VGI Protocol and Web Service for Historical Data Management

Rodrigo M. Mariano[1], **Karine R. Ferreira**[1], **Luis A. C. Ferla**[2]

[1] National Institute for Space Research (INPE)
São José dos Campos – SP – Brazil

[2]Federal University of São Paulo (UNIFESP)
Guarulhos – SP – Brazil

{rodrigo.mariano,karine.ferreira}@inpe.br, ferla@unifesp.br

Abstract. *Volunteered Geographic Information (VGI) is a phenomenon that uses the web to produce, assemble and disseminate geographic information provided by volunteers. VGI techniques generate detailed geographical data with low cost, taking advantage of citizens local knowledge. The definition of a VGI protocol is crucial to improve the quality of citizen-derived geographical data sets collected by a project. Protocols are also important to facilitate the reuse of VGI data for other projects and applications different from what was originally collected. This paper presents a VGI protocol that was defined for the Pauliceia 2.0 project and a web service that was built based on this protocol. Pauliceia 2.0 project aims to use VGI and crowdsourcing techniques to produce historical geographical data sets of São Paulo city from 1870 to 1940.*

1. Introduction

VGI, citizen science, crowdsourcing and collaborative mapping are examples of different terms used to refer to the general subject of collaborative work and citizen-derived geographical information. See et al. [See et al. 2016] present a good review of these terms and categorize them according to three main aspects: (1) information or process; (2) active or passive contributions; and (3) spatial or non-spatial user-generated information.

The term VGI was first defined by Goodchild [Goodchild 2007] as *"the harnessing of tools to create, assemble, and disseminate geographic data provided voluntarily by individuals"*. Goodchild and Li [Goodchild and Li 2012] define VGI as a version of crowdsourcing, focused on manipulating geographical information. Estellés-Arolas and Guevara [Estellés-Arolas and González-Ladrón-de Guevara 2012] define crowdsourcing as *"a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit."*.

There are many projects that use VGI techniques to collect data, such as OpenStreetMap, Wikimapia and Flickr. OpenStreetMap is the most well-known general platform that implements VGI successfully [Goodchild and Li 2012]. It is an editable map

of the world, provided by volunteers, being possible to handle free geographic data [OpenStreetMap 2017a]. It adopts the local expertise of the users to make updated maps.

Mapping agencies use robust protocols that drive the geographic data collection, while VGI projects regularly contain lack of standards or just supply vague instructions. The definition of a VGI protocol is crucial to improve the quality of citizen-derived geographical data sets collected by a project and to facilitate the reuse of these data sets for other projects and applications [Mooney et al. 2016].

Mooney et al. [Mooney et al. 2016] propose a generic protocol to drive VGI projects. This protocol establishes crucial issues that must be defined in the context of a VGI project in order to improve the understanding of volunteers and so the quality of the data produced by them. These issues include vector geographical data collection and management, user control, self-assessment and quality metrics and feedback to the community.

Pauliceia 2.0 project aims to build a computational platform for collaborative historical research based on VGI and crowdsourcing techniques [Ferreira et al. 2017]. Through this platform, citizens can contribute to produce historical geographical information of São Paulo city from 1870 to 1940. These contributions can be done in different ways, for example, by doing the vectorization of streets and buildings from historical maps or by uploading photos and information about historical places. Besides that, this platform allows historians to share data sets resulting from their researches.

This paper presents a VGI protocol for historical data that was defined for the Pauliceia 2.0 project and a web service based on this protocol, called VGI Management Web Service (VGIMWS), that was built in the Pauliceia platform. Ferreira et al. [Ferreira et al. 2017] introduce the Pauliceia project and its platform generally; while this paper presents a detailed description of the VGI protocol for historical data and VGIMWS service.

2. Related Work

This section presents projects that also use VGI and crowdsourcing techniques to produce historical geographical information, similar to Pauliceia 2.0 project.

OpenHistoricalMap [OpenStreetMap 2018] and HistOSM[1] projects are built on the OpenStreetMap (OSM) platform. OpenHistoricalMap is an effort to use the OSM infrastructure to produce a universal and historical map of the world. HistOSM is a web application to explore the historical objects of OSM, such as castles, ruins, monuments and memorials.

Building Inspector[2] is a web-based platform that allows citizens to produce, correct and analyze data from historical maps of New York city from 1853 to 1930. In this project, Budig et al. [Budig et al. 2016] propose a consensus polygon algorithm to extract a single polygon to represent each building from all polygons provided voluntarily.

ATLMaps[3] is an web portal of the Atlanta Explorer project that handles historical information of Atlanta city for post Civil War to 1940 [Page et al. 2013]. This portal

[1] http://histosm.org/
[2] http://buildinginspector.nypl.org/
[3] https://atlmaps.org/

allows user to visualize and explore historical maps, events and places. Users can produce their own projects using the layers that are ready to use in the portal and contribute with audios, annotations or images related to them.

Many VGI projects use the OSM infrastructure and its API (Application Programming Interface) to build their web platforms. In the beginning of the Pauliceia 2.0 project, our team evaluated if it was possible to build the Pauliceia 2.0 platform using the OSM infrastructure and its API. However, after some studies, we concluded that the OSM data model and operation are not suitable for the Pauliceia project due to the following reasons:

1. In the Pauliceia platform, historians can share data sets resulting from their researches and these data sets can not be edited by anyone. In OSM, data can be updated by anyone.
2. The historical features in the Pauliceia database are spatiotemporal, that is, they have a period to indicate when they existed. There are features that do not exist today anymore. In OSM database, the entities are not spatiotemporal. OSM considers that all entities stored in its database exist today.
3. In the Pauliceia platform, the historical data sets are organized in layers, while OSM data sets are not.
4. The community and domain of the Pauliceia project are very specific and structured, while the OSM is very general. The Pauliceia project has a specific domain with a particular spatial and temporal scope, generating a structured community.

Therefore, we decided not to use the OSM infrastructure and API. We defined a specific VGI protocol for the Pauliceia project and built a web service for VGI data management based on this protocol. These protocol and web service are crucial parts of the Pauliceia platform and they are described in the next sections.

3. VGI Protocol for Historical Data

Mooney et al. [Mooney et al. 2016] propose a generic protocol that organizes the issues related to citizen-derived geographical data management in five main stages, that are shown in Figure 1: (1) Initialisation; (2) Vector data collection; (3) Self-assessment and quality control; (4) Data submission; and (5) Feedback to the community.

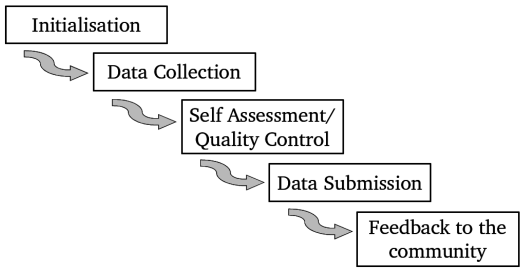

Figure 1. Main stages of VGI Protocol [Mooney et al. 2016]

A VGI protocol defines crucial issues that improve the understanding of volunteers about the project and all its mechanisms and methods to collect, manage and assess the quality of the citizen-derived geographical data. Thus, this helps to improve the quality of the data sets collected by volunteers in a project.

In this work, we define a VGI protocol specific for the Pauliceia 2.0 project following the guide proposed by Mooney et al. [Mooney et al. 2016]. The Pauliceia 2.0 VGI protocol is described in next sections.

3.1. Data types and initialization

In the Pauliceia 2.0 project, volunteers can upload and edit vector geographical data. The Pauliceia platform provides tools that allow citizens to include and edit geometries, such as points, lines and polygons, as well as textual and numerical values associated to geographical entities or features. The platform does not provide tools to edit and create raster data types.

One of the project goals is to use VGI and crowdsourcing techniques for the vectorization of features, such as streets and buildings, from historical raster maps. In this case, the data set gathered by volunteers can have a set of distinct geometries to represent the same feature. To extract the most accurate geometry to represent a single feature from this data set, we intend to employ methods that compute a single geometry that represents the majority opinion, as proposed by Budig et al. [Budig et al. 2016].

The users access the Pauliceia 2.0 platform through an online browser. Before starting the contribution, the collaborator needs to register himself/herself to the platform and accept a "Use Term" that describes mainly that the portal is not responsible for the collected data sets and that these data sets are public. The registration can be done by creating a new user or by using a social login through Google and Facebook accounts. Everybody can access the platform and visualise its data sets freely, however just registered volunteers can edit or add new historical data sets.

The Pauliceia 2.0 database is made available under the Creative Commons Attribution-ShareAlike 4.0 license (CC BY-SA)[4]. In a nutshell, this license authorises the people freely copy, share, modify and use the data for any propose, since the users cite Pauliceia 2.0 and its contributors. If the user reproduces the data, he or she must use the same license for the results.

To motivate volunteers to vectorize streets and buildings as well as historians to share their historical data sets, we intend to organize events oriented to this purpose, following the same idea of the mapathon events promoted by Google Maps [Tech2 2014] and OpenStreetMap [OpenStreetMap 2017b]. These events, called HistMapathon (Historic Mapping Marathon), will be organized in universities with historians and their students to promote the mass contribution of geographical data in the Pauliceia 2.0 platform.

3.2. Data model

Figure 2 shows the concepts of the Pauliceia 2.0 VGI protocol and their relationships, using an entity-relationship diagram. Its main concepts are: user, layer, reference, keyword and notification.

In the platform, the data sets are organized in layers as in GIS (Geographic Information System). A layer groups geographical features related to a subject that are described by the same set of properties. Each feature has spatial and non-spatial properties [Herring 2006]. The spatial properties are represented by geometric data types, such as

[4]https://creativecommons.org/licenses/by-sa/4.0/

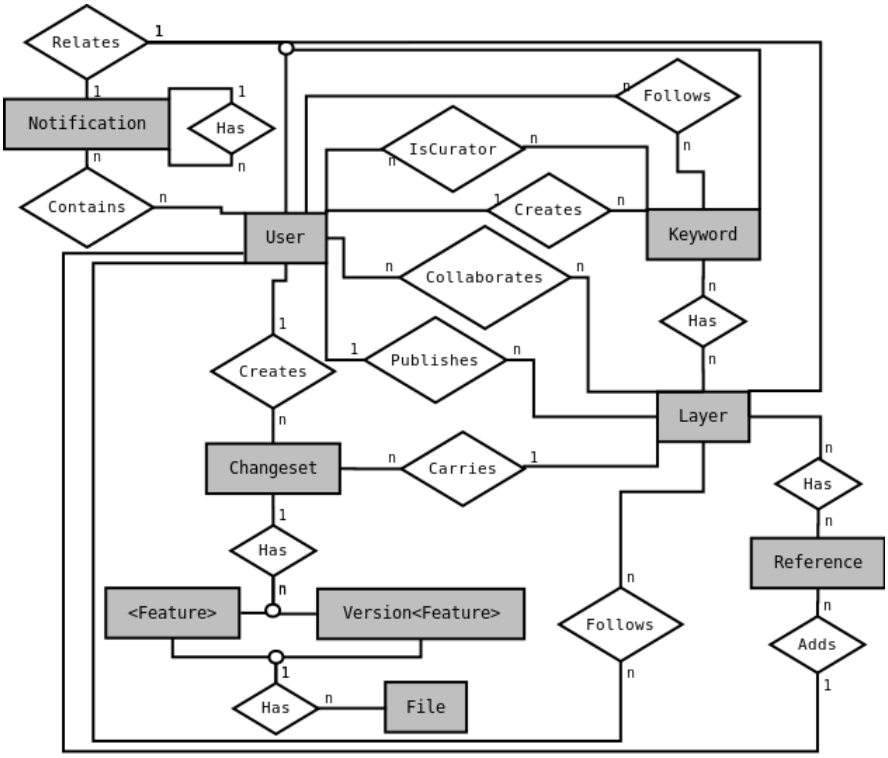

Figure 2. The Pauliceia 2.0 project data model

points, lines and geometries. The non-spatial properties are represented by alphanumerical data types, such as texts and numbers. A non-spatial property of a feature can contain links to media or documents, e.g photos and videos, that are stored in other repositories, like Google Drive, YouTube or Dropbox.

A layer contains a set of features and their versions along time. The changeset entity controls the features of a layer and their versions along time, keeping the history about when and what user realized each change. A changeset is a group of changes related to the features of a layer made by users in a period.

A layer can be associated to one or more keywords (e.g. crimes or factories) and to one or more bibliographical references, such as book, thesis or article. In the platform, the keywords are used in the search mechanism to select layers associated to specific themes. Each layer has an owner user who creates it in the platform and a set of collaborators. Collaborators are users that have permission to edit, delete and include new features into a layer. The collaborators of a layer are defined by its owner. Users can only edit data sets in layers where they are collaborators.

The communication among the platform users, called Pauliceia community, is done through notifications. Notifications can be reviews of data, comments or denunciation. Users can write notifications about a specific layer or about an another notification. Besides that, they can write general notifications for all Pauliceia community.

In the platform, there are two types of users: logged and unlogged ones. The un-

logged users can visualize, search and download the platform data sets as well as read all its notifications. The logged users can be of three types: normal, curator and administrator. The normal users can edit and contribute with new data sets, creating new layers and associating them with keywords and references. The curator users can edit all layers of the platform by by adding new keywords to them. The administrator users have permission to create, edit and remove all entities of the platform.

3.3. Data collection methods

There are two ways of data collection methods in the Pauliceia 2.0 platform: manual contribution and bulk import.

In the manual edition, users manually create and edit the spatial locations or boundaries of features by clicking on the historical maps presented in the web portal. Besides that, users can edit manually all attribute values associated to the features.

In the bulk importing, users can upload a group of features stored in well-known file formats of vector geographical data, such as shapefile or geojson. In the manual edition, users have to inform all metadada associated to the features. In the bulk importing, some types of metadada can be extracted automatically by the platform from the file content.

The shapefiles generated by the spatiotemporal geocoding tool of the Pauliceia platform can be directly used in the bulk import. Using the spatiotemporal geocoding tool, a user can upload a CSV (Comma Separated Values) file that contains a set of textual historical addresses and get a shapefile with all spatial locations of these addresses produced by the geocoder. So, this shapefile can be imported in the platform, creating a new layer in the Pauliceia database.

3.4. Quality control

In the literature, there are several proposals to evaluate VGI data quality. In a nutshell, these methods are described as quality measures, quality indicators and quality approaches. Quality measures verify the accuracy of VGI data in relation to the authoritative data provided by mapping agencies. Quality indicators measure the quality of VGI data in an abstract way when there is not authoritative data for comparison [Senaratne et al. 2017]. Quality approaches determine the degree of a fact, if it is possible to be true, and it can be automated or used of human intervention [Goodchild and Li 2012]. One proposal of VGI data quality introduced by Goodchild and Li [Goodchild and Li 2012] is the crowdsourcing approach. One interpretation of this approach is the use of the Linus Law[5]. Linus Law is the ability of using the people to verify the contributed data of VGI to converge to the truth.

To use quality measures, it is necessary to use authoritative data for comparison [Senaratne et al. 2017]. In the Pauliceia 2.0 project context, there is no authoritative data and so we can not use quality measures. Thus, the Pauliceia 2.0 team evaluated different types of quality indicators and approaches to evaluate the citizen-derived geographical data of the project, such as gamification (e.g. ranking) [Senaratne et al. 2017], trustworthiness and user reputation [D'Antonio et al. 2014]. Nevertheless, a consensus

[5]"Given enough eyeballs, all bugs are shallow" [Raymond 2017]

was reached that the best thing would be to adopt a crowdsourcing approach, using notifications and denunciations provided by the Pauliceia 2.0 community.

In the Pauliceia 2.0 platform, a notification is a comment from a user related to a layer or to another comment. A user can write a notification describing the positives or negatives points of a layer, warnings or suggestions to improve it, such as suggestions of new bibliographic references related to the layer. A denunciation is a special kind of notification made to alert administrators that a layer contains inappropriate data (e.g. copyright data or owned by another researcher). The administrators of the platform receive these reports, evaluate the layers associated to denunciations and can remove them from the platform as well as its owner user.

3.5. Feedback to the Community

A collaborative project become better as more users assist in it. Hence, it is important that users supply feedbacks about their experience with the project [Mooney et al. 2016]. In the Pauliceia 2.0 platform, volunteers are encouraged to make comments, give opinions and observations about their experiences on the platform, indicating the positive aspects and suggestions to improve it. This feedback can be done by the available mailing list and social networks that are managed by the Pauliceia 2.0 team. The feedback is important to improve the platform.

Using the Pauliceia 2.0 platform, researchers can disseminate and share their historical data sets as well as receive feedback from other researchers about them through notifications. Such data sets will be freely available on the portal and thus will achieve a greater visibility in the scientific community and dissemination.

Historians can write notifications on their layers or on the layers of other users, providing feedback on their status, such as remarks, praise or hints (e.g. indicate a new reference for that layer). Users can also write general notifications for all members of the Pauliceia 2.0 community, such as research event announcements. If a user finds unappropriated data in the platform, he or she can report it through denunciations. Besides that, users may follow layers of interest from other authors and receive notifications about them by e-mail.

4. VGI Management Web Service for Historical Data

Figure 3 shows the Pauliceia 2.0 platform architecture [Ferreira et al. 2017]. The platform is open source, online and service oriented. Service-oriented systems are well appropriate to supply a better interoperability among applications. The spatiotemporal data sets of the project are stored in a PostgreSQL database with spatial extension PostGIS (vector data) and in GeoTIFF files (raster data).

The platform architecture contains two groups of web services. The first group contains standards of geographic web services specified by the Open Geospatial Consortium (OGC), such as Web Map Service (WMS), Web Feature Service (WFS), Web Coverage Service (WCS) and Catalogue Service Web (CSW). The second group is composed of two web services designed and implemented to augment the functionalities of the OGC standard services, attending to specific and crucial demands of the Pauliceia 2.0 project.

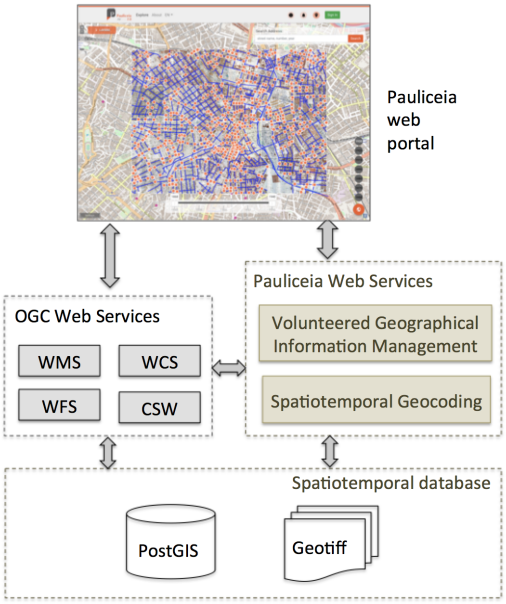

Figure 3. Pauliceia 2.0 platform architecture. [Ferreira et al. 2017]

This section describes the Volunteered Geographical Information Management Web Service (VGIMWS) that was designed and built based on the Pauliceia VGI protocol described in section 3. It provides all necessary functionalities for dealing with historical citizen-derived geographical information, such as user control, spatiotemporal features management as well as user edition of notifications and denunciations.

Figure 4 shows the architecture of VGIMWS. It is a RESTful web service developed in Python language. The chosen standard for data exchange is GeoJSON and JSON, that handle data with geographic information or without, respectively. It provides function to create, edit and remove all concepts described in section 3, such as user, layer, features and notifications.

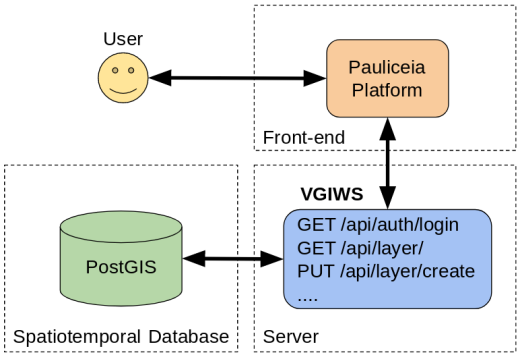

Figure 4. VGIMWS architecture.

All software codes of Pauliceia 2.0 project are free and open source and can

be found at the Github of the project [6]. This Github contains the source code of the VGIMWS[7], its documentation[8] and instruction about how to run the web service and to install its dependencies.

Figure 5 shows a sequence diagram of one function of the VGIMWS that creates a new layer. First, the user tries to log in the platform using one URL and the VGIMWS returns a HTTP status, success or error. If the user is able to log in the platform, he or she can enter on the dashboard, create a new layer and associate keywords and references to it. After that, the volunteer can import a shapefile using the bulk import or create an empty layer. Lastly, the user must inform other metadata about the layer, such as its temporal columns.

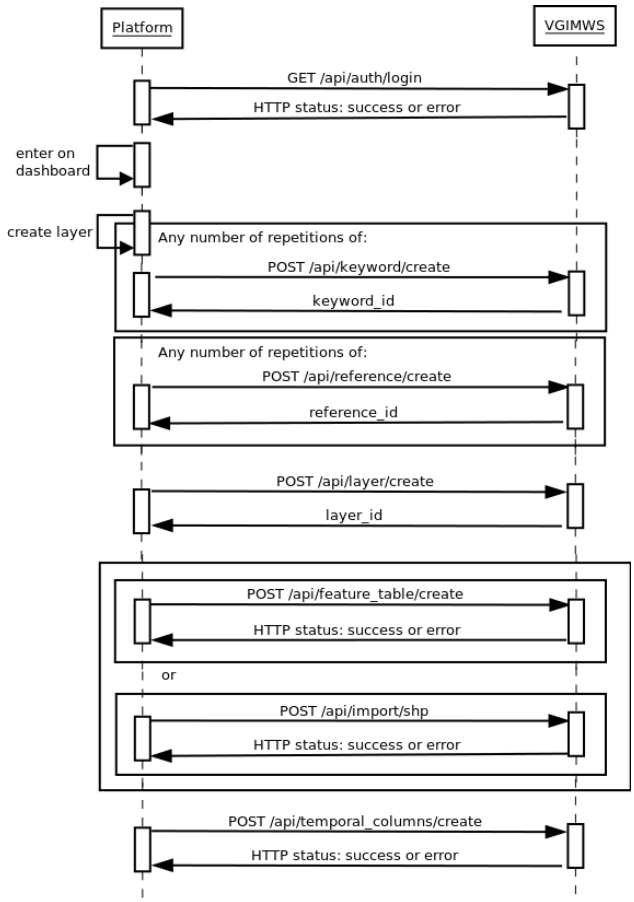

Figure 5. Add a new layer.

A set of tables is proposed to store metadata related to the temporal information and media attributes of the layers, shown in Figure 6. It is an extension of the Simple Feature Access Model proposed by OGC. The GEOMETRY_COLUMNS and the

[6]https://github.com/Pauliceia
[7]https://github.com/Pauliceia/vgiws
[8]https://github.com/Pauliceia/vgiws/blob/master/doc/README.md

FEATURE TABLE are tables defined by OGC, while the MEDIA_COLUMNS, TEMPORAL_COLUMNS and MASK are proposed in this work.

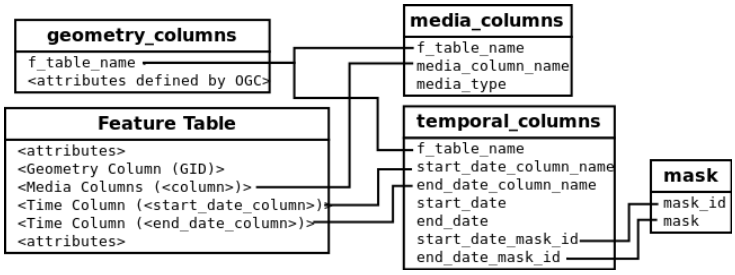

Figure 6. Feature metadata tables.

FEATURE TABLE is a table that stores features, where the columns are the attributes and the rows are the features. GEOMETRY_COLUMNS is a table that contains metadata about the geometry properties of a feature table [Herring 2006]. MEDIA_COLUMNS is a table that contains the metadata about the properties of a feature table that are links to media, such as photos or videos, that are stored in other repositories, as Google Drive, YouTube or Dropbox. TEMPORAL_COLUMNS is a table that defines the temporal attributes of the feature table. Its attributes are the start date, end date and the temporal bounding box of a feature. MASK is a table that saves the possible masks for the start date and end date, susch as "YYYY-MM-DD". Both MEDIA_COLUMNS and TEMPORAL_COLUMNS contain a reference to a feature table that is registered in the GEOMETRY_COLUMNS.

Figure 7 presents the complete spatiotemporal database model of Pauliceia 2.0 project. This model express the conceptual model described in section 3.2. It contains tables to store the concepts of the Pauliceia 2.0 project, such as user, layer, reference, keyword, notification, changeset, feature table, media, temporal information and followers. This database was built using PostgreSQL with the spatial extension PostGIS.

5. Conclusion

VGI has emerged with the purpose of collecting geographical data sets fast and with low cost. However, to improve the quality and the reuse of these data sets, it is necessary to define protocols to guide VGI projects.

This paper presents a VGI protocol for historical data that was defined in the context of Pauliceia 2.0 project. This project aims to develop an online platform for collaborative research of historical data, using VGI and crowdsourcing techniques. Besides that, this paper describes a RESTful web service, called VGIMWS, that was built in the Pauliceia platform based on the VGI protocol. VGIMWS manipulates all the protocol concepts through specific URLs.

The VGI protocol helps to increase the quality of the historical citizen-derived geographical data of the Pauliceia 2.0 platform. It defines crucial issues that improve the understanding of volunteers about the Pauliceia project and all its mechanisms and methods to collect, manage and assess the quality of the citizen-derived geographical data. The proposed VGI protocol and the VGI management web service are generic, so both can be used to other collaborative historical project.

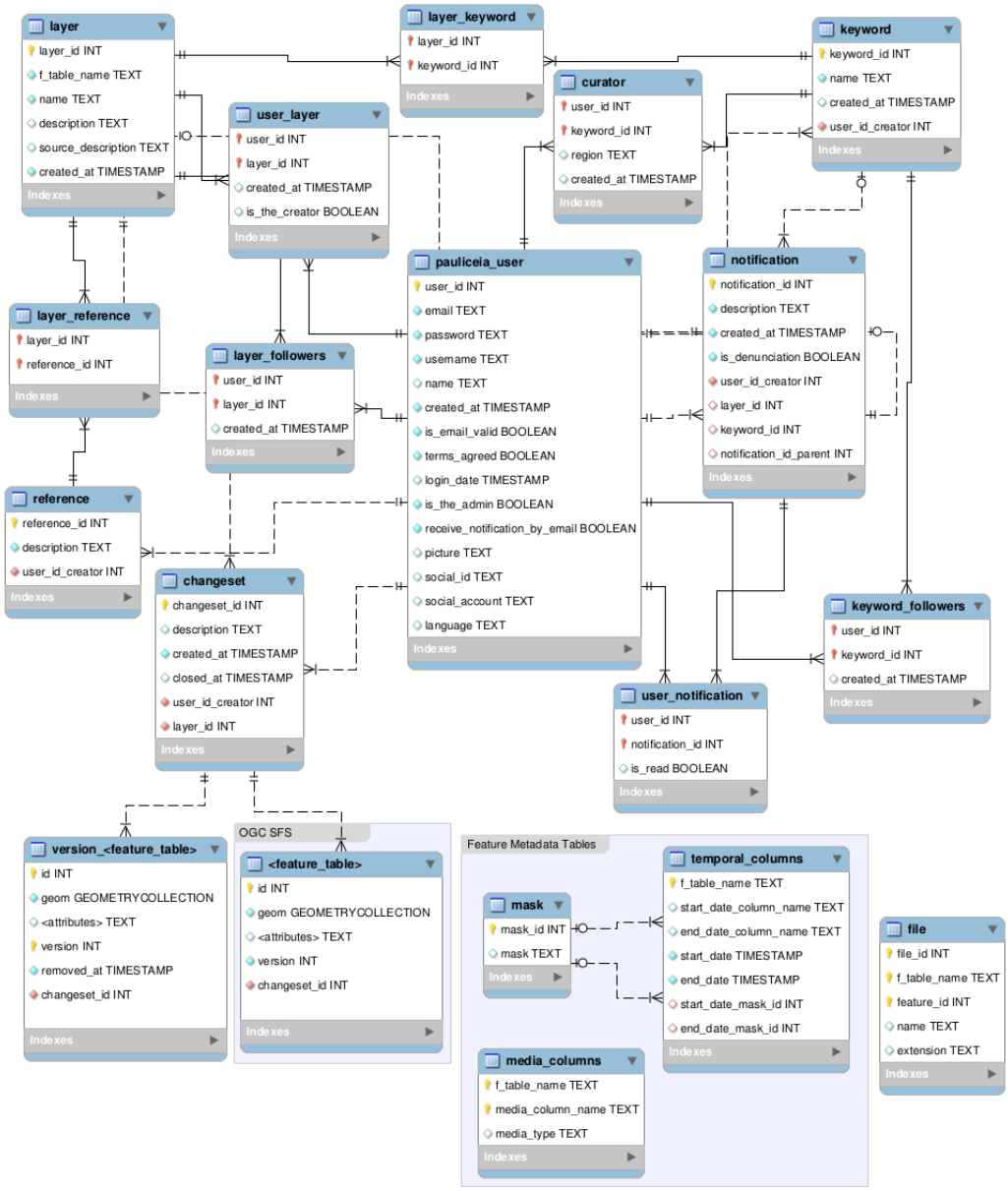

Figure 7. Database Model of the Pauliceia 2.0 project.

6. Acknowledgment

Pauliceia 2.0 project is funded by São Paulo Research Foundation (FAPESP) eScience Program, grant #2016/04846-0. The authors thank FAPESP for granting student scholarship #2017/03852-9.

References

Budig, B., van Dijk, T. C., Feitsch, F., and Arteaga, M. G. (2016). Polygon consensus: smart crowdsourcing for extracting building footprints from historical maps. In

Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, page 66. ACM.

D'Antonio, F., Fogliaroni, P., and Kauppinen, T. (2014). Vgi edit history reveals data trustworthiness and user reputation.

Estellés-Arolas, E. and González-Ladrón-de Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200.

Ferreira, K. R., Ferla, L., de Queiroz, G. R., Vijaykumar, N. L., Noronha, C. A., Mariano, R. M., Wassef, Y., Taveira, D., Dardi, I. B., Sansigolo, G., Guarnieri, O., Musa, D. L., Rogers, T., Lesser, J., Page, M., Britt, A. G., Atique, F., Santos, J. Y., Morais, D. S., Miyasaka, C. R., de Almeida, C. R., do Nascimento, L. G. M., Diniz, J. A., and dos Santos, M. C. (2017). Pauliceia 2.0: A computational platform for collaborative historical research. *Proceedings XVIII GEOINFO, December 04th to 06th, 2017*, pages 28–39.

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

Goodchild, M. F. and Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics*, 1:110–120.

Herring, J. (2006). Opengis implementation specification for geographic information-simple feature access-part 2: Sql option. *Open Geospatial Consortium Inc*.

Mooney, P., Minghini, M., Laakso, M., Antoniou, V., Olteanu-Raimond, A.-M., and Skopeliti, A. (2016). Towards a protocol for the collection of vgi vector data. *ISPRS International Journal of Geo-Information*, 5(11):217.

OpenStreetMap (2017a). About. `https://www.openstreetmap.org/about`. Accessed on 02/08/2018.

OpenStreetMap (2017b). Mapathon. `http://wiki.openstreetmap.org/wiki/Mapathon`. Accessed on 18/08/2018.

OpenStreetMap (2018). Open historical map. `https://wiki.openstreetmap.org/wiki/Open_Historical_Map`. Accessed on 05/08/2018.

Page, M. C., Durante, K., and Gue, R. (2013). Modeling the history of the city. *Journal of Map & Geography Libraries*, 9(1-2):128–139.

Raymond, E. S. (2017). Release early, release often. `http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/ar01s04.html`. Accessed on 10/08/2018.

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.

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.

Tech2 (2014). Why is google's mapathon in hot waters in india? all you need to know. http://www.firstpost.com/tech/news-analysis/why-is-googles-mapathon-in-hot-waters-in-india-all-you-need-to-know-3655197.html. Accessed on 18/08/2018.