Creating Municipal Databases from OpenStreetMap: The Conceptual Database Schema

Vinícius Garcia Sperandio[1], Vitor Eduardo Concesso Dias[1], Sérgio Murilo Stempliuc[2], Jugurta Lisboa-Filho[1]

Universidade Federal de Viçosa (UFV) - Viçosa, MG, Brazil[1]

Faculdade Governador Ozanam Coelho (FAGOC) - Ubá, MG, Brazil[2]

vinicius.sperandio@ufv.br, vitor.dias@ufv.br,
smstempliuc@gmail.com, jugurta@ufv.br

Abstract. *The systematic use of volunteered geographic information (VGI) has increased in face of the easy production of spatial data from several sources and devices. OpenStreetMap is a VGI platform that collects collaborative geographic data from any region on Earth and makes it openly available free of charge. This paper presents a method for automatic conceptual schema generation based on metadata extracted from the OpenStreetMap platform aiming to create a database to support decision-making in municipal administration. The experiments were carried out on a pilot area, which enabled testing the efficacy of the method to generate conceptual schemas.*

1. Introduction

The popularization of web 2.0 has favored the development of applications that allow users to share spatial information over the Internet. Hence, users become not only consumers, but also contributors and producers of information [Budhathoki 2007]. Ever since, platforms that allow users to be more than consumers and start acting as sensors and volunteered data producers have been standing out [Goodchild 2007]. Volunteered geographic information (VGI) has been the main source of data on platforms such as OpenStreetMap (OSM) [OSM 2018] and Wikimapia [Wikimapia 2018]. These systems allow users to map any part of the globe in a free, rapid, and intuitive manner [Haklay and Weber 2008].

According to Goodchild (2007), official geographic information production has decreased in recent decades especially due to the high cost of generating such information and the cuts in funding for cartographic services. This way volunteered geographic information has been used as a way to minimize this issue.

In Brazil, small municipalities are particularly impacted by restricted funds, besides the lack of qualified labor to produce and maintain geographic information. Such information is important for planning and decision-making toward the economic and social evolution of a city along with environmental matters. A municipality must have access to information on its territory to be administered with more quality and efficiency [Miranda et al. 2012].

These data are usually stored in spatial databases and are handled by geographic information systems (GIS) so that managers are able to carry out spatial analyses that

support decision-making. However, it is important that data modeling be accurate since that is when the real-world elements are transformed into database elements [Elmasri and Navathe 2011]. Consequently, good modeling reduces the need for corrective maintenance, contributes to better understanding of the data stored and their relationships, and allows for a greater number of spatial analyses. That prevents future expenses since maintenance is the software engineering activity that generates the most cost and a high volume of effort.

The OpenStreetMap platform aims to provide and collect collaborative geographic data free of charge from the local knowledge of its users. It thus becomes a free alternative for municipal mapping to which the population itself can contribute, despite lacking training in cartography[1]. In 4 years, its taxpayers mapped 29% of the English territory, achieving 80% of similarity with cartographic bases of national agencies [Haklay 2010]. Nonetheless, in order for data to be used in a decision-making process, it is not enough to directly export the data to a geographic database, but rather the data must be properly structured to be used in spatial analyses through GIS software. For example, QGIS is a free piece of software that is able to read the file exported by the OSM platform and allows data visualization and handling, but the tables of attributes are generated according to the geographic types of each element. Therefore, elements or type *point*, for instance, are added to the same table, causing issues with data normalization. A poorly structured database hampers the understanding of the data acquired and limits the generation of spatial queries.

This paper aims to describe the automated generation process of the conceptual schema from data extracted from the OpenStreetMap platform to create a database. From a conceptual schema, a well-structured database can be automatically generated using CASE tools. Section 2 describes the OpenStreetMap platform, particularly the OSM-XML data file. Section 3 cites some works related to the research. The method used to automatically generate the conceptual schema is described in Section 4. A case study carried out on a small pilot area is presented in Section 5. Section 6 presents some conclusions and the next steps of the project.

2. OpenStreetMap Platform

The OpenStreetMap platform, released on August 9th, 2004, is a collaborative project that aims to allow users to freely and voluntarily map any region in any country [OpenStreetMap 2018]. The platform makes its data available under the Open Database License, thus they can be exported and added to GIS and database management systems (DBMS) with support to spatial data, which enables their broad use [ODbL 2018].

In most collaborative systems, the user is able to create content, add some tags related to content, and share it with other users. The OpenStreetMap platform has a tag system for the mapping elements, where important characteristics to understand the elements can be added, such as informing that the item mapped is a five-story hospital with a helipad [Ballatore and Mooney 2015]. The map features[2] provided by OSM

[1] On the OpenStreetMap platform, volunteered data are approved by moderators, which ensures a minimum quality of data.

[2] https://wiki.openstreetmap.org/wiki/Map_Features

describes and illustrates each tag available so as to increase the odds that the data contributed can be properly rendered by map visualization tools that use the platform.

Users can export the data from any area selected in the platform. The data are extracted as an OSM-XML file with three types of fundamental objects: node, way, and relation [Mooney and Corcocan 2012]. The node represents a point defined by only a pair of coordinates. Nodes are used to represent point-like objects, such as a bus stop, a traffic light, or a monument. Objects of the type way are used to represent linear structures (polyline) such as streets, roads, water courses, or closed regions (polygons), such as buildings and borders. A relation represents the relationship between the previous elements and may be a restriction, such as informing places where vehicle access is restricted, or inform multi-polygons, such as indicating that a set of buildings are part of the same condominium. Besides the elements and their characteristics, an OSM-XML file stores the timeline of updates of each element, featuring the dates and users responsible for the changes.

According to Kitchin (2014), a large volume of data is produced by national censuses and governmental records on municipalities and their citizens. However, these data are based on sampling surveys and there is usually no continuity to these surveys, besides restrictions to data access. Consequently, these volumes of data must be complemented by what can be called small data studies. Questionnaire application, case studies, and interviews are used to capture specific details on issues related to the municipality. According to Miller (2010), much of what is currently known about cities has been obtained from studies characterized by data scarcity. On the other hand, the OpenStreetMap platform seeks to provide a broader understanding of urban control, often in real time, being characterized as a Big Data project for its characteristics such as large data volume, high rate of data generation, data often referenced in time and space, relationships, etc.

3. Related Works

On the OpenStreetMap platform, contributors are free to choose the tags they deem correct to characterize a site or geographic object. The OSM Wiki[3] website has a rulebook with suggestions and instructions on how to attribute a characteristic to an object during a contribution. Davidovic et al. (2016) verified the contributions on the OSM system from 40 cities in different continents to find out whether contributors in these areas are using the rules. After selecting and analyzing ten tags, they concluded that the use of the suggestions for most tags is poor. It is possible that some users do not understand the importance of some attributes and are concerned only with informing geometric aspects of the elements.

Pruvost and Mooney (2017) explored the data model exported by the OSM platform, particularly the relationships contained in it, such as logical clustering of object of types point, line, and polygon, responsible for representing geographic relationships in the real world. The study analyzed the relationships in four European cities and assessed their complexity, composition, and flexibility of the data model in OpenStreetMap.

[3] http://wiki.openstreetmap.org/wiki

The OSM platform uses urban crowdsourcing as a paradigm in the collection of spatial information on municipalities and has proven capable of providing data that can be compared to governmental sources, but data coverage may be low and unevenly distributed across the city. Quattrone et al. (2014) modeled the spontaneous growth of digital information in some areas so as to plan means of collecting content from areas with high chances of being neglected. The research proposed a digital growth model of volunteered spatial information based on urban physical growth models used by urban planners. In order to identify the factors responsible for influencing growth and how they can change with time, the tests were carried out with data from the city of London over five years.

Almendros-Jiménez and Becerra-Terón (2018) developed a framework to analyze the quality of tags applied through the OSM platform in Spain. The evaluation method examines quality measures such as integrity, reliability, and consistency using the website Taginfo[4] as reference. The main cities in Spain were selected to be compared with some European cities and a web tool was developed to enable this type of evaluation anywhere in the world with the same quality indicators.

4. Process of Generating the Conceptual Database Schema for a Municipality

The method proposed for the generation of the conceptual schema of a municipality, using volunteered geographic information, is split into two steps.

The first step made the conceptual modeling of all object classes in the real world that can be mapped in the OSM platform. The conceptual schema was created using the UML-GeoFrame model [Lisboa-Filho and Iochpe 2008]. Figure 1 shows a simplified diagram of themes generated to illustrate the understanding of the universe at hand (municipal base), abstracting the internal content of each theme. Eleven themes and their respective relationships were identified.

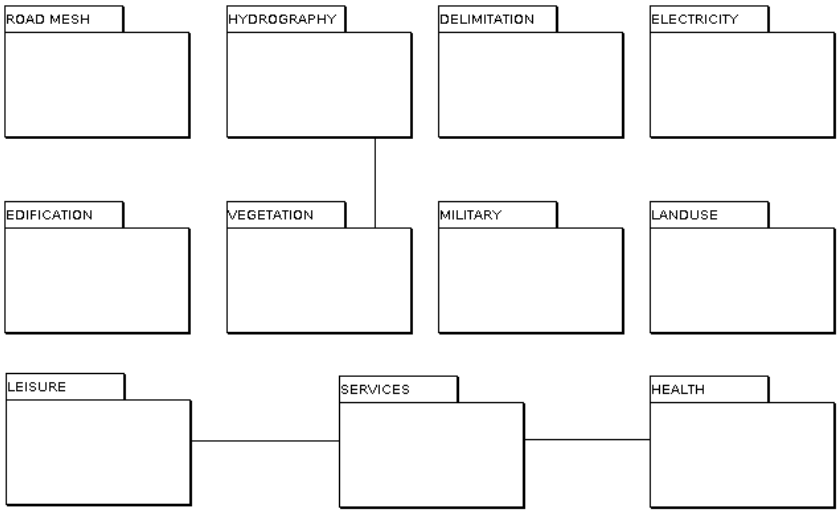

Figure 1. Overall diagram of themes of objects existing on the OSM platform

[4] https://taginfo.openstreetmap.org/

For each theme, the object classes that can be mapped on the OSM platform were modeled according to their level of affinity. Figure 2 illustrates a fragment of this schema and shows real-world object classes separated by themes. Moreover, examples of how the relationships occur among themes can be seen between themes “Leisure” and “Services” or between “Services” and “Health.” It is important to point out that each class has many subclasses, which cannot be exhibited due to space constraints. For instance, class “Transportation” is a subclass of class “Amenity” and has several subclasses (e.g., “taxi,” “parking,” “fuel,” “car_wash”) that were not included in the modeling.

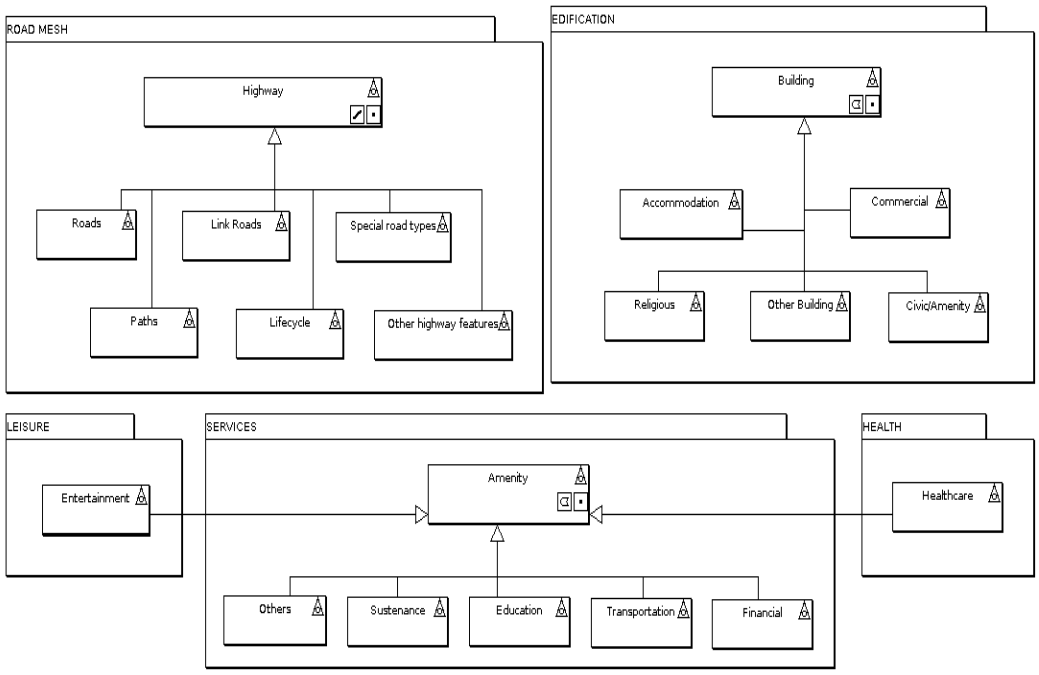

Figure 2. Examples of the class diagram of some themes modeled

The second step consists of reverse engineering from the OSM-XML metadata file extracted from the OSM platform that corresponds to a mapped area of a selected municipality for the creation of the conceptual data schema that contemplates the particularities of said municipality.

Figure 3 illustrates the simplified flowchart of the reverse engineering algorithm proposed whose input is the OSM-XML metadata file. The process starts with reading the OSM-XML file, from which the relevant information is extracted. The BeatifulSoap library [Beautiful Soap 2018] is used in this step to facilitate handling the XML file, which is parsed by the LXML library [Lxml 2018].

Next, the data obtained are tested to separate the mapped objects that have incomplete information from the others. During the separation, it is verified whether the object has a name. In case it does not, the object is stored on a list of nameless objects. This list is then saved as a text file in the format of a report containing the geographic stereotype, the coordinates, and the respective rectangle involving the incomplete objects.

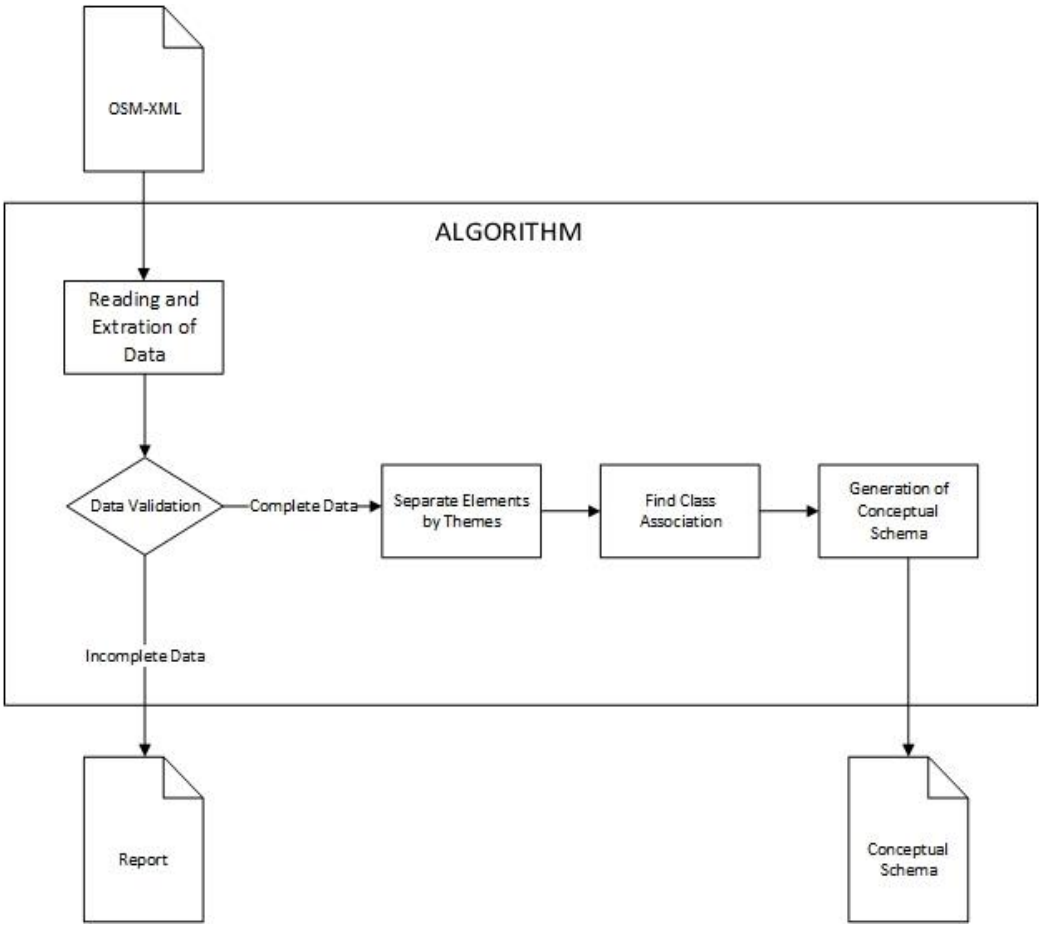

Figure 3. Simplified flowchart of the method proposed for generating the conceptual database schema

Valid objects are separated into lists according to the themes modeled (Figure 1) and the OSM-XM file provides the information regarding which superclass each object belongs to. For example, the part of the OSM-XML file that describes object "Hospital" contains a tag informing that it belongs to superclass "Amenity," hence object "Hospital" will be added to the list that contains objects of theme "Services."

Although the OSM-XML file provides a set of pieces of information (tags) on each object, that is still not enough to classify the exact subclass of objects. Therefore, a search must be performed among the subclasses of the superclass informed to find the exact class of the object. Taking object "Taxi" as an example, the OSM-XML file informs its superclass is "Amenity," thus the subclasses of "Amenity" must be searched (Figure 2) to find the one that best aggregates object "Taxi." That allows creating the relationship "Taxi" -> "Transportation" -> "Amenity."

During this phase of searching classes and relationships, a file is generated containing the layout of the conceptual schema such as color, font, and icons in addition to objects as classes and their connections. In the end, this file is processed by the software Graphviz [Graphviz 2018], which can transform a textual conceptual schema into a graphical schema to generate the UML class diagram.

5. Case Study: Pilot Area – Medicine Department of UFV

The case study consisted in delimiting an area on OSM to enable an in-depth study of the OSM-XML file and future controlled tests. Figure 4 shows the small pilot area chosen, which is located within the Federal University of Viçosa (UFV), in the state of Minas Gerais, Brazil, and illustrates some elements such as a state public school, the Medicine Department, a parking lot, a Health Division, and some roadways.

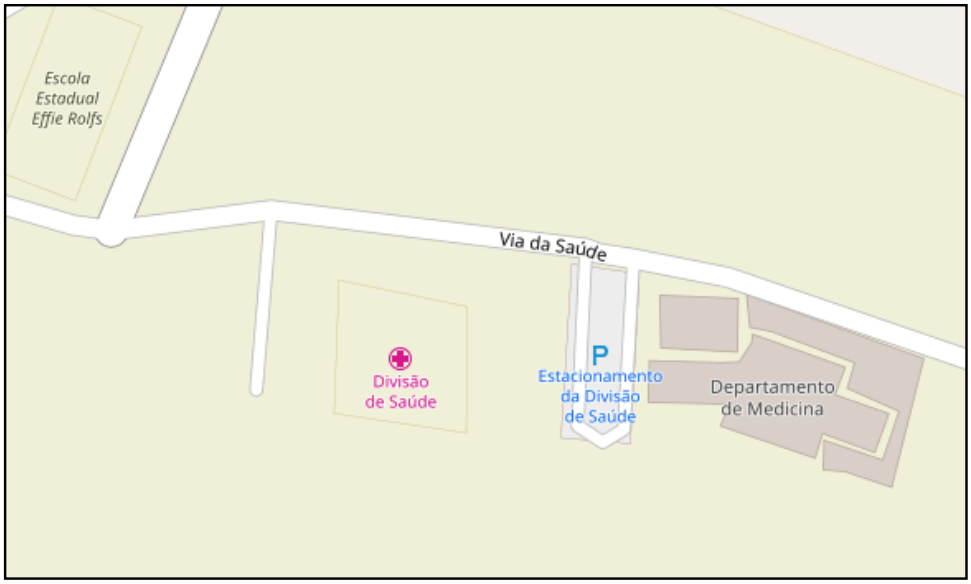

Figure 4. Image of the pilot area – Medicine Department of UFV

When the data of the area selected on the OSM platform is exported, the OSM-XML file is first ordered by all the *node* tags followed by the *way* tags and *relation* tags. All elements have an identifier (id) that can be used for relationships or dependencies. The *node* tag has only a single pair of coordinates (latitude and longitude), which will be used by some *way* tag, as shows Code 1. For example, the *node* tag of Code 1 refers to the coordinates of the upper left point of the Health Division and contains information on date and the volunteer user who created it.

```
<node id="2964381058" visible="true" version="2" changeset="56931769" timestamp="2018-03-06T11:15:43Z"
user="Valério Castro" uid="6237447" lat="-20.7618990" lon="-42.8619154"/>
```

Code 1. Example of node tag with metadata but with no attributes

A *node* tag may contain attributes, as shows Code 2. In this example, besides having a pair of coordinates, the *node* tag has the attribute “turning_circle” of the type “highway” specified in the pair <key k, value v>. This element corresponds to the small roundabout near the public school.

```
<node id="3907519974" visible="true" version="1" changeset="36137846" timestamp="2015-12-24T02:14:55Z"
user="Artur Vieira" uid="2180959" lat="-20.7624231" lon="-42.8655997">
 <tag k="highway" v="turning_circle"/>
</node>
```

Code 2. Example of *node* tag with attributes

The *way* tag has a similar structure to the *node* tag previously exemplified. The difference lies in *nd* tags, which reference *node* tags without attributes. Code 3 illustrates the Health Division element and its attributes. The first *nd* tag references code number 2964381058, which is the id of the node described in Code 1. The last pair of coordinates of the element is the same as the first to delimit a polygon, which is why the first *nd* tag is the same as the last.

```
<way id="292881345" visible="true" version="1" changeset="24162154" timestamp="2014-07-15T14:33:12Z"
user="Artur Vieira" uid="2180959">
 <nd ref="2964381058"/>
 <nd ref="2964381059"/>
 <nd ref="2964381060"/>
 <nd ref="2964381061"/>
 <nd ref="2964381058"/>
 <tag k="amenity" v="hospital"/>
 <tag k="name" v="Divisão de Saúde"/>
</way>
```

Code 3. Example of *way* tag

Running the reverse engineering algorithm with the OSM-XML file regarding the pilot area generated the conceptual schema illustrated by Figures 5 and 6, in which all elements mapped in Figure 4 are present with their respective relationships, themes, and attributes. It is noteworthy that the modeling returned two classes of the type "University," which corresponds to the Federal University of Viçosa (UFV) in Figure 5 and to the Medicine Department building in Figure 6. That occurred because the exported area contains part of the border of UFV, observed in the upper right corner in Figure 4 by a change in the background color of the image.

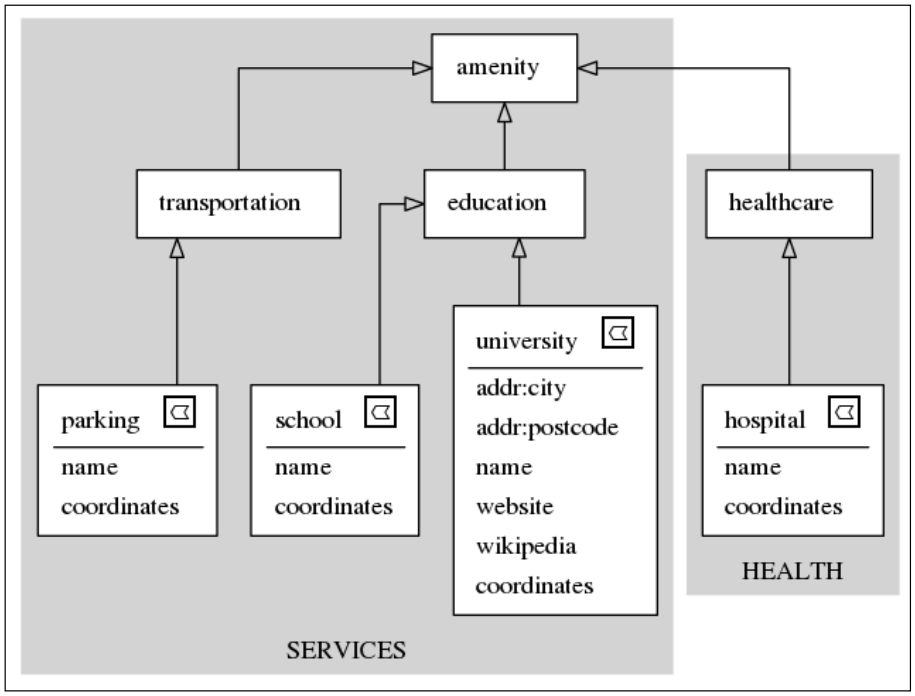

Figure 5. Conceptual schema generated from the pilot area of themes SERVICES and HEALTH

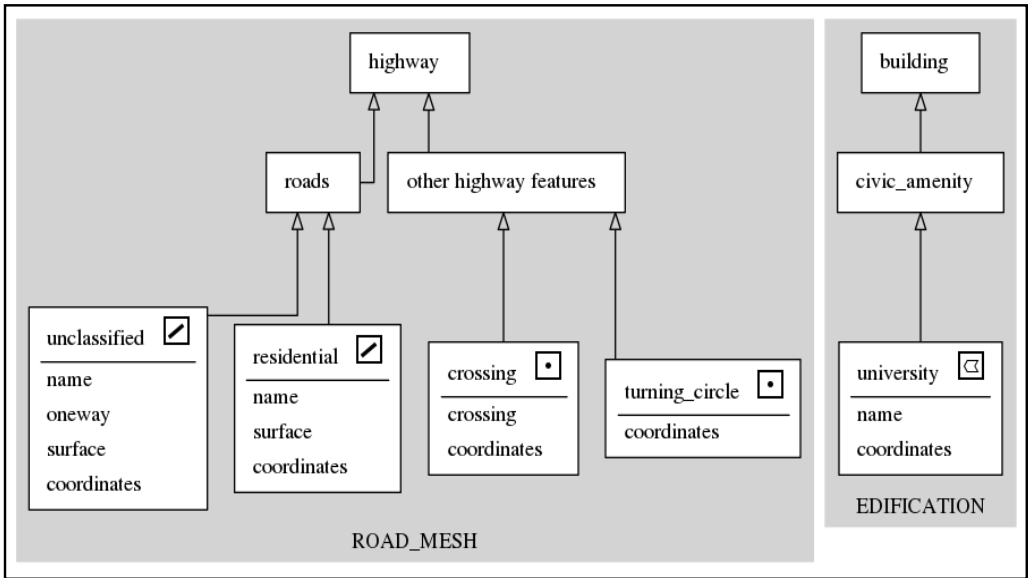

Figure 6. Conceptual schema generated from the pilot area of themes ROAD_MESH and EDIFICATION

Besides the conceptual schema, a file is generated with the elements that have no name. An excerpt of this file is illustrated by Figure 7, informing the object stereotype, its coordinates, and its involving rectangle to facilitate locating it on the map.

```
line
-20.7618025, -42.8611124
-20.7623424, -42.8611392
-20.7623925, -42.8610575
-20.7623528, -42.8609882
-20.7618343, -42.8609575
 -------------------------------------
|             -42.8609575             |
|-20.7623925              -20.7618025  |
|             -42.8611392             |
 -------------------------------------
```

Figure 7. Excerpt of the file generated with the unidentified tags

Finally, a second test was carried out with the aim of evaluating the effectiveness and efficiency of the algorithm. A load test was performed with all data available at the OSM of the city of Viçosa in Minas Gerais, with a total area of approximately 300km² and its demographic density of 241.20 inhabitants per square kilometer. The algorithm was running for 103 minutes on a personal computer, recognizing 73 entities belonging to the "ROAD MESH", "SERVICE" and "EDIFICATION" packages. The source code and the load test result are available through the link https://github.com/vinisperandio/OSM2Diagram.

6. Final Considerations and Future Works

This paper describes the process of automated generation of conceptual schemas from a file exported by the OpenStreetMap platform to create a geographic database for municipal administration based on volunteered geographic information (VGI).

The process is performed in two steps, the first consisting of creating a complete conceptual schema of the database available on the platform, i.e., contemplating all elements described in the specification of the OSM-XML file generated by this platform. The second step processes a reverse engineering algorithm that turns an XML file with data on a given selected area exported from the OpenStreetMap platform into a class diagram following the UML-GeoFrame model.

The results obtained in the pilot project allowed verifying that the work proposed presents a simple method to obtain volunteered data and an algorithm that, even in its initial version, proved capable of generating quality conceptual schemas. This is a good first step for municipalities with limited funds for cartographic services, because it allows to acquire knowledge of the features belonging to the municipality in an intuitive way, in order to facilitate administrative decision making. The method proposed is an initiative for researches that match free software and VGI systems in public administration. Silva et al (2018) show the potential of VGI such as a tool to support municipal managers in the decision-making process based on spatial analysis.

Future works include enhancing the performance of the algorithm presented and turning it into an application capable of yielding the script responsible for creating a NoSQL and relational geographic database in addition to the schemas.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

References

Almendros-Jiménez, J., and Becerra-Terón, A. (2018). Analyzing the Tagging Quality of the Spanish OpenStreetMap. *ISPRS International Journal of Geo-Information*, 7(8): 323.

Ballatore, A., and Mooney, P. (2015). Conceptualising the geographic world: the dimensions of negotiation in crowdsourced cartography. *International Journal of Geographical Information Science*, 29(12): 2310-2327.

Beautiful Soap (2018). Beautiful Soap 4.4.0 Documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Budhathoki, N. R. (2007). Reconceptualization of user is essential to expand the voluntary creation and supply of spatial information. In *Proceedings of Workshop on Volunteered Geographic Information*.

Elmasri, R. and Navathe, S. B. (2011). Sistema de Banco de Dados. Pearson AddisonWesley, 6th edition.

Goodchild, M. F. (2013). The quality of big (geo) data. *Dialogues in Human Geography*, 3(3): 280-284.

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

Graphviz – Graph Visualization Software, (2018). Welcome to Graphviz. http://www.graphviz.org/

Haklay, M., and Weber, P. (2008). Openstreetmap: User-generated street maps. *Ieee Pervas Comput*, 7(4): 12-18.

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*, v. 37, n. 4, p. 682–703, doi:10.1068/b35097

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal*, 79(1): 1-14.

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.

Lisboa-Filho, J. and Iochpe, C. (2008). Modeling with a UML profile. In Shashi Shekhar and Hui Xiong (Eds.) *Encyclopedia of GIS*, pages 691–700. Springer.

Lxml, (2018). Lxml – XML e HTML com python. https://lxml.de/

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, 50(1): 181-201.

Miranda, T. S., Lisboa Filho, J., Souza, W. D., Silva, O. C., and Davis Junior, C. A. (2011). Volunteered geographic information in the context of local spatial data infrastructures. In *Urban data management symposium* (UDMS), pages 123-138.

Mooney, P., and Corcoran, P. (2012). The annotation process in OpenStreetMap. *Transactions in GIS*, 16(4): 561-579.

Open Data Commons (2018). Open Data Commons Open Database License (ODbL). https://opendatacommons.org/licenses/odbl/

OpenStreetMap (2018). OpenStreetMap (OSM), https://www.openstreetmap.org , July.

Silva, L. P. et al. (2018). Bases cartográficas para municípios de pequeno porte geradas por informação geográfica voluntária. *Revista Brasileira de Cartografia* (no Prelo).

Pruvost, H., and Mooney, P. (2017). Exploring Data Model Relations in OpenStreetMap. *Future Internet*, 9(4): 70.

Quattrone, G., Mashhadi, A., Quercia, D., Smith-Clarke, C., and Capra, L. (2014). Modelling growth of urban crowd-sourced information. In *Proceedings of the 7th ACM international conference on Web search and data mining,* pages 563-572. ACM.

Wikimapia (2018). Wikimapia, http://wikimapia.org , July.