How do volunteer mappers use crowdsourced Mapillary street level images to enrich OpenStreetMap?

Levente Juhász
University of Florida
3205 College Ave.
Ft. Lauderdale, FL, USA
levente.juhasz@ufl.edu

Hartwig H. Hochmair
University of Florida
3205 College Ave.
Ft. Lauderdale, FL, USA
hhhochmair@ufl.edu

Abstract

An increasing number of crowdsourced geo-data repositories and their services allows volunteer mappers to utilize information from various data sources when contributing data to a crowd-sourced mapping platform. This study explores to which extent OpenStreetMap (OSM) contributors use the crowdsourced street level photo service Mapillary to derive mappable data for OSM during their editing sessions in the iD and JOSM editors. We cross-check the location of OSM edits with the geographic areas from which OSM contributors loaded Mapillary images into the editors to determine which OSM edits could have been based on information from Mapillary images. The findings suggest that OSM mappers are beginning to utilize information from street level images in their mapping workflow. This observed "cross-viewing" pattern between different datasets indicates that the use of data from one VGI platform to enhance that of another is a real phenomenon, leading to implications for VGI data quality.

Keywords: OpenStreetMap, Mapillary, VGI, data quality, cross-viewing, contribution behavior.

1 Introduction

Volunteered Geographic Information (VGI) (Goodchild, 2007) has been recognized as a valuable resource for the GI community to complement data from traditional sources, such as the census or aerial photographs. To assess the quality of heterogeneous VGI sources, studying contributor behaviour is essential (Elwood et al., 2012, Budhathoki and Haythornthwaite, 2013). (Bégin et al., 2013) incorporated editing sessions (OSM changesets) in their analysis to better understand characteristics and quality of collected VGI. Results show that new changesets of a contributor usually extend or overlap spatially earlier changesets and add lower priority features or new attributes.

OSM is arguably the most widely studied VGI platform. It was shown that OSM positional accuracy is better when more mappers edit the same area (Haklay et al., 2010) and that users are more likely to edit a greater variety of features in their home region than in external regions (Zielstra et al., 2014). Whereas VGI mappers rely primarily on their local knowledge for data contribution and editing, incorporating other data sources can improve data quality. Examples include tracing of features from high resolution aerial imagery (Haklay, 2010) or importing governmental data (Zielstra et al., 2013).

Mapillary's crowdsourced street level imagery is a unique addition to the list of available VGI sources. With the introduction of Mapillary in 2014 and the open license that it provides, OSM contributors can now use Mapillary image content to derive information that is not visible on aerial imagery (e.g. the type of a traffic sign) or to map features that would require in person exploration through field surveys (Juhász and Hochmair, 2016a). Evidence of OSM contributors that use Mapillary imagery to derive feature information was found by analyzing OSM edits that reference Mapillary in their tags (typically the source tags), which is referred to as cross-tagging (Juhász and Hochmair, 2016b). However, as tagging in OSM is inconsistent and contributors often follow tagging suggestions only poorly (Davidovic et al., 2016), any crowd-sourced data sets that were used for OSM edits (e.g. Mapillary, Flickr) may not be completely referenced in OSM tag content, calling for alternative methods to identify data use across different VGI platforms in data editing sessions.

This paper analyzes the viewing extents of the Mapillary image layer during OSM editing sessions with the iD and JOSM editors to estimate to what extent Mapillary images were likely used as a source for OSM edits.

2 Materials and methods

This section provides an overview of the data sources and the data processing methods used in this research. The goal of the data extraction was to identify individual OSM feature edits around the world that were likely based on Mapillary photos. Such edits would have to be made in the geographic area where and around the time when the user viewed the Mapillary image layer in one of the OSM editors.

2.1 Data sources

2.1.1 OpenStreetMap

Since this research studies the editing behavior of volunteer mappers, a full OSM history dump[1] was used which includes

[1] http://planet.openstreetmap.org/pbf/full-history/

all historical edits ever made to the database. Due to the large data volume, the pbf file was first split into world regions and then imported to a spatially enabled PostgreSQL database, using the osm-history-splitter and osm-history-importer tools.

2.1.2 Mapillary

Recent versions of the iD and JOSM editors are capable of loading Mapillary images into their editing environments. These requests which originate from the editors eventually leave footprints on Mapillary servers which can be expressed as a geographic area corresponding to the viewing extent of the editor. Mapillary provided us with a data dump of all viewing requests of the Mapillary layer together with their spatial extents and time stamps. For this analysis we used worldwide Mapillary viewing extent data that was collected between June 2015 and February 2016. In addition to this, another Mapillary data dump with individual photo locations was used to exclude OSM edits that are far from street level imagery and therefore probably not based on Mapillary imagery. Mapillary viewing and photo location data was stored in the same PostgreSQL database.

2.2 Data preparation and processing

2.2.1 Workflow

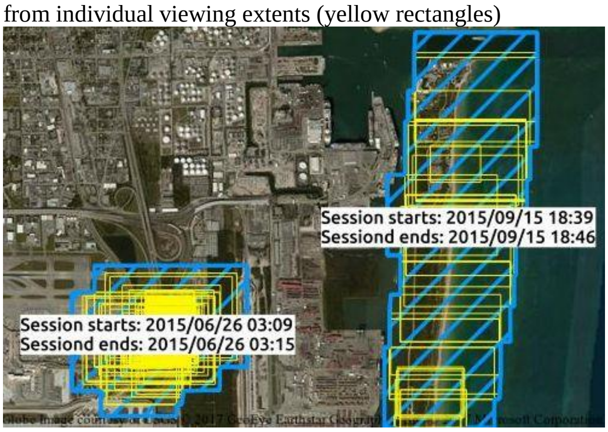
The size of a full OSM history dump of over 50 GB in the pbf format as well as millions of Mapillary viewing extents made it necessary to split the database into smaller tables corresponding to world regions. Also, custom indexes were constructed to speed up data extraction with SQL queries. The final database contained more than 5 billion rows (OSM edits and Mapillary viewing extents) and occupied approximately 1.7 TB of disk space.

Based on this customized database structure, a two-tiered data extraction approach was applied. The first step involved the extraction of OSM candidate features through a coarse spatio-temporal match between image viewing windows and OSM edits (Section 2.2.2 – 2.2.3). This reduced the database size for the second step, which involved more refined spatio-temporal overlay operations (Section 2.2.4).

2.2.2 Extraction of editing sessions

We use the term “editing session” to describe an uninterrupted time period within which Mapillary images are being loaded into the OSM editor from the same machine as part of the layer viewing request. Since the server-logged viewing data does not contain a unique OSM user identifier, we used the IP addresses associated with each request to aggregate the viewing data over a set time period. More specifically, for each IP address, a timeline was constructed that shows time stamps of user activities on the Mapillary layer, such as changing the viewing extent. An arbitrary one-hour threshold of idle time (i.e. no image requests) was used to construct separate editing sessions from the timeline. This one-hour threshold corresponds to the time period after which the OSM API closes changesets[2] if no more edits are made. The example in Figure 1 illustrates how numerous individual Mapillary layer viewing extents (yellow rectangles) were aggregated into two distinct editing sessions (blue polygons with hatched areas). These editing sessions have start and end timestamps and a geographic area that can be disjoint (which is not in the provided examples, though). This aggregation allows to reduce data volume without losing information about the editing activity.

Figure 1: Editing sessions (blue hatched polygons) aggregated from individual viewing extents (yellow rectangles)



2.2.3 Extraction of candidate features

Candidate features are OSM editing events (i.e., creation, modification) in the spatial and temporal proximity of Mapillary editing sessions. Since topological operations and comparison of event timestamps are resource intensive, identification of a coarse set of candidate features uses the spatial and temporal index constructed in the PostgreSQL database. That is, instead of checking the specific spatial relations between OSM editing events and Mapillary editing sessions, the database was instructed to utilize related spatial indexes to determine a potential spatial overlap between candidate features and Mapillary editing sessions. Similarly, instead of comparing specific timestamps (with the precision of milliseconds), an index built on the day of editing events was used to identify OSM editing events and editing sessions taking place on the same day. The extraction of candidate features uses therefore a coarse comparison of spatial extents and event times, which results in an overestimation of candidate features compared to the number of actual potential OSM edits based on Mapillary images.

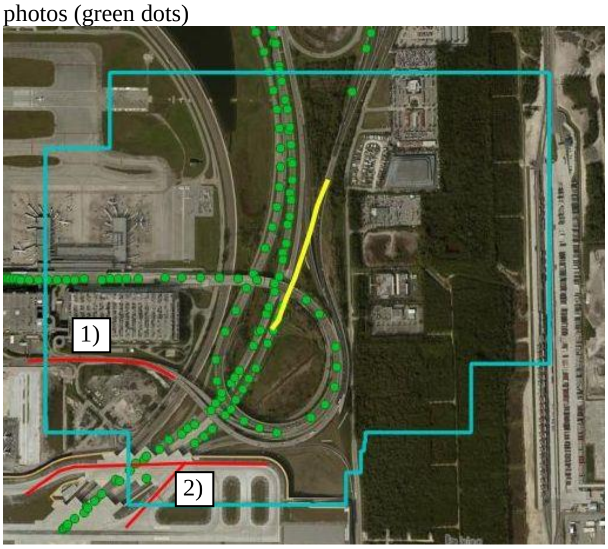
2.2.4 Extraction of OSM edits likely based on Mapillary

Next, a more refined filtering method was applied on candidate features for data extraction. Within this step, only those OSM editing events were retained for further analysis that were conducted after the start time of a Mapillary editing session and that were completed within one hour of the session end time. Since submission of an OSM changeset is

2 https://wiki.openstreetmap.org/wiki/API_v0.6

not automated (i.e. users need to send their changes in a separate step), this threshold is to allow some time in case users turned off the Mapillary layer before submitting their OSM changesets. In addition, candidate features further than 25m from the actual location of Mapillary photos were excluded. Figure 2 highlights the results of this filtering process. It shows an OSM edit that is considered to be Mapillary related[3] (yellow line) as well as other OSM candidate features (red lines) along with the location of Mapillary street level photos (green dots). In this example, the retained edit denotes a new highway exit added to OSM. The remaining candidate features overlapping with this session shown in red were excluded from the result set because they were either further than 25m from the imagery (see (1)) or they were added to OSM at a time that did not align with the editing session (see (2)).

Figure 2: Retained OSM edit based on Mapillary (yellow), excluded candidate features (red), and location of Mapillary photos (green dots)



3 Results

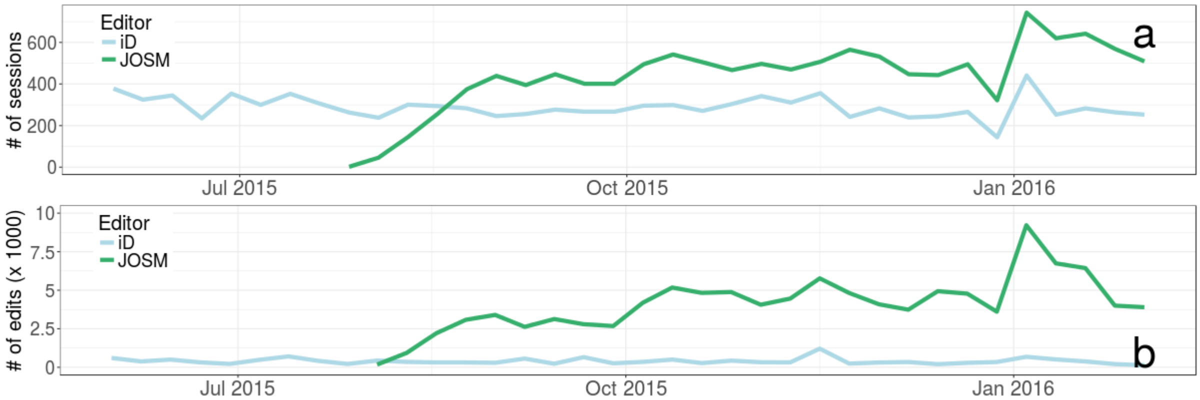
A total of 34,000 Mapillary editing sessions were identified between June 2015 and February 2016 out of which 8400 contained only a single Mapillary viewing request. The latter means either that (1) the user accidentally turned on the Mapillary imagery layer in the OSM editor, or (2) that there were no images available for that area so that the user turned off the layer immediately. Editing sessions with only request were therefore excluded from further analysis. A Mapillary viewing session lasted for 7 minutes and 39 seconds on average. This is the duration users spent on mapping OSM features while viewing Mapillary street level photos in the OSM editor. The longest observed session lasted for 5 hours and covered a large area along a highway in Belgium.

The popularity of Mapillary images used in the OSM editing workflow can be assessed from the number of editing sessions per week. Figure 3a shows this information for both analysed editors. As can be seen, at the beginning of the study period (June-July 2015), the Mapillary imagery layer was only accessibly within the iD editor. It became available in the JOSM editor in August 2015 as well. The average number of sessions peer week was 283 for iD and 441 for JOSM (after it became available).

With the extraction of OSM edits based on spatial and temporal constraints described in Section 2.2.4, the number of OSM edits per week for both editors can be computed as well (Figure 3b). The figure illustrates the higher popularity of the JOSM editor compared to iD when it comes to Mapillary image use for OSM edits. On average, 400 feature weekly edits originated from the iD editor as opposed to 4100 coming from JOSM. The clear preference for JOSM over iD was not expected, given that − at least − Novice users use iD more frequently than JOSM to edit OSM data (Yang et al., 2016). During the most active week (starting on January 4, 2016) almost 10,000 OSM map edits were identified.

Figure 4 plots the number of different OSM users who use the Mapillary layer function for OSM feature edits within a given week. The number clearly increases after the layer functionality became available in JOSM in August 2015. 980 unique users were found to contribute to OSM based on Mapillary layer views during the study period.

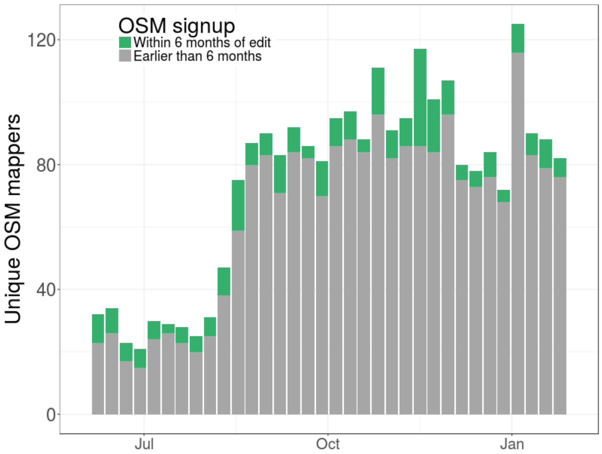
Figure 3: Number of OSM-Mapillary editing sessions per week, grouped by editor (a), and number of unique OSM mappers engaging in photo-mapping per week (b)



3 http://www.openstreetmap.org/way/356400079/history

To analyze the level of experience of OSM users who use the Mapillary layer service for feature editing, the sign up dates of these users were extracted from the main OSM API. Figure 4 shows the weekly number of analyzed OSM users by signup date. The bar chart suggests that novel users are quite active in utilizing the Mapillary layer feature. On average, 14% of weekly active users signed up to OSM within just six months before their editing activity. The proportion of weekly novice users ranged from 5% to 29%. When setting this limit to one month before editing based on Mapillary photos, this weekly rate is still 8% on average. More detailed analysis shows that almost 30% of all analyzed users created their OSM accounts after the introduction of Mapillary in 2014.

Figure 4: Weekly aggregated number of photo-mapping OSM users



The histogram of user sign up dates to OSM supports this general finding (Figure 5). A clear peak of new users signing up at the beginning of 2015 suggests that photo-mapping does not require one to be overly experienced with OSM. This peak could be the result of a special promotion activity conducted by Mapillary. Several meetups were organized to introduce Mapillary to wider audiences where Mapillary team members were present in multiple conferences and community events to promote the service. These promotions might have triggered a new crowd of mappers to sign up to OSM and to start with mapping from street level imagery information shortly after creating their OSM and Mapillary accounts.

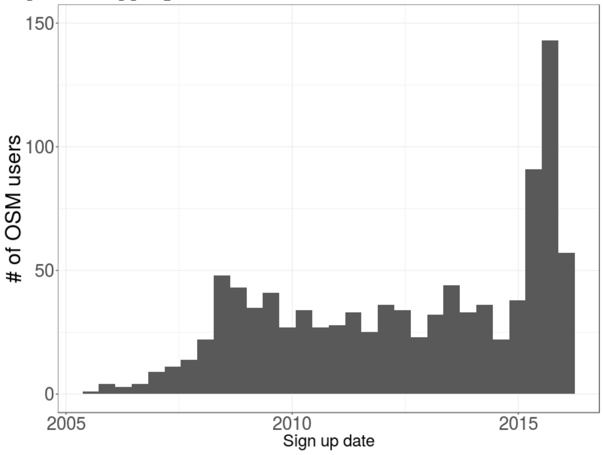
4 Discussion and future work

This paper examined to which extent OSM mappers use Mapillary imagery in their editing workflow. We used the viewing extents of Mapillary image requests submitted by the iD and JOSM editors, which provides a unique opportunity to study mapping behavior. These so-called editing sessions were spatio-temporally matched with a full history dump of the OSM database to extract those OSM edits that could be based on street level photos.

Although weekly counts of OSM feature edits based on Mapillary images are low compared to the number of all OSM feature edits submitted per week, our findings indicate that there is a certain group of OSM mappers who "cross-view" different VGI data sources for mapping purposes and, more specifically, use the crowdsourced Mapillary imagery service to do so. Studying the sign up date of those mappers who engage in this activity also indicate that, although this process is more complex than just drawing lines on top of aerial imagery, novice OSM mappers use Mapillary information, too, and provide valuable contributions by connecting these two VGI sources.

Our database contains more detailed information about the type of Mapillary related OSM edits, such as name changes. Therefore, we plan to extend the analysis to study what kind of information has been obtained from street level photos for subsequent OSM edits.

Figure 5: Histogram of sign up dates of OSM users engaging in photo-mapping



Since data quality is one of the most important aspects of VGI analysis, we plan for future work also to determine which improvements in VGI data quality can be associated with the re-use of VGI between multiple platforms.

References

Bégin, D., Devillers, R. and Roche, S. (2013) Assessing volunteered geographic information (VGI) quality based on contributors' mapping behaviours. In: *Proceedings of the 8th international symposium on spatial data quality ISSDQ*, Hong Kong. pp. 149-154, 2013.

Budhathoki, N. R. and Haythornthwaite, C. (2013) Motivation for open collaboration crowd and community models and the case of OpenStreetMap. *American Behavioral Scientist*, 57(5)**,** 548-575.

Davidovic, N., Mooney, P., Stoimenov, L. and Minghini, M. (2016) Tagging in Volunteered Geographic Information: An Analysis of Tagging Practices for Cities and Urban Regions in OpenStreetMap. *ISPRS International Journal of Geo-Information,* 5(12)**,** 232.

Elwood, S., Goodchild, M. F. and Sui, D. Z. (2012) Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the association of American geographers,* 102(3)**,** 571-590.

Goodchild, M. F. (2007) Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0 (Editorial). *International Journal of Spatial Data Infrastructures Research (IJSDIR),* 2**,** 24-32.

Haklay, M. (2010) How good is Volunteered Geographical Information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design,* 37(4)**,** 682-703.

Haklay, M., Basiouka, S., Antoniou, V. and 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.

Juhász, L. and Hochmair, H. (2016a) User Contribution Patterns and Completeness Evaluation of Mapillary, a Crowdsourced Street Level Photo Service. *Transactions in GIS,* 20(6)**,** 925-947.

Juhász, L. and Hochmair, H. (2016b) Cross-Linkage between Mapillary Street Level Photos and OSM Edits. In*:* Sarjakoski, T., Santos, M. Y. and Sarjakoski, L. T. (eds.) *Geospatial Data in a Changing World (Lecture Notes in Geoinformation and Cartography).* Berlin, Springer, pp. 141-156.

Yang, A., Fan, H. and Jing, N. (2016) Amateur or professional: Assessing the expertise of major contributors in OpenStreetMap based on contributing behaviors. *ISPRS International Journal of Geo-Information,* 5(2)**,** 21.

Zielstra, D., Hochmair, H., Neis, P. and Tonini, F. (2014) Areal delineation of home regions from contribution and editing patterns in OpenStreetMap. *ISPRS International Journal of Geo-Information,* 3(4)**,** 1211-1233.

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.