances stabilizes the performance on NLMaps v3 and improves the performance on NLMaps v4. Adding the instances from memorized feedback further improves the performance to an accuracy of 53.0%, which is still lower than the offline batch learning fine-tuning mentioned in the previous paragraph.

In conclusion, this work improves the existing NLMaps dataset and contributes two new datasets – one of which is especially valuable since it consists of real user queries – laying the groundwork necessary for further enhancing NLMaps parsers. The current parser – achieving an accuracy of 58.8% – can be used by OSM users via the new web interface [11] for issuing queries and also for correcting incorrect ones. Future work will concentrate on improving the web interface’s UX and enhancing the parser’s performance in terms of speed and accuracy.

References

[1] Hoffmann, S., Metten, M. T., & Quinion, B. (2021). Nominatim: Open-source geocoding with OpenStreetMap data. Retrieved from https://nominatim.org
[2] Olbricht, R. (2021). Overpass API. Retrieved from https://overpass-api.de

[3] Haas, C., & Riezler, S. (2016). A corpus and semantic parser for multilingual natural language querying of OpenStreetMap. In: *Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 740-750.
[4] Lawrence, C., & Riezler, S. (2018). Improving a neural semantic parser by counterfactual learning from human bandit feedback. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 1820-1830.
[5] Lawrence, C., & Riezler, S. (2016). NLmaps: A natural language interface to query OpenStreetMap. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*: System Demonstrations, 6-10.
[6] Will, S. (2018). Parsing NLmaps queries using adversarial neural machine translation. Unpublished bachelor's thesis, Heidelberg University.
[7] Staniek, M. (2020). Towards error-aware interactive semantic parsing. Unpublished master's thesis, Heidelberg University.
[8] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint*, arXiv:1409.0473.
[9] Topf, J. (2021) Taginfo. Retrieved from https://taginfo.openstreetmap.org
[10] Gwerder, S. (2014). Tag-Suchmaschine und Thesaurus für OpenStreetMa. Student research project, HSR University for Applied Sciences Rapperswil.
[11] Will, S. (2021) NLMaps Web. Retrieved from https://nlmaps.gorgor.de

Towards a framework for measuring local data contribution in OpenStreetMap

Maxwell Owusu[1,*], Benjamin Herfort[1] and Sven Lautenbach[1]

[1] Heidelberg Institute for Geoinformation Technology, Heidelberg, Germany
maxwell.owusu@uni-heidelberg.de, benjamin.herfort@heigit.org, sven.lautenbach@uni-heidelberg.de

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

OpenStreetMap (OSM) has proven to be a valuable source of spatial data for many applications, including humanitarian aid. Information on buildings and roads—that can be provided by remote mapping—is of highest concern for many humanitarian applications. However, further information that can only be mapped on the ground is of high importance for finer scale humanitarian action. Road surface information, type of material and information on the use of a building (health site, school, etc.) is highly relevant. OSM offers several possibilities of adding local knowledge [1]. Recent works deal with analyzing and classifying data production in OSM [2] and intrinsic analysis has gained popularity as an indicator for measuring quality of OSM data [3–6]. Nevertheless, relatively few scientific studies have touched on "local knowledge" and local data in OSM in sufficient detail.

The question of how much local knowledge is added and what kind of local data is added remains unanswered. Addressing this question is important since only local knowledge provides access to the plethora of contextual information that is necessary for many purposes. The term "local knowledge" is often debated in the OSM community due to its ambiguity. Consequently, it is hardly taken into account by researchers when evaluating OSM [1]. This study presents a metric to measure local data contributions in OSM and analyzes temporal patterns of local contributions at three case studies. The aim of the metric is to identify archetypes of places representing a variety of contextual information.

First, we evaluated Rebecca Firth's framework on OSM contribution types that focused on the humanitarian context—see the Twitter post at [7]. Second, we discussed with local community working groups how to measure local data contributions ("What exactly are local OSM data to you?"). The outcome of the community discussion provided valuable information to design a generalized workflow for measuring local data contribution in OSM. Subsequently, we identified aspects on which the local communities agreed with respect to perception of local data. Based on these first insights, we developed a classification schema for measuring local data in OSM that is "fit-for-purpose" for local OSM communities. This schema consists of four main levels and assigned OSM tags that could be used as indicators for each level. Third, we explored the temporal evolution of local data in OSM for three unique regions. These region's mapping activities are influenced by local mapping organizations. (i) Ramani Huria in Dar es Salaam, Tanzania, focusing on flood resilience (ii)

Owusu, M., Herfort, B. & Lautenbach, S. (2021). Towards a framework for measuring local data contribution in OpenStreetMap In: Minghini, M., Ludwing, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 16-18. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112234

Crowd2Map mainly operating in the Mara region, Tanzania and focusing on identifying features that can support the fight against girls and women at risk of female genital mutilation, and (iii) power mapping project by YouthMappers in the Koindugu District, Sierra Leone, focusing on mapping electrical grid infrastructure. We used the ohsome API to access the full history of OSM. We determined the density and the ratio (as the sum of all OSM tags to the number of OSM elements) per month for each region and localness level.

The outcome of the community discussion showed that local mappers/editors had different perceptions about local knowledge. The type of local data produced depends on: (1) the context within which the data is produced and (2) the character/interest of the individual performing the mapping. However, the local data produced could be broadly categorized as "core" or "specific". The "core" category consisted of the objects that cut across almost all projects or activities (e.g., buildings, roads, place names and administrative boundary) and the category "specific" were special elements mapped as a results of a particular interest or aim of the project (e.g., culvert, drains, access types, parking type).

The developed metric is illustrated in Figure 1, showing the four main levels. Level 1 consists of objects that can be derived easily by remote mapping from satellite images such as roads and building (this is information that does not require local knowledge), level 2 focuses on place names and administrative boundaries which are frequently imported, level 3 focuses on the presence of general (e.g., residential and commercial) or specific amenities (e.g., school, clinic, and point of interest) and level 4 focuses on micro-data that provides further contextual information about an object (e.g., road:maxspeed, surface condition). Level 1 and 2 mainly fall into the "core" category whereas level 3 and 4 mainly belong to the "specific" category (which will vary across different regions).

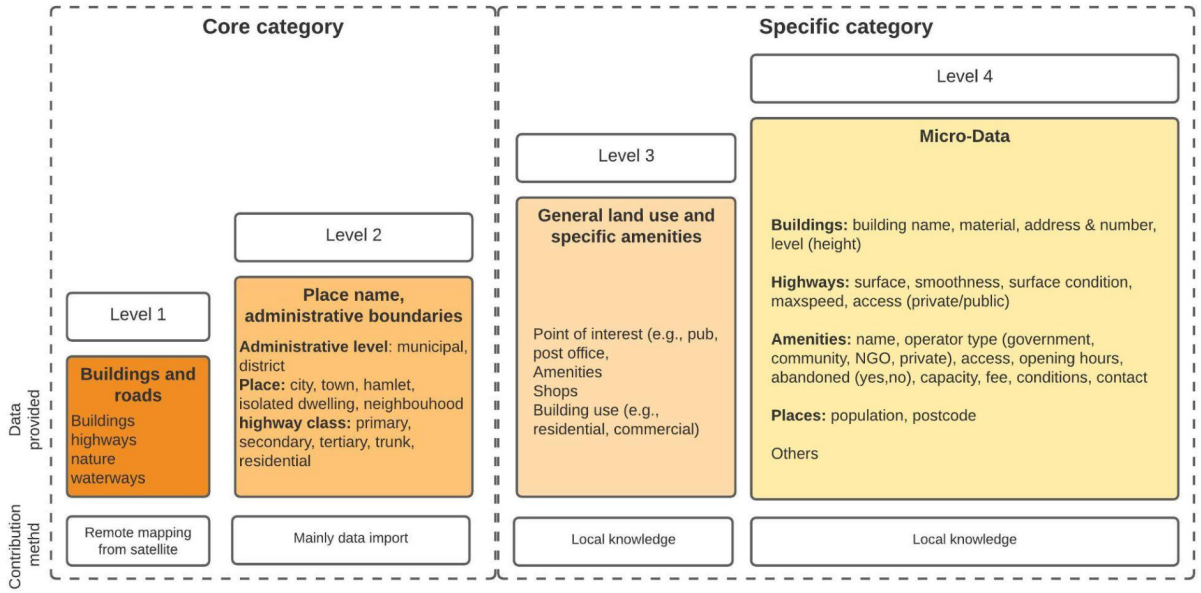

Figure 1. Classification schema for measuring local OpenStreetMap contribution.

From the temporal analysis, we observed that the amount of features in OSM decreased from level 1 to level 4. The ratio between level 1 and level 4 shows how widely local information is present in OSM at a specific location. Thereby, it provides insight on the quality of the OSM data and fitness-for-purpose for applications that need information beyond the existence of highways or buildings. Most of the mapping in the selected region

started in 2015. By digging deeper into the objects mapped, each selected region depicts unique characteristics which are largely shaped by the interest of contributors/ organizations. Mapping patterns are clearly distinct from each region with respect to the development of tags. For example, there was a high amount of local data regarding waterways, drainage, and solid waste in Dar es Salaam and very low in Mara region and the Koinadugu District. It reveals the distinct mapping stories of individuals or organizations. Our results show further that there is no common path from level 1 to Level 2 to level 3 among the different regions. For the case of Dar es Salaam, mapping of features of the three levels has happened more or less simultaneously. Mapping in the Mara region focused first on place names (level 2) and then on amenities (level 3) as well as buildings and roads (level 1). For the Koinadugu District, mapping of level 2 started in 2011 and was followed by mapping of buildings and roads (level 1) in 2014 and amenities (level 3) from 2017 onward.

The classification schema helps to conceptualize a metric to measure localness of OSM data at different levels of details. This metric can be easily used to group OSM data into the categories "core" and "specific". By analyzing the temporal patterns, we identified that the contribution of local data was highly unequal and largely depended on the interest of the mapper(s). The research shed light on the richness of contextual information in OSM as well as an indication for the quality of data. In future research we would like to extend the results presented here by including more regions and more perspectives from local OSM communities. By doing so, we hope to be able to extend the definition of local data by considering the editors' local knowledge as well.

References

[1] Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1), 139-167.

[2] Grinberger, A. Y., Schott, M., Raifer, M., & Zipf, A. (2021). An analysis of the spatial and temporal distribution of large-scale data production events in OpenStreetMap. *Transactions in GIS*, 25, 622-641.

[3] Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., & Zipf, A. (2021). The evolution of humanitarian mapping within the OpenStreetMap community. *Scientific Reports*, 11(1), 1-15.

[4] Barron, C., Neis, P., & Zipf, A. (2014). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, 18(6), 877-895.

[5] Klonner, C., Hartmann, M., Dischl, R., Djami, L., Anderson, L., Raifer, M., Lima-Silva, F., Degrossi, L.C., Zipf, A., & Porto de Albuquerque, J. (2021). The Sketch Map Tool Facilitates the Assessment of OpenStreetMap Data for Participatory Mapping. *ISPRS International Journal of Geo-Information*, 10(3), 130.

[6] Neis, P., & Zipf, A. (2012). Analyzing the contributor activity of a volunteered geographic information project - The case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1(2), 146-165.

[7] Firth, R. (2019). Twitter post. Retrieved from https://t.co/rDaSraiVZF

Towards understanding the temporal accuracy of OpenStreetMap: A quantitative experiment

Levente Juhász[1,*]

[1] GIS Center, Florida International University, Miami, FL, USA; ljuhasz@fiu.edu

This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

The ability to provide timely information compared to traditional collection methods of geographic information is generally considered as one of the main advantages of volunteered geographic information (VGI) since its emergence in the 2000s [1]. In addition to several anecdotal examples illustrating how VGI data can provide more up-to-date information than authoritative sources, the literature provides ample evidence on the usefulness of VGI in applications that require timely geodata, such as disaster management [2, 3]. For example, the Haiti earthquake relief effort in 2010 laid the foundations for how remote contributors of OpenStreetMap (OSM) and other platforms can make a difference and aid responding humanitarian agencies after a crisis [4]. The Humanitarian OpenStreetMap Team has made numerous contributions and helped save lives at numerous instances ever since [5]. However, apart from these examples, the temporal dimension of VGI has not received much research attention outside the application of disaster management, and there is a huge gap between assessing temporal accuracy and other factors of data quality, such as spatial accuracy [6, 7]. Abreicht et al. highlighted the lack of formal acknowledgment of temporal aspects in the concept of VGI and proposed a framework called 'Volunteered Geo-Dynamic Information' to fully integrate spatial and temporal aspects of VGI [8]. Other works utilizing the temporal component in VGI often focus on the behavior of contributors rather than the currency and temporal validity of map features they contributed [9–11], or studied the evolution of data over time [12, 13]. While these approaches are useful, by nature they cannot provide a quantitative measure of how current OSM (or VGI in general) is. [14] noted during their investigations that the temporal accuracy of OSM could not be measured using their traditional extrinsic method, because OSM data was compared to authoritative data that did not contain temporal information (i.e. most recent street configuration regardless of when road segments were built or renovated). Another project, 'Is OSM up-to-date?' recognizes the lack of information on temporal accuracy and developed a tool that uses an intrinsic approach to visually show features that potentially contain outdated information [15]. However, by nature, an intrinsic approach can also not provide an absolute measure of how up-to-date OSM is.

This research attempts to fill a gap in the literature by conducting an experiment on the currency of VGI. Using OSM data as a case study, it measures the temporal accuracy of

Juhász, L. (2021). Towards understanding the temporal accuracy of OpenStreetMap: A quantitative experiment
In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 19-22. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112236

selected map features. This research overcomes previous limitations by using official data provided by the Florida Department of Transportation (FDOT). The dataset contains details about state-funded highway construction projects, including the date these projects were completed, therefore, accurately measuring the temporal accuracy of OSM features is possible by comparing dates projects were finished with the time at which corresponding OSM edits in the database were made. This time difference describes how long it took the OSM community to adapt to real-world changes and update the map database accordingly.

The historical version of highway construction projects was filtered to projects completed between May 15, 2016 and April 1, 2021. Further, only a subset of projects were used, that resulted in either new infrastructure (new roadways, roundabouts or highway ramps) or widening of existing roadways (i.e. new lanes excluding bike lanes and turning lanes). Other construction projects, such as traffic improvements, road resurfacing, regular maintenance (e.g. bridge rehabilitation), bike infrastructure etc. were excluded, since a useful, high-quality road network database can be maintained without the addition of this information, therefore, they are less likely to migrate into OSM. The methodology uses augmented diffs from the Overpass API to find all changes that occurred on OSM highway features (creation, modification and deletion) and are spatially and temporally close to construction projects. These changes are then matched with a record from the highway construction dataset. Irrelevant changes (i.e. changes made to other highway features) are removed. This is done by manually interpreting and evaluating changes and construction projects using a description field (e.g. *"WAKULLA SPRINGS ROAD @ OAK RIDGE ROAD ROUNDABOUT"*). The data extraction algorithm initially queries the Overpass API for changes one week beyond the completion date of a particular project. In case no relevant change can be found, iterative queries for 7-day-long time slices are made until a relevant change is found, or until the current date is reached. Lastly, the time difference between the end date of construction projects and the first OSM change that introduced the change in OSM are calculated. For example, the example from above can be found with the following Overpass query (https://overpass-turbo.eu/s/16XV) that uses the location of the highway construction. The relevance of a change can be verified using changeset comments: (https://www.openstreetmap.org/changeset/87938707). In this example, the changeset comment *"Added new round about."* confirms that the OSM edit is related to the FDOT dataset. The difference between the construction end date (July 3, 2019) and the time when this change appeared in OSM (July 13, 2020) is 1 year and 7 days.

The FDOT construction dataset was manually checked and matched with OSM edits. The final dataset contains 23 new highways and roundabouts, and 44 road widening projects (lane additions). Only 3 out of 23 highway construction projects have not been added to OSM. However, these projects were recently finished (October 2020), therefore it is possible that they will be added to OSM at a later time. The remaining projects were mapped in OSM within 133 days on average (median=45, SD=262). Widening projects are less likely to be mapped in OSM in a timely fashion, as 23 out of 44 of those projects are not mapped. The remaining widening projects seem to be added at a slower rate (mean=158 days, median=147, SD=388). Boxplots of construction projects are shown in Figure 1. Interestingly, several projects of both types have been mapped in OSM ahead of the official construction end date. This shows the flexibility of the OSM, which is able to adapt to informal scenarios, where roadways are already being used while construction crews are still working.

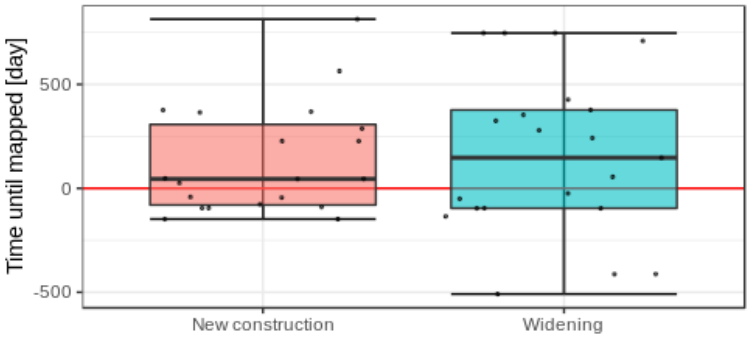

Figure 1. Lag between construction projects and corresponding mapping events

This experiment is the first attempt to investigate the timeliness and currency of VGI using OSM as a case study. The limitations of the study include the reference dataset, that does not contain federally or locally funded projects, therefore misses a large number of constructions, and the methodology, that cannot capture the diversity of the OSM community and also disregards changes beyond the transportation infrastructure. Future work will conduct analysis using more VGI data sources outside the domain of mapping applications (e.g. Points of Interest), new methodology using tile-reduce, OSM QA tiles and vector tiles built from other datasets. The new methodology will be scalable and will allow for analysis across world regions. Furthermore, a rule-based decisions approach based on tags and semantics will be used to eliminate the need for manually checking and verifying whether VGI updates correspond to the reference dataset or not.

References

[1] Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4), 211-221.

[2] Horita, F. E., Degrossi, L., Assis, L. F. de, Zipf, A., & de Albuquerque, J. P. (2013). The use of Volunteered Geographic Information (VGI) and Crowdsourcing in Disaster Management: A Systematic Literature Review. In: *Proceedings of the 19th Americas Conference on Information Systems*.

[3] Neis, P., & Zielstra, D. (2014). Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. *Future Internet*, 6(1), 76-106.

[4] Zook, M., Graham, M., Shelton, T., & Gorman, S. (2010). Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake. *World Medical & Health Policy*, 2(2), 7-33.

[5] Herfort, B., Lautenbach, S., de Albuquerque, J. P., Anderson, J., & Zipf, A. (2021). The evolution of humanitarian mapping within the OpenStreetMap community. *Scientific Reports*, 11(1), 3037.

[6] Antoniou, V., & Skopeliti, A. (2017). The Impact of the Contribution Microenvironment on Data Quality: The Case of OSM. In: Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C. C., & Antoniou, V. (Eds.) *Mapping and the Citizen Sensor*, 165-196. Ubiquity Press, London.

[7] Yan, Y., Feng, C.-C., Huang, W., Fan, H., Wang, Y.-C., & Zipf, A. (2020). Volunteered geographic information research in the first decade: A narrative review of selected journal articles in GIScience. *International Journal of Geographical Information Science*, 34(9), 1765-1791.

[8] Aubrecht, C., Aubrecht, D. Ö., Ungar, J., Freire, S., & Steinnocher, K. (2017). VGDI – Advancing the Concept: Volunteered Geo-Dynamic Information and its Benefits for Population Dynamics Modeling. *Transactions in GIS*, 21(2), 253-276.

[9] Begin, D., Devillers, R., & Roche, S. (2018). The life cycle of contributors in collaborative online communities -the case of OpenStreetMap. *International Journal of Geographical Information Science*, 32(8), 1611-1630.

[10] 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.
[11] Neis, P., & Zipf, A. (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project—The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1(2), 146-165.
[12] Girres, J.-F., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, 14(4), 435-459.
[13] Zielstra, D., & Hochmair, H. H. (2011). A Comparative Study of Pedestrian Accessibility to Transit Stations Using Free and Proprietary Network Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2217, 145-152.
[14] Arsanjani, J. J., Barron, C., Mohammed, B., & Helbich, M. (2013). Assessing the Quality of OpenStreetMap Contributors together with their Contributions. In: Vandenbroucke, D., Bucher, B., & Crompvoets, J. (Eds.) *Proceedings of the 16th AGILE Conference on Geographic Information Science*, 14-17.
[15] Minghini, M., & Frassinelli, F. (2019). OpenStreetMap history for intrinsic quality assessment: Is OSM up-to-date? *Open Geospatial Data, Software and Standards*, 4(1), 9.

Introducing OpenStreetMap user embeddings: Promising steps toward automated vandalism and community detection

Yinxiao Li[1,*] and Jennings Anderson[2]

[1] Facebook, Boston, USA; yinxiaoli@fb.com
[2] YetiGeoLabs, Montana, USA; jennings.anderson@gmail.com

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

With more than 11B edits from 1.6M unique mappers and openly editable by anyone, the OpenStreetMap (OSM) database inevitably contains vandalism. Our approach to detecting it leverages the analytical power and scalability of machine learning through OSM user embeddings. Embeddings are effective in capturing semantic entity similarities that are not explicitly represented by the data. Since word embeddings were first introduced based on the assumption that words adjacent to each other share similar meanings [1, 2], the concept of embeddings has been extended beyond word representations to any entity, so long as one can produce a meaningful sequence of the entities. Therefore, we build OSM user embeddings with mappers as entities by constructing sequences of mappers based on shared editing histories and similar behaviors.

Development of automated vandalism detection methods in OSM has been slow in part because there is no published corpus of bad or vandalized edits from which to train and validate [3]. Vandalized name attributes are especially problematic because this text is rendered on the basemap. The most infamous instance of this type of vandalism was the changing of "New York City" to an ethnic slur; this name attribute was subsequently rendered on maps drawing from OSM data [4]. As part of this work, we construct and make available the first OSM vandalism corpus for the name attribute of OSM features. Potential examples of vandalism are collected from the OSM Changeset Analyzer (OSMCha) web-based validation tool. These records are then manually reviewed by the Facebook mapping team to identify egregious name changes. Negative samples (non-vandalism) were randomly sampled from a previously validated vandalism-free snapshot of OSM. All of our examples are extracted from OSM data only, no external or additional conflated data sources.

To construct meaningful sequences of OSM users where adjacent users share similar mapping patterns, we analyzed the edit history of every OSM object and the temporal/semantic editing patterns of individual mappers. These sequences were then fed into a word2vec skip-gram model to train OSM user embeddings.

Li, Y. & Anderson, J. (2021). Introducing OpenStreetMap user embeddings: Promising steps toward automated vandalism and community detection
In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 23-26. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112241

Shared object editing histories are sequences of OSM users who have edited the same object, in chronological order of editing. These sequences represent mappers who share interest in the same objects on the map. This yields 2B sequences of mappers.

Semantic and temporal mapping patterns are sequences of OSM users that have shared editing characteristics with regard to how and when they edit the map. Starting with *changesets*, we extract the following keys for each OSM element edited in a given changeset: `addr:country`, `admin_level`, `amenity`, `building`, `highway`, `natural`, `place`, `source`. Each of these attributes are not always present, but were chosen based on empirically observed high-level editing patterns among users. Additionally, we extract the following metadata: the presence of `name` tag, the `version` number, the editing software (e.g. iD editor, JOSM), and any hashtags (possibly denoting specific mapping campaigns). Finally, we group all of these edits by two types of temporal patterns: first, the date of the changeset, and second, the hour of the week of the changeset, per year (with 168 hours in a week, we aggregate across each *week-hour* in a given year). This yields 30M sequences of mappers where each sequence describes a list of mappers sharing similar temporal editing patterns, editing tools, and types of edits.

To detect vandalism, we train a Gradient Boosting Decision Tree (GBDT) model with xgboost library [5], which consists of context and content features, similar to [6]. We applied OSM user embeddings into this model by creating two embedding features, `kmeans_cluster` and `cos_sim_last_5_users`. To create `kmeans_cluster`, we ran k-means clustering on OSM users, assigned a cluster to any user with an embedding, and then encoded the cluster based on the average number of edited changesets among each cluster. The idea behind `cos_sim_last_5_users` is that users who are similar to each other are more likely to edit the same objects. Starting with an edit to an OSM object, we compute the cosine similarity between the user responsible for the edit and the previous five mappers that edited the object.

Next, we trained a new model by injecting the embedding features, and we have seen a solid relative improvement of 1.3% in our primary metric, area under the receiver-operator curve (AUC-ROC), which is a widely used metric in vandalism detection research [6]. The feature importance of `kmeans_cluster` is ranked as high as 2/49, with a coverage of 99.9%, while `cos_sim_last_5_users` has an importance rank of 16/49, largely due to a relatively low coverage of 64%, meaning that the majority of edits in OSM create new objects, so there can be no editing history for these.

Because of the AUC improvements and high feature importance, Facebook has deployed this model in production to detect vandalism, as a part of the data validation in the Facebook Map and Daylight Map, a validated, vandalism-free distribution of OSM [7].

The accurately labeled dataset of vandalism to named elements in OSM is a tremendous asset to researchers hoping to further the work of automated vandalism detection. As part of this work, we are publishing this fully labeled vandalism corpus for others in the OSM research community to use [8].

In addition to vandalism detection, OSM user embeddings can be used for automated community detection within OSM. OSM is comprised of many distinct groups of mappers; considering each of these groups a different sub-community makes OSM a "community of communities" [9]. The creation of the temporal and semantic editing patterns were specifically designed to create sequences of mappers with high likelihood of belonging to

the same community, the resulting user embeddings then group these communities of mappers appropriately.

To apply the user embeddings to community detection, we constructed a network graph of OSM users in which two users are connected if they have a high cosine similarity as defined by the user embedding. For each mapper in our embedding, we compute the cosine similarity for the top 100 most similar other mappers. If the cosine similarity is above 0.95, then an edge is created between the two nodes with a weight equal to the actual cosine similarity value. Figure 1 shows a depiction of this network graph for the top 10,000 most prolific editors in 2020.

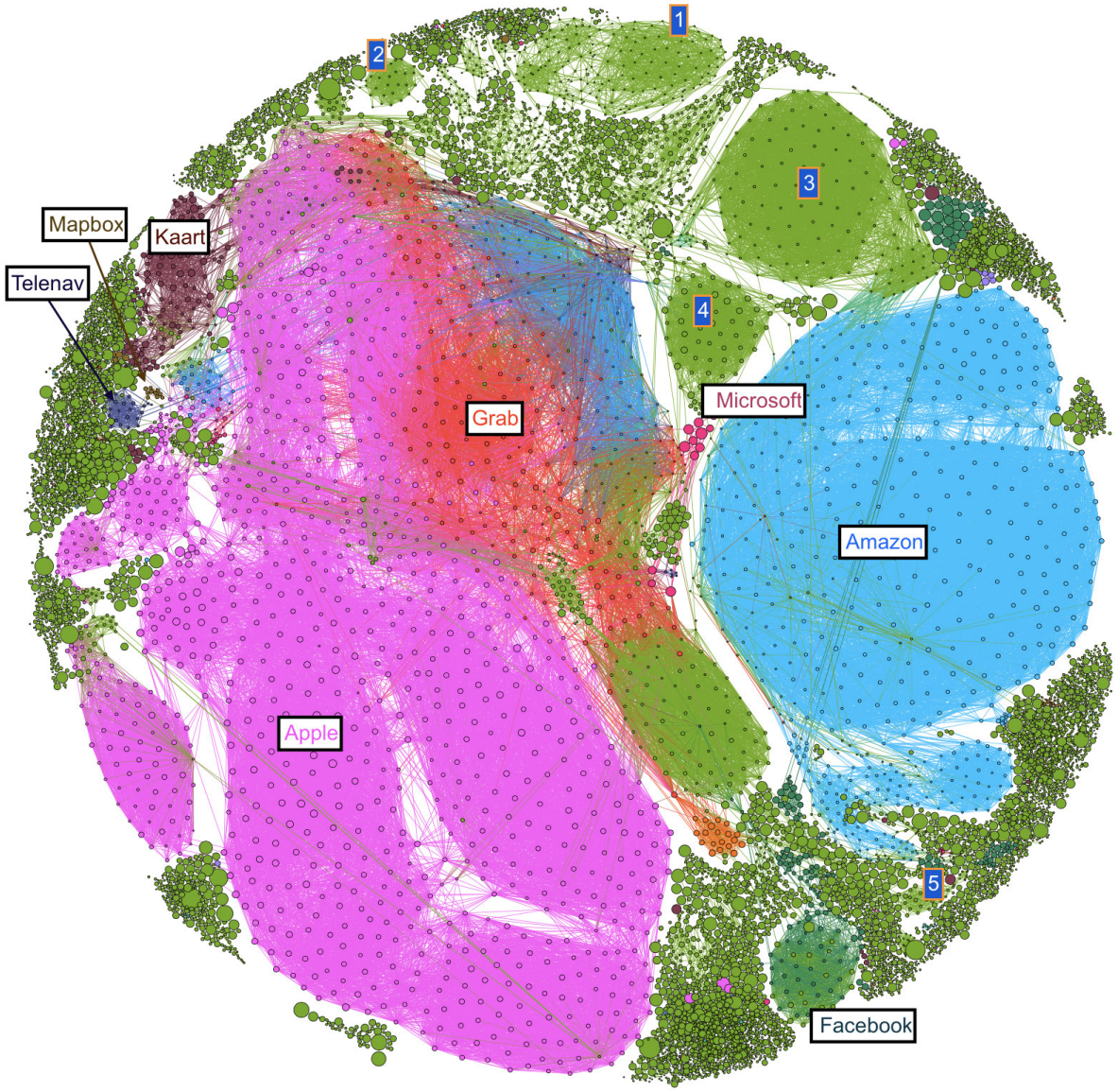

Figure 1. Network graph of the top 10,000 mappers in 2020 by changeset count. Nodes (mappers) are colored with regard to corporations they are known to be associated with. Green nodes are not known corporate editors.

Previous research on corporate editing in OSM has identified specific teams of paid mappers working for different corporations [10]. These paid mappers represent one type of editing "community" in OSM. Since these mappers publicly disclose their affiliations, the

members of this mapping community are known. Using the list of paid editors identified in [11], we colored the nodes in Figure 1 with their corporate association, labeled as such. The green nodes represent users that are not known corporate editors. The purpose of this labeling is not to identify corporate editors, but rather to validate the performance of our approach at identifying known communities: The obvious clusters between different colors show that these different corporate communities are successfully identified by the OSM User Embeddings.

Of interest, then, are the additional clusters appearing among the green nodes which may represent specific editing communities in OSM. For an illustrative example, we label 5 example community clusters in Figure 1. Manually investigating the mapping activity of the users associated in each numbered cluster reveals that the network graph successfully clusters other sub communities of editors in OSM. Cluster 1, for example, appears to be a group of Russian editors who joined OSM within the past two years and are heavily focused on editing buildings this past year in small cities near the Caspian Sea. Cluster 2 also appears to have many Russian mappers who joined OSM in the past four years and specifically clean up tags all over the world. Cluster 3 is a group of previously unidentified Amazon mappers, showing the abilities of this approach to potentially help identify more corporate editors. Clusters 4 and 5 represent two distinct groups of humanitarian mappers participating in Humanitarian OpenStreetMap Team (HOT) mapping tasks last year.

References

[1] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*, arXiv:1301.3781.

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint*, arXiv:1310.4546.

[3] Truong, Q. T., Touya, G., & Runz, C. (2020). Osmwatchman: Learning how to detect vandalized contributions in osm using a random forest classifier. *ISPRS International Journal of Geo-Information*, 9(9), 504.

[4] Zaveri, M. (2018). New York City is briefly labeled 'Jewtropolis' on Snapchat and other apps. Retrieved from https://www.nytimes.com/2018/08/30/business/jewtropolis-map-new-york-snapchat.html

[5] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

[6] Heindorf, S., Potthast, M., Stein, B., & Engels, G. (2016). Vandalism detection in wikidata. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 327-336.

[7] Facebook (2021). Daylight Map Distribution. Retrieved from https://daylightmap.org

[8] Facebook (2021). OSM Name Vandalism Corpus Released. Retrieved from https://daylightmap.org/2021/05/24/name-vandalism-corpus-release.html

[9] Solís, P., Anderson, J., & Rajagopalan, S. (2020). Open geospatial tools for humanitarian data creation, analysis, and learning through the global lens of youthmappers. *Journal of Geographical Systems*.

[10] Anderson, J., Sarkar, D., & Palen, L. (2019). Corporate editors in the evolving landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information,* 8(5), 232.

[11] Anderson, J. (2021). A 2021 update on paid editing in OpenStreetMap. Retrieved from https://www.openstreetmap.org/user/Jennings%20Anderson/diary/396271

A proposal for a QGIS plugin for spatio-temporal analysis of OSM data quality: the case study for the city of Salvador, Brazil

Elias Elias[1,*], Fabricio Amorim[1], Leonardo Silva[1], Marcio Schmidt[1], Silvana Camboim[1] and Vivian Fernandes[2]

[1] Programa de Pós Graduação em Ciências Geodésicas, Federal University of Parana, Curitiba, Brazil; elias_naim2008@hotmail.com, fabricioamorimeac@hotmail.com, scharth.leo@gmail.com, marcio.schmidt@ufu.br, silvanacamboim@gmail.com
[2] Programa de Pós Graduação em Engenharia Civil, Federal University of Bahia, Salvador, Brazil; vivian.fernandes@ufba.br

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

The development of methodologies to evaluate geospatial data quality is one of the most important aspects to be considered while obtaining this data. For developing countries, such as Brazil, the lack of investment for the maintenance of topographic mapping, especially on a big scale, is a recurrent challenge for the National Mapping Agencies [1]. For example, studies reveal areas in Brazil that have never been mapped and that the topographic mapping in the 1:25,000 scale is nearly 5% of its extension [2].

Technological advances have enabled a series of methodologies for obtaining geospatial data [3]. One example is presented as Volunteered Geographic Information (VGI) [4]. In this case, the update of information may occur faster and with a reduced cost in detriment to the traditional structures of topographic mapping [5]. A successful case of VGI is the OpenStreetMap (OSM) project, which presents the growth in the number of contributors and contributions or mapped features. To understand the quality of OSM features and their integration potential in topographic mapping, different surveys worldwide have put efforts to evaluate its quality, whether by its extrinsic [6, 7] or intrinsic [8] aspects. In this regard, some studies have evaluated the quality of OSM features by combining extrinsic and intrinsic aspects, like [9], which evaluated the positional accuracy of OSM based on the combination of edit history. Besides that, recent works have focused on comprehending spatial and temporal aspects of events in OSM contributions [10], as well developing add-ons for evaluating data quality, as presented by [11], where the authors developed a QGIS toolbox to evaluate parameters of the intrinsic quality of OSM features.

The literature identifies the heterogeneity of the data as one of the main challenges for the integration processes. The value of the quality parameters may vary according to the study area, the indicator used or even the temporal variations in the dynamics of the geographical space itself. In this context, to understand the integration of OSM data to the

Elias, E.N.N., Amorim, F.R., Silva, L.S., Schmidt, M.A., Camboim, S.P. & Fernandes, V.O. (2021). A proposal for a QGIS plugin for spatio-temporal analysis of OSM data quality: the case study for the city of Salvador, Brazil
In: Minghini, M., Ludwing, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 27-30. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112244

topographic mapping, it is crucial to connect aspects related to the quality and heterogeneity of data. Research work like [1] argues that, based on the obtained quality, the resources resulting from VGI may be used to integrate, detect changes or report errors. Therefore, classifying resources from OSM according to their usability in a certain region becomes essential, especially in developing countries like Brazil. Besides that, research that explores issues of quality, heterogeneity, and contributions patterns of OSM is still not widespread in developing countries [12].

Given the importance of classifying OSM features according to their usability for a given region, especially in developing countries, few researchers have explored quality, heterogeneity, and contribution pattern issues in OSM in Brazil. Based on the issues addressed, we propose a hypothesis that understanding aspects of the extrinsic and intrinsic quality of the quality of OSM features will help decision making regarding the integration of such data in topographic mapping. The main focus is on the spatio-temporal aspects of contributions in developing countries. Thus, this research has the objective to evaluate the extrinsic quality of OSM features for the county of Salvador-Bahia-Brazil (the northeast region of the country). Therefore, we investigated indicators of positional accuracy, thematic accuracy and completeness, the visualisation of heterogeneity of data, and the analysis of the edition history. To accomplish the evaluation of extrinsic quality, the OSM features were compared to the topographic mapping of the country from the Cartographic and Cadastral System of the County of Salvador (Sistema Cartográfico e Cadastral do Município do Salvador - SICAD, 2006) and features from the Urban Development Company of the State of Bahia (Companhia de Desenvolvimento Urbano do Estado da Bahia - CONDER).

The analysis of positional and thematic accuracy was made through procedures of feature sampling. The analysis of completeness occurred from comparing the total number of available features. The verified categories were features from the road system, religious, educational, and health buildings. We divided the municipality of Salvador into sub-regions to identify different local patterns of quality in the analysis of thematic accuracy and completeness. Visualisation allows obtaining the data's heterogeneity through a plugin developed in the software QGIS, making the planimetric positional evaluation for point and line features. The statistical procedures for developing the plugins were realised based on the Brazilian law to evaluate geospatial data quality analysis [13] and based on the method of double buffer proposed by [14]. The plugin is available, and it is possible to be accessed in the online repository [15]. Even though the final results comprehend aspects of Brazilian law, they can be replicated to obtain the discrepancies and posterior adjustments. We used the Ohsome Application Programming Interface (API) [16] to identify the patterns concerning the OSM editing history. Thus, from the adaptations performed in scripts given by researchers linked to Ohsome, it was possible to identify the aspects of OSM contributions between 2008 and 2020. We also tested the generation of regression curves and calculated the number of daily contributions to identify these patterns. These verifications were occasionally created through the generation of an evolving rectangle of 5x5 km in the study area. The disposition of the rectangle was given through a visual analysis with a larger quantity of OSM features.

The evaluation of extrinsic evaluation highlighted the variability of the results obtained in [17]. In analysing the positional accuracy, the scale found varied from 1:20,000 to 1:30,000, while the discrepancies between the mapped coordinates and the reference one

varied between 0.12m and 10.27m. In analysing completeness, it was observed that features that corresponded to the road system presented better results concerning the other categories. The road system presented a completeness percentage of 82%, while in the other features, the variation was from 29% to 46%. When analysing thematic accuracy, it turns out that the primary source of errors is related to the absence of names in editing. In the analysis of the OSM contribution history growth of represented features, it was possible to notice a near-linear function, with an R2 value of 0.94. As there is a finite amount of mappable elements at a given area, we can use this function to model the contributions patterns over time until the region is saturated. Besides that, it was possible to observe that the patterns of collaboration can be affected by different variables because it was noticed that in 2016, more than 800 features were added in a short period. These aspects can be related to events such as data importation or mapathons.

The development of add-ons for evaluating OSM data quality that departs from the making of statistical procedures up to visualising the heterogeneity of data will assist in the decision-making as to data quality. The development of QGIS plugins for OSM data quality assessment that execute from statistical procedures to visualisation of data heterogeneity will assist in decision making regarding data quality

From the add-on developed, it was possible to notice that the magnitude of discrepancies did not present patterns and that this may vary according to the period of editingand the database used for the contributions. Based on the obtained results, we noticed the relevance in identifying the aspects of quality and heterogeneity in OSM contributions. For Brazil, identifying these characteristics may numerally indicate the integration potential of these data to authoritative mapping. Besides that, it will estimate the influence of unusual agents, like it is the case of data import in the contributions. The continuity of the studies is recommended to identify the causes of different patterns of growth and the continuity of studies to automatise the quality procedures.

References

[1] Maulia, N. (2018). Development of an update procedure for authoritative spatial data by the combination with crowdsourced information. Master's thesis, Technische Universitat Dresden.

[2] Silva, L. S., & Camboim, S. P. (2020). Brazilian Nsdi Ten Years Later: Current Overview, New Challenges And Propositions For National Topographic Mapping. *Boletim De Ciências Geodésicas*, 26(4).

[3] Brovelli, M. A., Boccardo, P., Bordogna, G., Pepe, A., Crespi, M., Munafò, M., & Pirotti, F. (2019). Urban Geo Big Data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-4/W14, 23-30.

[4] Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211-221.

[5] Du, H., Alechina, N., Jackson, M., & Hart, G. (2017). A method for matching crowd-sourced and authoritative geospatial data. *Transactions in GIS*, 21(2), 406-427.

[6] Brovelli, M. A., & Zamboni, G. (2018). A new method for the assessment of spatial accuracy and completeness of OpenStreetMap building footprints. *ISPRS International Journal of Geo-Information*, 7(8), 289.

[7] Zhang, H., & Malczewski, J. (2017). Accuracy evaluation of the Canadian OpenStreetMap road networks. *International Journal of Geospatial and Environmental Research*, 5(2).

[8] Minghini, M., & Frassinelli, F. (2019). OpenStreetMap history for intrinsic quality assessment: Is OSM up-to-date? *Open Geospatial Data, Software and Standards*, 4(1), 9.

[9] Nasiri, A., Ali Abbaspour, R., Chehreghan, A., & Jokar Arsanjani, J. (2018). Improving the quality of citizen contributed geodata through their historical contributions: The case of the road network in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 7(7), 253.
[10] Grinberger, A. Y., Schott, M., Raifer, M., & Zipf, A. (2021). An analysis of the spatial and temporal distribution of large-scale data production events in OpenStreetMap. *Transactions in GIS*, 25(2), 622-641.
[11] Sehra, S. S., Singh, J., & Rai, H. S. (2017). Assessing OpenStreetMap data using intrinsic quality indicators: an extension to the QGIS processing toolbox. *Future Internet*, 9(2), 15.
[12] Camboim, S. P., Bravo, J. V. M., & Sluter, C. R. (2015). An Investigation into the Completeness of, and the Updates to, OpenStreetMap Data in a Heterogeneous Area in Brazil. *ISPRS International Journal of Geo-Information*, 4(3), 1366-1388.
[13] DSG. (2016). Norma da Especificação Técnica para Controle de Qualidade de Dados Geoespaciais (ET-CQDG). *Diretoria do Serviço Geográfico*, 1, 94, Brasília-DF.
[14] Santos, A. P. (2015). Controle de qualidade cartográfica: Metodologias para avaliação da acurácia posicional em dados espaciais. PhD thesis, Universidade Federal de Viçosa.
[15] Elias, E. (2019). AcuraciaPosicional_PEC-PCD. Retrieved from https://github.com/eliasnaim/AcuraciaPosicional_PEC-PCD
[16] Heidelberg Institute for Geoinformation Technology (2021). Ohsome. Retrieved from https://heigit.org/big-spatial-data-analytics-en/ohsome
[17] Elias, E. N. N. (2019). Qualidade de Dados Geoespaciais em Plataforma de Mapeamento Colaborativo. Master thesis, Universidade Federal da Bahia.

An automated approach to identifying corporate editing activity in OpenStreetMap

Veniamin Veselovsky[1]*, Dipto Sarkar[2], Jennings Anderson[3] and Robert Soden[1]

[1] Department of Computer Science, University of Toronto, Toronto, Canada; venia@cs.toronto.edu, soden@cs.toronto.edu
[2] Geography and Environmental Studies, Carleton University, Ottawa, Canada; diptosarkar@cunet.carleton.ca
[3] Department of Computer Science, University of Colorado Boulder, Boulder, USA; jennings.anderson@colorado.edu

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

In the past five years, the OSM community has seen a dramatic rise in organized editing, including corporate, humanitarian, and educational, on the platform. These new actors have continued the ongoing debate surrounding OSM's relationship with organized editing, with new rules and best-practices being implemented to align the interests of the organizations with those of the community.

We became interested to study how the editing habits of these new actors differed from the community as a whole, but were quickly confronted by the challenge of producing accurate measures of their activities. In this paper we aim to fill this gap by creating computational methods of understanding different editing behaviours on OSM to classify editors as being corporate or volunteer. Classifying individual editors has been done in the past, on a more local level, for example in the recent analysis on editing in Mozambique. [1]

Studying corporate editing behaviour, first requires a list of corporate editors. In the past, researchers have searched individual "organized editing team" webpages. Instead, our paper presents a novel method for classifying users on the platform, by scraping user profiles. There are two possible approaches to extract corporate mappers based on user profiles. The first approach uses a clustering of the keywords within the profiles. Though effective at uncovering relations between users (like students, programmers, Garmin editors, Colorado mappers), this method failed to properly capture all known corporate groups. Instead we did a keyword search for corporations listed on the Organized Editing List and classified similar users together. This included a list of 2,177 known corporate mappers with over 50 unique changesets.

Using this extracted list, we discern features that could act as "signals" for organized editors. Explicitly, which features from the changesets can point to an editor being corporate or volunteer. Do corporate editors edit specific types of items? Do their time series signatures differ?

Veselovsky, V., Sarkar, D., Anderson, J. & Soden, R. (2021). An automated approach to identifying corporate editing activity in OSM In: Minghini, M., Ludwig, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 31-33. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112248

For the creation of these features, we relied on Jennings Anderson's past work on corporate editing for inspiration [2]. The first set of features came from OSM changeset metadata which is rich with user descriptive data like the editor used, comments, and source. We find that most organizations use editors like JSOM and iD. Next, we attempted to model which objects corporations edit by finding descriptive words like "service", "road", and "building" in the comments of the changeset. We observed that most corporations focus on services and roads, as opposed to buildings which tend to be dominated by volunteer mappers.

The third feature was motivated by the observation that as the interests of a corporation change, the editing of its mapping team can also change. This has led to the documented phenomena of corporate mappers having a geographically dispersed editing pattern. This is markedly different from many volunteer mappers who often begin by mapping their local neighbourhoods. Using established metrics, we calculated the geographic dispersion for each user based on the latitude and longitude of their edits.

The metric we found most effective was the timeseries signature. Corporations have a traditional 9-5 mapping schedule, whereas non-corporate mappers tend to map far more haphazardly, including significant mapping on the weekend. When attempting to convert the time series signature into a usable metric, we came across a problem: time zones. All changesets in OSM are normalized to UTC time, this means that a user editing at 8am in Toronto, Canada and another user editing at 8pm in Beijing, China would in fact appear to be editing at the same time in OSM. Longitude and latitude data are not an effective method of extracting the mapper's time zone, since editing on OSM is increasingly done remotely, through "armchair mapping".

To utilize this strong signal, we developed a new method for normalizing a user's time signature, and it was based on the observation that individual corporations have several key editing patterns, depending on where their employees are located. For example, Facebook has two such patterns, each displaced by around 8 hours. This motivated us to create a "corporate editing signature" and translate user time signatures to find the minimal distance between the two. After using this method of adjustment, we were able to significantly improve the alignment of the time-series. In other words, we were able to recover the local time zone of most of these corporate editors. Figure 1 illustrates corporate mappers before and after adjustment.

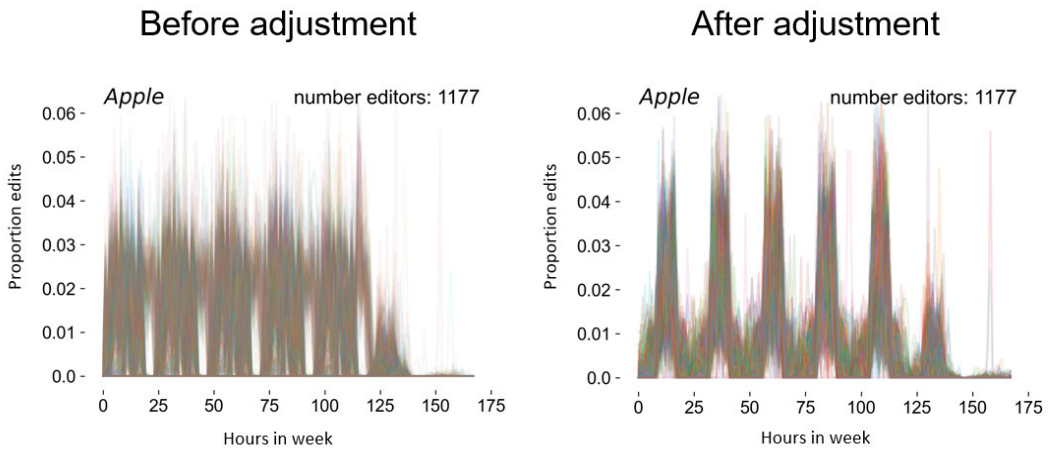

Figure 1. This plot shows how corporate time zones were recovered after minimizing distance between corporate actors and a "corporate mapping signature".

Once we realigned each user using this method, we calculated the distance between a user's adjusted time signature and the “corporate signature”. This feature ended up acting as a key determinant of the likelihood of a given editor being corporate. Out of the top 100 editors (who had the smallest distance to the corporate signature) all of them belonged to corporations.

Utilizing the user features we predict whether an editor is corporate or not. We experimented with several classification algorithms, including logistic regression, k-nearest neighbours, support vector machines, and neural networks. The four most important features in the prediction task, ordered by impact on model, were the geographic dispersion, time series score, first edit date, and the editor type. All models provided comparable results offering a high recall of 96%+ and predicting anywhere between 700 to 2,000 additional corporate mappers. Examining the newly predicted mappers reveals users that map for humanitarian groups like HOT, corporate mappers that the initial scrape didn’t pick up on, corporate mappers who reveal their association only in the hashtags, users who are likely corporate mappers with no ability to know for certain, and volunteers. After removing any “predicted mappers” who have known humanitarian associations from the most conservative model we arrived at a list of 500 newly identified corporate mappers. We are now entering the stage of further validating the different models based on a manually annotated set of users that any of the models predicted to be corporate.

References

[1] Madubedube, A., Coetzee, S., & Rautenbach, V. (2021). A Contributor-Focused Intrinsic Quality Assessment of OpenStreetMap in Mozambique Using Unsupervised Machine Learning. *ISPRS International Journal of Geo-Information*, 10(3), 156.

[2] Anderson, J., Sarkar, D., & Palen, L. (2019). Corporate Editors in the Evolving Landscape of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 8(5), 232.

Involvement of OpenStreetMap in European H2020 projects

Damien Graux[1,*] and Thibaud Michel[2]

[1] Inria, Université Côte d'Azur, CNRS, I3S, Sophia Antipolis, France; damien.graux@inria.fr
[2] Wemap, Montpellier, France; thibaud@getwemap.com

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2021 Conference after peer-review.

Since 1984, the European Commission has been supporting research through various successive programmes. Recently, from 2014 to 2020, the EU invested approximately 80 billion euros into its eighth programme, named Horizon 2020 (H2020) [1]. Among various focuses such as the excellence of science or industrial secondments, H2020 emphasised supporting an open access policy for all research results [2]. Moreover, H2020 projects were strongly encouraged to use open source software and tools.

Practically, all research domains were eligible to be supported by the H2020 programme, and therefore, the scopes of the projects vary from e.g. computer science, to philology passing by agriculture. Technically, as these projects are almost always involving several partners located in several European Member States joining forces from multiple institutions, there is often a need to deal with data coming from different places. And, more generally, geo-data are often involved to tag information which may be research data, meeting localisation, partner addresses, etc.

In such a context where open source tools and open access databases are recommended by the European Commission, we analyse the presence of OpenStreetMap (OSM) in H2020 projects. In addition, we also review the presence of other geographic services such as Google, Bing and Baidu maps, in order to better understand how researchers tend to choose one over the other. Furthermore, while OSM is a database, several services are created around it, including the basemap at openstreetmap.org, our study thereby reviews the use of OSM as an ecosystem of services.

Thanks to the open access policy, participants of H2020 projects had to make their results available. To do so, various types of materials were submitted to the European portal which then offers them publicly. As a consequence, for each project, one can access the articles (through DOIs), the blog posts, the slide decks, or also the deliverables. In particular, in our study, we decided to focus on the deliverables as they are accessible on the European Commission portal directly [3, 4] and as they are the common reports written by the partners to describe their approaches. Indeed, these deliverables (usually written on a regular basis during the course of the project) report on the findings and methodology set up to achieve the project's goals and the authors explain their architectural choices in depth such as describing the tools used. As a consequence, cartographic services, if involved at some

Graux, D. & Michel, T. (2021). Involvement of OpenStreetMap in European H2020 projects
In: Minghini, M., Ludwing, C., Anderson, J., Mooney, P., Grinberger, A.Y. (Eds.). Proceedings of the Academic Track at the State of the Map 2021 Online Conference, July 09-11 2021, 34-36. Available at https://zenodo.org/communities/sotm-2021
DOI: 10.5281/zenodo.5112252

stage in the project, are likely to be mentioned in these documents either as acronyms (e.g. OSM) or as website references (e.g. https://www.openstreetmap.org).

In order to obtain the deliverables together with projects' information, we combined two European sources of information to gather all of the facets we wanted to cover: CORDIS [3] and Data.Europa [4]. In particular, we extracted from CORDIS various high-level information about the projects themselves: from their names and acronyms to their durations passing by the specific European call-for-fundings they answered and obtained their money from. This latter category can be useful in order to have a finer-grained understanding of the domains which are prone to involved cartographic services. Next in order, Data.Europa was used to download the deliverables themselves, which required several days of computing resources.

Overall, during the course of the H2020 programme, 33636 projects were funded by the European Commission. Depending on the type of action which was set by the projects, not all of them had some open deliverables written (and thereby available on the Europa platform). Actually, a large part of these projects did not have deliverables per se but rather articles or web posts. We indeed counted 25157 projects without deliverables which restricted our study to the remaining 8479 projects. Out of them, we listed a total of 92612 distinct deliverables to be analysed, representing more than 260 GB.

Technically, once all of these deliverables were downloaded, we searched them for various terms to know if some cartographic services are involved in the text. We therefore set up several regex rules (e.g. 'open.?street.?map' or '[^a-z0-9]osm[^a-z0-9]') which were run over the 92000+ deliverables. This allowed us to systematically count all the occurences of the considered cartographic solutions. In the end, we found that 1840 deliverables (from 651 projects) mention OpenStreetMap. More precisely (see Figure 1), through all the H2020 deliverables, there are approximately: 18600 mentions to OSM, 2800 to Google Maps, 226 Bing Maps and 4 to Baidu Maps. Empirically, we notice that 1) one order of magnitude separates the occurrences of each cartographic service and 2) OpenStreetMap is by far the most represented solution and thereby the one on which public European researchers rely the most. Contextually, it is also interesting to note that not all the deliverables (1796 of them) mentioning "point of interest" refer to a cartographic service.

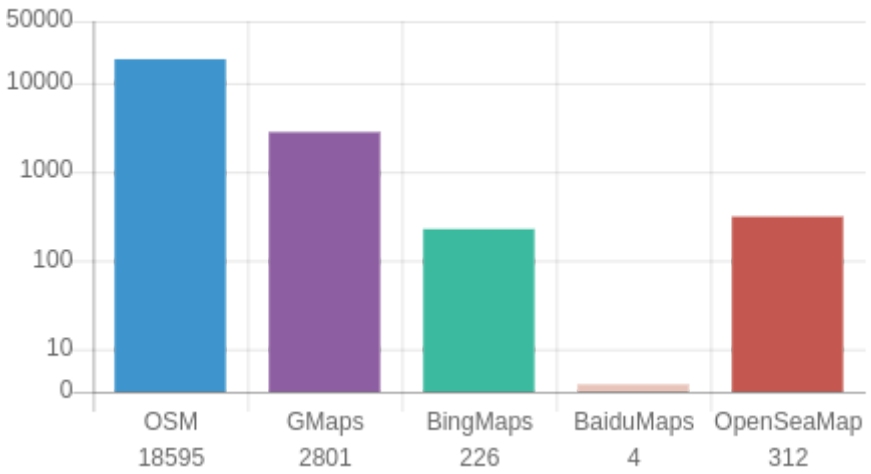

Figure 1. Number of occurrences over all H2020 open deliverables (logarithmic y-axis).

Moreover, we also analysed the co-occurence cases, where different cartographic providers are jointly mentioned within a single deliverable. Notably, there are not that many. Indeed, only 59 deliverables mention both OSM and Bing Maps, over the 226 occurrences of the latter; and only 291 deliverables mention both OSM and Google Maps, over the 2800 occurrences of Google Maps. Besides, only 39 deliverables mention OSM, Google Maps and Bing Maps. Such figures tend to suggest that once a group of researchers has chosen a cartographic solution, they tend to stick to it and do not try to compare them.

Furthermore, regarding OpenSeaMap, we counted 312 mentions from 27 deliverables, among which 20 ones mention both OSM and OpenSeaMap, showing how connected the two initiatives are.

In this study, we systematically analysed all the available H2020 deliverables, searching for cartographic service references, with a specific focus on OpenStreetMap. Our efforts show that OSM is the most used cartographic service in European H2020 projects in terms of mentions in the deliverable's texts, followed by Google Maps with one order of magnitude less mentions. It is worth noting that these projects involving OSM were backed by almost 4 billion euros of public money.

Based on these first interesting results, we plan to extend our scope of analysis following three axes. First, we think that it could be worth reviewing the other types of project's results such as the articles or the software source code bases. Second, we hope our approach paves the road to similar reviews of publicly-funded initiatives, and based on this observation we plan to apply our scripts to other European funding programmes. Third, additional cartographic services could also be integrated into our pipelines such as ApplePlans or other OSM-related initiatives like OpenCycleMap in order to extend the covered scope.

Finally, for reproducibility purposes, we also share on a public GitHub repository [5] all the necessary scripts to download the deliverables and generate the statistics. Furthermore, the webpage at [6] provides the reader with additional and detailed analyses together with visualisations, hoping these will help the community better understand the impact of OSM within the public European research landscape.

References

[1] European Commission (2021). Horizon 2020. Retrieved from https://ec.europa.eu/programmes/horizon2020

[2] European Commission (2020). Open access. Retrieved from https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

[3] European Commission (2021). CORDIS: EU research results. Retrieved from https://cordis.europa.eu/projects/en

[4] European Commission (2021). The official portal for European data. Retrieved from https://data.europa.eu/en

[5] Graux, D., & Michel, T. (2021). OSM in H2020. Retrieved from https://github.com/dgraux/OSM-in-H2020

[6] Graux, D., & Michel, T. (2021) OpenStreetMap in H2020 Projects: Reviewing the mentions of OpenStreetMap in H2020 European projects. Retrieved from https://dgraux.github.io/OSM-in-H2020