QualiOSM: An Architecture to Improve Data Completeness on OpenStreetMap

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

[1] Universidade de Brasília (UnB), Brazil
gabriel.medeiros93@gmail.com, mholanda@unb.br
[2] Fundação Getúlio Vargas (FGV), Brazil
liviadegrossi@gmail.com

Abstract. OpenStreetMap (OSM) is a large spatial database in which geographic information is voluntarily contributed by thousands of users. In Geographic Information Systems (GIS), and more specifically, in Volunteered Geographic Information (VGI), as in the case of OSM, the issue of data completeness is a constant concern, since users without technical knowledge actively participate in the processes of including, editing and excluding data. Also in the case of OSM, users can add information to the objects assigning special labels for them. These labels are popularly called tags, and the process of assigning them to objects contributes to improving the attribute completeness, an important metric of data quality. In this context, this article proposes the QualiOSM architecture, which generates an automatic tag adder with the purpose of improving the completeness of address information for OSM objects in Brazil, using the reverse geocoding tools Nominatim, CEP Aberto and the database from Correios. The QualiOSM architecture showed good results for improving the completeness of city, neighborhood and street information in OSM objects, especially in scenarios of large urban centers, where the level of mapping is usually better compared to scenarios in rural or peripheral environments.

Categories and Subject Descriptors: H. 2 [**Database Management**]: Miscellaneous; H. 2.8 [**Databases Applications**]: Spatial Databases and GIS

Keywords: Geographic Data, Geographic Information Systems, Volunteered Geographic Information, Quality Dimensions, Completeness, OpenStreetMap

1. INTRODUCTION

A Geographic Information System (GIS) consists of an aggregation of factors, such as hardware, software, data, people and institutions, which perform the collection, storage, analysis and dissemination of information related to areas of the Earth's surface [Tomlinson 2007]. From the 2000s, with the advent of the Web 2.0, new systems have emerged, which are capable of providing an ever greater interaction with their users, since they allow geographic information to be added, edited or deleted via forms filled dynamically or directly within of a map. These types of systems became known as Volunteered Geographic Information (VGI) [Goodchild 2007].

The term VGI was developed for the purpose of describing computer systems in which a large number of collaborating users are engaged in the creation or editing of geographic information. In the last decades, VGI has been used for countless goals, whether for environmental monitoring or for the management of natural disasters, in addition to assisting in mapping the most remote regions of the planet, where access to mapping is more difficult [Senaratne et al. 2017].

A successful example of VGI is the collaborative mapping tool OpenStreetMap (OSM), used in this work as a case study. Created in 2004 by a student of computing, Steve Coast, from University

College London (UCL), the initial idea of the project was to map only the United Kingdom region, but it soon aroused the interest of researchers from other countries [Ramm et al. 2010]. However, the data provided by volunteers requires special attention in relation to the data quality. One of the main reasons for this is the great heterogeneity among their users, since they use different tools and technologies, and have different levels of detail or precision [Senaratne et al. 2017].

One of the great challenges of collaborative tools is that they require the participation of users to include new information in the database. In this context, [Nielsen 2006] presents the Rule of 1%, also called Rule 90-9-1, which describes the inequality of participation present in this type of system. According to this rule, 90% of the users of these types of tools only use the service without any type of active collaboration, 9% collaborate sporadically and only 1% of the members effectively collaborate with the project, participating in the effective creation of content. Thus, maintaining data quality in these systems is a major challenge, since a small portion of the users is responsible for creating a large amount of data.

Since users of the OpenStreetMap tool actively participate in the processes of inclusion, editing and exclusion of data, the issue of data quality is always present, since it is necessary to verify whether the data is being inserted or modified in the correct way. There are countless studies in the literature that have analyzed the issue of data quality in collaborative tools, specially in OSM, but there are still few studies involving some type of implementation work in this area. Thus, the objective of this article is to carry out the development of the QualiOSM architecture, in order to assist in improving the completeness of the address information within the OSM platform.

The rest of this document is structured as follows: Section 2 provides the theoretical reference with the basic concepts that this research involved; Section 3 presents the related work; Section 4 describes the development of the QualiOSM architecture; Section 5 shows the results obtained; the results are discussed in Section 6; finally, Section 7 presents the conclusion and future work.

2. THEORETICAL REFERENCE

Geospatial data, also called geographically referenced data or geographic data, is data that can be displayed, manipulated and analyzed using a spatial attribute, which denotes a location on the Earth's surface [Yeung and Hall 2007]. In general, geographical features can be represented as a matrix (raster data) or as a set of coordinates (vector data). Thus, vector data usually works with three basic abstractions (points, lines and polygons), while in the raster representation, the data is represented as a set of horizontal and vertical lines forming a kind of grid [Monteiro et al. 2001].

The past few decades have witnessed a profound change in the way geographic data, geographic information and, more broadly, geographic knowledge are being produced and disseminated due to the phenomenal growth of a set of related technologies, which have become popularly known as Web 2.0. Although different concepts have emerged to describe this new trend, the general idea can be translated into the use of tools to create, share and analyze geographic information across multiple users and multiple computing devices/platforms [Sui et al. 2013]. In this way, the phenomenon of Volunteered Geographic Information is part of a profound transformation in the way geographic information is being produced and distributed in the world nowadays [Goodchild 2007].

The related term of crowdsourcing is used to describe a set of computational techniques and methods that depend on a large number of users to solve specific tasks [Doan et al. 2011]. An increasingly popular category of crowdsourcing is space crowdsourcing, in which tasks must be completed at a specific time and place. Space crowdsourcing has spurred a series of recent industrial successes, including urban shared economy services (Uber and Gigwalk) and space-time data collection tools (OpenStreetMap and Waze) [Tong et al. 2020].

The OpenStreetMap tool is a spatial crowdsourcing project created to build a free geographic database of the planet, with the objective of obtaining a record of all existing geographic features. In the beginning, the tool was used basically for street mapping, but currently encompasses trails, buildings, woods, beaches, among other objects. Along with the geographical characteristics, the project also includes administrative boundaries, details of land use, bus routes and other information that can be added through the use of tags, labels with information associated with the objects [Bennett 2010].

Quality is a key component of any data set. According to ISO 19157, which establishes the principles that describe the quality concept of geographic data, quality can be defined as the degree to which a set of characteristics meets a set of pre-established requirements [ISO 2013]. Specifically, in a set of geospatial data, decision making for a given purpose is strongly based on quality measures, named quality dimensions in literature, such as accuracy, completeness and logical consistency. This applies even more to VGI, since users actively participate in the processes of inclusion, editing and exclusion of information [See et al. 2017].

This article focuses on the quality dimension of completeness, more specifically in relation to the address information associated with the objects of the OpenStreetMap tool. Thus, the QualiOSM architecture was developed with the objective of improving the completeness of objects within the OSM platform, adding missing address information to the objects using the reverse geocoding tools Nominatim and CEP Aberto, and the official database from the post office service of Brazil (Correios database).

3. RELATED WORK

Several works were found that explored data quality within collaborative tools, addressing the process of adding tags associated with objects. For example, [Ames and Naaman 2007] explored the motivation for tagging images on Flickr, concluding that most users perform this process of tagging to make information more accessible to the general public. In addition, [Kennedy et al. 2006] assessed the performance of trained classifiers with photos from Flickr and their associated tags, demonstrating that tags provided by users contain a lot of incorrect information. Besides that, [de Oliveira et al. 2015] used machine learning techniques in order to produce VGI data from social media and then they measured the quality of the data automatically produced.

Regarding collaborative mapping tools, [Codescu et al. 2011] organized an ontology with the objective of standardizing and facilitating the hierarchy of tags within the OpenStreetMap tool, but concluded that the use of an ontology is only efficient if the users keep the tags constantly updated within the OSM platform. Still within the OpenStreetMap tool, [Mooney and Corcoran 2012] performed an analysis of more than 25,000 objects in the database in Ireland, United Kingdom, Germany and Austria. The results indicated that there are some problems arising from the way users tag objects in OSM. The study also showed that these identified problems are a combination of the flexibility of the labeling process and the lack of a more rigid mechanism to verify adherence to the OpenStreetMap ontology in relation to the tags added by its users.

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

Although many articles were found that explored the issue of data quality in collaborative tools, no work was found that proposed a tool that uses an automatic tag adder in order to improve data completeness. Thus, this article works as an extension of [de Medeiros et al. 2020] and proposes the implementation of the QualiOSM tool for the purpose of improving the quality of geographic information within OpenStreetMap, mainly with regard to the process of attributing address tags to objects. The intention of the implemented architecture is to contribute to the completeness of the address information of objects on the OSM platform.

4. QUALIOSM ARCHITECTURE

The proposed architecture for improving data quality within the OpenStreetMap tool is based on a three-tier model, in which business logic, data access and the user interface are developed and maintained as independent modules [Fowler 2002]. As can be seen in Figure 1, the QualiOSM architecture was divided into three distinct layers: the top layer is the Presentation Layer, responsible for providing the interface between the user and the Java OpenStreetMap (JOSM) data editor, since the architecture was implemented in the form of a plugin for that editor; the functionality of the tag adder was 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 on the basis of the OSM platform and interacting with the reverse geocoding tools Nominatim, CEP Aberto and the Correios database.

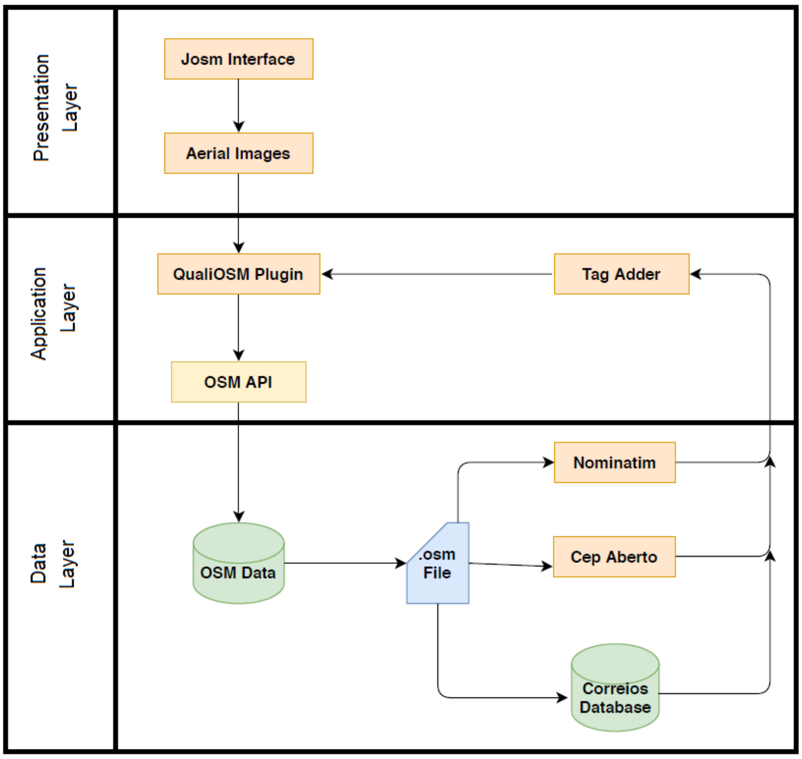

Fig. 1. QualiOSM Architecture.

In this architecture, the Presentation Layer is the outermost layer and is responsible for providing the interface for interaction with the user. It is in this layer that the user will be able to load the aerial images and view the loading of the exported data from the OpenStreetMap tool in the extension *.osm*. The JOSM data editor is the editor responsible for the largest number of edits to OSM objects[1], it has a friendly interface and, through the map view window, the user can load layers of aerial images, as well as view the data exported from the OpenStreetMap tool in *.osm* format. In addition, through

[1] https://en.wikipedia.org/wiki/JOSM [Accessed May 2021].

the implementation of a plugin, a developer will be able to customize the menus of the main JOSM screen according to the implemented features.

The Application Layer is responsible for containing the business logic and performing the interaction with the OpenStreetMap API. For the implementation of the tag adder inside the QualiOSM tool, the reverse geocoding technique was used, in which the extraction of textual information, such as name or address, is performed from a pair of geographic coordinates (latitude and longitude). This technique is common in many geographic application scenarios, such as free online mapping services [Kounadi et al. 2013]. In this work, the reverse geocoding tools Nominatim and CEP Aberto were used for the purpose of inserting missing address information into OSM objects. In addition, the list of postal codes in Brazil, which can be found in the official database of Correios, was downloaded in the form of a *.csv* file in order to insert more precise information in relation to the Postal Address Code (named *Código de Endereçamento Postal* - CEP - in Portuguese) associated with objects within the platform OSM. Thus, the tools Nominatim and CEP Aberto were used to insert the general addressing tags (addr:city, addr:building, addr:suburb and addr:neighborhood) and also to insert the code postal tag (addr:postcode), while the Correios database was used to insert the postal code tag and to validate the other tags.

The Data Layer is responsible for interacting with the reverse geocoding tools Nominatim, CEP Aberto and the Correios database. The two reverse geocoding tools search for the address information associated with the selected objects using their respective platforms on the Web, from a url address, which will be returned using the implemented method called *loadAddress*. For the Nominatim tool, the base url was `https://nominatim.openstreetmap.org/reverse?format=json` and for the CEP Aberto tool the base url was `https://www.cepaberto.com/api/v3/nearest?`. These urls were then concatenated with the latitudes and longitudes of the selected objects in the standard interface of the JOSM data editor.

The reverse geocoding API implemented within the Nominatim tool does not exactly calculate the address of the coordinate it receives, but finds the OpenStreetMap object closest to the requested coordinate and returns its address information. For the use of the reverse geocoding technique within the CEPAberto tool, it is necessary to use the API with the method nearest. Thus, given a pair of latitude and longitude, the zip code closest to the point corresponding to these coordinates is returned. The search is limited to a radius of 10km from the point referring to the coordinates passed as a parameter. The Correios database, on the other hand, is composed of a *.csv* file, with the coordinates of each object already associated. Thus, the postal code information was included based on the coordinate closest to the center of the object selected in JOSM. The distance was calculated according to the formula of the shortest distance between two points, expressed in Equation 1,

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

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

When analyzing statistics present on the website TagInfo[2], a system created with the objective of finding and aggregating information about the OSM tags, it was observed that among the five most used tags for OpenStreetMap points, four are address tags (addr:house-number, addr:street, addr:city and addr:postcode). It was also possible to observe that these four tags are among the ten most used both for lines and for OpenStreetMap objects in general. In addition, the most used address label, "addr:house-number", was associated with more than 51 million points on March 1, 2020, corresponding to more than a third of the total points contained in the OSM platform. Thus, the

[2] https://taginfo.openstreetmap.org/ [Accessed February 2021].

importance of improving the completeness of address information within the OpenStreetMap platform was clearly apparent.

In this context, the QualiOSM tool was developed with the objective of improving the completeness of the address information associated with objects on the OpenStreetMap platform. The application was written in the Java programming language and implemented as an extension (plugin) within the Java OpenStreetMap (JOSM) data editor. The decision to implement the QualiOSM application within the JOSM data editor was made for several reasons: (i) it is the data editor most widely used by users of the OSM; (ii) it is multiplatform, being written in the Java programming language; (iii) and it offers a plugin mechanism to extend its main functionality. With an easily understandable user interface, the proposed tool can enable any OpenStreetMap collaborator to enrich the map with address information, since no specific knowledge in semantic Web languages or underlying formalisms are required [Ruta and Di Sciascio 2012].

After adding the plugin QualiOSM to the JOSM editor, the user can make use of the functionality of the tag adder by loading the *.osm* file with the OpenStreetMap data to be edited on the map. The user must then 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 user interface of the QualiOSM tool can be seen in Figure 2.

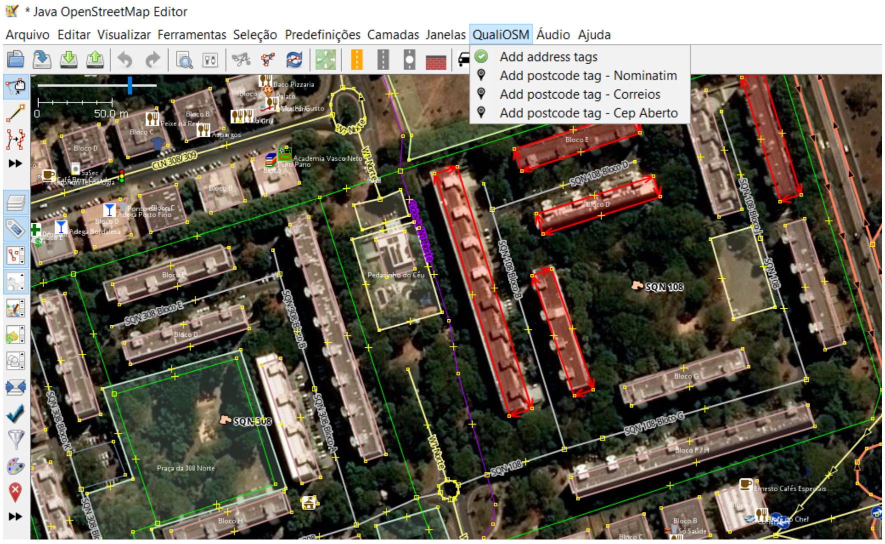
* Java OpenStreetMap Editor
Arquivo Editar Visualizar Ferramentas Seleção Predefinições Camadas Janelas QualiOSM Áudio Ajuda
Add address tags
Add postcode tag - Nominatim
Add postcode tag - Correios
Add postcode tag - Cep Aberto
50.0 m
SQN 108
SQN 308

Fig. 2. QualiOSM user interface.

5. RESULTS

The QualiOSM architecture was evaluated in different test scenarios by performing data extraction in the standard OpenStreetMap format in different areas of interest, described below:

- Scenario I: corresponds to the part of the administrative region of Plano Piloto, in the city of Brasilia (Distrito Federal - DF), the capital of Brazil. The center of the city is known for being well planned, the buildings are arranged in an organized way and not very close to each other. Data were collected within the following bounding box: minimum latitude = -15.7929; maximum latitude = -15.7322; minimum longitude = -47.9093; maximum longitude = -47.8561.
- Scenario II: corresponds to part of the city of Rio Branco, in the state of Acre (AC). This region was chosen based on the “Mapping Flood Prone Urban Areas in Brazil”, available at the Hot

Tasking Manager Project[3]. In this scenario the houses are arranged much closer to each other, making the task of mapping buildings more challenging. Data were collected within the following bounding box: minimum latitude = -9.9903; maximum latitude = -9.9733; minimum longitude = -67.8242; maximum longitude = -67.8021.

- Scenario III: Considers the peripheral part of the interior cities of Mogi das Cruzes and Suzano, in the state of São Paulo (SP). Data were collected within the following bounding box: minimum latitude = -23.6901; maximum latitude = -23.6253; minimum longitude = -46.3922; maximum longitude = -46.2971.
- Scenario IV: Considers the area belonging to the Rocinha community, in the state of Rio de Janeiro (RJ). In this environment the buildings are very close to each other, and there are many buildings not listed in the official government records. Data were collected within the following bounding box: minimum latitude = -22.9896; maximum latitude = -22.9866; minimum longitude = -43.2498; maximum longitude = -43.2397.

The data for each scenario was downloaded in accordance with the bounding boxes of each area. Next, simulations were realized in order to verify how the tags addr:city, addr:building, addr:suburb and addr:neighbourhood would be included using the Nominatim and CEP Aberto reverse geocoding tools. Later, the Correios database was used in order to insert postcode information and validate the other tags. To carry out the simulations, the objects characterized as “residential buildings” were selected, that is, in the context of this work, the objects whose building tag was associated with the following values: “yes”, “residential”, “house” or “apartment”. Subsections 5.1 to 5.4 describe the results observed in each scenario.

5.1 Scenario I - Brasilia (DF)

For Scenario I, part of the Administrative Region of Plano Piloto in Brasília (Distrito Federal - DF) was chosen. The results of the simulations can be seen from Table I, which shows that the Nominatim tool proved to be efficient for the inclusion of tags addr:city and addr:suburb, reaching all 210 buildings found in this scenario. Tags addr:building and addr:neighborhood were also improved, with the addition of information by 67.62% and 91.43%, respectively.

The CEP Aberto tool had a slightly lower performance compared to the Nominatim tool, leading an increase of 98.09% in buildings associated with the tag addr:city and a 94.76% increase in objects associated with tags addr:suburb and addr:neighborhood. The CEP Aberto tool did not contain data referring to the names of buildings separated from their respective streets and therefore, there was no change to the simulation by the inclusion of the tag addr:suburb.

Table I. Simulation of Inserting Address Tags in Scenario I.

	Before	**Nominatim**	**CEP Aberto**
addr:city	0,48%	100%	98,57%
addr:building	0%	67,62%	0%
addr:suburb	0%	100%	94,76%
addr:neighbourhood	0%	91,43%	94,76%

An analysis was also carried out in relation to the addition of the addr:postcode tag, since during the simulations, the inclusion of wrong postal codes was detected when using Nominatim and CEP Aberto. When using the Correios database, despite reducing the percentage of objects associated with the postal code tag, the tool guarantees that the inserted postcodes will be correct. As can be seen in Figure 5.1, after simulating the addition of the addr:postcode tag in buildings for Scenario

[3] https://tasks.hotosm.org/projects/6124 [Accessed in October 2020].

I, it was found that Nominatim and CEP Aberto were adding more wrong information than correct information. Using the Nominatim tool, 96.15% of the buildings were associated with incorrect postal code information and 3.85% were associated with the tag correctly. In the case of CEP Aberto, only 17.31% of buildings were correctly associated with the postal code tag, while 26.92% were associated incorrectly. Already using the post office database, 67.31% of the 210 buildings were correctly associated with the postal code tag and there was no addition of incorrect information. In addition, 32.69% of the buildings remained unchanged.

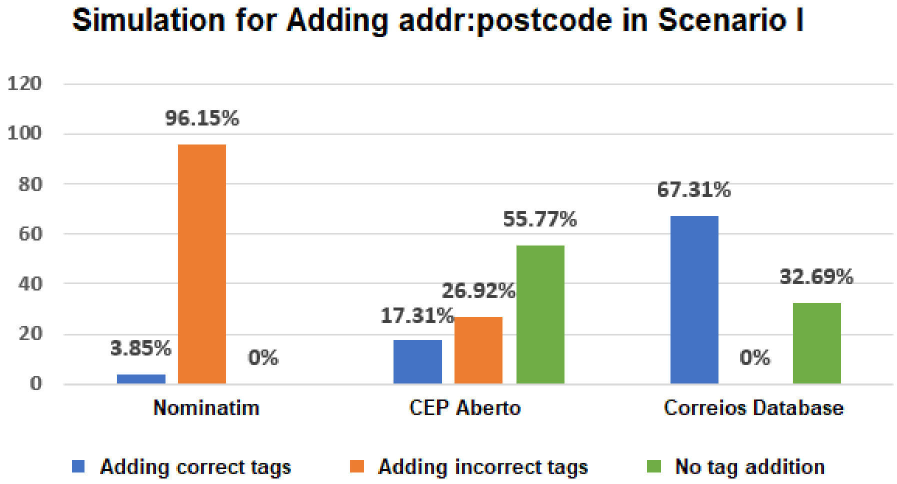

Fig. 3. Simulation for Adding Tag addr:postcode in Scenario I.

5.2 Scenario II - Rio Branco (AC)

For Scenario II, part of the city of Rio Branco, in the state of Acre (AC), was chosen. The result of the simulations can be seen from Table II which shows that the Nominatim tool proved to be efficient only for the inclusion of the tag addr:city, which was added in 100% of the 4,049 buildings residential properties found in the scenario. The tag addr:suburb was associated with 0.02% and the tag addr:neighborhood was associated with 4.57% of buildings. CEP Aberto did not have a good result either, adding the tags addr: city, addr:neighborhood and addr:suburb in approximately 12% of the buildings.

Table II. Simulation of Inserting Address Tags in Scenario II.

	Antes	**Nominatim**	**CEP Aberto**
addr:city	0,1%	100%	12,15%
addr:building	0%	0%	0%
addr:suburb	0,02%	0,02%	12,08%
addr:neighbourhood	0%	4,57%	12,05%

Performing the analysis in relation to the addition of tag addr:postcode, using the database of Correios, results are shown in Figure 4. After simulating the addition of the tag addr:postcode in buildings for Scenario II, it was found that the Nominatim and CEP Aberto tools were adding more incorrect information than correct information. With the Nominatim tool, 96.39% of buildings were associated with incorrect postal code information and only 3.61% of buildings were associated with the correct information. In the case of CEP Aberto, 0.2% of the buildings were correctly associated with the postal code tag, 0.1% of the buildings were associated with incorrect tags and 99.7% of the buildings remained unchanged. Using the post office database, 39.34% of the buildings were correctly associated with the postal code tag, 60.66% of the buildings remained unchanged and there was no addition of incorrect information.

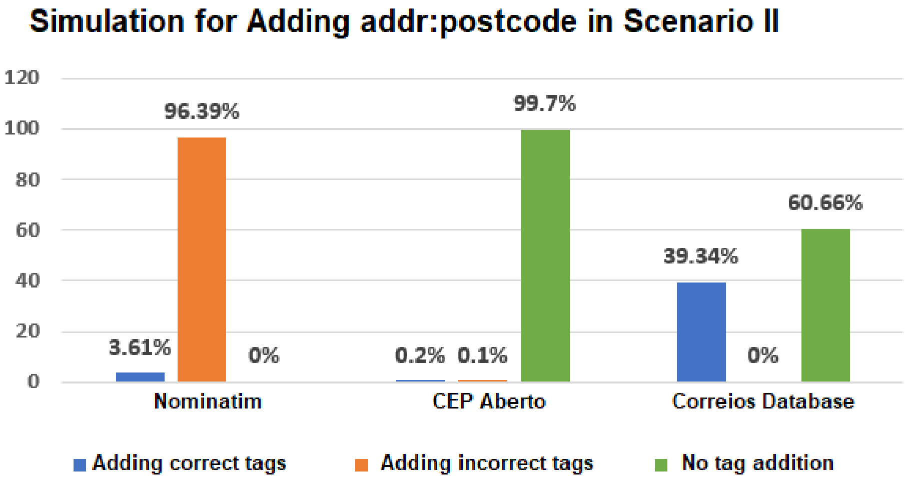

Fig. 4. Simulation graph for the addition of tag addr:postcode in Scenario II.

5.3 Scenario III - Mogi das Cruzes and Suzano (SP)

For Scenario III, part of the rural area of Mogi das Cruzes and Suzano, in the state of São Paulo (SP) was chosen. The results of the simulations can be seen from Table III which shows that the Nominatim tool proved to be efficient for the inclusion of the tag addr:city, which was associated to 100% of the 47 residential buildings found in the scenario. The tag addr:suburb was associated with 70.21% of buildings and the tag addr:neighborhood was associated with only 2.13% of buildings. CEP Aberto had an unsatisfactory performance, associating the address tags to only 4 buildings, corresponding to 8.51% of the buildings found in the scenario. There were no changes in relation to the tag addr:building due to the lack of this information both in the Nominatim tool and in CEP Aberto in relation to the observed scenario.

Table III. Simulation of Inserting Address Tags in Scenario III.

	Antes	Nominatim	CEP Aberto
addr:city	0%	100%	8,51%
addr:building	0%	0%	0%
addr:suburb	0%	70,21%	8,51%
addr:neighbourhood	0%	2,13%	8,51%

Performing the analysis in relation to the addition of the tag addr:postcode and using the Correios database, the results shown in Figure 5 were obtained. As can be seen, after simulating the addition of the tag addr:postcode in buildings for Scenario III, it was verified again that the Nominatim tool was adding more wrong information than correct information. With Nominatim, 70.21% of the buildings were associated with incorrect postal code information and 29.79% were associated with the correct information. In the case of the CEP Aberto, 14.89% of the buildings were correctly associated with the postal code tag, 85.11% of the buildings remained unchanged and there was no addition of incorrect information. Already using the post office database, 40.82% of the 47 buildings were correctly associated with the postal code tag, 59.18% of the buildings remained unchanged and there was also no addition of incorrect information.

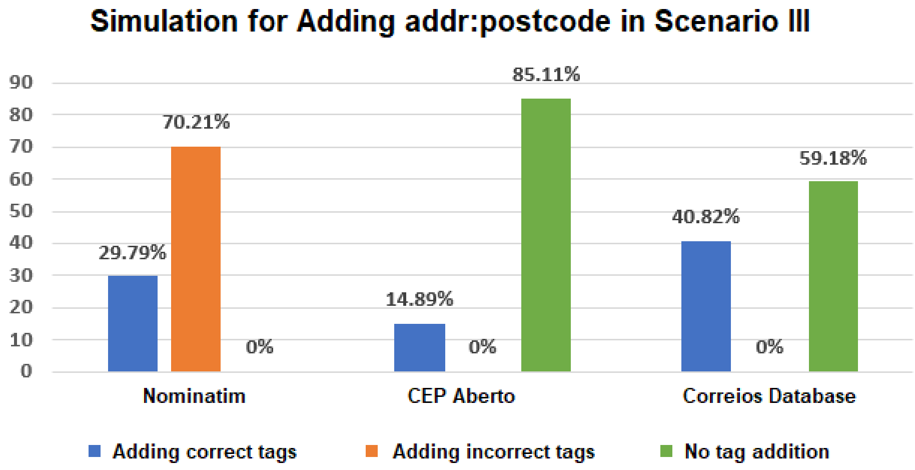

Fig. 5. Simulation graph for the addition of tag addr:postcode in Scenario III.

5.4 Scenario IV - Rio de Janeiro (RJ)

For Scenario IV, the region corresponding to the Rocinha community in the state of Rio de Janeiro (RJ) was chosen. As can be seen from Table IV, the Nominatim tool managed to contribute 100% to the inclusion of the tags addr:city and addr:suburb, reaching all 280 buildings found in this scenario. The tag addr:building also saw a good improvement, with the addition of tags in 67.62% of the associated buildings. Regarding CEP Aberto, there was an increase of 25.36% in buildings associated with the tag addr:suburb and 26.78% in buildings associated with the addr:neighbourhood tag. Thus, it appears once again that the Nominatim tool proved to be more efficient for the inclusion of address tags compared to CEP Aberto.

Table IV. Simulation of Inserting Address Tags in Scenario IV.

	Antes	Nominatim	CEP Aberto
addr:city	0.71%	100%	0.71%
addr:building	0%	67,62%	0%
addr:suburb	0.71%	100%	26.07%
addr:neighbourhood	0%	0%	26.78%

Performing the analysis in relation to the addition of the tag addr: postcode and using the Correios database, the results found in Figure 6 were obtained. As can be seen, after simulating the addition of the tag addr:postcode in buildings for Scenario IV, it was found that the Nominatim tool only inserted incorrect postal code information. After simulating the insertion of tags using CEP Aberto, only 0.36% of the buildings had their tags correctly associated; 26.07% of the buildings were associated with incorrect tags and 73.57% of the buildings remained unchanged. Using the Correios database, 9.29% of the buildings were correctly associated with the tags, 90.71% of the buildings remained unchanged and no incorrect information was added.

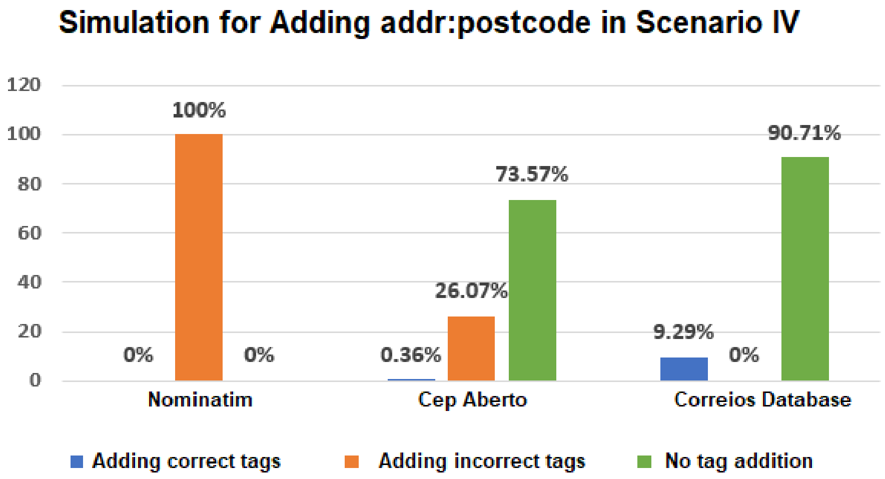

Fig. 6. Simulation graph for the addition of tag addr:postcode in Scenario IV.

6. DISCUSSION OF RESULTS

In order to improve data quality within the OpenStreetMap tool, it was necessary to use other tools, and in this way, the QualiOSM architecture became dependent on the quality dimensions existing in these tools. The Nominatim tool, despite having a good completeness and trying to completely fill in the postal code information, has a low accuracy, which causes incorrect information to be added to the OSM database. CEP Aberto, on the other hand, does not have a good completeness when compared to the Nominatim tool and it is noted that there is a lot of information missing from the tool, especially in relation to postal code information.

Among the three approaches used for the inclusion of the postal code tag, the one that proved to be the most efficient was the use of the Correios database, since the database is accurate and thus reduces the chances of adding wrong information within the OpenStreetMap platform. However, one of the existing difficulties for the application of this approach is the fact that not all public places located in Brazil are georeferenced with the provision of their respective geographic coordinates, which compromises the completeness of the database. Thus, a complete and accurate database of buildings would be of fundamental importance in order to improve the quality of collaborative geographic data in Brazil.

Besides that, some limitations were observed when using the QualiOSM architecture. The tool presented the best results for Scenario I, since the central area of the city of Brasilia corresponds to a fully planned region, in which the buildings are correctly associated with a well-defined zip code within the database of the Post Office. However, in relation to Scenario IV, it was observed that the post office database is not complete, as it is an area characterized by the existence of haphazard housing and few formal buildings.

7. CONCLUSION AND FUTURE WORK

The QualiOSM architecture was developed with the aim of improving the completeness of the address information within the OpenStreetMap tool. The tool, implemented in the form of a plugin for the JOSM data editor, showed good results in relation to the inclusion of new information regarding the name of the city, neighborhood and suburb in objects on the OSM platform, thus contributing to an improvement in the percentage of objects with associated labels. This fact was observed especially in relation to the city center scenarios used for testing, where the level of mapping is better when compared to the level of mapping in rural or peripheral regions.

As future work, it is intended to explore other tags in addition to the address tags used in this work, as well as to use other tools besides Nominatim and CEP Aberto to obtain new information from OSM objects. It is also intended to test the tool in other scenarios and to evaluate other dimensions of quality in collaborative systems, as well as to test the application of the architecture implemented in other editors outside of JOSM. In addition, the use of the Correios database can be used for other purposes, for example, for the identification and mapping of buildings within the OpenStreetMap tool.

REFERENCES

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

Bennett, J. *OpenStreetMap: Be your own cartographer.* Packt Publishing, 2010.

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

Davidovic, N., Mooney, P., Stoimenov, L., and Minghini, M. 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, 2016.

de Medeiros, G. F., Degrossi, L. C., and Holanda, M. Qualiosm: Melhorando a qualidade dos dados na ferramenta de mapeamento colaborativo OpenStreetMap. *Simpósio Brasileiro de Banco de Dados (SBBD)*, 2020.

de Oliveira, M. G., de Souza Baptista, C., Campelo, C. E., Acioli Filho, J. A. M., and Falcão, A. G. R. Producing volunteered geographic information from social media for lbsn improvement. *Journal of Information and Data Management* 6 (1): 81–81, 2015.

Doan, A., Ramakrishnan, R., and Halevy, A. Y. Crowdsourcing systems on the world-wide web. *Communications of the ACM* 54 (4): 86–96, 2011.

Fowler, M. *Patterns of enterprise application architecture.* Addison-Wesley Longman Publishing Co., Inc., 2002.

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

ISO, I. 19157: 2013: Geographic information—data quality. *International Organization for Standardization: Geneva, Switzerland*, 2013.

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

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

Monteiro, A. M., Camara, G., Fucks, S., and Carvalho, M. Spatial analysis and GIS: A primer. *National Institute for Space Research*, 2001.

Mooney, P. and Corcoran, P. The annotation process in OpenStreetMap. *Transactions in GIS* vol. 16, pp. 561–579, 2012.

Nielsen, J. The 90-9-1 rule for participation inequality in social media and online communities, 2006.

Ramm, F., Topf, J., and Chilton, S. OpenStreetMap: Using and enhancing the free map of the world. *UIT Cambridge*, 2010.

Ruta, M., S. F. I. S. L. G. and Di Sciascio, E. Semantic annotation of OpenStreetMap points of interest for mobile discovery and navigation. *IEEE First International Conference on Mobile Services*, 2012.

See, L., Estima, J., Pődör, A., Arsanjani, J., Bayas, J., and Vatseva, R. Sources of VGI for mapping. *Citizen Sensor*, 2017.

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

Sui, D., Elwood, S., and Goodchild, M. *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice.* Springer, 2013.

Tomlinson, R. F. *Thinking about GIS: geographic information system planning for managers.* Vol. 1. ESRI, Inc., 2007.

Tong, Y., Zhou, Z., Zeng, Y., Chen, L., and Shahabi, C. Spatial crowdsourcing: a survey. *The VLDB Journal* 29 (1): 217–250, 2020.

Yeung, A. K. W. and Hall, G. B. *Spatial Database Systems: Design, Implementation and Project Management.* Springer, 2007.