A Platform for Collaborative Historical Research based on Volunteered Geographical Information

Karine R. Ferreira[1*], Luis Ferla[2*], Gilberto R. de Queiroz[1], Nandamudi L. Vijaykumar[1], Carlos A. Noronha[1], Rodrigo M. Mariano[1], Denis Taveira[1], Gabriel Sansigolo[1], Orlando Guarnieri[2], Thomas Rogers[3], Jeffrey Lesser[3], Michael Page[3], Fernando Atique[2], Daniela Musa[2],Janaina Y. Santos[4], Diego S. Morais[4], Cristiane R. Miyasaka[2], Cintia R. de Almeida[2], Luanna G. M. do Nascimento[2], Jaine A. Diniz[2] and Monaliza C. dos Santos[2]

[1] INPE - National Institute for Space Research, São José dos Campos – SP – Brazil
[2] UNIFESP - São José dos Campos and Guarulhos – SP – Brazil
[3] Emory University - The Halle Institute for Global Learning, The Emory Center for Digital Scholarship, and The Department of History – Atlanta – United States
[4] Arquivo do Estado de São Paulo – São Paulo – SP – Brazil
[*]{karine.ferreira@inpe.br, ferla@unifesp.br}

Abstract. Digital humanities research promotes the intersection between digital technologies and humanities, emphasizing free knowledge sharing and collaborative work. Based on digital humanities features, this paper describes the architecture of a computational platform for collaborative historical research designed and developed in an ongoing project called Pauliceia 2.0. This project aims to produce historical data of São Paulo city from 1870 to 1940 and to develop a computational platform that allows researchers to explore, integrate and share urban historical data sets. The Pauliceia 2.0 platform main goal is to use volunteered geographical information (VGI) and crowdsourcing concepts to produce past geographical data and to allow historians to share historical data sets resulting from their researches. In this work, we present the Pauliceia 2.0 platform architecture and its underlying VGI protocol.

Categories and Subject Descriptors: CCS 2012 [**Information systems**]: Data management systems

Keywords: Digital humanities, Volunteered geographical information, Crowdsourcing, Spatiotemporal geocoding, historical data

1. INTRODUCTION

Digital Humanities (DH) research takes place at the intersection of digital technologies and humanities, producing and using applications and models that make possible new kinds of teaching and research both in humanities and in computer science [Terras 2012]. The digital humanities community is interdisciplinary, linking together the humanistic and computational approaches. This community includes people with different backgrounds that come together around values such as openness and collaboration [Spiro 2012].

DH have drawn increasing institutional support and intellectual interest among scholars working on historical research in universities throughout the world. Historians working within digital humanities promote the use of Geographical Information Systems (GIS), among other tools, to understand historical data. DeBats and Gregory [DeBats and Gregory 2011] argue that GIS has directly contributed to the advancement of knowledge in history, mainly in urban history.

Free knowledge sharing and collaborative work have indeed become core features of digital humani-

ties [Spiro 2012]. The role of the world network of computers, in particular web 2.0, has boosted these aspects of the field. It places value not only on the broad dissemination of studies and investigations, but also on opportunities for collaboration and putting into practice those theoretical values. Nowadays historians can benefit from a wide variety of technological options to disseminate their research widely and to participate in collaborative investigation across boundaries of time and space.

In the literature, there is a variety of terms used to represent the general subject of collaborative work and citizen-derived geographical information, such as volunteered geographical information (VGI), science 2.0, crowdsourcing and collaborative mapping. See et al.[See et al. 2016] present a good review of these terms, providing some basic definitions and highlighting key issues in the current state of this subject. The authors categorize these terms according to three main aspects: (1) information or process that can be used to generate it; (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". According to [Estellés-Arolas and González-Ladrón-de Guevara 2012], crowdsourcing is "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".

In GIScience, many efforts have been made to propose general frameworks and protocols that can be followed by the VGI projects [Davis Jr et al. 2013] [Mooney et al. 2016]. Davis Jr. et al. [Davis Jr et al. 2013] propose a general framework for VGI applications, based on coordinating both web-based tools and mobile applications. Mooney et. al. [Mooney et al. 2016] propose a protocol for the collection of vector data in VGI projects. Besides providing a standard for data collection, the protocol also guides users to contribute and to improve the overall data quality of the project, which positively impacts their motivation to keep on providing information. In the recent years, an increasing number of projects have used crowdsourcing and VGI concepts to produce historical geographic data. Examples of crowdsourcing projects to collect historical geographic information are shown in next section.

This paper describes the architecture of a computational platform for collaborative historical research designed and developed in an ongoing project called Pauliceia 2.0. This project has two main objectives. The first is to collect, select and digitize historical data of São Paulo city from 1870 to 1940. During this period the city went through a dramatic process of urbanization, almost unique in terms of contemporary history. This transformation was taken as a challenge by several historians to investigate a range of issues within this period. The second goal is to design and build a computational platform that allows researchers to explore, integrate and publish urban historical data sets. This platform will appeal to historians to not only explore historical data sets provided by the project, but also to contribute by including and sharing their own knowledge and data sets.

The Pauliceia 2.0 platform main goal is to use VGI and crowdsourcing techniques to produce past geographical data and to allow historians to share data sets resulting from their researches. In this work, we present the VGI protocol and the spatiotemporal geocoding method for historical data defined and developed in the Pauliceia 2.0 platform. This paper is an extended version of Ferreira et al. [Ferreira et al. 2017], published and presented in XVIII Brazilian Symposium on Geoinformatics (GeoInfo 2017) (http://www.geoinfo.info/geoinfo2017/). This paper presents more details about the Pauliceia VGI protocol, such as its data model, the spatiotemporal geocoding and the results of the project.

2. RELATED WORK

In this section, we present some projects that have similar features to Pauliceia 2.0 and highlight differences between the computational platform proposed in this paper and the ones provided by such

projects.

OpenStreetMap (OSM) is the most well-known general platform that implements VGI successfully. It allows users to edit and work with free geographical data, following an open content license. There are many applications that are built on top of the OSM database. Two examples of OSM applications that focus on historical data sets are HistOSM[1] and OpenHistoricalMap[2]. HistOSM is a web application to visually explore historic objects stored in the OpenStreetMap database, such as monuments, churches and castles. OpenHistoricalMap is an effort to use the OSM infrastructure as a foundation for creating a universal, detailed, and historical map of the world.

The Atlanta Explorer project creates historical geodatabases, geocoders and 3D models of Atlanta city post Civil War to 1940 [Page et al. 2013]. ATLMaps web portal[3] allows users to explore historical maps, Atlanta Explorer geodatabases, and students generated content of Atlanta city about different subjects, such as historical events, sites and land use. The project members argue that it presents a broad potential for using crowdsourced information about particular sites and structures. For now, the project portal does not allow citizen-derived geographical information.

The New York Public Library promotes a crowdsourcing project to create polygonal representation of building footprints and attributes from insurance atlases from 1853 to 1930 of New York City. This project provides a web-based application called Building Inspector[4] that allows citizens to extract, correct and analyze data from historical maps. The volunteered information is used in training machine learning algorithms to recognize building shapes and other data on digitized insurance atlases. 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 by citizens in this project.

The Digital Harlem website[5] is based on legal records, newspapers, archives and published sources, to provide information on everyday life in New York City's Harlem neighborhood in the years 1915 - 1930. The website enables users to look for events and places and create interactive web maps. The Digital Harlem historical database was created by the project members, without using crowdsourcing and VGI concepts.

The British Library has a project to employ crowdsourcing to georeference a vast collection of historical maps and to disseminate them through a web portal [Southall and Pridal 2012]. This project provides an online georeferencer tool [6] that enables overlaying historic maps with modern ones from which one may compare the past with the present and georeference these historical maps. The Library originally turned to crowdsourcing in 2011 and since then five releases of maps have been made public, with extremely successful results. Participants georeferenced 8,000 maps and after undergoing a check for accuracy they were duly approved. They also developed a portal Old Maps Online[7] with a geographic search interface to identify and view historic maps from a variety of available collections.

The project Lx Conventos aims to develop an online system with spatial and temporal navigation on data sets related to the dissolution of religious orders in the dynamics of urban transformation in nineteenth century Lisbon [Gouveia et al. 2015]. These data sets include a large set of multimedia information, such as historic and contemporary cartography and georeferenced photos, videos and 3D models, provided by the projects partners.

Perret et al. [Perret et al. 2015] describe a project that creates the roads and streets of France in

[1]http://histosm.org
[2]http://www.openhistoricalmap.org/
[3]https://atlmaps.org/
[4]http://buildinginspector.nypl.org/
[5]http://digitalharlem.org
[6]http://www.georeferencer.com
[7]http://www.oldmapsonline.org/

the 18th century by digitization of historical maps using collaborative methodology. However, it is not clear in the paper whether the collaboration is by trained operators or by the general public and specialists in history. Cura et al. [Cura et al. 2017] present another project from France that deals with collaborative geocoding in History. The authors propose a solution that is open source, open data and extensible for geocoding based on the building of gazetteers that have geohistorical objects collected from historical topographical maps. The case study was Paris in the 19th-20th centuries. The results can be visualized over modern or historical maps and even verified and/or edited in a collaborative manner. They store several instances of the same space at different moments in history that can be pictured as a snapshot of a given instant. The system enables collaborative editing, but the user profiles of those who collaborate and post content into the system are not clear.

ImagineRio[8] is another initiative from an American University that provides a platform to understand the evolution, both social and urban, of Rio de Janeiro, Brazil, looking into the entire history of the city. Several views from the perspective of artists, historical maps and architectural plans, in space and time have been organized. It is an open-access digital library. It is possible to relate elements within a web environment in which a streaming of data (vector, spatial and raster) is conducted. Such data can also be inspected, toggled, visualized and naturally queried. It is quite valuable for architects, urbanists, and scholars to consult or view some particular spatiotemporal aspects of the history of the city. An interesting aspect of this project is the availability of a mobile app to enable interested parties and tourists to explore the city.

Pauliceia 2.0 project has many similarities with these projects and has been influenced by most of them, following a strong trend towards urban history and its relationship with space. Several projects described in this section use crowdsourcing and VGI concepts to vectorize specific features from historical maps, such as building footprints and streets, to georeference historical maps as well as to geocode historical places. The main difference between the similar projects and Pauliceia 2.0 is that we are proposing a computational platform that allows historians to share geographical data sets resulting from their researches on São Paulo city. The related projects described in this section are not designed for research sharing.

Using crowdsourcing and VGI concepts, the Pauliceia 2.0 platform's main goal is to allow researchers to share their historical data sets, providing a proper environment for collaborative work. Besides that, this platform allows citizens to help the project team to vectorize streets and buildings from historical maps. Pauliceia 2.0 is a platform for digital humanities based on free knowledge sharing and collaborative work.

3. PLATFORM ARCHITECTURE

The Pauliceia 2.0 platform is open source, web-based and service-oriented. Its architecture is shown in Figure 1. It is implemented using the GIS library TerraLib and the web geoportal framework TerraBrasilis developed by INPE [Câmara et al. 2008]. Service-oriented architectures are suitable for data and functionality exchanging across systems, promoting a better integration and interoperability among technologies. The Pauliceia 2.0 spatiotemporal vector data is stored in a PostGIS database system and raster data in Geotiff files.

The architecture has two groups of web services. The first group is composed of geographical web services defined by the Open Geospatial Consortium (OGC): Web Map Service (WMS) for map images, Web Feature Service (WFS) for vector data, Web Coverage Service (WCS) for coverage data, and Catalogue Service Web (CSW) for metadata of spatiotemporal data, services and related objects [Open Geospatial Consortium 2017]. OGC has played a crucial role in geospatial data interoperability by proposing web services standards for visualizing, disseminating and processing geospatial data.

[8]http://hrc.rice.edu/imagineRio/home

The dissemination of the Pauliceia database through OGC web services is important for interoperability, integration with other applications and data sharing. The Brazilian National Infrastructure for Spatial Data (INDE) specification is based on OGC web services. The purpose of INDE is to catalog, integrate and accommodate the existing geospatial data produced and maintained by agencies of the Brazilian Government so that they are easily located, explored and accessed for a wide variety of uses through the internet. The Pauliceia 2.0 historical data sets are disseminated according to INDE specification.

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. The first service, called "Volunteered Geographical Information Management", provides all necessary functionalities for dealing with citizen-derived historical information. Such functionalities include user control; insertion, edition and deletion of spatiotemporal data sets as well as user notifications and reviews. The second service provides a spatiotemporal geocoding method for historical data. In the next sections, we detail these two web services.

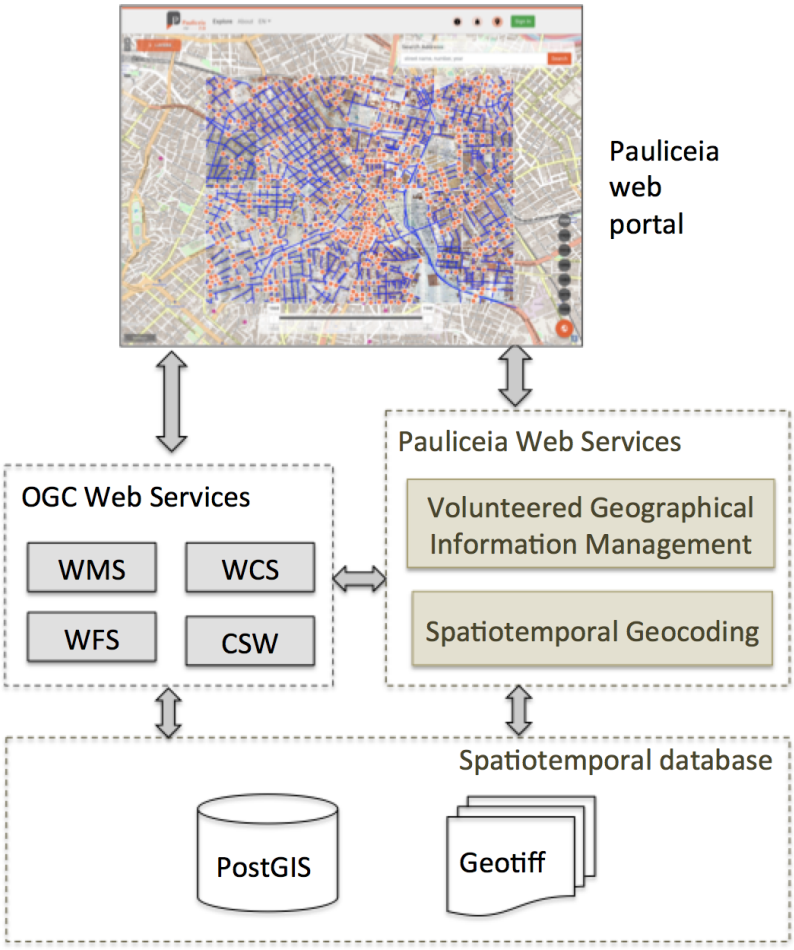

Fig. 1. Pauliceia platform architecture.

The Pauliceia web portal provides graphical interfaces that allow users to interact easily with the platform and its web services. Through this portal, users can select, visualize, filter, query and

download historical data sets, write notifications about these data sets as well as contribute with new historical data.

4. VOLUNTEERED GEOGRAPHICAL INFORMATION FOR URBAN HISTORY

In the recent years, an increasing number of projects have used crowdsourcing and VGI concepts to produce historical geographic data. In the Pauliceia 2.0 project, we intend to use VGI and crowdsourcing techniques for the vectorization of streets and buildings from historical raster maps and for the collection and sharing of historical data sets resulting from researches of historians. All these data sets should be restricted to the urbanization period of the São Paulo city from 1870 to 1940, which is the historical scope of the Pauliceia 2.0 project.

Even though it is possible to reach high standards of data quality with VGI projects, comparable to those collected by National Mapping Agencies (NMAs) and Commercial Surveying Companies (CSC) [Ludwig et al. 2011], [Graser et al. 2014], [Ciepłuch et al. 2010], the lack of a rigorous protocol is often a major source of errors and an obstacle to the wider dissemination of VGI initiatives [Mooney et al. 2016].

To ensure the data quality in VGI applications, it is necessary to establish a protocol that balances the need for meticulous data collection strategies and the motivation for contributors to follow its guidelines. Given the importance of creating a VGI protocol, Pauliceia 2.0 project defines its own protocol based on the proposal of Mooney et. al. [Mooney et al. 2016]. In this section, some topics stipulated in the Pauliceia 2.0 VGI protocol are described, such as data types, data collection methods, data model, quality control mechanisms and feedback to the community. The VGI Management web service shown in Figure 1 is implemented based on this protocol.

4.1 Data types and collection methods

The Pauliceia 2.0 platform provides tools that allow volunteers to create and edit spatial locations and boundaries of features using vector data types, such as points, lines and polygons. Besides geometries, volunteers can create and edit attribute values associated to features using textual or numerical data types as well as links to photos, videos and documents that must be stored in other platforms such as YouTube and Dropbox. 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].

All data sets in the Pauliceia 2.0 platform will be available under the Creative Commons Attribution-ShareAlike 4.0 license (CC BY-SA)[9]. Basically this license allows people freely copy, share, adapt and use data for any propose, even for commercial purposes, since users properly credit the Pauliceia project and its contributors. Besides that, if users create new information from the available data sets, they must use the same license for the results.

In the platform, there are two types of collection methods: manual edition and bulk importing. In the manual edition, users create and edit the spatial locations or boundaries of features by clicking on the on the historical maps presented in the web portal drawing area. Besides that, users can edit all attribute values associated to these 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.

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

In the manual edition, users have to provide all metadada associated to the features. In the bulk importing, some types of metadada can be automatically extracted by the platform from the file content.

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. These events will explore tutorials about the platform and how to contribute, such as mapathons promoted by Google Maps and OpenStreetMap. These events can be organized in universities with historians and their students to promote the mass contribution of geographical data in the Pauliceia 2.0 platform.

4.2 Data model

The data model of the Paulicea 2.0 VGI protocol, shown in Figure 2, contains four main concepts: user, layer, keyword and notification. Everyone can visualize and access the project data sets through the web portal, but only registered users can add contributions and edit data sets. Users can register in the platform using social login from Facebook and Google, after accepting the project Use Policy terms. These terms specify that the data provided by the platform is public under the CC BY-SA license and the platform is not responsible for any issues that may arise from users providing copyrighted data.

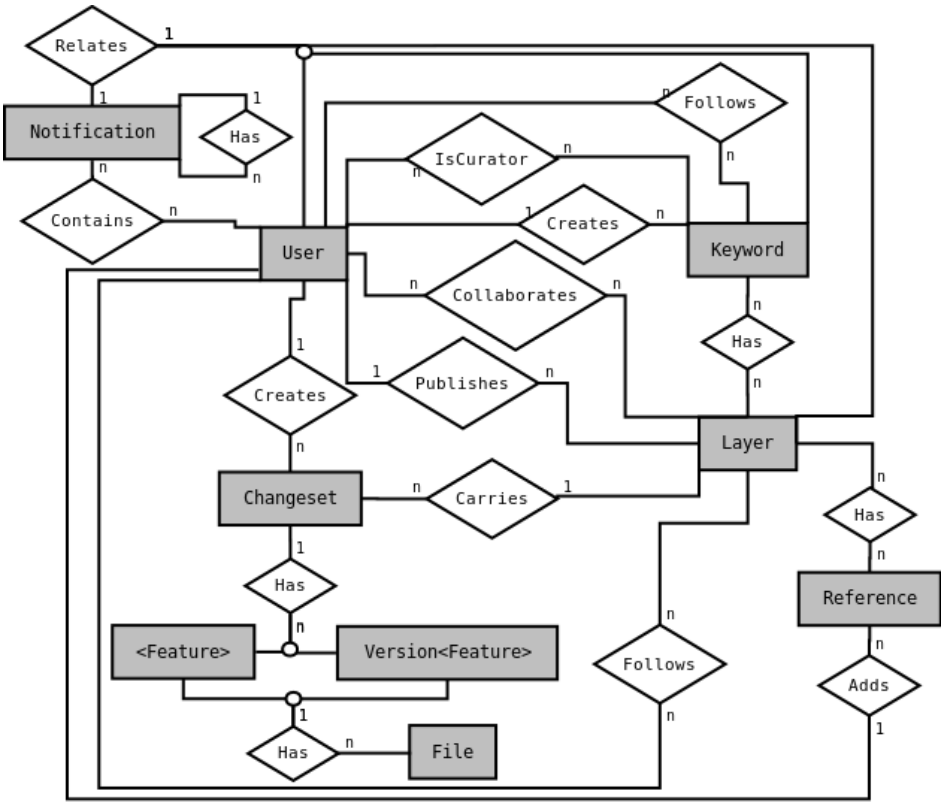

Fig. 2. Pauliceia 2.0 VGI protocol data model

In the platform, the data sets are organized in layers as in GIS. A layer groups geographical entities or features related to a subject that are described by the same set of properties. A layer contains a set of features and their versions along time. Besides that, a layer can be associated to one or more bibliographical references.

The features of a layer and their version along time are controlled by the entity Changeset of the model. It contains the history of the features, when and what user updated them. A changeset is a group of changes related to the features of a layer made by users in a period.

A layer has an owner user who creates it in the platform and a group of users that are their collaborators. A user can own one or more layers and a layer has only an owner user, represented by the "Publishes" relationship. 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. If a user wishes to be part of a layer as collaborator, it is necessary to make a request to its owner. Users can only edit data sets in layers where they are collaborators.

Each layer is associated to one or more keywords, such as public health and cultural places. The keywords can be defined by any user and are managed by the users that are curators in the platform. Keywords are used to search layers associated to specific themes in the platform.

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

4.3 Quality control and feedback

With respect to quality control, users are expected to self-assess the data, checking for coherence, adequate quality and correctness of the attributes, before submitting it to the platform. Once the data is submitted to the platform, all users of the Pauliceia community can access it and write reviews or comments about it using notifications. Taking into consideration the fact that the target audience of this platform is people with prior knowledge of the field (historians and students), it is expected that they use their own knowledge to point out errors or inconsistencies in the project database.

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. Thus, notifications and denunciations work as a mechanism of quality control maintained by the Pauliceia 2.0 community.

A collaborative project progresses as more users participate in it. Therefore, it is important to improve the user experience as a means of encouraging more contributors to join the platform. The user will be encouraged to provide feedback about his experience with the platform, commenting about the positive aspects and what needs improvement, giving opinions, making observations and suggesting changes. This feedback can be provided via mailing lists or social media, and will be used as an important base for improving the platform.

5. SPATIOTEMPORAL GEOCODING FOR HISTORICAL DATA

A crucial feature of the Pauliceia 2.0 platform is to provide functionalities that allow historians to share historical geographic data resulting from their researches. In this case, most historical data sets have textual addresses to indicate spatial locations in the past. Thus, it is necessary to provide a geocoding algorithm able to transform historical textual addresses into geographical coordinates.

Geocoding is the process of transforming textual data into geographical information. Obtaining coordinates from textual addresses is one of the most important functionalities provided by Geographical Information Systems (GIS) [Martins et al. 2012]. Address geocoding has to deal with challenges related to variations in textual addresses, such as abbreviations and missing parts. Martins et. al [Martins et al. 2012] propose a geocoding method for urban addresses whose output includes a geographic certainty indicator, which provides the expected quality of the results.

In the literature, there are many proposals of efficient geocoders for current addresses, but they do not deal with historical data [Martins et al. 2012] [Asher et al. 2009] [Lee 2009]. A geocoder for historical information must operate on spatiotemporal data sets, that is, spatial entities whose geometries and attributes vary over time. The challenges of creating an address geocoding system for historical data are mainly related to the variation of names, geometries and numerations of streets and buildings over time. In the Pauliceia 2.0 database, every spatial entity, such as a street segment and a place with an address, has an associated period that indicates when it is valid. Thus, the geocoding method for this database has to take into account all valid periods associated to spatial entities.

Cura et al [Cura et al. 2017] argue that historical geocoding requires dedicated approaches and tools due to three reasons. The first is that existing geocoding services do not consider the temporal aspect of the data sets they rely on. They implicitly work on a valid time that is the present. The second reason is that traditional geocoders are based on a complete and strict hierarchy, such as city, street, and house number, which is verifiable. Historical data, however, are full of uncertainties and are not directly verifiable. One has to check possibly incomplete and conflicting available historical sources and, very often, make assumptions or hypotheses. The third is that the historical sources available to construct a geocoding database are sparse (both spatially and temporally), heterogeneous, and complex. Based on these reasons, Cura et al [Cura et al. 2017] propose an open source solution for geocoding that is based on gazetteers of geohistorical objects extracted from historical maps.

The geocoding web service that is being designed and implemented in the Pauliceia 2.0 project has to consider all these particularities of historical data sets. Using this service, historians can geocode a single address or a set of addresses via CSV files. Each address has to contain its street name, number and year. The service computes geographical coordinates associated with the addresses using the historical places and street segments stored in the project database. Besides the geographical coordinates, the service returns a degree of certainty associated with each coordinate. This value indicates how confident the geographical coordinate generated by the service is, based on the number of available historical entities that were used in the geocoding process.

To populate the project database, we designed and implemented a web portal for historical address edition [10] shown in Figure 3. This portal provides functionalities to insert, delete, edit and search historical addresses. Through this portal, project members are collecting and inserting historical addresses of São Paulo city from 1870 to 1940, such as houses, buildings, churches and squares, into the project database. Each address has a street name, a location number, a period when it is valid and a geographical coordinate that is informed by clicking on the historical map in the portal. The stored addresses can be viewed in an intuitive and simultaneous way by several users registered in the system.

The geocoding service relies on the historical addresses and street segments stored in the project database. So, the construction of a good quality database is crucial to the success of the geocoding process. The greater the number of addresses identified, the greater the accuracy of the resulting base. The project members are using different types of historical sources, such as legislative documents, newspapers, license plate books and advertisement leaflets, to collect these addresses.

The address geocoding process is depicted in Figure 4. In the first frame (Figure 4A) the streets are shown as linear features, with a name and an interval of validity. The historical places are features represented by points (spatial locations) and several attributes, such as name, location, and the validity interval. The relationship between each historical place and its street is also materialized in the database. Therefore, in order to geocode an address such as "Augusta Street, 50, 1937", a street in 1937 named "Augusta" is searched first. Then the places associated with this street, whose time interval contains the year of "1937" are recovered (Figure 4B). Finally, a linear interpolation of the searched address is computed taking the linear geometry of the found street (Figure 4C).

[10] http://www.pauliceia.dpi.inpe.br/edit

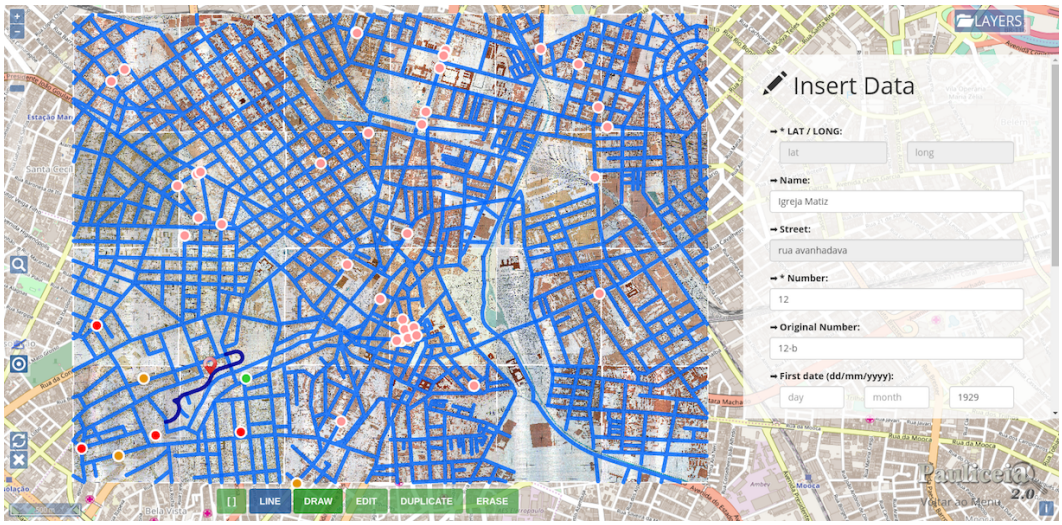

Fig. 3. Web portal for historical address edition.

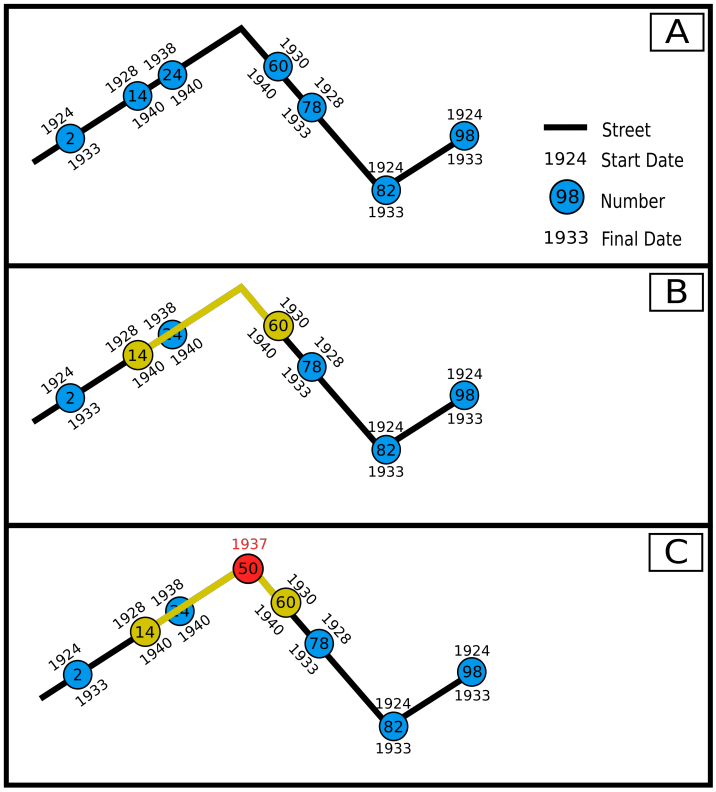

Fig. 4. The geocoding process for the address "Augusta Street, 50, 1937": (A) The street "Augusta" and its historical places (blue circles); (B) Its historical places whose time interval contain the year of "1937" (yellow circles); (C) the geocoding result (red circle).

6. RESULTS

Figure 5 shows the Pauliceia web portal available at http://www.pauliceia.dpi.inpe.br. Using the component "Search Address" in this portal, users can use the geocoding service described in Section 5

for any historical address represented by its street name, number and year. The component "Layers" organizes the layers of the platform that were provided by researchers and volunteers. Through this component, users can select layers that they are interested in, visualize, analyze and download them.

The component "Edit" contains tools to create and edit points, lines and polygons as well as attribute values associated to them. This component is under development and will be used by volunteers to create and edit features, such as street vectors from historical maps. There is a slider on the bottom of the portal that is responsible for the temporal filtering of the layer features. Using the component "Sign In", users can register and contribute to the platform, following the VGI protocol described in Section 4.

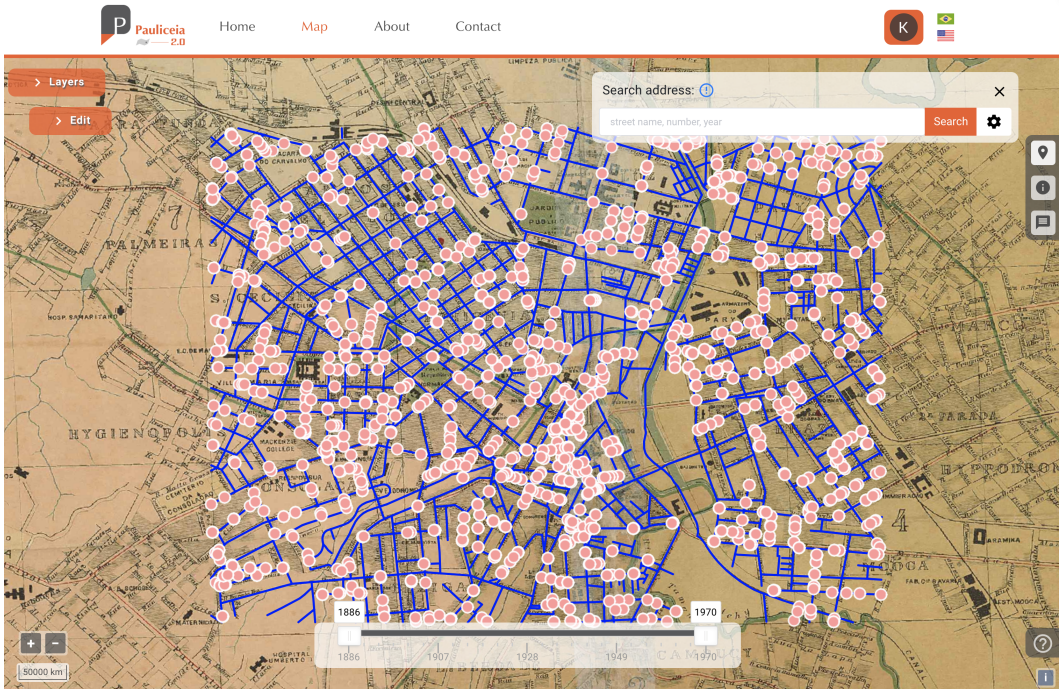

Fig. 5. Pauliceia web portal.

The Pauliceia web portal is implemented based on the Single Page Application (SPA) concept using the framework Vue.js [11] developed in JavaScript language. Besides this framework, we are using the tools jQuery[12] and OpenLayers [13]. The web services "VGI Management" and "Spatiotemporal Geocoding" shown in Figure 1 are developed using the languages Python and Node.js.

6.1 Historical data sets

The historical data sets of the Pauliceia 2.0 project are created by the following processes:

(1) **Digitization, mosaic creation and georeferencing of historical maps.** This process produces raster data as geotiff files and is done by the project members.

[11] https://vuejs.org/
[12] https://jquery.com/
[13] https://openlayers.org/

(2) **Vectorization of streets and buildings from historical maps.** This process produces vector data to represent the historical streets and buildings, based on the historical maps generated in item (1). Nowadays, the project members are vectorizing the streets from such maps manually using the gvSIG software [14]. However, after launching the final version of the Pauliceia 2.0 platform, this task will be done online through the component "Edit" of the platform.
(3) **Gathering and georeferencing of historical addresses.** The project members are collecting and georeferencing historical addresses through the web portal shown in Figure 3. These addresses are crucial for the geocoding processes, as described in Section 5.

At the moment, all these processes are done by the project members. However, after finishing the platform, we intend to use VGI and crowdsourcing techniques for vectorization of streets and buildings from historical maps as well as for the gathering and georeferencing of historical addresses, as described in Section 4. We also intend to evaluate the use of automatic methods in these processes.

Based on these processes, the current version of Pauliceia 2.0 database contains the following data sets:

(1) Seven mosaics of georeferenced historical maps of São Paulo city for the years 1868, 1870, 1880, 1890, 1910, 1920 and 1930. They are stored as geotiff files and are available through WMS OGC web service.
(2) Streets extracted from the historical raster maps described in item (1). The project members vectorized streets for the years 1920, 1930 e 1940. They are stored in the PostGIS database as vector data and are available through WMS and WFS OGC web service.
(3) Historical addresses of different types of places, such as churches, houses and buildings, that were collected by project members through the web portal shown in Figure 3. They are stored in the PostGIS database as vector data and are available through WMS and WFS OGC web service.

6.2 Event to promote citizen-derived historical information

In order to test the VGI and crowdsousing concepts for gathering historical data sets, the project team held an event with volunteers at UNIFESP Digital Humanities Laboratory from 19 to 27 April 2018. Eight volunteers, history students and researchers, attended this event to collect historical addresses using the web portal for historical address edition shown in Figure 3.

In the first part of the event, the project members presented the Pauliceia project, its web portal and the goals of the event. After that, each volunteer received a set of historical paper documents that were used to select historical addresses. In the final part of the event, the volunteers created and edited these selected addresses in the web portal, providing their spatial locations by clicking on the historical maps presented in the portal drawing area. Figure 6 illustrates the process to gather historical addresses by volunteers in this event.

The event results were very positive. The volunteers collected historical addresses of 146 streets, which mean a third of the project pilot area streets. Besides that, the project members identified issues in the web portal to be improved in order to meet VGI requirements for further events.

7. FINAL REMARKS

This paper presents the architecture of a computational platform that contains crucial modules to build an environment for collaborative historical research. This platform is being developed in an ongoing project called Pauliceia 2.0, and its final version will be launched in December 2018. The fundamental concepts and structure of this platform are described in this paper.

[14] http://www.gvsig.com/pt/produtos/gvsig-desktop

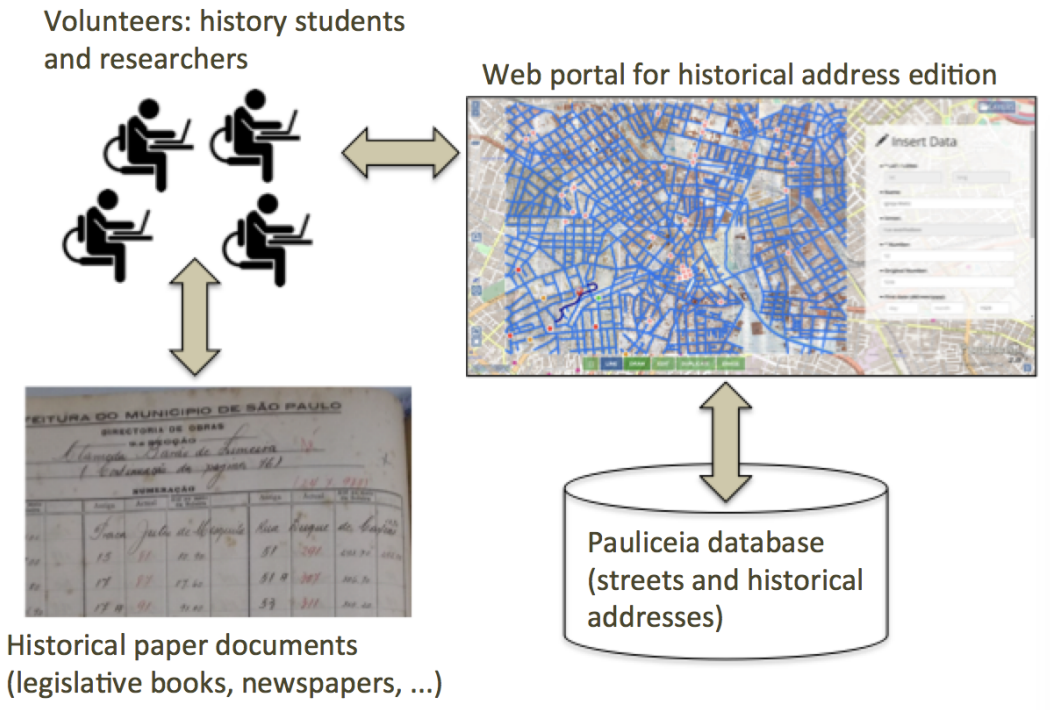

Fig. 6. Event to promote citizen-derived historical information in the Pauliceia project

It is important to emphasize that the platform architecture, its modules and web services, were designed and developed independently of the Pauliceia 2.0 project scope. Therefore, they can be used for other projects that aim to promote collaborative historical research in different regions or periods.

The two web services, "VGI Management" and "Spatiotemporal Geocoding", are RESTful application program interface (API) developed in Python language. Thus, they can be accessed externally by other software tools which support this kind of interface. The chosen standard for data exchange is JSON and its geographic counterpart GeoJSON. All source codes are open source and are available in the github link www.github.com/Pauliceia and all the data sets are available as OGC web services, as presented in Figure 1, through the link http://www.pauliceia.dpi.inpe.br/geoserver.

8. ACKNOWLEDGMENT

Pauliceia project is funded by FAPESP eScience Program (Grant 2016/04846-0). We are also grateful to FAPESP for granting students scholarships: #2017/03852-9, #2017/11637-0, #2017/11625-2 and #2017/11674-3.

REFERENCES

Asher, M., Giddens, C., Eslambolchi, H., and Stewart, H. J. Geocoding method using multidimensional vector spaces, 2009. US Patent 7,627,545.

Budig, B., van Dijk, T. C., Feitsch, F., and Arteaga, M. G. 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*. ACM, pp. 66, 2016.

Câmara, G., Vinhas, L., Ferreira, K. R., Queiroz, G. R. D., Souza, R. C. M. D., Monteiro, A. M. V., Carvalho, M. T. D., Casanova, M. A., and Freitas, U. M. D. Terralib: An open source GIS library for large-scale environmental and socio-economic applications. *Open source approaches in spatial data handling*, 2008.

Ciepłuch, B., Jacob, R., Mooney, P., and Winstanley, A. C. Comparison of the accuracy of openstreetmap for ireland with google maps and bing maps. In *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences 20-23rd July 2010*. University of Leicester, pp. 337, 2010.

Cura, R., Dumenieu, B., Perret, J., and Gribaudi, M. Historical collaborative geocoding. *Working paper submitted to Humanities. arXiv preprint arXiv:1703.07138*, 2017.

Davis Jr, C. A., de Souza Vellozo, H., and Pinheiro, M. B. A framework for web and mobile volunteered geographic information applications. In *Proceedings of XIV Brazilian Symposium on Geoinformatics (GeoInfo 2013)*. pp. 147–157, 2013.

DeBats, D. A. and Gregory, I. N. Introduction to historical GIS and the study of urban history. *Social Science History* 35 (4): 455–463, 2011.

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

Ferreira, K. R., Ferla, L., Queiroz, G. R., Vijaykumar, N. L., Noronha, C. A., Mariano, R. M., Wassef, Y., Taveira, D., Dardi, I., Sansigolo, G., Guarnieri, O., Musa, D., Rogers, T., Lesser, J., Page, M., Britt, A., Atique, F., Santos, J., Morais, D., Miyasaka, C., Almeida, C., Nascimento, L., Diniz, J., and Santos, M. Pauliceia 2.0: A computational platform for collaborative historical research. In *Proceedings of XVIII Brazilian Symposium on Geoinformatics (GeoInfo 2017)*. pp. 28–39, 2017.

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

Gouveia, J., Branco, F., Rodrigues, A., and Correia, N. Travelling through space and time in lisbon's religious buildings. In *Digital Heritage, 2015*. Vol. 1. IEEE, pp. 407–408, 2015.

Graser, A., Straub, M., and Dragaschnig, M. Towards an open source analysis toolbox for street network comparison: Indicators, tools and results of a comparison of OSM and the official austrian reference graph. *Transactions in GIS* 18 (4): 510–526, 2014.

Lee, J. Gis-based geocoding methods for area-based addresses and 3d addresses in urban areas. *Environment and Planning B: Planning and Design* 36 (1): 86–106, 2009.

Ludwig, I., Voss, A., and Krause-Traudes, M. A comparison of the street networks of navteq and OSM in germany. In *Advancing geoinformation science for a changing world*. Springer, pp. 65–84, 2011.

Martins, D., Davis Jr, C. A., and Fonseca, F. T. Geocodificação de endereços urbanos com indicação de qualidade. *Proceedings of XIII Brazilian Symposium on Geoinformatics (GeoInfo 2012)*, 2012.

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

Open Geospatial Consortium. OgcÂő standards and supporting documents. `http://www.opengeospatial.org/standards/`, 2017. Accessed on 2017-09-20.

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

Perret, J., Gribaudi, M., and Barthelemy, M. Roads and cities of 18th century france. *Scientific Data* 2 (150048): 1–13, 2015.

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

Southall, H. and Pridal, P. Old maps online: Enabling global access to historical mapping. *e-Perimetron* 7 (2): 73–81, 2012.

Spiro, L. This is why we fight: Defining the values of the digital humanities. *Debates in the digital humanities*, 2012.

Terras, M. Quantifying digital humanities. *Melissa Terra Blog*, 2012.