Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

Article

Historical collaborative geocoding

Rémi Cura [1,4,†], Bertrand Dumenieu [3,4,†], Nathalie Abadie [2], Benoit Costes [2], Julien Perret [2,3,4], Maurizio Gribaudi [3,4]

1 IGN; remi.cura@gmail.com
2 IGN; first.last@ign.fr
3 EHESS; last@ehess.fr
4 GeoHistoricalData; first@geohistoricaldata.org
* Correspondence: julien.perret@ign.fr
† These authors contributed equally to this work.

Abstract: The latest developments in digital humanities have increasingly enabled the construction of large data sets which can easily be accessed and used. These data sets often contain indirect localisation information, such as historical addresses. Historical geocoding is the process of transforming the indirect localisation information to direct localisation that can be placed on a map, which enables spatial analysis and cross-referencing. Many efficient geocoders exist for current addresses, but they do not deal with temporal information and are usually based on a strict hierarchy (country, city, street, house number, etc.) that is hard, if not impossible, to use with historical data. Indeed, historical data are full of uncertainties (temporal, textual, positional accuracy, confidence in historical sources) that can not be ignored or entirely resolved. We propose an open source, open data, extensible solution for geocoding that is based on gazetteers composed of geohistorical objects extracted from historical topographical maps. Once the gazetteers are available, geocoding an historical address is a matter of finding the geohistorical object in the gazetteers that is the best match to the historical address searched by the user. The matching criteria are customisable and include several dimensions (fuzzy string, fuzzy temporal, level of detail, positional accuracy). As the goal is to facilitate historical work, we also propose web-based user interfaces that help geocode (one address or batch mode) and display over current or historical topographical maps, so that geocoding results can be checked and collaboratively edited. The system has been tested on the city of Paris, France, for the 19th and the 20th centuries. It shows high response rates and is fast enough to be used interactively.

Keywords: Historical dataset; geocoding; localisation; geohistorical objects; database; GIS; collaborative; citizen science; crowd-sourced; digital humanities

---

1. Introduction

1.1. Context

In historical sciences, cartography and spatial analysis are extensively used to reveal the spatial organisations that hide within data with textual indirect spatial references like placenames or postal addresses. Mapping such data requires each item to be geocoded, *i.e.* assigned with coordinates on the Earth surface by matching the indirect spatial reference with entities identified in a reference geographical datasource (*e.g.* a topographic map georeferenced in a well-known coordinate reference system) [1]. Problems emerge when such spatial references are obsolete due to the temporal gap between the data to be geocoded and the reference datasource: locating the London Crystal Palace (destroyed by fire in 1936) on a today map would be rather tricky. Worse still, it might create ambiguities and possibly lead to erroneous geocoding, as the Crystal Palace refers nowadays to a South London residential area. Whereas manual geocoding can deal with such cases, the constantly growing volume of historical data, which results from the multiplication of initiatives in the digital humanities, calls for automatic approaches. Despite the existence of highly efficient (atemporal) geocoding tools and API, a truly historical geocoder is still at stake [2,3].

1.2. Related work

Geocoding is an inevitable step in any spatially-based study with considerable bodies of data, which makes it a critical process in various contexts: public health, catastrophe risk management, marketing, social sciences, etc. Many geocoding web services have been developed to fulfil this need, originating from private initiatives (Google Geocoding API, Mapzen[1]), public agencies (the French National Address Gazetteer[2]) or from the open-source community (OSM Nominatim[3], Gisgraphy[4]). These services can be characterized in terms of their three main components [1,4]: input/output data, reference dataset and processing algorithm. The *input* is the textual description the user wants to refine into coordinates (*e.g.* "13 rue du temple, Paris, France"). The *reference dataset* contains geographic features associated with textual descriptions of addresses. The *processing algorithm* consists in finding the best match between the latter and the former input description. Finally, the output usually contains a geographical feature and its similarity score (perfect match or approximate for instance).

These tools answer millions of geocoding queries each day with great success. Indeed, the quality of geocoding services can be estimated via two very important criteria [5]. The first is the database quality: how complete and up to date is the reference database? The second is the results characterization: how spatially accurate is each result and what is its associated confidence?

Despite their quality, such geocoding approaches can not be used for (geo)historical data for three main reasons. The first is that existing geocoding services do not take into account the temporal aspect of the query or the dataset they rely on. Indeed, they usually rely on current data, such as *OpenStreetMap*[5] data, continuously updated. As such, they implicitly work on a valid time that is the present (or possibly the interval between the beginning of the database construction and the present time). The second reason is that they rely on an exhaustive, strongly hierarchical database whose accuracy can be checked against ground truth (*i.e.* there is always a way to check the actual localization of an address). Unfortunately, historical data are not easily verifiable: one has to check them with difference available (geo)historical sources (possibly incomplete and conflicting) and, often, making assumptions or hypotheses. Such hypotheses are in their turn continuously challenged and updated by new discoveries, and there is no way to give a truly definitive answer. Indeed, primary sources may also be wrong or misleading. The third reason is that historical sources available to construct a gazetteer are sparse (both spatially and temporally), heterogeneous, and complex. We believe that all these specificities call for a dedicated approach. Similar observations have already been made in the context of archival data by the UK National Archives for instance [6]. Large historical event gazeteers already exist [7] and provide an important basis for the construction of the reference dataset. Nevertheless, we found few related articles [2,8], and for all of them geocoding was not a focus.

We could not find an historical geocoding approach that considers the characterization of geocoding results. Yet, this aspect is essential for historical geocoding because of the very unprecise and sparse nature of geohistorical data. Indeed, geocoding results have to be validated and/or edited manually. Considering the large amount of addresses ($> 100000 for Paris$) and the potential complexity of the task, this is clearly a lot of work. Fortunately, several projects such as *OpenStreetMap* have lead the way for what is usually called *Volonteered Geographical Information* (VGI) [9] of *crowdsourcing geospatial data* [10]. This approach consists in using a collaborative approach to solve the problem collectively, usually by implying citizens in the process. As suggested in a recent typology of participation in

1 mapzen.com
2 adresse.data.gouv.fr/api/
3 nominatim.openstreetmap.org
4 gisgraphy.com
5 openstreetmap.org

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

citizen science and VGI [11], different levels of participation can be defined. These levels go from "crowdsourcing", where the cognitive demand is minimal, to "extreme citizen science" or "collaborative science", where citizens are involved in all stages of the research (problem definition, data collection and analysis). In the rest of this article, we propose a collaborative historical geocoding approach that opens a way for a simpler participation of citizens in geohistorical research (thanks to dedicated interactive tools), but also for a more collaborative geohistorical science (thanks to a reproducible research approach [12–15], open source tools and open data).

1.3. Approach and contributions

In this article, we focus on the historical geocoding problem. Following the classical approach, we will present in particular the construction of a geohistorical database and the development of matching (data linkage) methods that fully use the temporal aspects of the geohistorical data and the input query. The main contributions of this article consist in (1) a formalisation of the historical geocoding problem, (2) a minimal model of geohistorical objects that can be easily re-used and extended, (3) an open source geocoding tool that is powerful, easy to use and can be extended with any geohistorical data, (4) a graphical tool to control and edit the geocoding results, which can then optionally be used to enrich the geohistorical database, (5) qualifications of geocoding results in term of textual, spatial, temporal aspects.

2. Approach overview

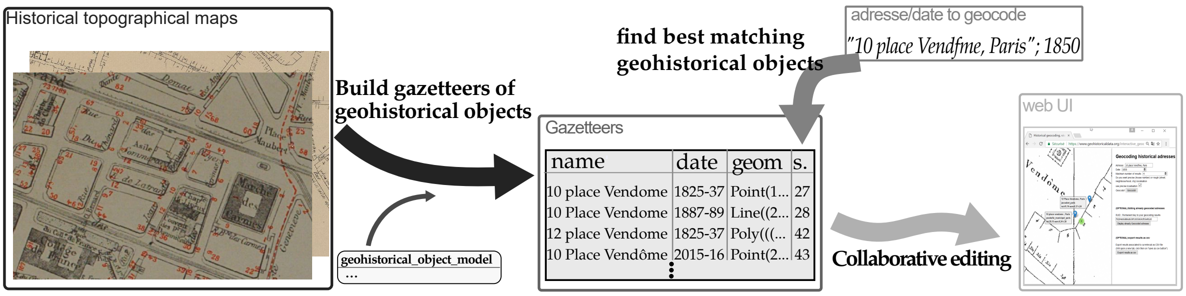


Figure 1. Gazetteers of geohistorical objects are created based on information extracted from georeferenced historical maps. Geocoding an historical address is then finding the best matching object based on a customised function (semantic, temporal aspect, spatial precision, etc.). Results can be displayed through a dedicated web interface for collaborative editing.

Based on historical sources and historical topographical maps, we extract geographic features that are gathered into gazetteers. These (geo) historical feature are modeled in a generic way into a Relational Database Management System (RDBMS). Geocoding an historical address is then finding the geohistorical object in the gazetteer that best match this historical address, thus may be done by means of distances that can be customised by the user. Lastly the results can be displayed via a web mapping interface over current or historical topographical maps, and further checked and edited collaboratively. Figure 1 illustrates this approach.

2.1. Building a geohistorical gazetteer: Extracting geohistorical objects from historical topographical maps

The starting point to build gazetteers is information extracted from historical topographical maps. The first part of extraction is to digitize the maps and georeferencing the map in a defined coordinate reference system. This maps are historical sources, and as such an historical analysis is performed to estimated the probable valid time (temporalization), positional accuracy, completeness, confidence, relation to other historical maps, etc. The whole process is carefully designed and explained in detail

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

in [16]. Then geohistorical objects are extracted from the referenced historical maps, manually (in a collaborative way), or with the help of computer vision techniques.

2.1.1. General consideration about building a spatio-temporal database

Extracting information from topographic historical maps amounts to building a spatio-temporal database. There are several approaches to do so, and we stress that we do not attempt to create a continuous spatio-temporal database. Instead, we store representations of the same space at multiple moments in history, the well-known snapshot model [17]. The main advantage is that for a given moment in time we can have several conflicting snapshots coexisting. This is essential, as solving the conflict may not be possible, and reporting these several conflicting geocoding results to historians may help appreciate these results. The drawbacks of this model, *i.e.* information redundancy and its inability to store the changes themselves, can be overcome during the geocoding process.

2.1.2. Historical topographic maps as geohistorical sources

In our approach, we focus on historical topographic maps as the main sources for two main reasons:

- the way they portray spatial information is close to today topographic mapping, making the integration of the information they convey in a Geographical Information System (GIS) easy;
- the main goal of topographic maps is to provide a reliable depiction of the shape and location of geographical features.

Although this choice seriously reduces the number of possible sources and therefore lessens the quantity of accessible spatial information, it aims at efficiency. Indeed, topographic maps are a good compromise between their reliability, the quantity of spatial information they contain and the complexity of extracting information.

We create a snapshot for each historical topographic map by relying on three steps:

1. georeferencing the map in a well-defined coordinate reference system supported by common GIS tools,
2. assigning the map a *valid time*,
3. extracting geographical features from the map.

2.1.3. Georeferencing topographic historical maps

We have to establish a correspondence between each pixel of the historical maps and geographical coordinates. To do so, we first choose a common spatial reference system (SRS). Then we identify common geographic features between historical maps and current maps: so-called ground control points (GCP). Last, we compute a warping transform that will respect at best the matching points. Finding GCPs between current maps and historical maps can be increasingly difficult as we go back in time, because there are less and less perenial GCPs. Consider for instance the city of Paris, where the French Revolution and its consequences combined with the 19th century transformations (including the so-called Haussmannian transformations) resulted in massive changes in the shape of the city. To this end, we can start by georeferencing *e.g.* 20th maps to current maps, then georeference *e.g.* 19th maps to 20th maps, and so on for even older maps.

Common spatial reference system (SRS)

The choice of a SRS is not easy, as each SRS induces projection errors that depend on the covered area. Whatever the choice of SRS, it is essential that the implied accuracy is well-known and documented in order to qualify the absolute accuracy of each geo-referenced map. We restrain ourselfs to SRS using meters as base units (opposed to SRS using degrees), as they are much closer in nature to those used in historical topographical maps.

Selection of ground control points

The identification of pairs of GCPs is a critical step because the number, distribution and quality (*i.e.* positional accuracy, reliability, confidence) of the points strongly influence the quality of the georeferencing. While the quality of the selected points depends on each map, a simple rule of thumbs is to select, as much as possible, homogeneously distributed points to go further [18]. Three parameters have to be considered: the geometric type of the features carying the ground control points, their nature and the method used to identify them. Usually, features chosen as ground control points are represented by 2D points; lines or surfaces may also be used, and possibiliy even curves [19].

On historical maps, the positionnal accuracy of mapping themes can vary greatly either due to the purpose of the map, or the mapmaking process itself. Optionally, one can rely on geodetic features drawn on the map such as meridian or parallels provided one can fully characterised the geodetic characteristics of these lines.

The actual identification of GCPs can be achieved by automatic or manual processes. Automatic approaches are notably used for historical aerial photographs, where feature detection and matching algorithms are well fitted [20]. Common GIS tools offer georeferencing softwares allowing to manually select pairs of *ground control points* identified in both the input and the reference maps. Such tools are often used for historical maps georeferencing because: (1) they are easy to utilize and (2) they allow historians to control the quality and reliability of the identified points using co-visualization between both maps.

Choosing a geometric transformation model

Once an acceptable set of paired features has been identified, the last step is to compute the transformation from the input map to the reference. Several transformation models have been proposed: global transforms (affine, projective), global with local adaptations (polynomial-based) and local transforms (rubbersheeting, Thin Plate Spline, kernel-based approaches, etc.). Studies have been conducted to assess the relevance of these transformation for historical maps [18,21,22]. They show that choosing a model is mostly a matter of compromise between the final spatial matching between the feature pairs (*i.e.* the expected residual error) and the acceptable distortion of the map regarding its legibility. Exact or near-perfect matching between features can be achieved with local transforms and high order polynomials, whereas the internal structure of the map is most preserved by global transformations. Low order polynomials offer a compromise between both constraints.

2.1.4. Temporalization: locating geohistorical sources in time

Georeferencing is a way of locating multiple maps in the same reference space. Similarly, *temporalization* is the process of locating each geohistorical source in time. When building spatio-temporal snapshots from historical maps, the key problem is to determine the moment where the map is representative of the actual state of the area it portrays, *i.e.* the *valid time* of the map. We considered the valid time of each map as the period starting with the beginning of the topographic survey and ending with the publication of the map, which are often uncertain. Representing uncertain or imprecise periods of time is a common issue when dealing with historical information and many authors relied on the fuzzy set theory to represent and reason on imperfect temporal knowledge [23,24]. We model imprecise valid times as trapezoidal fuzzy sets. They are trapezoidal functions of time with values ranging from 0 (the source provides no information at this time) to 1 (geographical entities portrayed in the map are regarded as existing and tangible at this time). We rely on the pgSFTI[6] postgres extension to store and manipulate such temporal fuzzy information. For instance, Figure 2

[6] https://github.com/OnroerendErfgoed/pgSFTI

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

illustrates the *valid time* of a map whose topographic survey started in year 1775, ended between 1779 and 1780 and which was engraved late 1780.

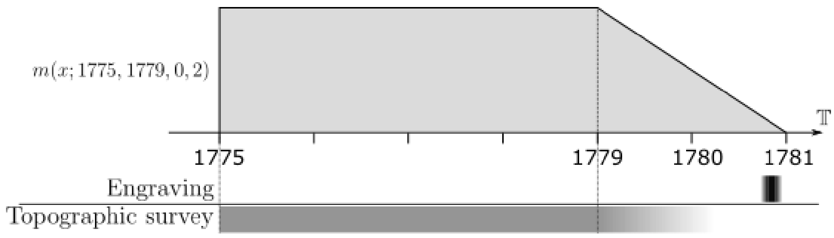


Figure 2. An uncertain valid time modelled as a trapezoidal fuzzy set function

2.1.5. Extracting information from maps

Once the historical maps have been georeferenced and temporalized, their cartographic objects can be extracted to produce geohistorical objects. The most classical way to extract information from maps is by human action with a classical GIS software (*e.g.* QGIS). However each one historical map of Paris contains a large amount of information to be extracted (*e.g.* tens of thousands of street names, hundreds of thousands of building number, etc.). A first solution is then to use computer vision and machine learning methods to create automatic extraction tools. These tools can process the whole map in a few hours. Regrettably, such tools are difficult to design, are very specific to each historical map, and may produce low quality results (see Figure 3). Recently, collaborative approaches have shown to be very efficient for building large geographical databases in a relatively short period (OSM[7], NYPL[8]).

Figure 3. In this example, hand written text is automatically detected and extracted (red) from an historical map.

2.2. Modelling geohistorical objects

Information extracted from historical maps is used to create gazetteers. Those are made of geohistorical objects. To this end, we design a geohistorical objects model with all necessary attributes and also flexibility to adapt to the great variety of geohistorical object types and sources. Our goal is to provide a generic minimal (geo)historical object model that can be used by other and easily extended when necessary.

---

7 http://www.openstreetmap.org
8 http://buildinginspector.nypl.org/

2.2.1. modelling choices

Geohistorical data are extremely diverse, both in terms of historical sources and in terms of how the sources were dealt with by historians. As such, historians use complex tailored models. We do not aim at modelling all geohistorical data in all their complexity. Instead, we propose to model the bare minimal common properties of all geohistorical objects, and offer mechanisms so this model can be easily extended and tailored to the specificities of the data. To define the bare minimal model, we start from the very nature of a geohistorical object, that is both an historical object and a geospatial object. The extension mechanism is provided via a database-object oriented design using table inheritance, and is packaged into a PostgresSQL extension[9].

2.2.2. geohistorical objects model

Geohistorical objects have both an historical and a geospatial part. We stress that modelling historical source and numerical origin process of a geohistorical object is an essential part. The detail of the model are illustrated in Figure 4.

Historical aspect

In our model, an historical object is defined by its name, source and temporalization.

- *Name.* By name, we mean the historical name that was used to identify the object in the historical source, and the current name that is used by historians to identify the object in the current context. For instance, the historical name of the Eiffel tower in Paris may be "tour de 300 mètres", but, today, it is referenced as "tour Eiffel".
- *Source.* A historical object is defined by a primary historical source (document), where the object is referenced. Beside the historical source, the way the object was digitized in this source is also essential. For instance, a street name may have the Jacoubet topological map as historical source, and would have been digitized via collaborative editing on the georeferenced map.
- *Temporalization.* Any historical source is associated with temporal information (fuzzy dates), which is the period during which the source is likely relevant. Beside the historical source temporal information, a historical object can also have its own temporal information. For instance, a street may have been extracted from a historical map having been drawn between 1820 and 1842. Besides this information, using other historical documents allow to narrow the probable existence of this street to 1824-1836.

Geospatial aspect

A geohistorical object is also defined by geospatial information: a direct spatial reference (geometry) and its positional accuracy metadata.

- *Geometry.* A feature has a geometry which follows the OGC standard[10]. It may be a point, polyline, polygon, or a composition of any number of those, in a specified SRS. The geometry is extracted from the historical source (in a manual or automatic way).
- *Positional accuracy.* Historical features have positional accuracy information. This precision expresses the spatial uncertainty of the historical source (the person drawing the map may have made mistakes) and the spatial imprecision of the digitizing process (the person editing the digitised map may have made a mistake). One historical source may contain several accuracy metadata, one for each geohistorical object type it contains. For instance, a historical map may contain buildings and roads. Buildings may have a different positional accuracy (5 metres) than road axis (20 metres). Besides, the digitising process precision may have been of 5 metres.

---

[9] https://github.com/GeoHistoricalData/geohistorical_objects
[10] http://www.opengeospatial.org/standards

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

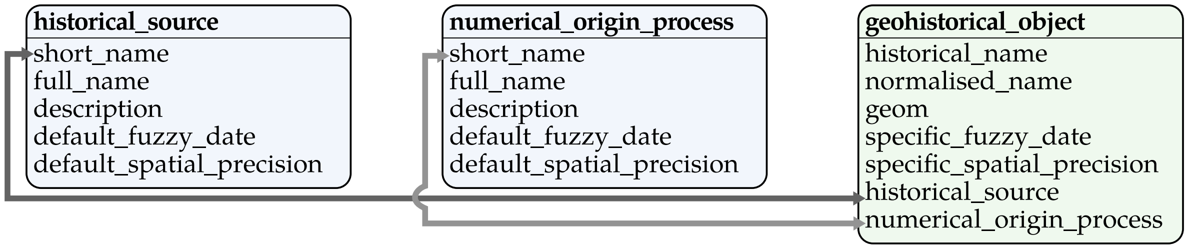


Figure 4. The geohistorical object model, where each object is characterized by its historical source (for instance the historical map the object was described in) and a numerical origin process, which is the process through which the object was digitized. Besides source and origin process, an object is also described by a fuzzy date, a text and a geometry.

2.2.3. A database of geohistorical objects

We define a conceptual schema for geohistorical objects, which is based on two names, a source, a capture process, fuzzy dates and a geometry. This defines the core of a generic geohistorical object. Yet this geohistorical object model is easily extendible using the table inheritance mechanism, an object-oriented design mechanism that is available in PostgreSQL (see Figure 5).

Table inheritance

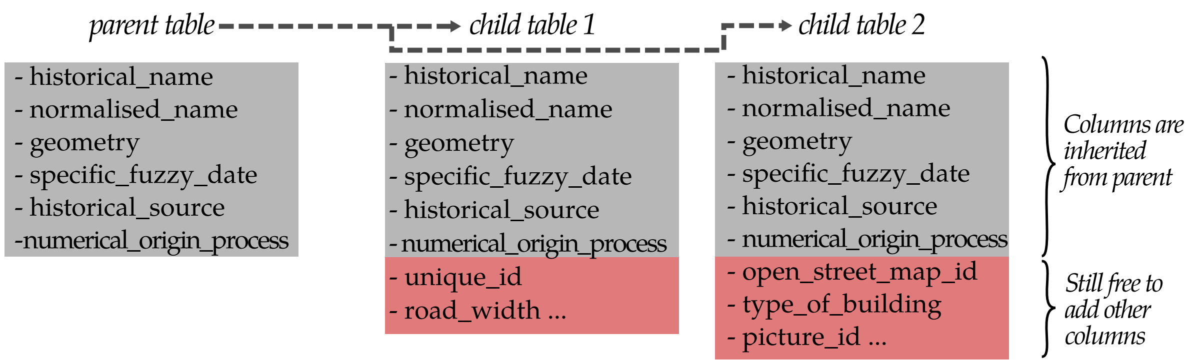


Figure 5. The table inheritance mechanism: a child table inheriting from a parent table inherits all the parent columns, and can also have its own.

The concept of table inheritance is simple. When a table *child* is created as inheriting from a table *parent*, *child* will have at least the columns of *parent*, but can also have other columns (provided there is no name/type collision). This means in our case that a table of geohistorical objects will inherit from the main geohistorical object table, *i.e.* will have all the core columns of geohistorical objects (names, sources, temporal aspect, spatial aspect), but can also have its own tailored column, providing the necessary flexibility.

Another key aspect of table inheritance is that the *parent* table is queried, the query will be executed on not only the rows of *parent* table, but also on the rows of all *child* table. This means that all tables using the geohistorical object model will be virtually grouped and accessible from one table.

Simulated inheritance of index and constraints

The PostgreSQL table inheritance mechanism is however limited in some aspects, because constraints and index can not be inherited. Constraints are essential, because they are used to guarantee that any geohistorical object will correctly use existing sources from the sources tables ("historical_source" and "numerical_origin_process"). Indexes are also essential, because when using

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 262; doi:10.3390/ijgi7070262

hundred of thousand of geohistorical obejcts, they are needed to help speed the queries. We index not only names, but all geohistorical object core columns (names, sources, temporal aspect, spatial aspect). We propose a registering function that the user can execute only once when creating a new geohistorical object table.

Modelling a geohistorical object from the user perspective

The practical steps to create geohistorical objects are simple:

1. Add the historical source and numerical origin process in the source and process tables.
2. Create a new table inheriting geohistorical objects and containing your additional custom columns
3. Use the registering function with this table name
4. Insert your data in the table.

2.3. Geocoding historical addresses with geohistorical object gazetteers

In the previous section, we explained how we create gazetteers of geohistorical objects from maps.

1. an historical map is scanned,
2. scans are georeferenced using hand picked control points,
3. historical work allow to estimate temporal information and spatial precision of the map,
4. roads name and axis geometry is extracted from the scan (manually or automatically),
5. building numbers are extracted from the scan (manually or automatically),
6. in some cases, building numbers can be generated from the available data,
7. normalised names are created from historical names,
8. geohistorical objects are created.

The next step is to use these gazetteers to geocode historical addresses.

2.3.1. Historical geocoding concept

In our method, geocoding something is finding the most similar geohistorical objects within the available gazetteers, which then provides the geospatial information. This approach relies on two key components: gazetteers of geohistorical objects, and a metric to find the best matches. This approach allows to perform geocoding in a broad sense, as it does not rely on a structured address (number, street, city, etc.), but rather on a non constrained name.

2.3.2. Creating geohistorical object gazetteers for geocoding

geohistorical object gazetteers are key for the geocoding. These objects are extracted from topographical historical maps and inserted into geohistorical objects tables. Each table form a gazetteer.

Database architecture for geocoding

We again use the PostgreSQL table inheritance mechanism. To this end, we create two tables dedicated to geocoding. Now gazetteers tables that will be used in geocoding must inherits from these two tables. "precise_localisation" table is for building number geohistorical objects, *e.g.* "12 rue du temple, Paris". "rough_localisation" table is for road axis, neighbourhood, cities geohistorical objects. We chose to have two separate tables for ease of use and performance. Geocoding queries are then performed on the two parents tables, but thanks to inheritance, these parents tables virtually contains all the gazetteers table containing the actual geohistorical objects, as illustrated in Figure 6.

2.3.3. Finding the best matches

Once geohistorical objects gazetteers describing precise and rough localisation are available, geocoding is finding the best match between the input query and the objects.

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

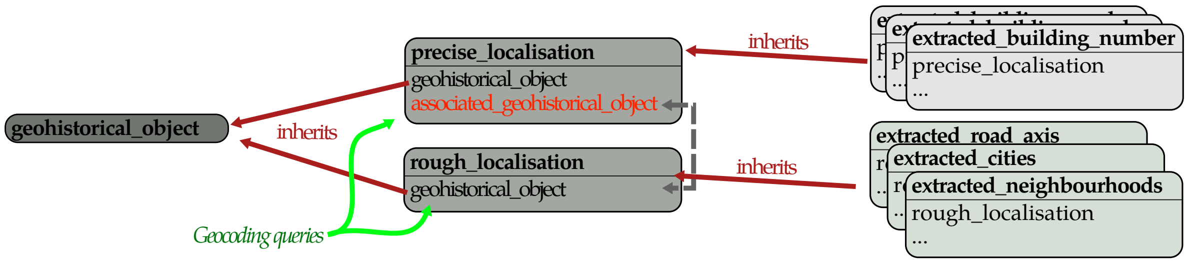


Figure 6. Geocoding table architecture. Two tables of geohistorical_object are the support for geocoding queries. Because all extracted geohistorical objects tables inherits from these two tables, they both virtually contains all the objects.

Concept

We call the potential matches "candidates", and the problem is then to rank the candidates from best to worst. The user can chose how many candidates he wants, depending on the application. For an automated batch geocoding, the best match (top candidate) is optimal. For a human analysis of data, several matches may be more interesting (top 10 candidates for instance). What can be qualified as "best" depends on the user expectations. We provide a number of metrics than can be combined by a user into a tailored ranking function. The function is expressed in SQL, with access to all postgres math functions. We describe the available metrics and give example of such function.

Example

For instance when a user geocodes the address "12 rue de la vannerie, paris" in 1854, user may be more interested into geohistorical objects that are textually close (*e.g.* a geohistorical object "12 r. de la vannerie Paris", 1810), or maybe geohistorial objects that are temporally close (*e.g.* "12 r. de la *T*annerie Paris",1860).

Metric: string distance w_d

We use the string distance provided by the PostgreSQL Trigramm extension (pg_trgm[11]), which compares two strings of characters by comparing how many successive sets of 3 characters are shared. For instance "12 rue du temple" will be farther away from "12 rue de la paix" than from "10 r. du temple".

Metric: temporal distance t_d

Both the address query and the geohistorical object are described by fuzzy dates. In order to compare such temporal information, we propose a simple fuzzy date distance that casts fuzzy dates into polygons. The x axis is the time, and the y axis is the probability of existence of the object. Then the distance between twon dates A and B is computed as shortest_line_length(A,B) + Area(A) - Area(A $\cap$ B). Note that this distance is asymmetric.

Metric: building number distance b_d

To get building number distance, a function first extracts the building number both from the input address query (b_i) and from the geohistorical object (b_d). If b_i and b_d have same parity, the distance is $\mid b_d - b_i \mid$. If parity is different, the distance is $\mid\mid b_d - b_i \mid +10 \mid$. In France, building numbers have in general the same parity on each side of the street (*e.g.* Left : 1,3,5,.. ; Right: 2,4,6..). We analysed

[11] https://www.postgresql.org/docs/current/static/pgtrgm.html

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

current building number in Paris and determined that on average, given a building number b_i, the closest building number with a different parity has a 10 number difference.

Metric: positional accuracy s_p

Another way to rank the geohistorical objects is to use their positional accuracy. The positional accuracy of a geohistorical object is either the positional accuracy computed for this objects when it is available, or the default positional accuracy of its geohistorical source.

Metric: level of detail distance s_d

Providing localisation information at different level of detail, depending on the user requirement is an important quality issue for our geocoder. For instance if the level of detail of the user's query data is the city, there is no need to perform a more precise geocoding. Therefore the user can specify a target scale range (S_l, S_h). Then given a geohistorical object whose geometry is buffered $(geom_b)$ with its spatial precision, the scale distance is defined by $least(\mid \sqrt{area(geom_b)} - S_l \mid, \mid \sqrt{area(geom_b)} - S_h \mid)$. The formula $\sqrt{area(geom_b)}$ gives an idea of the spatial scale of the geohistorical object.

Metric: geospatial distance g_d

The user may provide an approximate position for the area he is interested in. For instance in France both city "Vitry-le-François" (East) and "Vitry-sur-Seine" (near Paris) exist, but are very spatially far away. A user expecting results in the Paris area may provide a geometry (a point for instance) near Paris. Then the classical geodesic distance is computed between the provided geometry and the candidates geohistorical object.

Example of matching function

The different metrics can be weighted and combined depending on the user needs. Equation 1 gives an example that favour good string similarity, but not at the price of a large temporal distance.

$$100 * w_d + 0.1 * t_d + 10 * n_d + 0.1 * s_p + 0.01 * s_d + 0.001 * g_d \quad (1)$$

2.4. Collaborative editing of geohistorical objects

The geocoding approach we have presented in the previous section works inside a PostgreSQL database. Given an input address and fuzzy date, plus a set of parameters, it returns the geohistorical objects that matches the input the best. Yet the geocoding results are only as good as the gazetteers used. The geohistorical objects within the gazetteers may be spatially imprecise, mistakenly named or simply missing. Given that the volume of geohistorical objects is large (for Paris, approximately 50 k building number per historical map), we create a collaborative platform to facilitate geocoding, visualising the results and editing the geospatial objects when necessary. To this end, we create a dedicated web application so collaborative editing is possible without having to install specific tools.

2.4.1. About collaborative editing

Given the complexity of calibrating automatic extraction tools on specific maps and their relative reliability, the collaborative digitisation of vector objects from maps is a safe alternative. For instance, we used such an approach in order to extract the main feature of the Cassini maps (18^{th} century France) [25]. Furthermore, the results of the collaborative extraction of features can then be used to test, calibrate or train automatic extraction algorithms.

2.4.2. Collaborative editing architecture

Figure 7 outlines the architecture used for collaborative editing.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

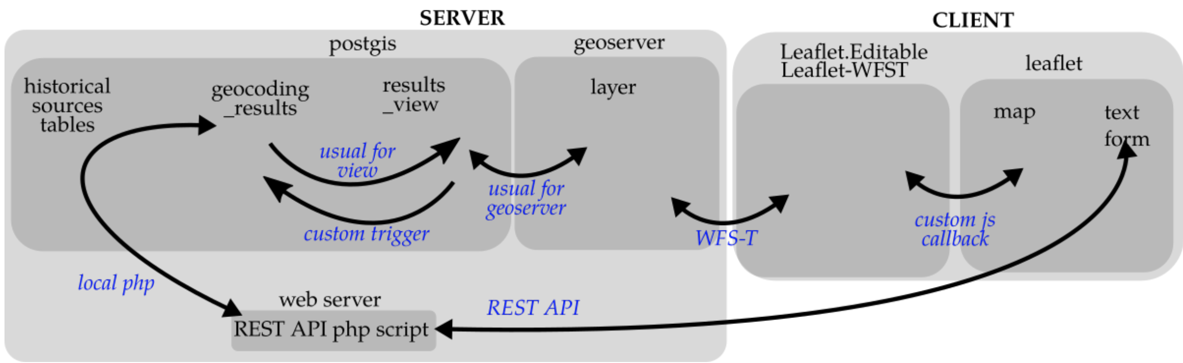


Figure 7. Conceptual architecture for interactive display and edit of geocoding results. The stack ocntains only standard components.

Architecture

The hearth of the architecture is a PostgreSQL database server, which contains the geohistorical objects gazetteers that will be used for geocoding as well as the geocoding function. A webserver can geocode addresses and return results via a REST API. However, the webserver has another option where the results are not returned, but instead written in a result table along with a random unique identifier (RUID). The RUID is then the key that permit to display and edit the results. To this end, a geoserver can access (read and edit) the result table via the WFS-T protocol. A web application based on Leaflet then acts as a user interface to display and edit the results via the geoserver.

Persistence of geocoding results and edits

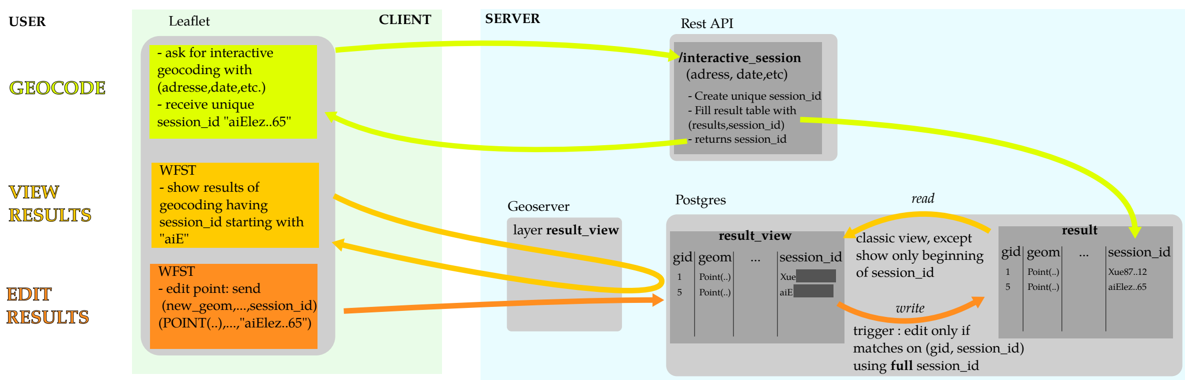


Figure 8. Collaborative display and edit is achieved through a mix of standards (REST , WFS-T) and custom solution (triggers) that enable the sharing of a basic public/private key.

The architecture that allows persistence of results is illustrated in Figure 8. When using the RUID mechanism, each geocoding result (that is the found geohistorical object from the gazetteers) is associated to this RUID. That way the user can always access its results, regardless of the computer session or browser cache issues.

To edit, a specific mechanism is used. The user does not directly edit the result table, as he could potentially edit other people results. Instead, the user edit a dedicated result_view that acts like a bouncer. It allow edit only if the edit is occurring on a row that has the user RUID. User edit of the geospatial objects do of course not affect the source data, for a tracking purpose.

Instead, a new user edit automatically creates an edited copy of the geohistorical object in a dedicated table "user_edit_added_to_geocoding" that is a gazetteer and is used by the geocoding

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

process. In this table are inserted the edited geohistorical objects. The objects retain their "historical_source", but their "numerical_origin_process" is changed to properly document the fact that they are the result of a collaborative editing.

2.4.3. Collaborative editing user interface

We consider that building an efficient user interface is very important for historical geocoding. In particular, many end users are specialised on history rather than on computer science, and thus an easy access to geocoding is essential. All our interfaces are web-based for a maximum of compativility. We propose three interfaces whose results are shared.

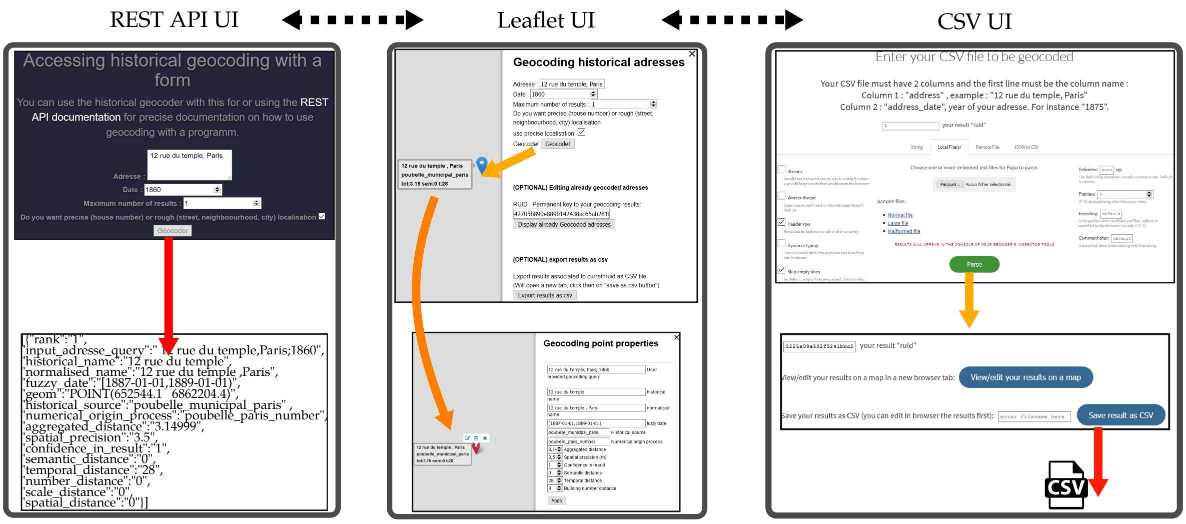


Figure 9. Various Web User Interface for use of historical geocoding.

Interface for a REST API.

The simplest interface we propose is a form that helps build the necessary REST API parameters. Indeed, a REST API works via URL containing precise parameters, and it can be tedious to manipulate. For instance:
https://www.geohistoricaldata.org/geocoding/geocoding.php?adresse=2012ruedutemple,Paris&date=1860&number_of_results=1&use_precise_localisation=1
This interface is designed to be used in an automated way, for batch geocoding.

Interface for batch geocoding via CSV files.

In our experience historian often work with spreadsheet files, where each line will be a potential historical object, along with an address and a date. To facilitate the geocoding of these addresses, we propose a User Interface that can read Coma Separated Value (CSV) files (which is a standard spreadsheet format), and geocode the address and date within. This interface is built around PapaParse[12] Javascript framework. Then, the geocoding results can be either downloaded as a CSV file, or displayed and edited in a web application.

[12] http://papaparse.com

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

Interface for display and edit of results.

The most complex interface we propose is based on the Leaflet[13] Javascript framework. There, the user can geocode an address, or use already geocoded address via the RUID mechanism (see Section 2.4.2.1), be it from previous sessions or from geocoded CSV files. The geocoding results are displayed on top of a relevant historical map, and can be edited. User can edit results geometry as well as results names (historical and normalised). We stress that although such edit are stored in the database, and used by further geocoding queries, they do not affect source data, by design.

3. Results

We perform several experiments to validate our approach. First we use the geohistorical model to integrate objects extracted from historical topographical maps from the 19th century for the city of Paris, and the current OpenStreetMap road axis and building numbers for Paris city surroundings. We successfully integrate the road axis, building numbers, and neighbourhoods to the geocoder sources. We then perform multiscale geocoding of dozens of thousand of historical addresses extracted manually by historians and extracted automatically by automatic process. For one of our datasets, we ask the historian to manually correct the automated geocoding results, so as to evaluate the quality of our method. Last, we test the collaborative editing of geohistorical object in two scenarios: analysis (several results for one address), and edit (efficiency of check/edit top results for several addresses).

3.1. geohistorical objects sources

We mainly use three historical sources of geohistorical objects to perform geocoding. The first two are Historical topographic maps of Paris from the 19th century. These maps are georeferenced then street axis (and possibly building numbers) are manually extracted. The third historical sources are road axis and building number for Paris surrounding extracted from current Open Street Map data.

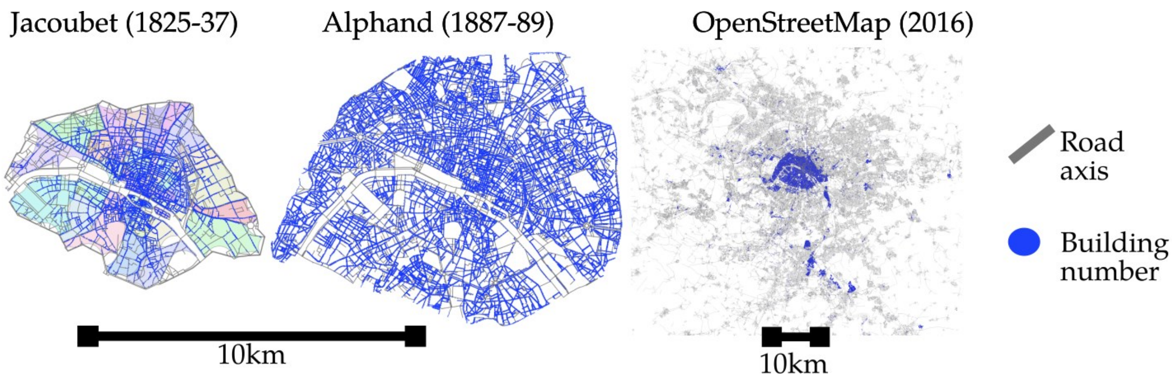


Figure 10. geohistorical objects used from geocoding extracted from the source maps.

3.1.1. Historical topographic maps used

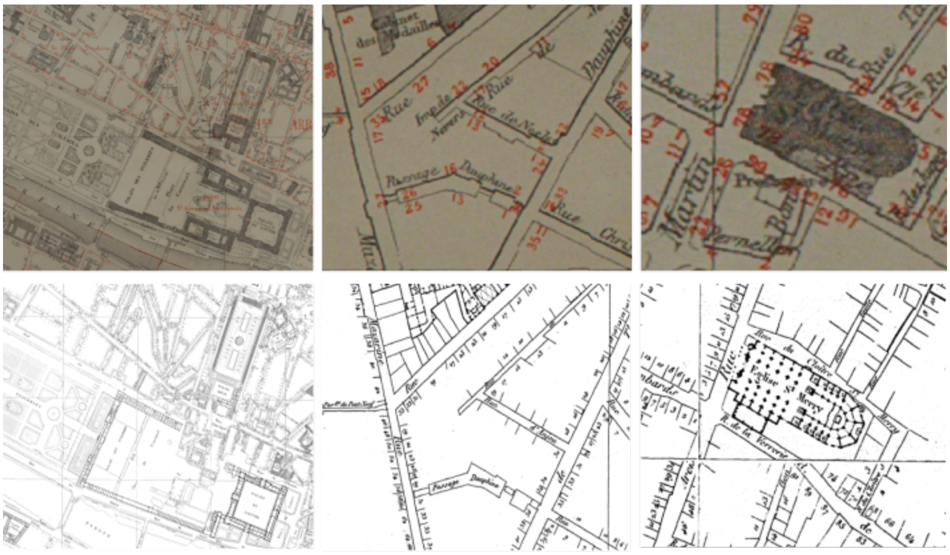
We integrated two major French atlases of Paris from the 19th century as geohistorical sources. The first one is the "Atlas municipal de la Ville, des faubourgs et des monuments de Paris"[14] created at the scale of 1 : 2000 between 1827 and 1836 by Theodore Simon Jacoubet, an architect who was working for the municipal administration of Paris. The second atlas is the 1888 edition of the "Atlas municipal des vingts arrondissements de la ville de Paris"[15]. For legibility reasons, we refer to the first

13 http://leafletjs.com
14 Municipal atlas of the city, suburbs and monuments of Paris.
15 Municipal atlas of the 20 districts of Paris

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

atlas as the "Jacoubet atlas" and the second as the "Alphand atlas"[16]. The Jacoubet atlas depicts a city standing between the housing development following the sale of the properties confiscated during the French Revolution and the majors changes in the urban structure arising from the emergence of the fist train stations in 1837-1840 and the so-called Haussmannian transformations.

The Alphand atlas is a portray of Paris at the scale of 1 : 5000, after most of the Haussmannian transformations (major rework of Paris urbanism in the 19th century) and after the city was merged with 11 of its neighboring municipalities in 1860. Both atlases contain large scale topographic views of Paris, separated in several sheets (54 and 16 respectively) and portray the urban street network with each street named, building of public purposes and religious buildings (see Figure 11). In addition, the house numbers are specified for most of the streets in the city, although the Alphand atlas pictures only the numbers at the start and end of each street section. Both atlases are also built upon triangulation canvas covering the entire city, allowing us to expect a high positional accuracy of the geographical features they contain.



(a) (b) (c)

Figure 11. Samples of the georeferenced Alphand Atlas (2nd row) and Jacoubet Atlas (1^{rst} row) at different scales: district (a) and urban islet (b). Column (c) shows how buildings are portrayed in the maps.

We georeferenced the two atlases using the grids drawn on the maps, which are aligned on the Paris meridian, as a pseudo-geodetic objects to identify feature pairs. The dimensions of the grid cells also appear on the maps, allowing us to reconstruct the grids in a geographic reference system. We have chosen to georeference the maps in the Lambert I conformal conic projection, which uses the Paris meridian as prime meridian and rely on the NTF (Nouvelle Triangulation Française) geodetic datum. The main advantage of this projection is that it is locally close to the planar triangulation of Paris used in the atlases. Thus, the projection of the maps can be reasonably approximated by the Lambert I projection, making the reconstruction of the grids in the target coordinate reference system straightforward. In addition, since both maps are at high scale and are reliable because they are official maps with high positional accuracy, we used rubbersheeting as the geometric transform model. The georeferencing process applied for each atlas was the following:

[16] From the name of Jean-Charles Alphand who was at the time the director of the department of public works of Paris.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 262; doi:10.3390/ijgi7070262

- reconstruct the meridian-aligned grid with Lambert I coordinates;
- in each sheet, mask the non-cartographic parts out (cartouche, borders,etc.);
- for each sheet, set pairs of ground control points at each intersection between the vertical and horizontal lines of the grids in the map and in the reconstructed grid;
- transform each sheet with a rubbersheeting transform based on the ground controls points previously indentified on the grids.

geohistorical objects extraction

Based on these atlases, vectorial road axis are manually drawn and the road name inputed For Alphand map, the building number at the beginning and end of ech street segment is also inputted. For Jacoubet, the building numbers from a previous map (Project Alpage, Vasserot map, [26]) are adapted to fit the Alphand map. Multiple series of successive checking and editing are performed using ad hoc visualisations and tools.

For Alphand, building numbers are then generated based on the available information (for each street segment, for each side, beginning and ending number) by linear interpolation, and an offset. The size of the offset is estimated by using current Paris road width when the road has not changed to much.

3.1.2. Other geohistorical sources

We also use current data from OpenStreetMap. We use the version of the data that has been transformed to be used by the Nominatim geocoder. Custom scripts extract road axis and building numbers. The dataset covers Paris city and its surroundings, and is dated to 2016.

3.2. Geocoding of Historical datasets

One of the end goal of our geocoding tool is to be useful for historians. Therefore, we contacted several historians working on Paris (19th century). They had been collecting historical addresses, which we geocoded by importing their data into the geocoding server. Figure 12 shows an extract of the thousands of geocoded addresses, while table 1 gives an overview of the number of successes and timing.

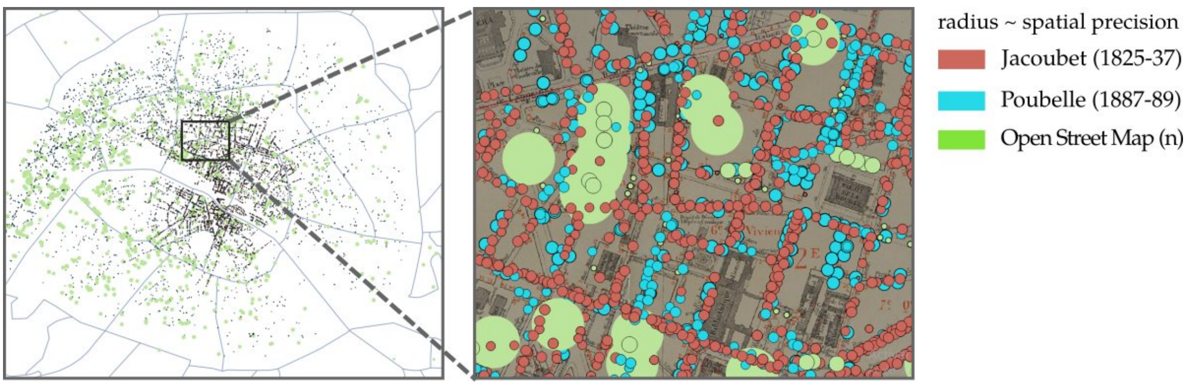


Figure 12. All geocoded historical datasets. Size is proportional to spatial precision.

3.2.1. Manually collected dataset

South Americans: South America immigrants living in Paris in 1926, manually input from census, collected by Elena Monges (EHESS).
Textile: Professionals of textile industry in Paris, manually input from the "Almanachs dy Commerce de Paris", from 1793 to 1845, collected by Carole Aubé (EHESS).
Artists accommodations: Addresses of artist studios and artists accommodations between 1791 and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

Table 1. All geocoded historical datasets facts.

Dataset name	input addresses	response rate (rough)	secs/1000 addresses
South Americans	13991	13743 (250)	138
Textile	5777	5688 (16)	135
Textile 2	3070	3053 (2)	110
Artists accommodations	13907	10215 (2955)	244
Health administrators	1887	1698 (171)	316
Belle epoque (0.3)	6467	3880(337)	280
Belle epoque (0.5)	6467	6000	351

1831, collected by Isabelle Hostein (EHESS) to study their impact on Paris development.
Health administrators: Addresses of public health and hygiene administrators in Paris between 1807 and 1919 ([27]), collected by Maurizio Gribaudi and Jacques Magaud (INED-EHESS).

3.2.2. Belle epoque

We geocode another set of addresses that are automatically extracted from directory of Paris financial societies between 1871 and 1910. Directories are books referencing address of company (and name and other information). The process of automatic extraction is complex in itself (Project Belle Epoque, [28]), and is out of scope of this article. We only describe it briefly here.

First, each page of the directories of Paris for specific years are photographed. Pictures are then straightened, and information is extracted via an OCR software which has been configured for the directory specific layout. Further rule based processing parse the text into address fields. As a result of this automatic process, the quality of addresses is often significantly lower than manually edited addresses. Therefore, we test two settings by allowing a greater maximum string distance from 0.3 to 0.5 (over 1).

3.3. Manual editing of the geocoding results for evaluation

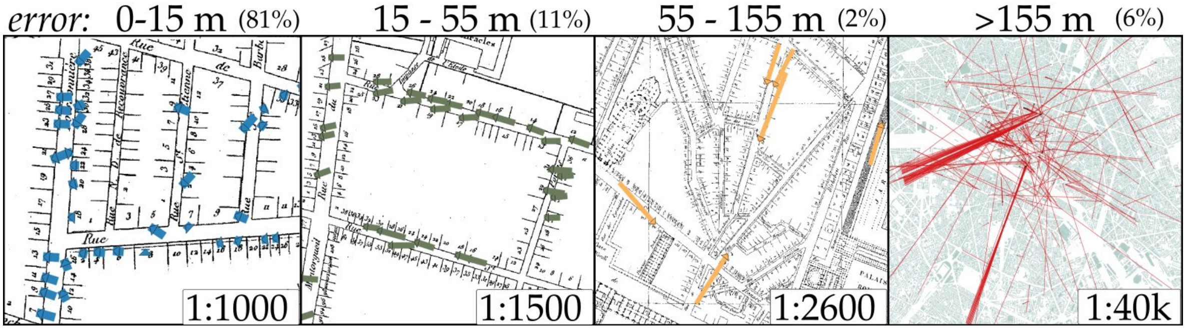


Figure 13. An historian manually move the geocoded addresses.

For one of the data set (Textile 1 and 2), the historian manually correct the geocoded results. We then plot the segment between address point resulting of automated geocoding and address point after manual editing. Results are presented in the table 2 and in Figure 13 We classify the results based on the length of this segment (*i.e.* the error in meter the geocoding method made).

- When the edit move the adress point less than 15 meters, we can consider that the edit is mostly about small moves , for instance centering the point on the building limit.
- Between 15 and 55 meters, the correct street has been found, but the building numbers are slightly misplaced (a few numbers).
- Between 55 and 155 meters, in most cases the street is correct, but the building numbers are far from their correct position.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

Table 2. Evaluating the error of geocoded results, via the dist. (geographic distance) of edit, the percentage of the total 8823 addresses, the average aggregated distance score, the average string distance w_d, the average temporal distance t_d, and the subjective most common edit reason we encountered while browsing the data

dist. (m)	%	avg(agg)	avg(sem)	avg(tempo)	main edit cause (subjective)
0 - 15	81 %	9.4	0.07	19.5	moving point on building limit
15- 55	11 %	12.4	0.09	27.2	small numbering editing (same street)
55 - 155	2 %	23.7	0.14	41.2	large numbering editing (same street)
155 - 7.2k	6 %	26.9	0.18	49.1	editing street

- Above 155 meters, streets are wrong in most of the case.

We stress that given Paris building average size, and the lack of precise definition of an address (is it the position of the door, of the center of the building,...?) results up to 55 meters could be considered as very close to ground truth.

3.4. Collaborative editing

We propose several User Interface for easy geocoding, and collaborative editing of the geocoding results. We informally tested the interfaces and found that they facilitate geocoding, especially for the batch mode. We also test the collaborative editing in two use cases. In the first use case a specialised user geocodes a single address and display the top 3 results corresponding to this address. The user is expert and its goal is geocode an address and assess the reliability of the result at the same time. In the second use case, a user batch geocode several addresses (30), looking of the best result for each adress. Then the user display the results on the map and check/edit the adresses.

Figure 14. Two use cases: First use case, an expert geocodes an address and analyse the top 3 results to assess the reliability of the result. Second use case: a user batch geocodes 30 addresses (1 result per address) in Paris and chek/edit the results.

3.4.1. Use case 1: top 3 results for one address

Using the web application, we geocode the address "10 rue de vaugirard, paris" for the date 1840, and ask for the top 3 results, as shown in first part of illustration 14. A matching building number geohistorical object exists in the three gazetteers extracted from the three maps. Based on the results, we can safely assume that this building number has not changed during the last 2 centuries.

3.4.2. Use case 2: batch geocoding of 30 addresses and check/edit

In this use case, a regular user is to check/correct 30 random addresses from the Jacoubet map using the web application. The task is performed quickly, the check and edit of each address is a matter of a few seconds. The main time consuming task is the loading of the background historical map, due to unfortunate hardware limitations. The edit speed seems to be on par with a desktop based edit solution (using QGIS).

4. Discussion

4.1. Genericity

Reaching a more generic geocoding service is important if we want to make it usable in other contexts and to profit from the various sources of knowledge on past spaces.

4.1.1. Geohistorical sources and data

Using external resources from the Web of data as new sources

Besides features representing address points and streets, georeferenced features of other types could be used with benefits by the geocoding service. As a matter of fact, people often refer to places of interest, such as famous buildings, monuments like statues or fountains or even named neighbourhoods to describe their position in space. We thus consider adding data about places of interest to improve our geocoding service. Like the data that was used to build the geocoder, such data could be gathered from ancient maps. But they may also come from existing gazetteers and knowledge bases published on the Web of data, such as DBpedia[17], Yago[18], the Getty Thesaurus of Geographical Names [19] or the gazetteer of place names published by the French National Library[20].

Widen the spectrum of cartographic sources

We exploit Jacoubet and Alphand maps, yet there are several more to be exploited toward the end of the 19th century, and in the beginning of the 20th century. From the beginning of the 20th century, Paris city administration produced a map per year. Of course, the main improvement direction would be to add maps of other cities/countries! For France at least, major cities have often been mapped starting from 1900.

Before the beginning of 19th century, the address system was very different in Paris. in mid 18th century, the address system was in fact that each building would have a specific name (no number, no notion of street name) in its neighbourhood. Our geocoding system has also been designed with this type of addressing but it has not been tested yet. More generally, this type of indirect localisation is very close to the field of web of knowledge.

Diversity in geohistorical objects natures

In this article several type of geohistorical objects were used for geocoding: building numbers, streets axis, neighbourhood. Other datasets were investigated as well, such as the city limits extracted by the project Geo Historical Data in a collaborative way from the Cassini maps [25]. In fact, a compiled version of city limits (GeoPeuple project [29]) from 1793 to 2010, created by EHESS, has also been tested. But building cadastres could also be integrated so as to have a building layout associated to an address rather than a point, which would solve an old problem of address points. Indeed, there is currently no consensus as to where a building number address point should be positioned: on the entry door, on the letter box, etc. More excitingly, in some cases, more precise data is available, giving the layout of apartments in buildings.

---

17 http://wiki.dbpedia.org/
18 http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/#c10444
19 http://www.getty.edu/research/tools/vocabularies/tgn/
20 http://data.bnf.fr/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at ISPRS Int. J. Geo-Inf. 2018, 7, 262; doi:10.3390/ijgi7070262

4.1.2. Genericity in usages

Named Entity Linking

As we mentionned in section 4.1, people often refer to place names to describe their position in space. The task of retrieving place names in a gazetteer or in a knowledge base, also known as (Spatial) Named Entity Linking or toponym resolution, is a widely used way of desambiguating mentions of spatial named entities extracted from texts by means of natural language processing approaches for information retrieval, information extraction or document indexing purposes [30]. As we plan to upgrade our geohistorical database with data about places of interest, we also have to adapt our geocoding service in order to make it retrieve reference data stored in the database and corresponding to place names mentions proposed by the users. Spatial Named Entity Linking implies solving issues related to places names inherent ambiguity [31], such as the fact that a place may have several names or the fact that several places may be designated by the same name. For each spatial named entity mention to be disambiguated, unsupervised state of the art approaches first select candidates from the gazetteer based on character string similarity. Then, they introduce additional criteria in order to decide which candidate is the best reference for a given place name, usually taken from the textual context of the mention [32,33]. In cases where textual context is very limited, like in tweets or location descriptions extracted from directories, this step of candidate ranking reveals even more challenging [34].

Analysis tool of the cartographic sources content

It is interesting to look at what historical sources were the most used for geocoding, although the historical source are chosen based on a complex ranking function. If we take the example of the over 10k geocoded addresses from the "Artists accommodations" dataset, we could expect all of the results to be drawn from the Jacoubet map, as the dataset is between 1793 and 1836, and the Jacoubet map is also in this range. Yet, analysing the results shows that if Jacoubet was used for 80% of the addresses, Alphand was used for 15%, although the map comes 30 years after. More surprisingly, the OpenStreetMap current data is still used for 5% of addresses, although it is about 2 centuries after the dataset.

Similar analysis on other datasets show similarly that all maps are always used, with of course a focus on the temporally closest map. Interestingly, these results are in agreement with similar work as presented in [35], chapter 4, where a prototype of multi-temporal geocoding is proposed. The approach shows that for different datasets, all references maps (Jacoubet, Alphand and BDAdresse (2010)) are used, with proportions depending on the parameters considered and the weights of each criteria. We think that this results are explained by the fact that historical maps miss some information, contain errors, and do not have the same geographical coverage.

4.2. Quality of the geocoding

4.2.1. Increasing the quality of the gazetteers

Collaborative enrichment

We propose several ways to use the geocoding capabilities in an easy way through web based User Interfaces. As we propose prototypes, the experiments are merely proofs of concepts for the moment. For a real validation, a complete user study would be required, which is outside of the scope of this article.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 April 2018 doi:10.20944/preprints201804.0088.v1

Peer-reviewed version available at *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 262; doi:10.3390/ijgi7070262

Cross-referencing historical topographic maps

One way to improve quality of available historical data is by using advanced crossreferencing. Indeed, the process of linking and merging similar data from heterogeneous datasets, which is called data conflation, enables to transfer information from one feature to the another, and thus may brings additional knowledge about data imperfections without using ground truth data which are non-existent for geohistorical data. For instance, [35,36] proposed an aggregated spatio-temporal graph to merge and confront historical road networks. This process can reduce data heterogeneity and allow to detect aberrations such as toponymic or numbering errors, or doubtful temporal trajectories of objects like short disappearances, thereby leading to better data quality. Advanced cross-referencing also makes it possible the construction of a genealogy of adresses by considering temporally linked addresses, that can deal with toponymic evolution or changes in adressing systems or numbering of buildings, thus paving the way for better spatio-temporal geocoding result.

4.2.2. Communicating the reliability of a geocoding

Geocoding qualification and quality measures

Modern geocoders are evaluated by how often they find a localisation, and how precise is the localisation they return (see [37] for instance). The first criterion shows how able to retrieve an address the geocoding algorithm is and also how exhaustive the gazetteer is. The second criterion refers to the positional accuracy of the gazetteer. Using such quality evaluation measures that encompass both the algorithm results and the gazetteer completeness makes the evaluation of their respective quality impossible. Contrary to that, in the field of named entity linking, distinct quality evaluation measures have been proposed for to the entity retrieval algorithm, like the measures proposed by [33] and completed by [38], and for the reference knowledge base (see [39] for knowledge bases general quality measures and [40] to evaluate the fitness of some knowledge bases for a given named entity linking task).

Geovisualisation

The prototype of graphical user interface we propose could be improved in several ways. The goal would be to efficiently provide information to user about the quality of geocoding, and the context of results. First, the point displayed to represent the result could have a size proportional to estimated spatial precision. This would help to visually assess the relevance of information. Second, the result could be colour-coded to represent the temporal proximity witht he input date. In a similar spirit, when multiple results are proposed, a time slider would be most useful to graphically disambiguate between result candidates. Third, the background historical map displayed in the prototype is currently set. Yet, the most appropriate background map could be automatically displayed based on the input addresses date provided by the user. Last, the current prototype becomes easily cluttered whend displaying a great amount of labels. Several strategies could be used, such as a better clustering of spatially close results, shorter labels, or better labels placement.

4.2.3. Integrating user correction into historical sources

In collaborative editing, edit come from untrusted sources. Validating edits and solving conflicts is then a classical problem. In our prototypes, every user edit is potentially used by the geocoder (they are added to a dedicated gazetteer). We could use a voting scheme where edits are only taken into account when a sufficient number of user have made them. However, we stress that due to the number of data to edit (several hundred thousands building numbers), we prefer to rely on the user benevolence, by considering that user spending time editing centuries old historical data are committed to accurate editing.

4.2.4. Scalability

The main design choice of our geocoding architecture is to use a flat model for the address (an address is any set of characters), as opposed to current geocoder which are highly hierarchical (an address refers to a street, that refers to a neighbourhood, etc.). This modelling choice gives the freedom that is necessary for data as incomplete as the historical ones, but also comes with a tradeoff regarding scaling capabilities. Indeed, for strongly hierarchical data, it is possible to have separate databases for each city for instance, thus preventing one database to grow too much, and ensuring a nice scaling capability.

This is not however the case with our architecture. By using database indexes, we can theoretically guarantee a fast geocoding time for up to few dozen of millions of geohistorical object used as sources. The main bottleneck in this case is not the temporal aspect (it relies on PostGIS geometry, which enable multiple theoretical solution for scaling), but the textual aspects (*i.e.* the address string itself). To scale over dozens of million of addresses, specific architectures may be used to deal with the text search, for instance distributed database (database sharding), in a similar spirit to the current software Elastic Search. We stress however that given the current available amount of historical sources, such scaling problem should not be an issue before a long time.

5. Conclusion

This article tackles the historical geocoding problem. As shown throughout the article, the historical aspects bring major complications to the geocoding problem. The main difficulties come from the nature of historical data (uncertainty, fuzzy date, precision, sparseness), which prevents the use of current-address geocoding methods based on strong hierarchical modelling. Instead, we propose a historical geocoding system based on a sound geohistorical object model. This model is designed to cover the minimal features, and, by its genericity, modularity, and open source nature, can easily be extended to feat other historical sources. Geohistorical objects from several historical sources have been integrated into the database and coherently georeferenced and edited to form gazetteers. Geocoding an address at a given time is then a matter of finding the best matching geohistorical object in the gazetteers, if any. Our simple, coherent, historical geocoding system has been tested on several real-life datasets collected by historians and can be easily used for other places/times/types of localisations. Diverse historical sources covering two century for the city of Paris have been integrated into the geocoder. The proposed geocoder is able to localise a large percentage of addresses with a fast speed (about 200ms per address). Finally, the article describes a prototype of web-based User Interface that demonstrates the interest of collaborative editing of localisation of addresses, and helps historians and other digital humanities researchers use geocoding services.

Supplementary Materials: All the code and additional documentation are available on the project websites http://geohistoricaldata.org and its associated code repository https://github.com/Geohistoricaldata. The code for the geocoder itself it available here: https://github.com/GeoHistoricalData/historical_geocoding.

Acknowledgments: Thanks to the Belle Epoque project and Angelo Riva and Thierry Géraud, the Institut Louis Bachelot for funding. Thanks to historians who contributed to creating the datasets, especially Benoit Costes for the edit of Alphand map.

1. Goldberg, D.W.; Wilson, J.P.; Knoblock. From Text to Geographic Coordinates: The Current State of Geocoding. *Journal of the Urban and Regional Information Systems Association* **2007**, *19*, 33–46.
2. St-Hilaire, M.; Moldofsky, B.; Richard, L.; Beaudry, M. Geocoding and Mapping Historical Census Data: The Geographical Component of the Canadian Century Research Infrastructure. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* **2007**, *40*, 76–91. doi:10.3200/HMTS.40.2.76-91.
3. Daras, K.; Feng, Z.; Dibben, C. HAG-GIS: A spatial framework for geocoding historical addresses **2014**.
4. Hutchinson, M.J.; Veenendaal, B. An agent-based framework for intelligent geocoding. *Applied Geomatics* **2013**, *5*, 33–44. doi:10.1007/s12518-011-0063-z.

5. Roongpiboonsopit, D.; Karimi, H.A. Comparative Evaluation and Analysis of Online Geocoding Services. *International Journal of Geographical Information Science* **2010**, *24*, 1081–1100. doi:10.1080/13658810903289478.
6. Clough, P.; Tang, J.; Hall, M.M.; Warner, A. Linking archival data to location: A case study at the UK national archives. *Aslib Proceedings* **2011**, *63*, 127–147. doi:10.1108/00012531111135628.
7. Mostern, R.; Johnson, I. From named place to naming event: creating gazetteers for history. *International Journal of Geographical Information Science* **2008**, *22*, 1091–1108, [http://dx.doi.org/10.1080/13658810701851438]. doi:10.1080/13658810701851438.
8. Smith, D.A.; Crane, G. Disambiguating Geographic Names in a Historical Digital Library. International Conference on Theory and Practice of Digital Libraries. Springer, 2001, pp. 127–136.
9. Goodchild, M.F. Citizens as sensors: the world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. doi:10.1007/s10708-007-9111-y.
10. Heipke, C. Crowdsourcing geospatial data. *{ISPRS} Journal of Photogrammetry and Remote Sensing* **2010**, *65*, 550 – 557. {ISPRS} Centenary Celebration Issue, doi:http://dx.doi.org/10.1016/j.isprsjprs.2010.06.005.
11. Haklay, M., Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*; Sui, D.; Elwood, S.; Goodchild, M., Eds.; Springer Netherlands: Dordrecht, 2013; pp. 105–122. doi:10.1007/978-94-007-4587-2_7.
12. Fomel, S.; Claerbout, J.F. Guest Editors' Introduction: Reproducible Research. *Computing in Science Engineering* **2009**, *11*, 5–7. doi:10.1109/MCSE.2009.14.
13. Aruliah, D.A.; Brown, C.T.; Hong, N.P.C.; Davis, M.; Guy, R.T.; Haddock, S.H.D.; Huff, K.; Mitchell, I.; Plumbley, M.D.; Waugh, B.; White, E.P.; Wilson, G.; Wilson, P. Best Practices for Scientific Computing. *CoRR* **2012**, *abs/1210.0530*.
14. Wilson, G.; Bryan, J.; Cranston, K.; Kitzes, J.; Nederbragt, L.; Teal, T.K. Good Enough Practices in Scientific Computing. *ArXiv e-prints* **2016**, [arXiv:cs.SE/1609.00037].
15. Marwick, B. Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory* **2016**, pp. 1–27. doi:10.1007/s10816-015-9272-9.
16. Duménieu, B. Un Système d'information Géographique Pour Le Suivi d'objets Historiques Urbains à Travers l'espace et Le Temps. PhD thesis, Ecole des Hautes Etudes en Sciences Sociales, 2015.
17. Armstrong, M.P. Temporality in spatial databases. GIS/LIS'88, 1988, pp. 880–889.
18. Herrault, P.A.; Sheeren, D.; Fauvel, M.; Monteil, C.; Paegelow, M. A comparative study of geometric transformation models for the historical "map of france" registration. *Geographia Technica* **2013**, pp. pp–34.
19. Fabbri, R.; Kimia, B. 3D curve sketch: Flexible curve-based stereo reconstruction and calibration. Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 1538–1545.
20. Cléri, I.; Pierrot-Deseilligny, M.; Vallet, B. Automatic Georeferencing of a Heritage of old analog aerial Photographs. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **2014**, *2*, 33.
21. Bitelli, G.; Cremonini, S.; Gatta, G. Ancient map comparisons and georeferencing techniques: a case study from the Po River Delta (Italy). *E-perimetron* **2009**, *4*, 221–228.
22. Boutoura, C.; Livieratos, E. Some fundamentals for the study of the geometry of early maps by comparative methods. *e-Perimetron* **2006**, *1*, 60–70.
23. De Runz, C.; Desjardin, E.; Piantoni, F.; Herbin, M. Anteriority index for managing fuzzy dates in archæological GIS. *Soft Computing* **2010**, *14*, 339.
24. Kauppinen, T.; Mantegari, G.; Paakkarinen, P.; Kuittinen, H.; Hyvönen, E.; Bandini, S. Determining relevance of imprecise temporal intervals for cultural heritage information retrieval. *International journal of human-computer studies* **2010**, *68*, 549–560.
25. Perret, J.; Gribaudi, M.; Barthelemy, M. Roads and cities of 18^{th} century France. *Scientific Data* **2015**, *2*. 10.1038/sdata.2015.48.
26. Noizet, H.; Bove, B.; Costa, L. *Paris de Parcelles En Pixels*; Presses universitaires de Vincennes, 2013.
27. Gribaudi, M.; Magaud, J. L'action publique et ses administrateurs dans les domaines sanitaires et social en France, 1800 à 1900, 1999.
28. Lazzara, G.; Levillain, R.; Géraud, T.; Jacquelet, Y.; Marquegnies, J.; Crépin-Leblond, A. The SCRIBO Module of the Olena Platform: A Free Software Framework for Document Image Analysis. Proceedings of

the 11th International Conference on Document Analysis and Recognition (ICDAR); , 2011; pp. 252–258. doi:10.1109/ICDAR.2011.59.

29. Plumejeaud-Perreau, C.; Grosso, E.; Parent, B. Dissemination and Geovisualization of Territorial Entities History. *Journal of Spatial Information Science* **2014**, *8*, pp. 73–93. doi:10.5311/JOSIS.2014.8.119.
30. Shen, W.; Wang, J.; Han, J. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering* **2015**, *27*, 443–460.
31. Overell, S. The problem of place name ambiguity. *SIGSPATIAL Special* **2011**, *3*, 12–15.
32. Mihalcea, R.; Csomai, A. Wikify!: Linking Documents to Encyclopedic Knowledge. Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management; ACM: New York, NY, USA, 2007; CIKM '07, pp. 233–242. doi:10.1145/1321440.1321475.
33. Hachey, B.; Radford, W.; Nothman, J.; Honnibal, M.; Curran, J.R. Evaluating Entity Linking with Wikipedia. *Artif. Intell.* **2013**, *194*, 130–150. doi:10.1016/j.artint.2012.04.005.
34. Zhang, W.; Gelernter, J. Geocoding location expressions in Twitter messages: A preference learning method. *Journal of Spatial Information Science* **2014**, *2014*, 37–70.
35. Costes, B. Vers la construction d'un référentiel géographique ancien. Un modèle de graphe agrégé pour intégrer, qualifier et analyser des réseaux géohistoriques. PhD thesis, Université Paris-Est, 2016.
36. Costes, B.; Perret, J.; Bucher, B.; Gribaudi, M. An aggregated graph to qualify historical spatial networks using temporal patterns detection. 18th AGILE International Conference on Geographic Information Science, 2015.
37. Zimmerman, D.L.; Fang, X.; Mazumdar, S.; Rushton, G. Modeling the Probability Distribution of Positional Errors Incurred by Residential Address Geocoding. *International Journal of Health Geographics* **2007**, *6*, 1.
38. Brando, C.; Frontini, F.; Ganascia, J. REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets. *CSIMQ* **2016**, *7*, 60–80. doi:10.7250/csimq.2016-7.04.
39. Zaveri, A.; Rula, A.; Maurino, A.; Pietrobon, R.; Lehmann, J.; Auer, S. Quality assessment for linked data: A survey. *Semantic Web* **2016**, *7*, 63–93.
40. Brando, C.; Abadie, N.; Frontini, F. Linked Data Quality for Domain Specific Named Entity Linking. Proceedings of the 1st Atelier Qualité des Données du Web, 16ème Conférence Internationale Francophone sur l'Extraction et la Gestion de Connaissances, Reims, France, 2016.