QualiOSM: Improving Data Quality in the Collaborative Mapping Tool OpenStreetMap

Gabriel F. B. de Medeiros[1], Lívia C. Degrossi[2], Maristela Holanda[1],

[1]Departamento de Ciências da Computação – Universidade de Brasília (UnB)
Brasília – DF – Brasil

[2]Fundação Getúlio Vargas (FGV)
São Paulo – SP – Brasil

{gabriel.medeiros93,liviadegrossi}@gmail.com, mholanda@unb.br

Abstract. *The collaborative mapping tool OpenStreetMap (OSM) has a large database in which thousands of users are able to insert, edit and delete geographic data from the Earth's surface. As evidenced in multiple studies, collaborative tools tend to have a lack of data quality, since the information is often provided by inexperienced users. Due to its complexity, the quality of geographic data can be measured based on different aspects, which have been called quality dimensions in literature. In this context, this paper proposes the implementation of the QualiOSM tool in order to improve the quality dimension of attribute completeness within OpenStreetMap platform, increasing the address information associated with objects. The tool was tested in two different scenarios in Brazil: the city center of Brasilia, capital of the country, and part of the city of Rio Branco, in the state of Acre.*

1. Introduction

The activities of mapping and spatial data collection have undergone drastic changes in recent decades, due to factors such as the use of georeferencing, the emergence of devices with integrated GPS, the improvement of broadband internet and the development of high quality graphics. These new technologies have given rise to systems in which users are able to generate geographic information on a voluntary basis, thus the information contained in these types of systems has become popularly known as Volunteered Geographic Information (VGI) [Goodchild 2007], or more broadly, Crowdsourced Geographic Information (CGI) [See et al. 2016].

The increasing availability of CGI has drawn the attention of authors in the search for methods to assess data quality in collaborative activities [Degrossi et al. 2018], which were divided into three different categories: social media, collaborative mapping and crowd sensing [de Albuquerque et al. 2016]. In recent years, the proliferation of social computing practices has increased the amount of content generated by users online. This fact has brought positive and negative effects in relation to the study of geographic data and CGI [Meng et al. 2017]. On the one hand, the use of volunteers has enabled mapping of the most remote areas of the planet, where access is more difficult. On the other hand, collaborative data has brought difficulties regarding the degree of veracity of geographic information [Flanagin and Metzger 2008].

One of the challenges that researchers have in discussing, evaluating and measuring data quality is that it depends on different factors, like the characteristics of the

volunteer and the type of information collected [Bordogna et al. 2016]. Thus, the concept of quality was divided into different aspects, which were called quality dimensions. In this way, some quality dimensions explored in the literature are accuracy, completeness, logical consistency and reliability [Firmani et al. 2016]. This work focuses on the dimension of completeness, represented as the proportion between the presence of metadata associated with a set of objects compared to the total number of objects in that set [Sehra et al. 2017].

In a CGI, the metadata is usually stored in the format of tags, which are treated as a key-value pair associated with the object in order to add new information. In most collaborative systems, users create or send content, make the notes they want using tags and share this information with other users, who can make any edits they deem necessary. The process of adding tags, also called tagging, has been described as one of the dilemmas associated with the behavior of users on Web 2.0, since incorrect tagging leads to unsatisfactory results in relation to the completeness of the information [Liu et al. 2011].

A successful example of CGI is the collaborative mapping tool OpenStreetMap (OSM), used in this paper as a case study. The data provided by the volunteers, as in OSM, requires special attention regarding the quality of the information, since users actively participate in the processes of editing, inclusion and exclusion of objects. One of the main reasons for the lack of data quality in these types of tools is the great heterogeneity observed in relation to its users, as they use different technologies and have different levels of knowledge [Senaratne et al. 2017].

In this context, this paper presents the QualiOSM tool, in order to improve the completeness of objects within the OpenStreetMap tool through the implementation of an automatic tag adder for adding address information to objects. The tests were carried out based on data collection in two different scenarios in the country of Brazil, taking into account the urban centers of the city of Brasilia, in the Federal District and Rio Branco, in the state of Acre.

The rest of this paper is structured as follows: Section 2 presents a set of works related to the theme of this research; Section 3 describes the implemented tool QualiOSM, as well as the methodology and architecture used for its development; Section 4 describes how data from Brazil was collected and later divided into the two test scenarios for using the tool; Section 5 presents the results obtained from the use of the tag adder implemented within the QualiOSM tool; finally, Section 6 presents the conclusion and future work.

2. Related Work

There are several studies in the literature that explored the process of adding tags in collaborative tools. For example, [Ames and Naaman 2007] explored the motivation for attributing tags to images on Flickr, concluding that most users tag objects to make information more accessible to the general public. In addition, [Kennedy et al. 2006] evaluated the performance of trained classifiers with photos from Flickr and their associated tags, demonstrating that tags provided by users contains a lot of misinformation.

In relation to the collaborative mapping tools, [Codescu et al. 2011] organized an ontology in order to standardize and facilitate the hierarchy of tags within the OpenStreetMap tool, but concluded that the use of an ontology is only efficient if users keep the tags constantly updated within OSM platform.

Still within OpenStreetMap, [Mooney and Corcoran 2012] carried out the analysis of more than 25,000 objects in the database of Ireland, United Kingdom, Germany and Austria. The results indicated that there are some problems arising from the way users assign tags to objects in OSM. The study also showed that these identified problems are a combination of the flexibility of the tagging process and the lack of a more rigid mechanism to verify the adherence to the OpenStreetMap ontology in relation to the tags added by its users.

Besides that, [Davidovic et al. 2016] used the recommendations provided on the "Map Features" page from the Wiki of the OpenStreetMap project[1] and analyzed the OSM database in forty cities around the world to see if contributors in these urban areas were using the guidelines in their tagging practices. The study concluded that compliance with the suggestions and guidelines is generally average or poor, since users in these areas do not always have the same level of knowledge.

Differently from the works mentioned above, this work proposes the implementation of the QualiOSM tool in order to improve the quality of geographic information within OpenStreetMap, especially with regard to the process of assigning address tags to objects. Thus, the intention of the tool is to contribute to the completeness of address information of objects in the OSM platform, assisting in automating the insertion of this information in the OSM platform.

3. QualiOSM

The QualiOSM tool was developed with the purpose of improving the completeness of address information associated with objects on the OpenStreetMap platform. Implemented as an extension (plugin) within the Java OpenStreetMap Editor (JOSM)[2], responsible for the largest number of object edits within the OSM platform, the application was written in Java programming language and can be downloaded from a public repository in Github[3].

Analyzing statistics present on the website TagInfo[4], it was observed that among the five most used tags for OpenStreetMap points, four are address tags ("addr:housenumber", "addr:street", "addr:city" and "addr:postcode"). It was also possible to observe that these four tags are included among the ten tags most used both for lines and for OpenStreetMap objects in general. In addition, the most used address tag, "addr:housenumber", was associated with more than 51 million points on March 1st, 2020, corresponding to more than a third of the total points contained in the OSM platform. In this context, the purpose of this paper is to implement the QualiOSM tool in order to generate the key-value pair for address tags within OSM, thus contributing to the improvement of information completeness in the OSM tool.

For the implementation of the tag adder within the QualiOSM application, the reverse geocoding technique was used, in which the extraction of textual information, such as name or address, is performed from a pair of geographical coordinates (latitude and longitude). This technique is common in many geographic application scenarios,

[1]https://wiki.openstreetmap.org/wiki/Map_Features [Accessed in May 2020.]

[2]https://josm.openstreetmap.de/ [Accessed in May 2020.]

[3]https://github.com/gmedeiros93/josm/tree/master/josm/plugins/Quali_OSM [Accessed in October 2020.]

[4]https://taginfo.openstreetmap.org/ [Access in May 2020.]

for example, free online mapping services [Kounadi et al. 2013]. In this work, the tool Nominatim[5] was used, looking for names and addresses in OSM data from a pair of geographic coordinates and generating the address data in Extensible Markup Language (XML) or Javascript Object Notation (JSON) format.

After verifying the presence of much incorrect information in relation to the tag "addr:postcode" for objects in Brazil within the Nominatim tool, it was decided to use the reverse geocoding tool CEP Aberto[6] in order to complement the postal code information of OSM - Brazil objects. Besides that, the list of postal codes in Brazil, called in Portuguese "*Código de Endereçamento Postal*" (CEP), which is presented in the database of Correios (Post Office service in Brazil), was downloaded in the form of a *.csv* file to check the accuracy of the postal code information entered in the platform.

Figure 1 presents the architecture used to implement the QualiOSM tool. As can be seen, the architecture was divided into three layers: the outermost layer is the Presentation Layer, responsible for providing the interface between the user and the JOSM data editor, in addition to providing the loading of aerial images; the QualiOSM plugin and the functionality of the tag adder were developed within the Application Layer, in which it is also possible to see the interaction with the OpenStreetMap tool API; finally, the Data Layer is responsible for providing data management in the OSM Database and interacting with the tools Nominatim, CEP Aberto and Correios Database.

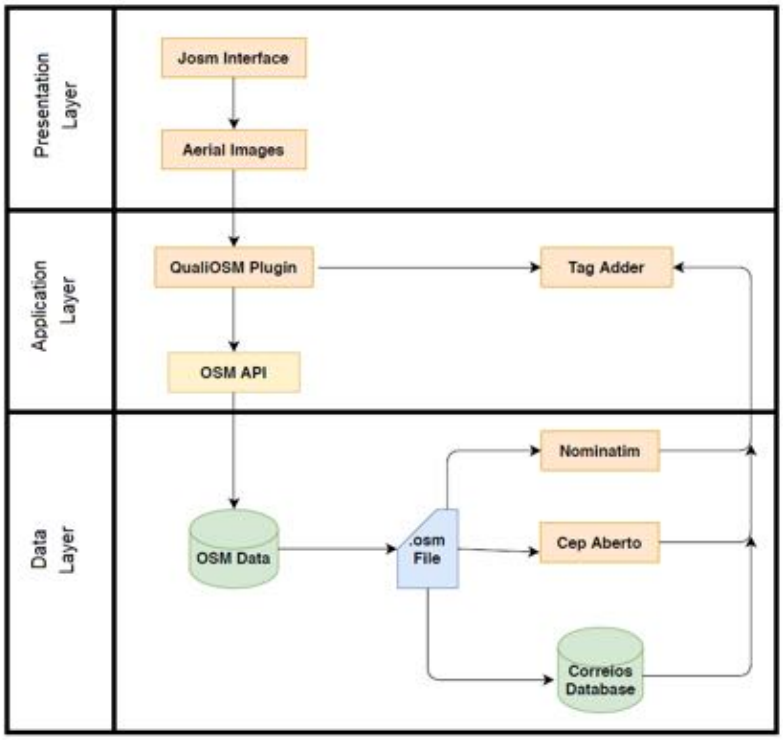

Figure 1. Architecture for implementing the QualiOSM application.

[5]https://nominatim.openstreetmap.org/ [Accessed in May 2020.]
[6]https://cepaberto.com/ [Accessed in August 2020].

The decision to implement the QualiOSM application within the JOSM data editor was reached for several reasons: (i) it is the data editor most widely used by OSM users [Ruta et al. 2012]; (ii) it is multiplatform, being written in the Java programming language; (iii) it offers a plugin mechanism to extend its main functionality. With an easily understandable user interface, the proposed tool can enable any OpenStreetMap user to enrich the map with address information, since no specific knowledge of semantic web languages or underlying formalisms is necessary.

After adding the plugin QualiOSM to the JOSM editor, the user can enjoy the functionality of the tag adder by loading the .osm file with the OpenStreetMap data to be edited on the map. Then, the user must select the objects and click on the "Add address tags" button. To insert the postal code information, the user can click on the options to use the tools Nominatim, CEP Aberto or Correios database. The interface of the QualiOSM tool can be seen in Figure 2.

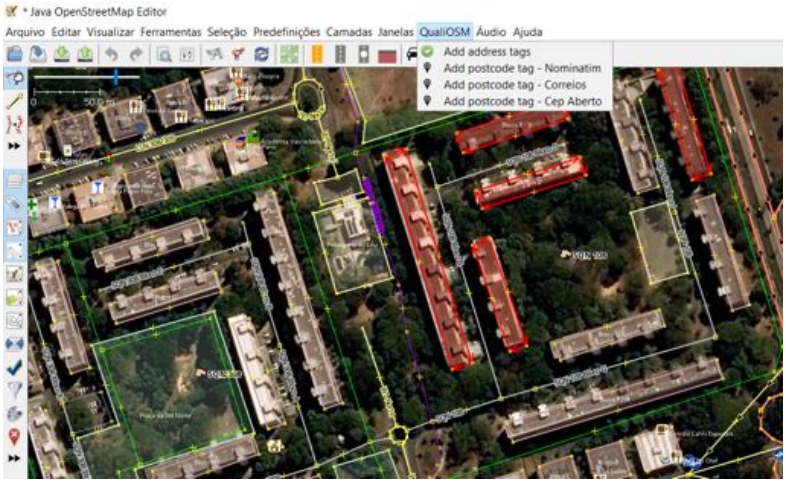

Figure 2. QualiOSM tool interface.

4. Geographic Data in OSM - Brazil

In order to carry out some analysis in relation to address tags in the entire territory of Brazil, the data from OpenStreetMap - Brazil was downloaded following an architecture containing two main layers: a layer for collecting data and a layer for viewing and analyzing data. As can be seen in Figure 3, two files were used to collect OpenStreetMap data: the file FullHistory.osm [7], containing the history of OpenStreetMap tool data corresponding to the entire planet until October 31, 2019; and the Brazil.poly file, containing the outline of the Brazil region, made available on the Geofabrik project website[8]. Then, these two files were processed with the osmconvert tool[9] for the creation of the BrazilHistory.osm file, containing the history of OpenStreetMap data in Brazil. Next, the BrasilHistory.osm file was processed in the osm2pgsql tool [10] with the purpose of importing the

[7]https://planet.osm.org/planet/ full-history / [Accessed in May 2020.]
[8]https://download.geofabrik.de/south-america/brazil.html [Accessed in May 2020.]
[9]https://wiki.openstreetmap.org/wiki/Osmconvert [Accessed in May 2020.]
[10]https://wiki.openstreetmap.org/wiki/Osm2pgsql [Accessed in May 2020.]

data into the PostgreSQL database. In addition, PostGIS extensions were used to treat spatial data, and Hstore to capture tags of OpenStreetMap objects.

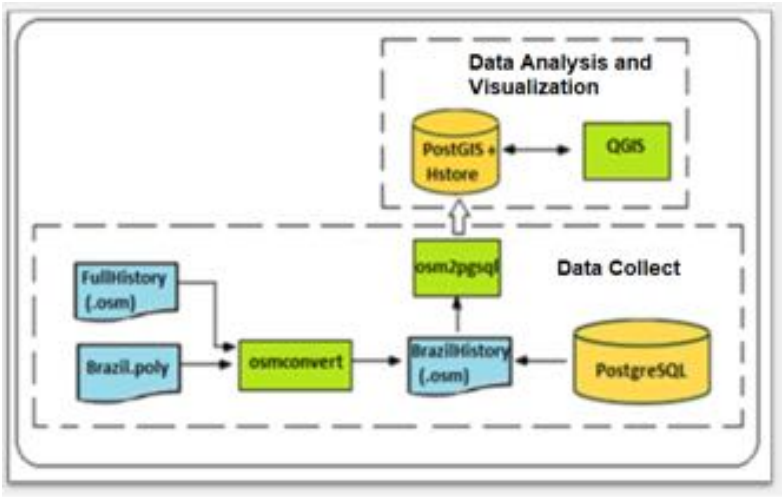

Figure 3. Architecture for collect and view data in OSM - Brazil.

The complete OpenStreetMap data in Brazil was downloaded so that analysis could be carried out in relation to the main labels used by mapping users in the country. The results showed that the most used tags were as follows: “addr:street”, “addr:city”, “addr:suburb” and “addr:postcode”. For this reason, these tags were chosen to integrate the tag adder implemented through the QualiOSM tool. From the OpenStreetMap data in Brazil, two different test scenarios for the QualiOSM application were considered:

- Scenario I - administrative region of Plano Piloto, in the city of Brasília. The center of the capital of Brazil is known for being a planned city, in which the buildings are arranged in an organized way and not very close to each other. Data were collected within the following bounding box: min latitude = -15.7929; max latitude = -15.7322; min longitude = -47.9093; max longitude = -47.8561.
- Scenario II - part of the city of Rio Branco, in the state of Acre (AC). This region was chosen based on the project “Mapping Flood Prone Urban Areas in Brazil”, available on the Hot Tasking Manager tool[11]. As can be seen, in this scenario houses are arranged much closer to each other, making the task of mapping the buildings more challenging. Data were collected within the following bounding box: min latitude = -9.9903; max latitude = -9.9733; min longitude = -67.8242; max longitude = -67.8021.

Since OpenStreetMap is a collaborative tool, it is natural that there is a great heterogeneity in the distribution of information mapping in relation to different regions, such as urban, rural and peripheral regions [Vargas-Muñoz et al. 2019]. Although mapping information on buildings and various other human constructions is widely available for urban areas, a significant number of buildings are not mapped in rural, peripheral regions or cities with less than 500,000 inhabitants, such as the city of Rio Branco.

[11]https://tasks.hotosm.org/projects/6124/ [Accessed in August 2020].

5. Results

Within OpenStreetMap, buildings are objects that often need associated address information, since users want to increase data about the location of points of interest, adding data such as the postal code, neighbourhood or building name. In this way, an analysis was carried out on the number of buildings that currently have address tags associated in Brazil and how these inclusions were made over time.

Thus, Figure 4 shows the evolution in relation to the inclusion of address tags in OpenStreetMap buildings in Brazil between the years 2009 and 2019. In this figure, it is observed that the inclusion of this type of tag has grown since 2015, but there is still a small number of buildings with associated address tags (in 2017, there were more than 860,000 buildings mapped, but only slightly more than 100,000 had associated address tags). In Figure 4 a peak of inclusion of these types of tags in 2017 is highlighted, mainly in relation to the tag "addr:street", corresponding to the street names. The predominance of this tag is because the OpenStreetMap tool has specialized in road mapping and information on names of roads near the buildings can facilitate routing mechanisms.

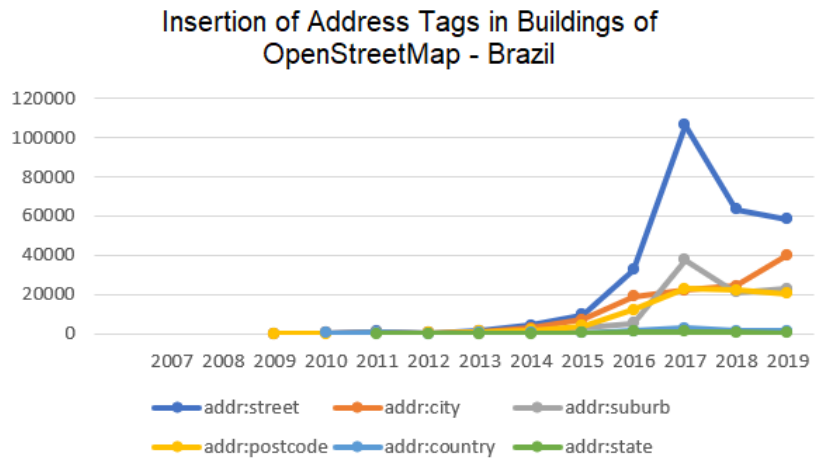

Figure 4. Insertion of address tags in buildings of OpenStreetMap - Brazil.

Within OpenStreetMap each user can also create their own labels to improve the map or to allow analysis of previously unmapped features. This feature can cause problems, such as the presence of misspelled tags which will be associated with a single object within the database. For the OpenStreetMap data in Brazil, 77 different address tags were identified, 27 of these tags (35%) were associated with only one object, and another 26 tags were associated with less than 10 objects.

Initially, the tests were performed with the addition of tags using the Nominatim reverse geocoding tool. Regarding the data from the first test scenario, a file in .osm format was downloaded containing the data of the administrative region of "Plano Piloto", a central district of the city of Brasília, corresponding to the data of July 31th, 2020. Regarding the second scenario, corresponding to the data in the city of Rio Branco in the state of Acre, data contained in the project "Mapping flood prone urban areas in Brazil" were used, through the Hot Tasking Manager tool. The tag adder was activated by select-

ing the preset "Human Construction/Edificaction" within the JOSM editor and then the adder was applied in the two different regions. After that, the tags associated with the selected objects were analyzed before and after the action of the tag adder.

The results obtained by adding tags in Scenario I can be seen in Table 1. In relation to the tag "addr:suburb", there was an increase from 1.76% to 41.83% of associated buildings; regarding the tag "addr:city", there was an increase from 2.12% to 45.46% of associated buildings; regarding the tag "addr:building", there was an increase from 0% to 10.21% of associated buildings. There was no change in relation to tags "addr:street"(5.28% of associated buildings) and "addr:housenumber" (1.59% of associated buildings) due to the lack of this information in the database of the Nominatim tool.

Table 1. Inclusion of address tags in Scenario I: Brasília - DF.

Tag	Before	After
addr:building	0%	10.21%
addr:city	2.12%	45.46%
addr:housenumber	1.59%	1.59%
addr:street	5.28%	5.28%
addr:suburb	1.76%	41.83%

Table 2 presents the results of applying the tag adder in the city of Rio Branco. As can be seen, the result was more satisfactory in relation to the inclusion of the tag "addr:city", in which there was a jump from 0.1% of associated buildings to 100% of associated buildings. However, there were no significant changes in relation to the other tags, "addr:building", "addr:housenumber", "addr:street" and "addr:suburb".

Table 2. Inclusion of address tags in Scenario II: Rio Branco - AC.

Tag	Before	After
addr:building	0%	0.40%
addr:city	0.1%	100%
addr:housenumber	0.03%	0.07%
addr:street	0.07%	0.07%
addr:suburb	0.03%	0.03%

When verifying the insertion of incorrect information in relation to the tag "addr:-postalcode", two more approaches were taken into account to include postal code information: using the reverse geocoding tool CEP Aberto and using the Correios database.

Cep Aberto acts similarly to the Nominatim tool, that is, from a geographic coordinate pair (latitude and longitude), it is able to search for information on that object and return this information in the form of a *.json* file. The Correios database, on the other hand, consists of a *.csv* file, with the coordinates of each object already associated. Thus, the postal code information was entered based on the coordinate closest to the center of the selected object in JOSM.

The distance was calculated according to the formula of the shortest distance between two points, expressed in the equation 1.

$$distance = \sqrt{(lat2 - lat1)^2 + (lon2 - lon1)^2} \quad (1)$$

Where (lat1, lon1) corresponds to the coordinates of the center of the selected object and (lat2, lon2) corresponds to the coordinates of the object within the Correios database. The algorithm finds the postal code of the selected object when this calculated distance is less than 10^{-4}.

An analysis was then carried out in relation to the addition of postal code tags in the tool, based on the reverse geocoding tools Nominatim and CEP Aberto, in addition to the use of the Correios database. The result for Scenario I (city of Brasilia) can be seen in Figure 5, in which it is observed that despite the fact that the Nominatim tool inserts postal code information for all selected objects, this tool adds lots of wrong information, having an error index of 96.15% and a hit rate of only 3.85%. The CEP Aberto tool obtained a hit rate of 17.31%, an error index of 26.92% and there was no addition of tags for 55.77% of the objects. Finally, the use of the Correios database led to a hit rate of 67.31% and did not add tags for 32.69% of the objects. One advantage of this approach is not adding wrong information to the OSM database.

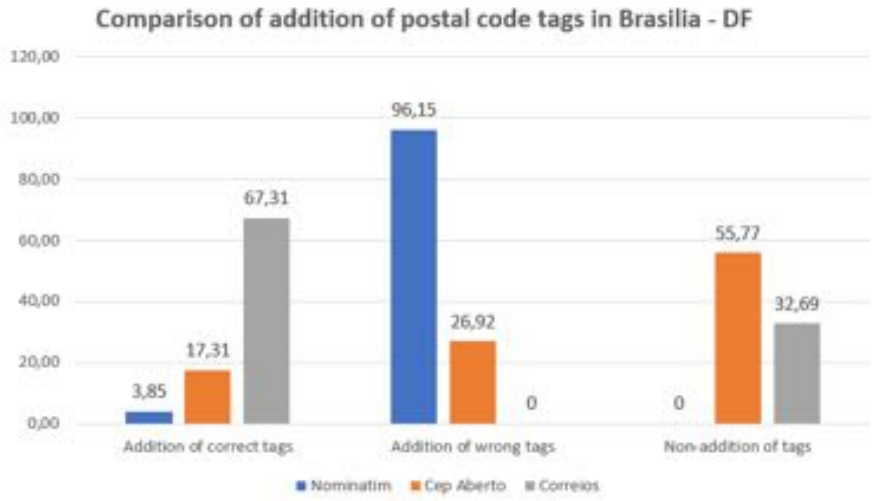

Figure 5. Comparison of addition of postal code tags in Scenario I.

Regarding Scenario II, in the city of Rio Branco, not much improvement was observed in relation to the completeness of the information using the CEP Aberto tool or the post office database. In addition, it should be noted that the Nominatim tool only inserted erroneous information in the QualiOSM tool, as can be seen in Figure 6.

An analysis was also carried out in relation to the time spent by the QualiOSM application for the inclusion of address tags in order to measure the performance of the tool. The tests were performed using a machine with 8.00 GB of RAM, Intel Core i7-9750H 2.60 GHz processor and Windows 10 operating system, 64 bits. For each selection of

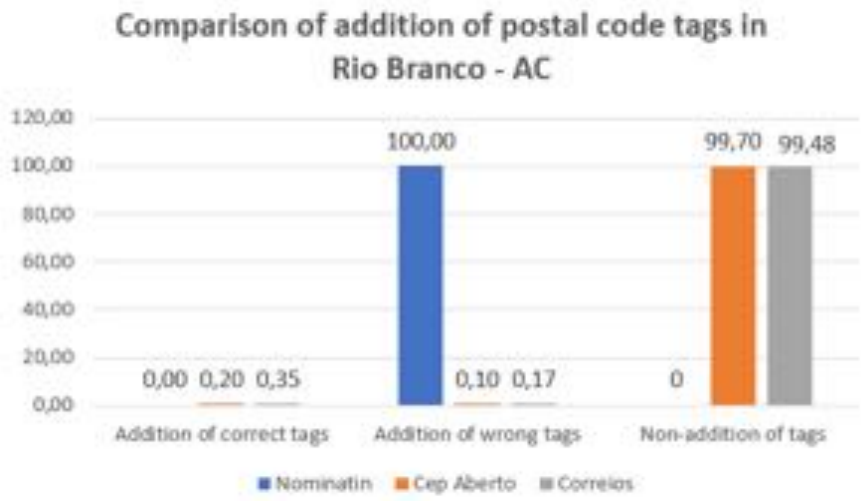

Figure 6. Comparison of addition of postal code tags in Scenario II.

objects, time was measured 10 times and the arithmetic mean was calculated. In this way, the results for Scenarios I and II are shown in Figure 7. Measuring the tool's execution time by adding ten more selected objects to each test, it can be seen that the time followed an approximately linear trend. Thus, the tool took an average of 500 milliseconds to include address tags per object.

Figure 7. Time to insert address tags in the QualiOSM tool.

6. Conclusion

The tag adder implemented in this work has shown potential for improving the completeness dimension for object information within the collaborative OpenStreetMap tool. In large urban centers that are well mapped within OpenStreetMap, as is the case in the city of Brasilia, the developed tool QualiOSM improved by almost 70% the addition of

postal code tags, which is an important tag for locating addresses, especially residential buildings.

It has been observed that the dimensions of completeness and accuracy often conflict with each other. The Correios database, despite having a good accuracy, still has many objects that are missing, especially when it comes to smaller urban centers, such as the city of Rio Branco, in Acre. On the other hand, with the Nominatim tool there was a greater increase in information, but there was also a greater increase in erroneous information, particularly, postal code information.

As a future work, we intend to explore other tags in addition to the address tags in this work, using other tools besides the Nominatim or CEP Aberto for finding information. It is also intended to test the tool in other scenarios and to evaluate other dimensions of quality in collaborative systems, such as logical consistency and accuracy.

References

Ames, M. and Naaman, M. (2007). Why we tag: Motivations for annotation in mobile and online media. *ACM SIGCHI Conf. Human Factors in Computing Systems*, page 971–980.

Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., and Rampini, A. (2016). On predicting and improving the quality of volunteer geographic information projects. *International Journal of Digital Earth*, 9(2):134–155.

Codescu, M., Horsinka, G., Kutz, O., Mossakowski, T., and Rau, R. (2011). Osmonto-an ontology of OpenStreetMap tags. *State of the map Europe (SOTM-EU)*, 2011.

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.

de Albuquerque, J. P., Eckle, M., Herfort, B., and Zipf, A. (2016). Crowdsourcing geographic information for disaster management and improving urban resilience: an overview of recent developments and lessons learned. *European handbook of crowdsourced geographic information*, pages 309–321.

Degrossi, L. C., Porto de Albuquerque, J., Santos Rocha, R. d., and Zipf, A. (2018). A taxonomy of quality assessment methods for volunteered and crowdsourced geographic information. *Transactions in GIS*, 22(2):542–560.

Firmani, D., Mecella, M., Scannapieco, M., and Batini, C. (2016). On the meaningfulness of "Big Data quality". *Data Science and Engineering*, 1(1):6–20.

Flanagin, A. and Metzger, M. (2008). The credibility of volunteered geographic information. *GeoJournal*, 72:137–148.

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

Kennedy, L., Chang, S.-F., and Kozintsev, I. (2006). To search or to label? predicting the performance of search-based automatic image classifiers. *ACM Workshop Multimedia Information Retrieval*, page 249–258.

Kounadi, O., Lampoltshammer, T. J., Leitner, M., and Heistracher, T. (2013). Accuracy and privacy aspects in free online reverse geocoding services. *Cartography and Geographic Information Science*, 40(2):140–153.

Liu, D., Wang, M., Hua, X.-S., and Zhang, H.-J. (2011). Semi-automatic tagging of photo albums via exemplar selection and tag inference. *IEEE Transactions on Multimedia*, 13:82–91.

Meng, Y., Hou, D., and Xing, H. (2017). Rapid detection of land cover changes using crowdsourced geographic information: a case study of beijing, china. *Sustainability*, 9(9):1547.

Mooney, P. and Corcoran, P. (2012). The annotation process in OpenStreetMap. *Transactions in GIS*, 16(4).

Ruta, M., Scioscia, F., Ieva, S., Loseto, G., and Di Sciascio, E. (2012). Semantic annotation of OpenStreetMap points of interest for mobile discovery and navigation. In *2012 IEEE First International Conference on Mobile Services*, pages 33–39. IEEE.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., et al. (2016). Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, 5(5):55.

Sehra, S. S., Singh, J., and Rai, H. S. (2017). Assessing OpenStreetMap data using intrinsic quality indicators: an extension to the QGIS processing toolbox. *Future Internet*, 9(2):15.

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

Vargas-Muñoz, J. E., Lobry, S., Falcão, A. X., and Tuia, D. (2019). Correcting rural building annotations in OpenStreetMap using convolutional neural networks. *ISPRS journal of photogrammetry and remote sensing*, 147:283–293.