Proceedings of the "Geographical and Cultural Aspects of Geo-Information: Issues and Solutions" AGILE 2019 Workshop, June 17th 2019, Limassol, Cyprus

The Institutional Contexts of Volunteered Geographic Information Production: A Quantitative Exploration of OpenStreetMap Data

A. Yair Grinberger[1] yair.grinberger@uni-heidelberg.de
Moritz Schott[1] M.Schott@stud.uni-heidelberg.de
Martin Raifer[1] martin.raifer@uni-heidelbeg.de
Rafael Troilo[1] rafael.troilo@uni-heidelberg.de
Alexander Zipf[1] zipf@uni-heidelberg.de

[1] GIScience Research Group, Heidelberg University,
Im Neunheimer Feld 348, 69120, Heidelberg, Germany

Abstract

The original notion of volunteered geographical information (VGI) offers a vision of democratizing geographical information systems (GIS) via the contributions of non-expert individuals, replacing authoritative episetemologies with more open and local geographical representations. Recent studies have questioned this vision, with empirical and conceptual investigations pointing to the effects of data production procedures on the resulting representation. In practice, many organizations and social institutions hold important roles in the production of VGI, thus integrating institutional epistemologies into VGI. This paper explores the role of such institutions in the production of OpenStreetMap (OSM) data by identifying and analysing large-scale contribution events, such as data imports or organized mapping efforts. The paper deploys a global event-identification query on the historical OSM database. The results show that large-scale events are responsible for a significant portion of OSM activities, especially in relation to the creation of data. The procedure identifies several event hotspots, prevalent in either highly developed regions or developing ones. Characterizing the events according to the institutional context that drives them, the paper suggests a relation between socio-economic contexts and the integration of specific institutional perspective into local representations. Hence, the paper contributes to our understanding of VGI as a product of complex interactions of social and institutional perspectives and offers a method towards considering these in research and practice.
Keywords: OpenStreetMap, VGI, Context, Data Imports, Institutions, Remote Mapping.

1 Introduction

From their early days, online geographical information systems (GIS) were hailed as a means towards "democratizing GIS" (Butler, 2006), visioning systems based on individuals of varying skills and perceptions contributing VGI (Goodchild, 2007). Recent studies however point to conceptual and empirical issues that subvert this individual-based vision (Byrne & Pickard, 2016; Haklay, 2013, 2016; Sieber & Haklay, 2015; Stephens, 2013). According to some of these, it is impossible to understand VGI without considering contribution procedures and the technical and institutional framework that they rely upon (Fast & Rinner, 2014; Sieber & Haklay, 2015). This is especially true when large volumes of data are contributed over a short time period, termed here large-scale data production events. Such events require the cooperation of multiple individuals via some kind of organization. Given their volume and impact on data, a possible implication is significantly biasing representation towards the institutional contexts through which they emerge.

One example of this are bulk imports of ready-made datasets into OSM, events reflecting the work of certain (usually governmental) institutes and their employees. While increasing coverage, these events carry with them institutional conceptual and epistemological baggage that, when producing data not fitting well to the project's structure, may lead to representation issues (Zielstra et al., 2013). Hence, imports can enforce institutional perspectives into OSM on the expanse of more local and individual epistemologies.

OSM, a collaborative mapping project that makes a prominent VGI example, also includes other event types. For example, local chapters organize 'field mapping parties' or 'mapathons' and organizations such as the Humanitarian OSM Team (HOT) mobilize different communities to make large-scale contributions from afar. Such institutions, while operating within the OSM framework, still hold their own epistemology and enforce it through guidelines and control structures (Palen et al., 2015). These epistemologies may still be different from the ones emerging via the individual-based process initially imagined in VGI.

Hence, the existence of large-scale contribution events in OSM, while adding much to the data, still subvert the initial VGI vision in general. This paper quantitatively explores this issue by studying the spatial distribution of large-scale events and relating these to institutional and social contexts. Below, we detail the data and procedure used for identifying events, the emerging results, and their implications.

2 Methodology

2.1 Event Identification

In this paper, we base our analysis on an assumption that a generic development of OSM data for a specific area would follow three stages, similar to the model described by Gröching et al. (2014): (a) initial interest from a small number of mappers, leading to low contribution numbers; (b) an increasing interest and awareness leading to a rise in the number of mappers and/or contributions; (c) saturation of the data leading to a decrease in the number of mappers and contributions. Over time, the number of contributions will create a normal-like distribution, meaning the cumulative function would take an S-shaped form (Figure 1). Large-scale events disrupt such developments, leading the process to continue as if it jumped forward in time (see cumulative curve w/ event in Figure 1).

Based on this conceptualization, the analysis here relies on fitting a logistic curve describing the development of the cumulative number of contributions C_t over time t (equation 1; α, β, ρ and μ are scaling coefficients) to observed data within a given region. Cases when the curve underestimates actual contribution volumes are indications of events, hence we use estimation errors to identify events. However, time series errors tend to be non-stationary showing a non-random temporal pattern in errors (see errors in Figure 1). We neutralize this by using time-lagged errors to identify events, i.e. error in time *t* minus error in time *t-1*, assuming a normal distribution of lagged errors. We define here events as periods with positive and significant errors at 95% confidence.

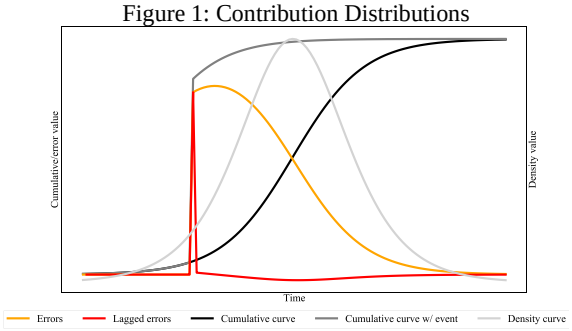
$$C_t = \frac{\alpha}{1+\rho * e^{-\beta(t-\mu)}} \quad (1)$$

2.2 Data extraction and processing

The above procedure requires producing time series data on cumulative contributions for a given spatial division and temporal resolution. For this, we have utilized the OSM History Database (OSHDB; Raifer et al., 2019) tool, which allows querying and aggregating OSM history data in a flexible way on a global scale using custom spatial divisions. The spatial division we used is based on the number of existing OSM entities – a quad-tree-like procedure starting from dividing the world into quadrants and continuing to divide each quadrant as long as the number of entities in one of its sub-quadrants is larger than 50,000. The resulting spatial system thus presents cells of varying sizes and number of entities[1]. The analysis did not consider cells with less than 20,000 entities (see Figures 2 and 4 for the resulting division). The temporal resolution we used is of one month, thus reducing the procedure's sensitivity to smaller events, and the temporal extent included all data since the beginning of the OSM project and up to April 2019.

The query designed for this research extracted for each spatio-temporal unit (i.e. for each cell and month combination) the total number of contribution actions by breaking down each contribution made during a specific month into basic operations. The number of operations in a contribution of the 'creation' type was defined to be the number of added nodes plus the number of created tags. Edit actions considered the total number of changes, i.e. the number of new nodes/tags plus the number of deleted nodes/tags. Deletion contributions were treated as one operation, since such edits can usually be carried by one click of a mouse. These operations were then aggregated to compute the monthly total. This query related to tagged nodes and ways only, excluding relations as they are responsible for only a small fraction of the data yet greatly increase computational load.

Figure 1: Contribution Distributions



Accumulating the monthly total of contribution operations for each cell over time creates the basic time-series data for the analysis detailed above (the time cumulative curve). The query also produced additional information for each spatio-temporal unit for post-processing, such as the number of active users (Users), the relative change in the number of contributions from *t-1* to *t* (Change), the maximal share of contributions made by one user (Max. Actions), the number of edited entities (Entities), the average number of geometry and tag actions per entity (Geometry Actions, Tag Actions), and the share of each contribution type out of all contributions (Deletions, Creations, Tag Changes, Geometry Changes). Notice that the choice of temporal resolution holds an implication for these statistics, meaning they may include non-event activities.

3 Results

3.1 The weights of events within OSM data

Out of 10,136 cells, 494 (4.9%) produced errors during the curve fitting procedure. For the remaining 9,642 cells, the procedure identified 56,578 events (5.9 events per cell, maximum of 19 events in one cell). These events produced 808,117,670 contributions and 6,318,493,481 actions, i.e. 14,283 contributions and 111,677 actions per event (maximum of 2,064,875 contributions and 12,851,643 actions).

To understand the impact of events on OSM, these figures were compared with the total number of contributions and actions in the history of OSM (Table 1). The weight of events is significant, with more than 40% of actions and contributions originating from events. Events especially dominate data creations with more than half of the data ever created in OSM attributed to events. While these results surely include some overestimations relating to the temporal resolution of the analysis, the volume of these events and the lack of results for 4.9% of the cells due to error probably compensate for this. Even so, eliminating the lower decile of events from the analysis (i.e. treating these as false positives) still results in

Table 1: Events' weight in OSM data

Measure	Entire OSM History	Events	% in Events	Median % per Cell	Interquartile Range
Total actions	$1.3*10^{10}$	$6.3*10^{9}$	46.7%	45.7%	26.2%
Geometry actions	$9.5*10^{9}$	$4.2*10^{9}$	44.1%	43.4%	26.9%
Tag actions	$3.9*10^{9}$	$2.1*10^{9}$	53.4%	46.9%	33.8%
Total contributions	$1.9*10^{9}$	$8.1*10^{8}$	41.5%	39.5%	25.6%
Creation contributions	$9.5*10^{8}$	$5.0*10^{8}$	52.4%	50.1%	35.9%
Deletion contributions	$1.3*10^{8}$	$4.3*10^{7}$	33.0%	25.0%	35.9%
Tag change contributions	$4.7*10^{8}$	$1.7*10^{8}$	36.4%	20.6%	29.8%
Geometry change contributions	$4.0*10^{8}$	$9.7*10^{7}$	24.4%	22.7%	27.3%

events representing 41.0% of contributions and 45.9% of actions. Hence, events are a significant driver of OSM data.

Breaking down the share of events in contributions by cell (Figure 2), exposes an uneven distribution with hotspots of event impacts existing in areas such as western and eastern Africa, Indonesia and the Philippines, Nepal, U.S.A, Canada, and to a certain extent Japan, France, Poland, Norway, and Italy. This uneven distribution of institutionalized contributions and hotspots within very different regions suggests the impact of other contextual influences the pattern of events.

3.2 Types and distributions of events

As a means towards exploring such influences and the different characteristics of events (as mentioned in the introduction), we have used the k-means clustering procedure to group events. The variables used for this were the maximal share of actions by one user (Max. actions, percentage) and the share (in percentage) of each type of contribution type out of all contributions, as these represent how centralized this contribution was and on what kind of themes/operations it focused. The procedure clustered events into six groups. To determine the number of clusters, we have computed several cluster separation measures (Davies-Bouldin index, the silhouette coefficient, and the Calinski-Harabasz score) for a range of *k* values. While these produced the optimal values for k=4, this result was judged as too restrictive in terms of representing the diversity of events. The separation measures did not agree on which *k* makes the second-best choice (ranging from 6 to 8) and thus we based our decision on a visual analysis of clustering results.

Figure 3 shows for each cluster the average values of the clustering variables and other available data using parallel coordinates. These allow distinguishing and labelling clusters. Four clusters show high Max. Actions values, meaning one user made most of the contributions, i.e. pointing to a bulk data

Figure 2: Events' share in OSM contributions by cell

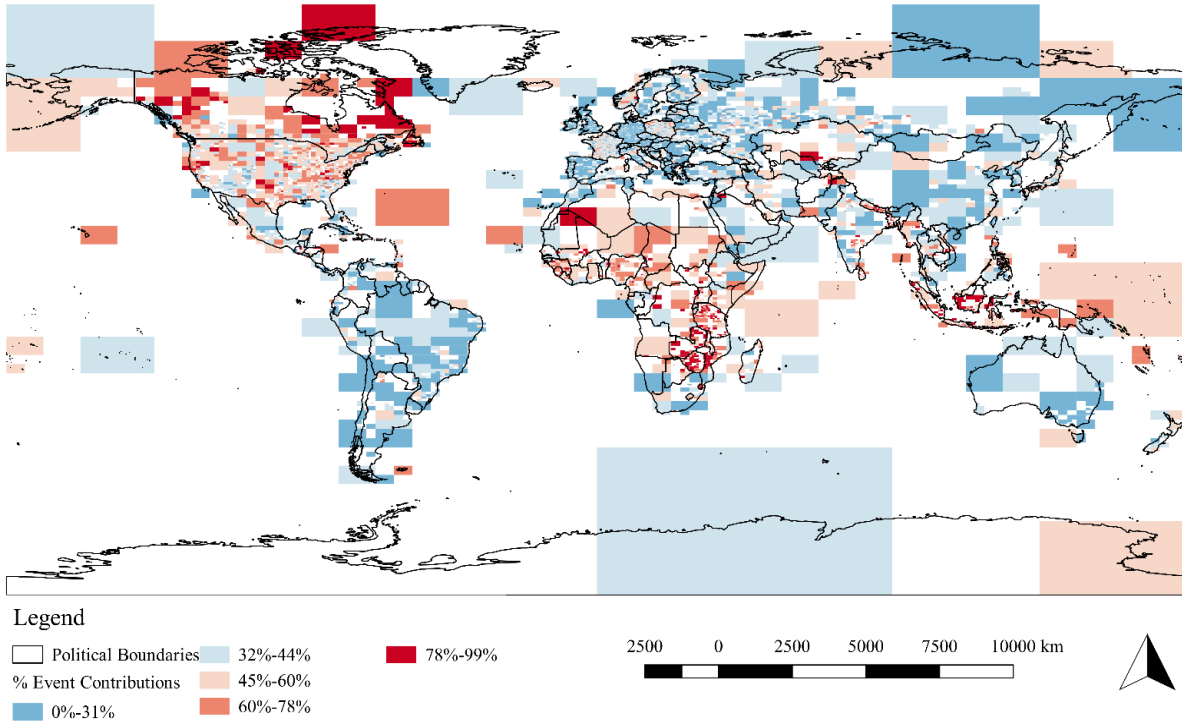
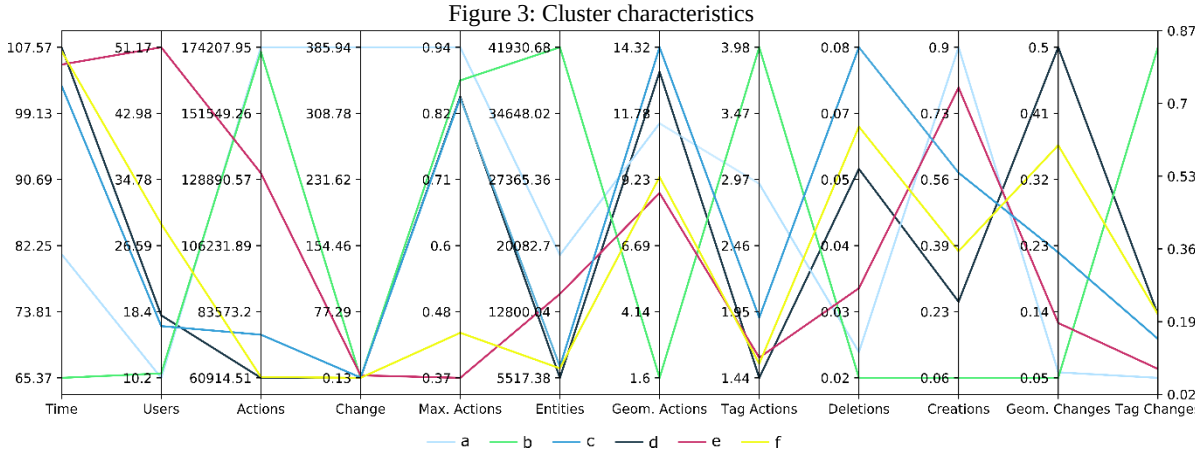


Figure 3: Cluster characteristics



import. Variables such as the share of contribution types and time (number of months since the first contribution to the area) differentiate between these imports (see Fig. 3):

(a) Early imports – the term early refers here both to chronology (*t* value) and to the event's timing – these events take place relatively early and create a very large effect (average change value of 386%), pointing to an underdeveloped database. Not surprisingly, these events mostly add new data, with creations making 90% of all contributions on average.

(b) Tag imports – another type of early imports including mostly tag operations (more than 85% of contributions, almost 4 tag actions per entity). Despite having high contribution volumes on average, these events do not affect geometry much. Incidentally, these take place mostly in the U.S.A.

(c) Late imports – these are bulk imports taking place in a more mature data region, hence change values are low, creations shares are still high, but geometry and tag changes become more prevalent.

(d) Data updates – this may represent the most 'mature' import, where creations receive less weight and the primary activity is updating of geometries, as evident also in the average number of geometry actions per entity.

The two other types present a more distributed kind of large-scale contributions, with actions spread across more users:

(e) Remote mapping event – representing the kind of practices common within HOT tasks, such events include high creation volumes but less tagging activity, indicative of little local knowledge. The average number of users however is very high, thus producing large contribution volumes.

(f) Local mapping event – while similar to remote mapping events in many aspects, these events still show much more focused work and local knowledge, as evident in the relatively high shares of tagging and geometry update contributions and low average number of edited entities.

In the context of institutional epistemologies, event types a-d conceptually seem to represent the same phenomenon – an import of a governmental/external epistemology into OSM. These make the majority of events (70.8% of all events; Table 2) with early and late imports being the most common types. The last two, representing the 3rd and 4th most common types (Table 2), do show difference, as the first represents the epistemological stance of the institute mobilizing the global community, mostly HOT, while the other represents more local epistemologies.

Identifying the most common event type for each cell (Figure 4) and comparing with Figure 2 suggests a pattern. Visually, there seems to be a correlation between event hotspots and event types, mediated by the socio-economic status of the region: late imports dominant the more affluent countries (Japan, France, Poland, Norway, Canada, with the U.S.A. dominated by tag imports) while remote mapping events being more common in the more developing economies (e.g. Indonesia, Eastern and Western Africa). Interestingly, many areas presenting lower event impacts are ones where early imports are most common. These include highly developed economies (e.g. Germany, Spain, the U.K., the European part of Russia, and most major urban areas of Australia), along with some emerging economies (e.g. eastern parts of China and parts of India).

Comparing events discussed in previous studies to the results here validates our results, showing these events were identified and correctly classified for the most part (Table 3). The exceptions are the 2009 Gaza Strip event, caused by multiple local contributions aggregated into one contribution, and some cases of the May 2015 event in Nepal, perhaps pointing to the fieldwork of the Katmandu Living Labs organization and the volunteers it attracted.

Table 2: Events by type

Event type	Frequency	Percentage
Early imports	15,852	28.0%
Tag imports	3,218	5.7%
Late imports	13,901	24.6%
Data updates	7,090	12.5%
Remote mapping events	7,244	12.8%
Local mapping events	9,273	16.4%

Figure 4: Most common event type, by cell

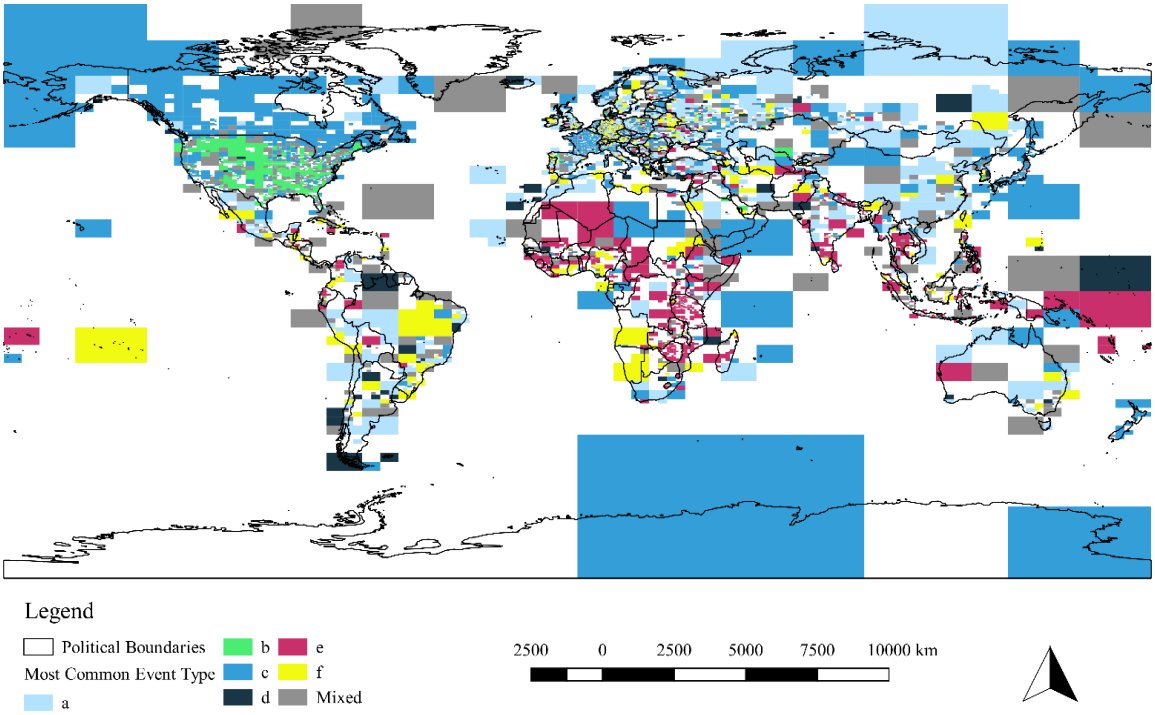


Table 3: Validation of events

Event location and time	Source	Details	Classification by the procedure
Gaza Strip, September 2009	Grinberger, 2018	Bulk import of the work of multiple local mappers	Early import
Gaza Strip, Summer 2014		HOT project	Remote mapping event
Tel Aviv, December 2012		Bulk import of official data	Early import
Tel Aviv, January 2013		Deletion of redundant data and tags after import	Tag import
Nepal, April and May 2015	Poiani et al., 2016	HOT project	Remote mapping wvent; the May 2015 portion of the event was classified as a local mapping event for several cells

4 Discussion and Conclusions

In this paper we have set out to evaluate the individual-driven vision of VGI by investigating large-scale contributions to OSM. The results here allow quantitatively assessing the relevance of this vision, showing that a significant share of the activity in OSM relies on some form of organized contribution, either that of an external data-collecting agency imported into OSM or of organizations operating within this project's framework. Hence, OSM data relies very much on, or is a product of, the work of institutional mediators that are not included in the original vision.

While such a pattern is not inherently problematic, it does hold the potential for introducing bias into representation in OSM. In the case of bulk imports, this may be caused when the workings of a small group of experts (those who created the data and those importing them) replace the democratic concept of crowdsourced contribution. Mapping events organized by local chapters or HOT, on the other hand, enforce epistemologies derived from these institutes' agendas via the organization and direction of data collection efforts. These epistemologies may be different than those emerging otherwise, e.g. when remote mapping events increase the involvement of non-local mappers in an area.

The results pertaining to the spatial patterns and types of events expose such potential impacts, also pointing to their complex relations to geo-social contexts. The negative correlation between the frequency of early import events and the weights of events in total data found for affluent and emerging economies[2] suggests that socio-economic context is both the driving force behind the 'problem' (institutional epistemologies dominating the data) and the 'solution' (an active local community reshaping the data). Imports require a minimal population of educated, skilled, and engaged mappers,

the kind of mappers that also make more competent individual contributors. In less developed economies, such mappers are harder to come by, meaning that the impacts of remote mapping events, typical of such regions, tend to last. Hence, while such events rely more on the contributions of individual mappers, they seem to fossilize an institutional perspective which was originated outside of these areas and do not necessarily reflect local views, needs, and perspectives.

With these results and the ability to compare trends across regions, this paper contributes to our understanding of the social, geographical, and institutional contingency of OSM data and procedures. The question remains whether this phenomenon is endemic to OSM, or whether it is common within VGI. In principle, even projects such as citizen reports on vandalism or biodiversity have parallel institutional databases that could be imported, yet such occasions may still be rare. Even so, as OSM makes perhaps the most celebrated and widely utilized VGI project, this issue requires further attention, especially given the increasing impact of corporate mappers on the data (Anderson et al., 2019). Future steps of the analysis would include looking at individual events, measuring their specific impacts and studying the development of data after these. Doing so would allow producing a deeper understanding of the interplay between local communities, institutions, social contexts, and data, pointing towards possible steps and interventions to institutional practices in OSM.

Endnotes

[1] While not considering human perceptions or administrative borders, this spatial division still captuers in most cases regional differences, at least at the national scale (see figure 2).

[2] Using the following definition: affluent economies - western Europe, U.S.A, and Australia; emerging economies - China and India; least developed areas - Sub-Saharan Africa and parts of the south-east Asia and Oceania.

Acknowledgements

We thank the two anonymous reviewers for their useful comments. This research was supported by the Humboldt Research Fellowship for Postdoctoral Researchers.

References

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

Butler, D. (2006) Virtual globes: The web-wide world. *Nature*, 439(16), 776-778.

Byrne, D. and Pickard, A. J. (2016) Neogeography and the democratization of GIS: A metasynthesis of qualitative research. *Information, Communication & Scoeity*, 19(11), 1505-1522.

Fast, V. and Rinner, C. (2014) A systems perspective on volunteered geographic information. *ISPRS International Journal of Geo-Information*, 3(4), 1278-1292.

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

Grinberger, A. Y. (2018) Identifying the effects of mobility domains on VGI: Towards an analytical approach. *Short Paper presented at VGI-ALIVE Workshop at AGILE 2018 Conference*, June 2018, Lund.

Gröching, S., Brunauer, R. and Rehrl K. (2014) Digging into the history of VGI data-sets: Results from a worldwide study on OpenStreetMap mapping activity. *Journal of Location Bassed Services*, 8(3), 198-210.

Haklay, M. (2013) Neogeography and the delusion of democratization. *Environment and Planning A: Economy and Space*, 45(1), 55-69.

Haklay, M. (2016) Why is participation inequality important? In: Capineri, C., Haklay, M., Huang, H., Antoniou, V. Kettunen, J., Ostermann, F. & Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*. London, Ubiquity Press, pp. 35-44.

Palen, L., Soden, R., Anderson, T. J. and Barrenechea, M. (2015) Success & scale in a data-producing organization: The socio-technical evolution of OpenStreetMap in response to humanitarian events. In: Mayer, T. & Do, E. Y.-L. (eds.) *Proceedings of the 33rd Annual CHI Conference on Human Factors in Computing Systems.* New York, The Association for Computing Machinery, 4113-4122.

Poiani, T. H., Rocha, R. d. S., Degrossi, L. C. and Albuquerque, J. P. d. (2016) Potential of collaborative mapping for disaster relief: A case study of OpenStreetMap in the Nepal earthquake 2015, *2016 49th Hawaii International Conference on System Sciences (HICSS)*, Koloa, HI, pp. 188-197.

Raifer, M., Troilo, R., Kowatsch, F., Auer, M., Loos, L., Marx, S., Przybill, K., Fendrich, S., Mocnik, F.-B. and Zipf. A. (2019) OSHDB: A framework for spatio-temporal analysis of OpenStreetMap history data. *Open Geospatial Data, Software and Standards*, 4(3), 1-12.

Sieber, R. E. and Haklay, M. (2015) The epistemology(s) of volunteered geographic information: A critique. *Geo: Geography and Environment*, 2(2), 122-136.

Stephens, M. (2013) Gender and the GeoWeb: Divisions in the production of user-generated cartographic information. *GeoJournal*, 78(6), 981-996.

Zielstra, D., Hochmair, H., H. and Neis, P. (2013) Assessing the effect of data imports on the completeness of OpenStreetMap – A United States case study. *Transactions in GIS*, 17(3), 315-334.