OpenStreetMap: Quality assessment of Brazil's collaborative geographic data over ten years

Gabriel Franklin Braz de Medeiros[1], Maristela Holanda[1], Aleteia Patrícia Favacho de Araújo[1], Márcio de Carvalho Victorino[2]

[1]Computer Science Department – University of Brasilia (UnB)
Brasilia – DF – Brazil

[2]Faculty of Information Science – University of Brasília (UnB)
Brasilia – DF – Brazil

gabriel.medeiros93@gmail.com, mholanda@unb.br, aleteia@unb.br, mcvictorino@unb.br

Abstract. *OpenStreetMap is a collaborative mapping tool in which users actively include, transform and exclude geographic data. Consequently, the quality and consistency of the information made available in the tool is of constant concern. To address this issue, this work performs an analysis of some of the quality parameters within OpenStreetMap, with the data referring to the region corresponding to Brazil, over a ten year period, as a source. Analyzing the parameters of Completeness, Logical Consistency and Temporal Accuracy, some basic characteristics of this type of tool can be observed, such as heterogeneity, since mapping does not occur uniformly.*

1. Introduction

With the advent of Web 2.0, in the early 2000s, Internet users were provided with the capability of creating, changing, and deleting site content in a very dynamic way [Goodchild, 2007]. This event led to the emergence of new techniques and computational methods, which depend on many users for the completion of specific tasks – described as crowdsourcing tools [Tapscott and Williams, 2007]. Some crowdsourcing tasks have customarily been carried out on traditional desktops. However, this method does not always work due to requirements involving the actual physical locations of specific objects. For this reason, a new paradigm called space crowdsourcing has emerged [Zhao and Han, 2016].

Subsequently, the development of smartphone devices with integrated GPS contributed significantly to the emergence of space crowdsourcing, since it allows users to complete tasks according to their physical location. In this context, the OpenStreetMap crowdsourcing tool (OSM), created in 2004 by computer student, Steve Coast, of University College London (UCL), aimed to create a free and editable world map built by volunteers, and released with an open content license [Mark, 2006].

All data from the OpenStreetMap tool can be downloaded for free in vector format, which leads to a widespread use of this data. Given the possible applications for the use of spatial data, such as region mapping, geographic analysis and risk prevention, the issue of information quality is fundamental [Girres and Touya, 2010]. Aside from this paper, few works have analyzed Brazilian OSM data. Thus, this paper performs an

analysis on the quality of the data inserted in the OpenStreetMap tool in the region corresponding to Brazil over a nine-year period, between the years of 2007 and 2016.

This paper is structured in the following sections: in Section 2 related works are presented. In Section 3, the quality parameters to be analyzed with the completion of this work. Section 4, presents the methodology for the development of the work. In Section 5, the obtained results are presented; Section 6 presents the conclusion and future work.

2. Related Work

In recent years, several researchers have already proposed different investigations into the quality of data in the OpenStreetMap tool. One of the precursors of these researches was Mordechai Haklay (2010), who conducted a study comparing the database of the OSM tool with the database of official agencies of London in the year of 2008.

Girres and Touya (2010) analyzed quality parameters such as geometric accuracy, semantic accuracy, completeness, logical consistency, and temporal accuracy within the OpenStreetMap database in France. Mondzech and Sester (2011) performed a data quality assessment of the OpenStreetMap tool in Germany, comparing the OSM database with the ATKIS software base. Barron et al. (2014) developed a framework for analyzing parameters, such as road network completeness and positional accuracy, comparing data from the cities of San Francisco (USA), Madrid (Spain) and Yaoundé (Cameroon).

Following the related work, though differing contextually, by focusing on the region corresponding to Brazil, this paper presents analysis on some quality parameters within the OSM tool in relation to the Brazilian collaborative geographic data.

3. Quality Parameters

Several elements (or components) have been proposed with the aim of describing and measuring the quality of geographic databases. These elements are called quality parameters, and some of these parameters are described below [Girres and Touya, 2010]:

- Completeness - Measures the relationship between absence (omission) and the presence of data or attributes in a database;
- Logical Consistency - Evaluates the degree of internal consistency, analyzing modeling rules and specifications;
- Temporal Accuracy - Evaluates the updating of the database as time passes.

This paper presents the evaluation of these quality parameters, beginning with the completeness of the name attribute, present in the OpenStreetMap tool objects. Next, the parameter of the logical consistency was evaluated when the presence of buildings modeled as point and polygon was verified, which could suggest a duplication of the data. Finally, the parameter of temporal accuracy was also evaluated when performing an analysis of the data insertion between the period of 2007 and 2016.

4. Methodology

This paper used the same methodology presented by [Medeiros and Holanda, 2017]. Thus, the file FullHistory.osm, containing the history with all the objects inserted in the

tool OpenStreetMap was downloaded. Next, the osmconvert[1] tool was used, to extract the data referring only to Brazil. For this, the file Brazil.poly was utilized, since it contains the polygon with the respective geographical delimitations of the country, Brazil.

After processing FullHistory.osm and Brazil.poly files in the osmconvert tool, a new file was generated, which was called BrazilHistory.osm. This new file was processed in the osm2pgsql[2] tool, which was responsible for importing the data into the PostgreSQL Database Management System (DBMS). PostGIS and Hstore extensions were used to manipulate spatial data, and to transform the metadata contained in the BrazilHistory.osm file into tags in the key-value format, respectively. The data visualization was done by QGIS software. Figure 1 presents an abstract architecture of the tools used in the analysis of this paper, divided into two layers: one layer for data collection and another for visualization and analysis of the data.

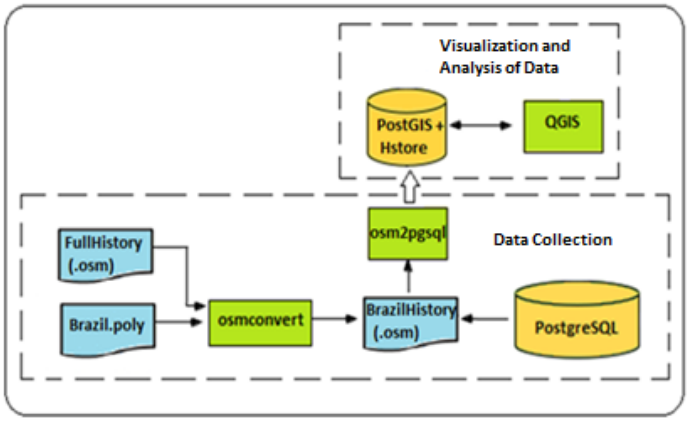

Figure 1. Architecture for data collection and visualization.

5. Results

OpenStreetMap works with three basic primitive types: nodes, ways, and relations. However, the osm2pgsql tool, when importing the data into the PostgreSQL DBMS, performs the conversion of these primitive types to the basic types used to represent data in vector format: points, lines, and polygons.

In this way, points symbolize objects whose location is relevant in the representation, but whose area can be disregarded; lines are used in the representation of paths between points, and the polygons represent objects whose area is relevant in the representation, marking a well-defined region [Monteiro *et al*., 2001].

Since the name attribute is commonly used for identifying the objects in the OpenStreetMap tool, regardless of whether it is referencing a point, line or polygon, this attribute was chosen to indicate data completeness. Thus, to analyze the completeness of the name attribute in the OSM tool, queries were made in the SQL language to search for the percentage of unnamed objects in OpenStreetMap - Brazil, in other words the

[1] http://wiki.openstreetmao.org/wiki/Osmconvert [Accessed in September 2017].
[2] http://wiki.openstreetmap.org/wiki/Osm2pgsql [Accessed in September 2017].

percentage of objects that were not associated with the attribute name. The results are depicted in Figure 2.

As illustrated in Figure 2, the percentage of unnamed points varied considerably between 2007 and 2016, reaching a level close to 20% in the year 2009, and it followed a growing trend until the year of 2013, when it reached the level of 83% of unnamed points. Also through Figure 2, the percentage of unnamed lines varied little, being constantly in the range between 70% and 90%.

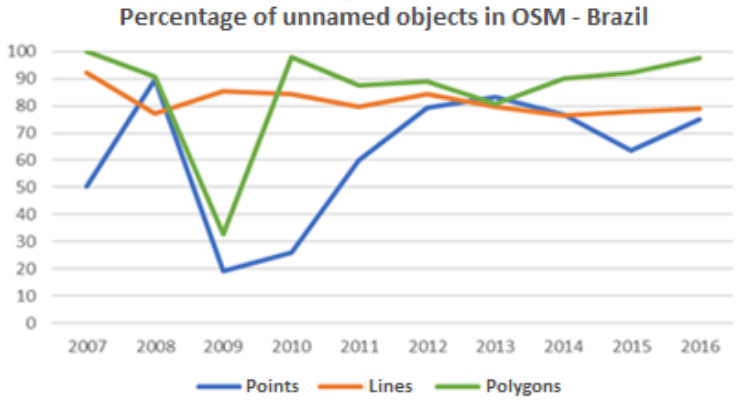

Figure 2. Percentage of unnamed objects in OSM – Brazil.

Since the OpenStreetMap tool is suitable for routing, some analyses have been carried out regarding the insertion of Brazilian road data. Thus, Figure 3 illustrates the evolution of named highways in OSM - Brazil in relation to the years 2008, 2010, 2012 and 2014. Figure 3 shows that many highways were only partially named, having a few stretches named, forming several disjointed sections on the maps. In the first years, the participation of the states of Goiás, São Paulo and Rio Grande do Sul is particularly noteworthy, especially in the year 2012. In the North Region, the Transamazon Highway is almost entirely named within the OpenStreetMap tool. However, there is a greater concentration of designated highways in the Southeast of the country.

Figure 3 also reveals an important aspect of the Temporal Accuracy attribute, which is the heterogeneity in the users' collaboration with the tool over time. Figure 3, clearly shows that the OpenStreetMap objects underwent more major changes between the years 2008 and 2012, than between the years 2014 and 2016, even though both periods were made up of a four-year interval.

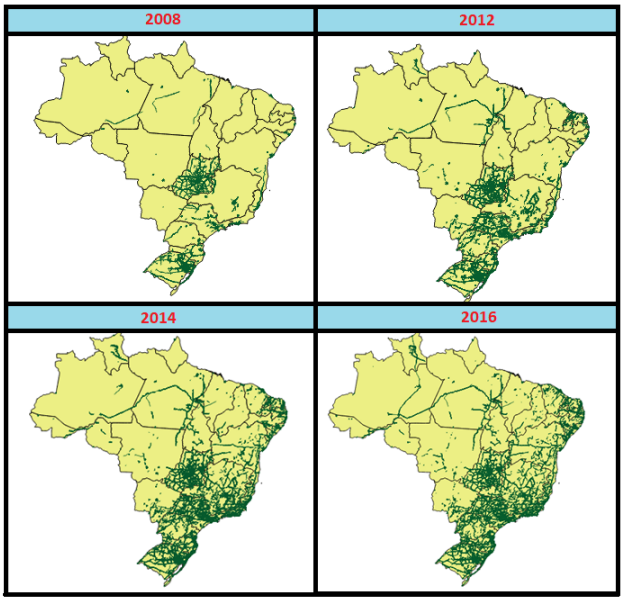

Figure 3. Insertion of named highways in OSM – Brazil.

For the analysis of the Logical Consistency parameter, it is important to observe the rules of modeling and specifications. For example, Figure 4 illustrates the number of buildings that were modeled as both point and polygon in OpenStreetMap – Brazil.

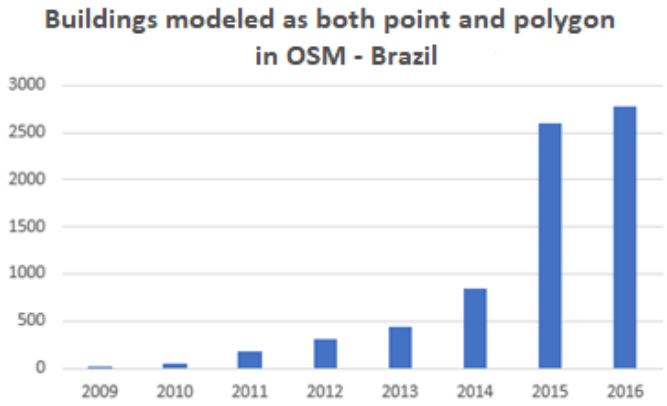

Figure 4. Buildings modeled as both point and polygon in the OSM - Brazil.

Figure 4 shows that there was an increase in the number of buildings modeled as both point and polygon in the OSM - Brazil, highlighting that the jump occurred between the years 2014 and 2015. This fact could indicate a duplication of data within the OpenStreetMap tool.

6. Conclusion

This work presented a set of analysis regarding the quality of collaborative data of the OpenStreetMap tool in Brazil, checking the parameters of completeness, consistency, and temporal accuracy. It was possible to identify that in the OpenStreetMap tool there is still a large amount of incomplete information (e.g., name attribute) and some errors (e.g., buildings modeled both as point and polygon sometimes can represent a duplication of the data). It was also possible to notice that there is a greater concentration of objects named in the South and Southeast regions of the country, and few data in the North region.

As a continuation of this project, further analysis is planned in relation to other quality parameters within the OpenStreetMap tool, as well as some analysis on typical errors found in this type of tool.

References

Barron, C., Neis, P. and Zipf, A. (2014). A comprehensive framework for intrinsic OpenStreetMap quality analysis. Transactions in GIS [1361-1682], 18:877–895

Girres, J. F. and Touya, G. (2010). Quality assessment of the french OpenStreetMap datasettgis. Transactions in GIS, 14(4): p. 435–459.

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

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, 37:682-703.

Mark, A. (2006). "Global Positioning Tech Inspires Do-It-Yourself Mapping Project". National Geographic News.

Medeiros, G. F. B. and Holanda M. T. OpenStreetMap: An analysis of the evolution of geographic data in Brazil. In: 2017 12th Iberian Conference on Information Systems and Technologies (CISTI), 2017, Lisbon. 2017. p. 1

Mondzech, J. and Sester, M. (2011). Quality analysis of OpenStreetMap data based on application needs. Cartographica. 46, 2, 115-125.

Monteiro, A. M., G. Camara, S.D. Fucks e M.S Carvalho (2001): Spatial analysis and gis: A primer. National Institute for Space Research.

Tapscott D and Williams A. D. (2007). Wikinomics: How mass collaboration changes everything. New York, Portofolio Hardcover.

Zhao, Y. and Han, Q. (2016). Spatial Crowdsourcing: Current state and future directions. IEEE Communications Magazine, 54(7): 102–107.