PeGazUs: A knowledge graph based approach to build urban perpetual gazetteers

Charly Bernard[1 [0009−0003−8170−3671]], Solenn Tual[1 [0000−0001−8549−7949]], Nathalie Abadie[1 [0000−0001−8741−2398]], Bertrand Duménieu[2 [0000−0002−2517−2058]], Joseph Chazalon[3 [0000−0002−3757−074X]], and Julien Perret[1 [0000−0002−0685−0730]]

[1] LASTIG, Université Gustave Eiffel, IGN-ENSG, 73 Avenue de Paris, 94165 Saint-Mandé Cedex, France
{charly.bernard, solenn.tual, nathalie-f.abadie, julien.perret}@ign.fr
[2] Centre de Recherches Historiques, École des Hautes Études en Sciences Sociales
bertrand.dumenieu@ehess.fr
[3] LRE, EPITA
joseph.chazalon@lrde.epita.fr

Abstract. Gazetteers, as compilation of named places, are central resources on the Web of data, as they provide a common ground to link and integrate many textual or structured resources on the Web. Gazetteers usually categorise and associate places names with geospatial coordinates. In more recent times, historical gazetteers, which aim to represent places from the past, have received increasing attention. The creation of these gazetteers poses specific challenges, including the definition of the identity of evolving places, the representation of their evolution through time (how they change, when the changes happen), and the population of the gazetteer based on scarce and heterogeneous historical sources.
We propose an approach to create an urban historical gazetteer on the evolution of two major urban large-scale types of places, namely addresses and land plots. Our proposal is inspired by approaches for creating knowledge graphs and takes advantage of the knowledge representation and reasoning possibilities offered by Semantic Web standards to address the aforementioned challenges. The approach was applied to the *Butte aux Cailles* district of Paris, for which a variety of contemporary and historical sources were used. The resulting knowledge graph can be used for a variety of purposes, including historical geocoding of old documents, identifying the use of a plot of land at a given date, and recording the events that led to its current state.

Keywords: Historical urban gazetteer · Knowledge graph · Addresses and land plots evolution.

1 Introduction

Spatial indexing of digitised archival records is a key issue to help scholars easily retrieve documents about a place of interest. State-of-the-art approaches generally compare places mentioned in documents with those in gazetteers, in order

to disambiguate homonyms and get absolute or relative location information. Gazetteers are place names repositories which serve two purposes: linking place names to locations, and describing the places they list [17]. Thus, they usually gather for each place at least one name, one type and one location represented most of the time by coordinates. But old documents are very likely to mention places that have changed or even disappeared through time. In recent years, many projects have been carried out to create historical gazetteers, but rarely at urban level.

Ducatteeuw [15] defines an urban historical gazetteer as an information resource representing places on the street level and their evolution through time. This kind of resource is meant to spatially index historical sources like censuses, tax registers, directories and so on. In this article, we aim at creating a gazetteer which represents not only streets, but the addresses and land plots located on the streets, as they are key spatial entities for the fine-grained geolocation of very large historical corpora. Addresses and plots are *fiat* geographical objects, designated by phrases or identifiers based on a spatial hierarchy. They are social constructs, generally created either through peoples' practices or by an administrative authority, whose use may persist over time, even after they have been officially cancelled by the authorities. Representing such kind of spatial entities in a historical gazetteer therefore poses specific challenges.

First, representing old geographical features, for which there is no ground truth anymore, implies relying on historical sources to extract useful data. Depending on their type and their valid time, these will provide very different descriptions of the geographical features. It will therefore be necessary to link the different representations of the same geographic entity across the available sources to leverage their complementary contributions. Another difficulty is that some sources do not directly describe the state of geographical entities, but the events that happen to them. For example, a record of municipal administration decisions could provide information on street name changes. It is therefore important to take advantage of this change information to infer new facts about the state of geographical features; and conversely, to use the data available about the successive states of geographical features to infer the events between them. These events can have consequences for the the properties of geographical entities or for their identity itself. It is therefore essential to have a model that can represent different successive property values for geographical entities, depending on the period of their existence and the events that have affected them.

In this article, we first present related works on the challenges posed by historical gazetteers. Then we present an ontology for representing an urban historical gazetteer containing descriptions of old addresses and plots and their evolution. We also present a generic strategy to automatically populate this ontology from scarce historical sources, that represent either the states of geographical entities or the events that affect them, and infer missing information. We evaluate the ontology and our populating strategy by constructing a geohistorical knowledge graph on the addresses and plots of land in the 19^{th} century Paris district of *La Butte aux Cailles*.

2 Challenges for a urban historical gazetteer

From early historical Geographic Information Systems ([12] or [16]) and geospatial standards for gazetteers ([7], [8]), to recent works based on semantic Web standards, the question of the most suitable data model to represent historical named places dataset has been widely addressed in the literature ([28], [17], [27]). GIS-based solutions require precise geometries to represent the shape and location of places and are often not suitable for old places for which such information is rarely available. As pointed out by Berman [9], knowledge graphs are well suited to represent fragmented, incomplete data scattered across multiples sources.

2.1 Ontologies of named places for historical gazetteers

The World Historical Gazetteer platform and the Pelagios project have jointly developed the Linked Places JSON-LD format for historical gazetteers [18]. In its underlying ontology, a place attestation can be described by several names, types, relations, locations and temporal information. This temporal property can be represented by timespans represented according to the OWL-Time ontology [5], periods identified by URIs pointing to time gazetteers like PeriodO [3] or even labels and duration values. Besides, it can be associated either to the place attestation itself or to its name, type, relation and location properties, so as to provide versions for them. [15] builds upon this format to propose an ontology for urban gazetteers, representing streets, at a high level of detail. It is based on the classes and properties used in the Linked Places format, but they are replaced by their equivalent classes and properties from the upper level ontology designed for cultural heritage data, namely the CIDOC Conceptual Reference Model.

The two previous models have the advantage of representing successive versions of places and their properties. However, they do not include the events that cause places to change. Finally, they are intended for named places, not geographical entities designated by complex statements that include references to other named places, like old addresses or land plots.

2.2 Representing data changes and real world events

One of the first attempts to represent evolving geographical entities and the changes they undergo has been proposed by [20]. This work uses geospatial ontologies representing Finnish municipalities and their part-of relationships for successive time spans. Changes are also represented to link the successive states of the municipalities. Five types of changes are considered: establishment, merge, split, name change and part-of relationship change.

[10] proposes an ontology called TSN to represent Territorial Statistical Nomenclatures with their territorial units and their successive versions. The change bridge approach proposed by [20] to represent the changes that territorial units undergo over time is reused and extended in the TSN-Change ontology.

For each change, it explicitly represents the upstream and downstream territorial units involved, the type of change (using the same types as [20]), the subchanges potentially induced by the current change and the real-world event responsible for this change at data level (designated by the *isCausedBy* predicate).

The Hierarchical Historical Territory ontology (HHT) proposed by [13] differs from the TSN/TSN-Change ontologies by focusing on representing multiple hierarchies between territories while TSN is designed to represent a single nomenclature. The chosen approach is also based on versions of territorial units, but with a temporal partitioning based on the territorial changes rather than on systematic snapshots, which are better suited to statistical data than to historical data. The change bridge principle is also adopted, but two types of changes are considered: those affecting a single territorial unit and those affecting several.

2.3 Populating an ontology of historical places and their changes

The approach proposed by [20] includes a methodology to construct the ontology time series from metadata tables describing changes and locations (current and historical). All the data is therefore prepared in advance, possibly manually, to match the ontology's expected content. [10] proposes to populate the TSN and TSN-Change ontologies using an algorithm that takes temporal snapshots of geographical data as input, populates the TSN ontology from this data and compares the geometries of administrative units to create links between successive versions of these units. Changes are then inferred automatically by interpreting the different configurations of links between versions of territorial units. [13] propose a rule-based algorithm to automatically link each territorial unit version, already represented according to the HHT ontology, to its chronological successor and detect and classify changes between them.

2.4 Attestations and historical sources

As stated in [17], gazetteers do not represent places, but attestations of places: Each resource representing a place should therefore be modelled as an aggregate of sourced assertions about that place. For historical gazetteers, where ground truth is no longer available, this is of particular importance.

Historical documents are the only available sources to report traces of the past. As secondary sources of information, their content is the result of interpretations and observations whose quality and reliability are variable and often difficult to assess. It is therefore necessary to use a variety of complementary sources to populate a historical gazetteer. Finally, as the information available is incomplete, it is often necessary to infer the missing data from the knowledge and facts available. It is thus essential to document the provenance of inferred facts, to enable users to distinguish them from those based on historical sources.

Ontologies such as Prov-O [4] or the Factoid Prosopography Ontology [1] have been proposed to describe the provenance of data and can be used in conjunction with ontologies to describe historical sources such as RiC-O [2].

3 Data sources on land plots and addresses

Land plots and addresses are typical cases of geospatial entities described in multiple, fragmentary and heterogeneous sources of information that have different temporal validities, and different ways of describing geographic entities.

3.1 Contemporary geographic data

The municipality of Paris publishes two main geospatial datasets on the city's thoroughfares and addresses. The first one named *Dénominations des emprises des voies actuelles*[4] describes all thoroughfares with their names, geometries, dates of creation along other secondary metadata. The second one, the *Dénominations caduques des voies*[5], describes ancient thoroughfares with their date of deletion, and is structured in the same way, but no geometry is provided.

The *Base Adresse Nationale* (BAN)[6] contains all the postal addresses registered in France, each address being structured as a list of housenumber, street, city, and zip code, with a geographical position.

Lastly, volunteered geographic information is also used. OpenStreetMap is a geographic database that tends to represent the current state of geographic entities, whereas Wikidata provides the history of these entities.

3.2 Historical large scale city plans

Multiple large-scale topographic maps of Paris have been produced since the 18^{th}. They can be leveraged as valuable sources of structured geohistorical data at the cost of extensive operations of vectorization. Several digital humanities projects have carried out such works and released open datasets[7]. As an example, the *Atlas National de la ville de Paris* finely depict the streets of Paris at the scale of 1:1720 and was published between 1791 and 1799. Another example is the *Atlas municipal des vingt arrondissements de la ville de Paris*, which represents the city of Paris at the end of the 19^{th} century. These sources have similarities with OpenStreetMap or BAN since their main goal is to describe geographical entities at a given point in time without taking into account their evolution.

3.3 Street dictionaries

Unlike maps, Paris street dictionaries are not a snapshot of the city, but instead provide a historical descriptions of every streets. Such sources contain indirect spatial references: the district to which the street belongs, addresses giving the beginning and the end of the lane. The *Dictionnaire administratif et historique*

[4] https://opendata.paris.fr/explore/dataset/denominations-emprises-voies-actuelles
[5] https://opendata.paris.fr/explore/dataset/denominations-des-voies-caduques
[6] https://adresse.data.gouv.fr/
[7] E.g. Projets Time Machine (https://ptm.huma-num.fr/), ALPAGE (https://alpage.huma-num.fr) or SODUCO (https://soduco.geohistoricaldata.org).

des rues de Paris et de ses monuments by Félix and Louis Lazare published in 1844 is a typical example of this type of document. Another major street dictionary is the *Dictionnaire historique des rues de Paris* by Jacques Hilairet is published in 1960. Unlike the previous dictionary, it includes streets in the outer districts of Paris, which became part of the capital in 1860.

3.4 Cadastral maps and registers

The Napoleonic land registry is the first land registry of the entire French territory. It was created between 1808 and 1850, depending on the departement and commune. Its goal was to make the system of land taxation more efficient and to make tax collection more equitable. The Napoleonic land registry consists of two types of documents: maps and registers. Index maps represent plot division on a very large scale. Each plot is delimited and associated with a number. The initial registers are the legend of these maps at the time of their production. Maps and initial registers were not updated after their creation. The mutation registers contain all the plot updates (taxpayer, land use, tax value) over time. Plots are grouped by taxpayers in folios (numbered page or part of page describing the properties of a taxpayer) and sometimes by thoroughfares. Each table line describing a plot is a version of this plot at a given time.

4 An ontology for historical urban gazetteers

The first contribution of this work lies in the PeGazUs[8] ontology. Like a perpetual calendar that represents the day of the week on any given date, it is intended to represent the address or plot number of a location on any given date. Like those proposed by [20], [10], and [13], it is based on the *Change Bridge* concept. But unlike them, it does not impose a hierarchy between them, and above all it allows versions of geographical entities to be represented whose property values can evolve over time. This is particularly useful when the identity of geographical entities cannot always be identified *a priori* in historical sources.

4.1 Ontology documentation and competency questions

To build the ontology, we followed the method called *Simplified Agile Methodology for Ontology Development*, also known as SAMOD [25]. This consists in separating a complex modeling problem into sub-problems called modelets, which are easier to process. A modelet begins with a natural-language argument describing the sub-problem to be addressed, along with a glossary defining the main terms involved. Each modelet comes with a set of informal competence questions, also in natural language, which represent the questions to be answered by the knowledge base. We defined these questions, by interviewing historians, archivists, and

[8] The PErpetual GAZetteer of approach-address UtteranceS ontology, its documentation, the data, the scripts used to build the data are available on this repository: https://github.com/umrlastig/pegazus-ontology

archaeologists about their needs regarding old addresses and plots. Finally, each question is associated with a set of example answers, which serve to validate the modelet once implemented.

The SAMOD method enables us to operate in rapid cycles, with regular testing of the ontology under construction. Different modelets were identified: address, temporal evolution, sources, land registry documents use and taxpayers.

4.2 Modelets structure

An address is an indirect spatial reference described by a structured statement that unambiguously designates a place [14]. The way addresses are structured is through an ordered sequence of spatial relationships between geographical entities (also called landmarks) [11]. Several models for addresses exist, such as *locn* [24] or ISO 19133 [6], but they focus on postal addresses and are not suited for less structured utterances typically found in historical sources, like *In the center of the capital city, between the Palais-Royal and the Tuileries, close to the main theaters.*

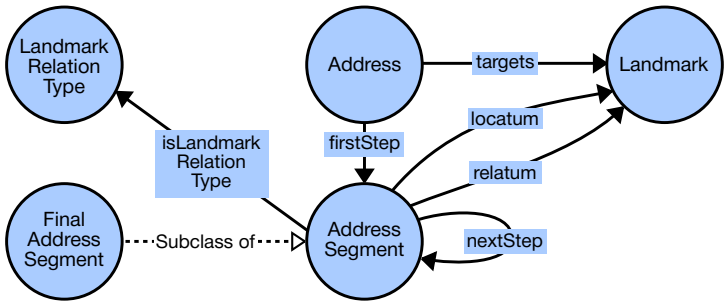

Fig. 1. Part of the ontology to describe addresses.

Addresses We proposed a common model for historical and postal addresses in [11] (see Fig. 1). An instance of `Address` corresponds to a structured statement. As it designates a place, it targets a `Landmark` which is a geographical entity (administrative unit, thoroughfare, building number, building...). An instance of `AddressSegment` is a spatial relationship between multiple instances of `Landmark` and its nature is given by `LandmarkRelationType`. To define the roles of landmarks for this relationship, `locatum` and `relatum` predicates are used [29]. In the spatial relation "Rue Gérard **is in** Paris", Rue Gérard is the locatum, Paris is the relatum, and "is in" is the landmark relation type. These segments form an ordered sequence which is described here by `firstStep` and `nextStep` predicates.

To validate this modelet, we selected the following competency questions: (1) What addresses are listed along a given street? (2) What are the coordinates of the target of a given address? (3) Which addresses are located in a given area?

An additional set of classes and properties are defined to describe plots depicted in the land registry. To validate this extension of the modelet, we defined the following questions: (4) Which are the plots located in a given commune or section of a commune? (5) What are the values of attribute `X` (nature, taxpayer or address) associated with a given plot ?

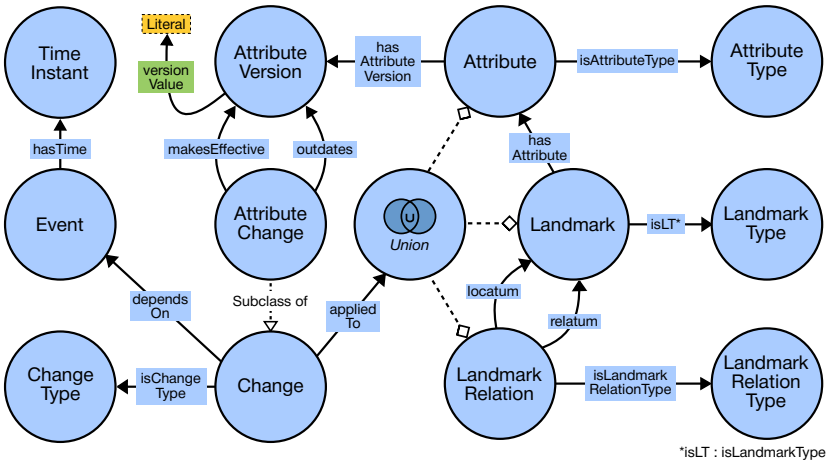

Fig. 2. Part of the ontology to describe geographical entities and their evolution.

Temporal evolution The landmarks mentioned in the addresses may change through time, which means that the addresses also change. To capture both changes, landmarks have to be modelled so that we can apply changes to them (see fig. 2). For landmark evolution, [10] and [13] use an approach based on the representation of states (also called versions). Each landmark has a set of its temporally ordered versions, each of which has a valid period. This implies that the attributes of the territorial units must have constant values for each version. But when landmarks evolve, changes do not always apply to all their attributes. For example, the extension of a street only affects its geometry but has no effect on its name. This is why we opt for a modelling approach in which each attribute is represented independently of the identity of the landmark to which it is related. Each attribute has a type (name, geometry, length, plot nature...) and may have versions, like in the Linked Places format. `LandmarkRelation`, a super-class of `AddressSegment`, allows to describe relations between landmarks such as spatial one. Fig. 3 shows an example of this modelling for the street called *Rue Gérard*.

Events that cause changes on the territory are also represented. These may involve one or more changes, each of which describes the evolution of a resource (`Landmark`, `LandmarkRelation`, `Attribute`). For example, Fig. 4 models an event that happened on August 30, 1978, whose textual description is "the eastern part of Rue Gérard located in Paris is now named Rue du Père Guérin". This event is composed of several changes: new geometry for the street called *Rue*

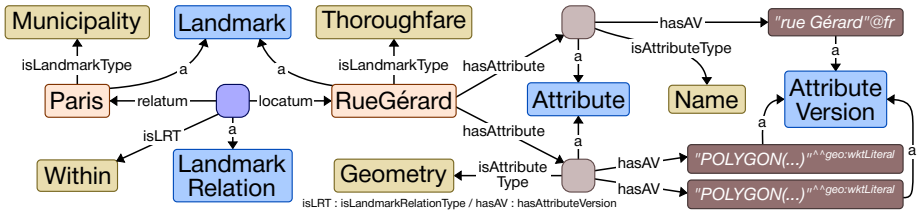

Fig. 3. *The Rue Gérard* modelled according the ontology.

Gérard, creation of the street called *Rue du Père Guérin*, and the appearance of a version for the name and geometry attributes of this street.

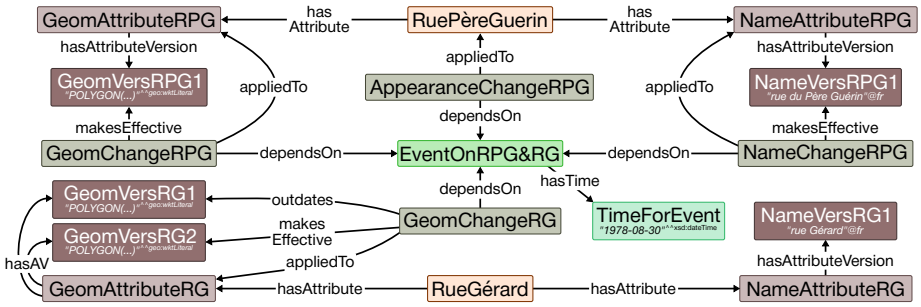

Fig. 4. Modelling of the event "the eastern part of Rue Gérard located in Paris is now named Rue du Père Guérin".

Four competency questions have been pointed out for this modelet: (1) what landmarks exist at a given time? (2) in what time interval(s) is an address of a given name valid? (3) what is the history of a landmark? In other words, what events are associated with it? (4) What states and events are missing from the history of an address? In the case of plots, we added the following questions: (5) What is the nature (or taxpayer or address) of a given plot at a given date? (6) What are the successive natures (or taxpayers or addresses) of a given plot?

Sources Each version of an attribute of a landmark can originate from one or more sources. Those sources might contain contradictory and/or false information. Thus, the link between data and primary sources must be preserved to be able to detect and explain these inconsistencies. Each source has to be described according to its nature. This description might also include its author, its creation date, its valid period (e.g. the period during which a register is updated). In the case of archival records made up of several parts, the articulation of these parts must also be described. Furthermore, if relevant, the tools and processes used to transform primary sources into structured data should be documented. The main competency questions for this modelet are: (1) From which source does information `X` come from? (2) What is the description of source `Y`?

Land registry documents This modelet has been developed to deal with land registry documents specificities. The Napeoleonic land registry is a sort of old fashion spatio-temporal database, printed and filled in manually. This modelet defines the capture rules that enable to follow one plot from one page to another within a register and from one register to another. It also lists the special values that appear in table cell of the registers (described in the `SpecialCellValuesList` skos concept list). These values are used to infer events (like construction, destruction) and changes, to follow the lifeline of each plot and to keep information like typography metadata (e.g. strikethrough) that provide a better understanding of how these document were filled out. The competency questions for this modelet are: (1) Which folios mention plot (or taxpayer) `X`? (2) Which resource have been created using a crossed-out table line?

Taxpayers Taxpayers are associated with the land they own or exploit. They can be natural or legal persons and are designated with a name, one or more given names, an activity and an address. The identity of the taxpayer is crucial information to link the plots mentioned in the different registers. It is also a relevant domain to study the land registry with a socio-economic point of view and to link the land registry with other significant archival resources like census or civil registration. The associated competency questions are as follow: (1) Who are the taxpayers of a given area? (2) What are the plots associated with a given taxpayer? (3) Who are the taxpayers living in a given commune? (4) Who are the taxpayers of a given commune with a given activity?

4.3 Alignments with existing vocabularies

According to good practices for the development of ontologies, existing vocabularies are reused. To describe the documents of the Napeolenic land registry, we integrate concepts of the *Records in Contexts Ontology* [2]. This ontology has been developed to describe archival records, including their structure, relations between their sub-parts, distinction between concepts of records and their instanciation. It also enables the description of derived data from primary sources. Combined with the Prov-Ontology [4] as described by [19], we can precisely detail the treatments that are used to create these new instanciations. Valid periods of attributes versions and temporal relationships are represented with the OWL-Time Ontology [5].

5 The *Butte aux Cailles* neighbourhood geohistorical KG

The Butte aux Cailles neighbourhood is located in the southeast of Paris. Formerly part of the neighbouring commune of Gentilly, it was incorporated into Paris in 1860. This event led to a major transformation of the local urban fabric and is one of the reasons why this study area was chosen.

The second contribution of this work lies in the algorithm proposed to populate the ontology. It is based on the factoid approach [23]. For each source,

its associated data is structured according to the ontology to form a so-called factoid graph. A factoid graph contains versions of landmarks describing them according to the source, possibly with various valid time intervals. Since factoids graphs are the base of construction of the knowledge graph, SHACL rules check if there are inconsistencies within its explicit triples [21]. Once all factoid graphs are built, we build a unique graph of facts whose goal is to rebuild each geographic entity identity from the factoid versions, and integrate all the attribute values from factoids graphs. We develop a six-steps process to reach this objective. This algorithm is iterative: some steps can be executed many times, while new knowledge is discovered at each step and is added to the final graph of facts.

5.1 Identity-based landmark versions rooting

The first step of the process aims to root similar landmarks from all the factoid graphs to a specific resource based on an identity criterion. Similar landmarks are linked to a root landmark using the `hasRoot` object property. Root landmarks can be landmarks from a given factoid graph (for plots) or a new empty resource initialised from one of these graphs (for addresses).

Districts and thoroughfares are linked based on name equality. In addition to having the same value, house number are equivalent if they are the locatum of a `Belongs` landmark relation whose relatum is the same thoroughfare (or district). Plots are linked if they have the same number and are in the same section of a commune. In the Napoleonic cadaster, when plots are split or merged, the new plots keep the same number as the previous objects, so that plot numbers are only a pseudo-identification criterion. In this situation a disambiguation step has to be performed later in the process (see section 5.4).

5.2 Ordering landmark versions and their attributes versions

Landmark versions linked to the same root are ordered using their valid time. For each version noted `?lv`, pairs of landmarks are formed with all versions that match this requirement : start($\texttt{?lv}_\texttt{i}$) $\geq$ start(`?lv`). Then, gaps between their valid times are calculated (i.e. difference between the end of the interval of $\texttt{?lv}_\texttt{i}$ and the start of the interval of `?lv`). If the gap is negative (not null), `?lv` `hasOverlappingVersion` $\texttt{?lv}_\texttt{i}$ relation is inferred. In case of positive gap (or null), `?lv hasNextVersion` $\texttt{?lv}_\texttt{i}$ is inferred if this gap is equal to the smallest gap (positive or null) related to `?lv`.

This step can be repeated with attribute versions. Indeed, the algorithm deals with heterogeneous sources that might not represent the same attributes (e.g. some have a geometry attribute and others do not). Thus, using global ordering of landmark versions could create disruptions in the sequence of attribute versions.

5.3 Inferring changes and events related to landmarks

This step aims at inferring `LandmarkAppearance` and `LandmarkDisappearance` changes depending on real-world `Event`s. In the case of addresses, changes and

events associated with the root landmarks are inferred using statements of the related factoids. For plots, this step is first done to infer these changes and events on the factoids. This first iteration of the method is mandatory to detect changes and events that impact landmark identity, and consequently to disambiguate versions that have the same root, the same plot number but might not be the same real-world object. A real-world plot is considered as a piece of land with a stable geometry between two events of type `Split` or `Merge`. Changes of type `LandmarkAppearance` or `LandmarkDisappearance` are detected using registers capture rules and the associated `Event` are created. As an example, two or more `Folio`s in a cell of the *Next folio* column of the mutation register tables are interpreted as a `Split` event.

5.4 Inferring landmarks identity

If the identity of landmarks is already known at the beginning of the process, as for the addresses, this step aims only at adding `hasTrace` relations between landmark versions and their root landmark in order to retrieve the information that composed them. If the identity is not known, as for the plots, this step aims to create `Landmark` resources that aggregate landmark versions that are likely to describe the same real-world object. Additional ordering relations between landmark versions are created based on the order of their mentions in the documents. The position of the line in the document used to create the landmark version is taken into account. `?lv1 hasNextVersionInSRCOrder ?lv2` and `?lv1 hasOverlappingVersionInSRCOrder ?lv2` are created, as well as their inverse properties.

According to these properties, ordered series of landmark versions without a version associated with a `LandmarkAppearance` or a `LandmarkDisappearance` change in the middle of their lifeline are merged to create a new `Landmark` resource corresponding to a real-world geographic entity. This new resource is linked to the versions that make it up with the `hasTrace` relationship.

5.5 Inferring attribute versions

The previous steps of the method lead to the linking of factoids resources describing the same real-world landmark (called fact landmark) and their temporal ordering. This fifth step aims to build attribute versions of facts landmarks using the attribute versions from the associated factoids. To do so, their values need to be compared and they have to be ordered temporally. Then, successive similar versions are aggregated.

Versions of the same attribute are ordered temporally using the strategy described in section 5.2. Their value are compared using different criteria according to the type of attribute. This results in creating `sameVersionValueAs` and `differentVersionValueFrom` relations between attribute versions. For instance, the comparison of two geometry attribute versions of a landmark has two be adapted to two situations. If geometries are lines or polygons, areas

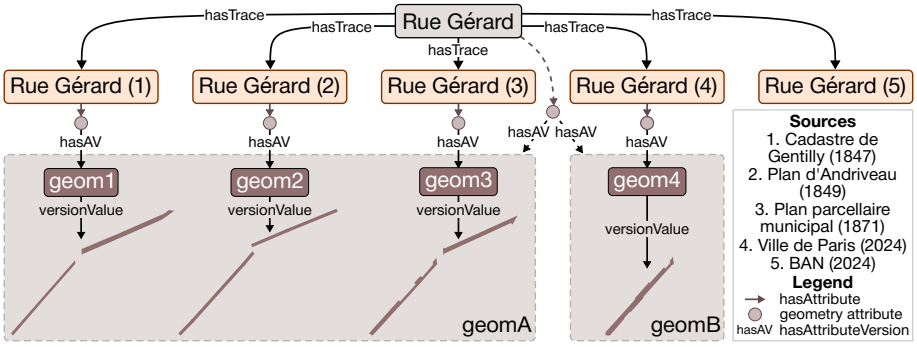

Fig. 5. Aggregation of geometry attribute versions from several sources for the street *Rue Gérard*. The dotted elements are those that have been inferred.

of the intersection and of the union between their bounding box can be computed. If the ratio between these areas is greater than a given value, they are considered as similar. In the case of points, they are considered as similar if the distance between them is below a threshold value. Taking the example of `?geom1` described in Fig. 5 : the following relations are infered `?geom1 sameVersionValueAs ?geom2, ?geom3 ; differentVersionValueFrom ?geom4`.

Finally, similar attribute versions can be merged. This is done according to two criteria: value similarity and valid time order. In other words, attribute versions are aggregated if they form a continuous sequence of versions with the same value and that follow each other in time. By taking the example for Fig. 5, `?geom1`, `?geom2` and `?geom3` are similar and the follow each other so they are aggregated in a root attribute called `?geomA`. On the other hand, `?geom4` is only similar to itself so it forms an aggregation of one version noted `?geomB`.

5.6 Inferring changes and events related to the attributes

Finally, changes and events that affect the attribute versions are inferred. They are ordered based on attribute valid time values. When successive versions with different values are detected, then changes are created. Each change depends on an event whose time can be derived from the validity of the versions.

6 Evaluation

6.1 Ontology evaluation

We checked the consistency and compliance with good design practices of the proposed ontologies using the HermiT reasoner integrated into Protégé[9] [22] as well as OOPS! tool[10] [26].

[9] http://protege.stanford.edu/
[10] https://oops.linkeddata.es/

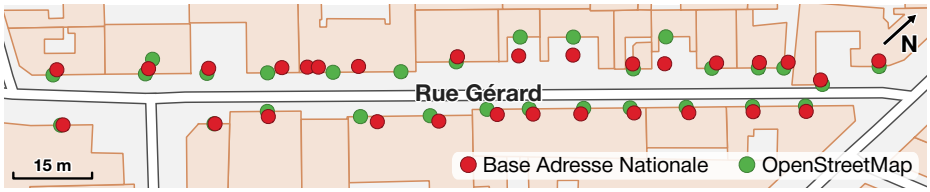

Fig. 6. Map of current addresses whose target is along the street *Rue Gérard*.

Besides, the evaluation of the ontology consists in checking if we are able to answer competency questions of each modelet. Fig.6 represents the answers of multiple questions: coordinates of addresses along a defined street during at a given instant (here 2024) with the source of each house number.

6.2 Knowledge graph evaluation

First of all, to evaluate the knowledge graph, we tested the internal consistency of the data using SHACL rules or SPARQL queries to see if the iterative construction had generated any inconsistencies. These inconsistencies can be the description of the appearance of a landmark that occurs after its disappearance. The only detected inconsistencies deal with attribute versions having different values but whose valid time intervals overlap. In this case, the method presented in section 5.3 induces that an event occurs in $[a; b]$ where $a > b$.

The consistency of the graph with the truth on the ground is particularly difficult to assess because the truth is no longer accessible. Dictionaries cited in section 3.3 provide a basis to estimate the quality of the graph. Indeed, we can check our graph (restricted to thoroughfares) built from different sources describing the state of the territory is coherent with the truth. By taking the example of section 5 with the geometry of the street *Rue Gérard*, we deduced a change appeared between 1888 and 2024 while the truth is this change appeared on August 30, 1978. It means what we deduced is coherent with the truth, the lack of precision is only due to the lack of sources describing the neighbourhood during the XX$^{\text{th}}$ century. Eventually, comparisons say the graph does not contain many contradictions with the ground truth.

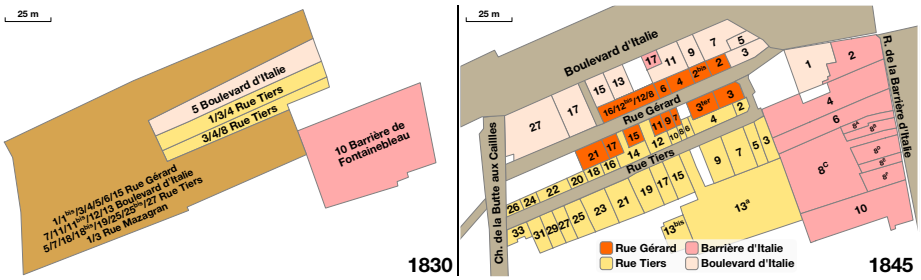

Fig. 7. Snapshots of the *Butte aux Cailles* district in 1830 and 1845 built from the graph representing the relation between plots and addresses.

In order to qualitatively assess the temporal and spatial coherence of the graph, two temporal snapshots representing the district of *Butte aux Cailles* in 1830 and 1845 were created (Fig. 7). These maps describe the addresses, roads and the associated plots. In order to produce these maps, we searched for all the landmarks that were valid in 1830 and 1845 in the sources. In the Napoleonic land registers, each plot is associated with a location (a place, a street or a full address). The locations associated with a given plot change over time for various reasons, such as road construction or operator error. In the first snapshot (1830) there are no geometries for thoroughfares, as historical sources only provide geometries from the 1840s. Some plots are associated with more than one address. In fact, the plot geometries were extracted from the 1810 cadastral index maps that were not updated until 1845 (split and merged plots have no geometric representation). In the 1845 snapshot, the data shown on the map represent a coherent urban structure of the district. The house numbers are placed along the streets and follow the rules of Parisian house numbering. Although there were no metrics, we can assume that the map represents plots with consistent addresses. Furthermore, the approximate street locations provided by the plot addresses are consistent with the 1845 snapshot.

7 Conclusion and future work

In this article, we presented an approach to construct a knowledge graph to create a multi-scale historical gazetteer from multiple and heterogeneous sources. The first contribution of this work is the PeGazUs ontology, which was proposed to model addresses, land plots, other geographical entities and their evolution, as well as sources. The second contribution is the algorithm defined to populate this ontology with data from different sources describing the Butte aux Cailles district from the end of the 18^{th} century to the present day. It enables to automatically link, temporally order and merge data fragments extracted from different sources and representing different attributes of the same real-world geographical entity. It also infers change and events from the available state data and vice-versa.

Although there was no ground truth with which to compare our graph, we were able to assess the quality of the graph qualitatively. The integration work presented in this article is restricted to a small district of Paris. The subsequent aim is to extend the spatial coverage of the graph and publish it. There are other interesting data to integrate, particularly those from directories containing a large number of addresses. In addition, some of these sources, such as very complete old address data (e.g. notaries' minute books), could be leveraged to assess the exhaustivity of the graph. Both contributions are available on the repository *https://github.com/umrlastig/pegazus-ontology*.

8 Acknowledgements

This work is supported by the French Ministry of the Armed Forces - Defence Innovation Agency (AID).

References

1. The factoid prosopography ontology, https://www.kcl.ac.uk/factoid-prosopography/ontology, last accessed 2024/07/12
2. International council on archives records in contexts ontology, https://www.ica.org/standards/RiC/ontology, last accessed 2024/07/12
3. Periodo, a gazetteer of periods for linking and visualizing data, https://perio.do/, last accessed 2024/06/24
4. Pro-o, the prov ontology, https://www.w3.org/TR/prov-o/, last accessed 2024/07/12
5. Time ontology in owl, https://www.w3.org/TR/owl-time/, last accessed 2024/06/24
6. International standards organisation. iso 19113: Geographic information — location-based services. tracking and navigation (2005)
7. Gazetteer service - application profile of the web feature service implementation spacification (2006)
8. International standards organisation. iso 19112: Geographic information — spatial referencing by geographic identifiers (2019)
9. Berman, M.L., Mostern, R., Southall, H.: Placing Names: Enriching and Integrating Gazetteers. Indiana University Press (2016). https://doi.org/10.2307/j.ctt2005zq7
10. Bernard, C., Villanova-Oliver, M., Gensel, J., Dao, H.: Modeling changes in territorial partitions over time: Ontologies tsn and tsn-change. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing. pp. 866–875 (2018). https://doi.org/10.1145/3167132.3167227
11. Bernard, C., Abadie, N., Perret, J., Duménieu, B.: Création d'un référentiel géo-historique d'adresses à partir de sources multiples. In: GAST - Gestion et l'Analyse de données Spatiales et Temporelles. Dijon, France (Jan 2024), https://hal.science/hal-04490732
12. Bol, P.K.: The china historical geographic information system (chgis). choices faced, lessons learned. In: Conference on Historical Maps and GIS. vol. 23 (2007)
13. Charles, W., Aussenac-Gilles, N., Hernandez, N.: Hht: an approach for representing temporally-evolving historical territories. In: European Semantic Web Conference. pp. 419–435. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-33455-9_25
14. Coetzee, S., Cooper, A.K., Ditsela, J.: Towards good principles for the design of a national addressing scheme. In: 25th International Cartographic Conference (ICC 2011). French Committee of Cartography, Paris, France (2011), http://hdl.handle.net/10204/5101
15. Ducatteeuw, V.: Developing an urban gazetteer: A semantic web database for humanities data. In: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities. pp. 36–39. Beijing, China (2021). https://doi.org/10.1145/3486187.3490204
16. Gregory, I.N., Bennett, C., Gilham, V.L., Southall, H.R.: The great britain historical gis project: from maps to changing human geography. The Cartographic Journal **39**(1), 37–49 (2002). https://doi.org/10.1179/caj.2002.39.1.37
17. Grossner, K., Janowicz, K., Keßler, C.: Place, period, and setting for linked data gazetteers. In: Berman, M.L., Mostern, R., Southall, H. (eds.) Placing names: Enriching and integrating gazetteers, pp. 80–96. Indiana University Press (2016). https://doi.org/10.2307/j.ctt2005zq7

18. Grossner, K., Mostern, R.: Linked places in world historical gazetteer. In: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities. pp. 40–43. Beijing, China (2021). https://doi.org/10.1145/3486187.3490203
19. Hersent, M., Abadie, N., Duménieu, B., Perret, J.: Modèles et outils pour la publication de métadonnées d'archives géographiques et de leurs données dérivées. In: Humanistica 2023, Association francophone des humanités numériques. Geneva, Switzerland (2023), https://hal.science/hal-04110787
20. Kauppinen, T., Väätäinen, J., Hyvönen, E.: Creating and using geospatial ontology time series in a semantic cultural heritage portal. In: Proceedings of the 5th European Semantic Web Conference, ESWC 2008. pp. 1–5. Tenerife, Canary Islands, Spain (June 2008). https://doi.org/10.1007/978-3-540-68234-9_11
21. Knublauch, H., Kontokostas, D.: Shapes constraint language (shacl) (2017), https://www.w3.org/TR/shacl/
22. Musen, M.A.: The protégé project: A look back and a look forward. AI matters **1**(4), 4–12 (June 2015)
23. Pasin, M., Bradley, J.: Factoid-based prosopography and computer ontologies: towards an integrated approach. Digital Scholarship in the Humanities **30**(1), 86–97 (2015). https://doi.org/10.1093/llc/fqt037
24. Perego, A., Lutz, M.: Isa programme location core vocabulary (2015), https://www.w3.org/ns/legacy_locn
25. Peroni, S.: A simplified agile methodology for ontology development. In: Dragoni, M., Poveda-Villalón, M., Jimenez-Ruiz, E. (eds.) OWL: Experiences and Directions – Reasoner Evaluation. OWLED ORE 2016. Lecture Notes in Computer Science, vol. 10161. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54627-8_5
26. Poveda-Villalón, M., Gómez-Pérez, A., Suárez-Figueroa, M.C.: Oops ! (ontology pitfall scanner !): An on-line tool for ontology evaluation. International Journal on Semantic Web and Information Systems (IJSWIS) **10**(2), 7–34 (2014). https://doi.org/10.4018/ijswis.2014040102
27. Schneider, P., Jones, J., Hiltmann, T., Kauppinen, T.: Challenge-derived design practices for a semantic gazetteer for medieval and early modern places. Semantic Web **12**(3), 493–515 (2021). https://doi.org/10.3233/SW-200394
28. Southall, H., Mostern, R., Berman, M.L.: On historical gazetteers. International Journal of Humanities and Arts Computing **5**(2), 127–145 (2011). https://doi.org/10.3366/ijhac.2011.0028
29. Tenbrink, T., Kuhn, W.: A model of spatial reference frames in language. In: Egenhofer, M., Giudice, N., Moratz, R., Worboys, M. (eds.) Spatial Information Theory, pp. 371–390. Springer Berlin Heidelberg, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23196-4_20