SURE-KG: A Knowledge Graph to Represent Real Estate and Uncertain Spatial Data from Advertisements

Lucie Cadorel[1,2], Andrea G. B. Tettamanzi[1], and Fabien Gandon[1]

[1] Université Nice Côte d'Azur, Inria, CNRS, I3S
[2] Septeo Proptech
lucie.cadorel@inria.fr, andrea.tettamanzi@univ-cotedazur.fr, fabien.gandon@inria.fr

Abstract. Real estate advertisements are a great source of information to analyze a territory and its real estate market. They are a fairly exhaustive and updated, and give a description of a property and its location. In this paper we present SURE-KG, a new knowledge graph built from a real dataset at the core of the industrial application of Septeo Proptech[3], a company operating in the real-estate domain. The knowledge graph relies on natural language processing and machine learning methods for information extraction, and semantic Web frameworks for representation and integration. It describes more than 100K real estate ads and 6K place-names extracted from French real estate advertisements from various online advertiser and located in the French Riviera. It can be exploited by real estate search engines, real estate professionals, or geographers willing to analyze local place-names. Fully in line with the open science and FAIR dynamics, the presented work is available under an open license with all the accompanying documents necessary to facilitate its reuse. The knowledge graph produced is compliant with common linked open data best practices.

Keywords: Real Estate · Spatial Data · Vague Places · Linked Data · Knowledge Graph.

1 Introduction

In recent years, many applications, on the Web, based on user-generated free text have arisen since the significant evolution in the field of the Natural Language Processing. Especially, textual data have played an important role in many geographic applications including Geographic Information Systems (GIS) enrichment or the location of events (e.g., natural disasters). These data might come from different sources such as travel blogs, social media or real estate advertisements, but they all qualitatively refer to locations. However, descriptions are often vague and uncertain (e.g., "Nearby the city center", "West of Nice,

[3] https://septeo-proptech.fr/

France") and pose a challenge to geocode places. For instance, in real estate advertisements, real estate agents often exaggerate the limits of a spatial entity in order to sell a property [5]. Therefore, exploiting real estate advertisements means dealing with uncertainty and fuzzy representation.

In this paper we present SURE-KG, a new knowledge graph built from a real dataset at the core of the industrial application of Septeo Proptech[4], a company operating in the real-estate domain. The company aims to use real estate data to give more insights to real estate agents to sell a property (e.g., by comparing a property for sale to similar ones). Septeo Proptech seeks to analyze the real estate market through the real estate advertisements. Indeed, the advertisements are a fairly exhaustive and updated source of data, and are published online which facilitates data collection (e.g., by crawling housing websites). Moreover, the property's attributes and its location are often described by the advertisers and could be extracted thanks to language models [3,**?**]. Nevertheless, these data are not structured and it might be difficult to exploit them, in particular to reason over them. Despite the ontologies and the knowledge graphs that can help to design and represent the extracted information and facilitate their interoperability, the description of the location and the environment is often uncertain and vague (e.g., "nearby the city center") and needs a suitable representation. Therefore, in this work, we propose to build a new knowledge graph that is, to the best our knowledge, the first one to represent, query and reason over uncertain and vague spatial data. Particularly, the contributions of this paper may be summarized as follows:

- we define a new ontology to represent real estate and uncertain spatial information;
- we build an extraction pipeline to process the advertisements;
- we generate a RDF dataset and publish it according to the standards and best practices of the linked open data.

The rest of this paper is organized as follows. Section 2 draws a review of and comparison with related works. In Section 3 we explain the ontology and extraction pipeline set up to process the initial corpus and generate the RDF dataset. Then, Section 4 details the characteristics of the dataset and services made available to exploit it. Section 5 illustrates the current exploitation and discusses future applications and potential impact of the dataset.

2 Related Works

The study of the Real Estate domain often involves other aspects such as Finance, Legal or Geography, and could focus on different levels (e.g., land or buildings). Previous works have shown that ontology formalization and knowledge graphs of the Real Estate domain depend on the data and use cases. In [9], the authors compare several Real Estate ontologies focusing on different aspects and levels: the land with cadastral data [10], the legal domain [12] and

[4] `https://septeo-proptech.fr/`

the transactions [11]. The ontology *proDataMarket* [13] gathers these three subdomains and studies the Real Estate market through the land and the transactions. However, this ontology is not based on up-to-date data and does not study the building and its environment. In other words, it is not possible to search for a real property for sale according to its attributes (floor size, floor level, etc.) and its location. The ontology *NAREO* [14] tries to answer one of this challenge by describing the neighborhood and the proximity to amenities to recommend a neighborhood according to location and environment criteria. Nevertheless, the authors do not represent the real property itself. Also, they only use official data such as *OpenStreetMap* and the French national institute for statistics and economic studies *INSEE* that do not contain local and vernacular places and the real estate agents' point of view on the environment (e.g., "residential", "quiet", etc.). The NAREO ontology is close to our approach but has not been populated and only focuses on one of our use cases (i.e., retrieving a place according to its proximity to facilities).

A real property is also a spatial object located in the territory and, although the exact position is not often given in the advertisements, the real estate agents describe its location mentioning spatial entities. Indeed, the location is one of the major factor in the purchasing decision. Various knowledge graphs have already incorporated spatial entities such as *DBPedia* [18], *Yago2Geo* [17], *WorldKG* [16] or *KnowWhereGraph* [15]. The spatial entities mainly come from government agencies (e.g., *INSEE* and *IGN* in France) or volunteered geographic information (VGI) such as *Wikipedia* or *OpenStreetMap*. However, their use is limited in our application since the real estate agents mention vernacular places that are not always included in these graphs (e.g., "city center"). That is why different approaches ([19],[20],[21]) have been developed to harvest the Web and extract vernacular places to enrich the gazetteers. Nevertheless, there are no standards to represent the concept of place. In [22], the authors discuss requirements to build the next generation gazetteers that should contain vernacular places. They point out that existing gazetteers such as *GeoNames*[5] have built their own collection of place names, spatial references, and ontology, that might not be suitable to maintain and extend to new places. They suggest a semi-automatic approach to include Web-mined information in gazetteers. Nevertheless, they do not mention how to deal with non named place (e.g., "city center", "pedestrian area") and vague location (e.g., "nearby the Promenade des Anglais"). In [24], the author reviews the different conceptualizations of place in vocabularies and outlines some guidelines to design a place ontology. The author mentions the plurality of the definition of a place according to the domain (e.g., place cognition vs place engineering) that makes difficult to design a single ontology. Therefore, he highlights the importance of formalizing the provenance of a place concept by citing the community that generated it. In [23], the authors describe the Linked Places format to represent a place in historical data. A place is defined by required (e.g., ids, names, geometry, when) and optional (e.g., types) but its use has limits to be applied to cognitive places. Finally, to gather Semantic Web

[5] `http://geonames.org/ontology/ontology_v3.0.rdf`

and Geospatial communities, *GeoSPARQL*[6] has been developed to represent and query the spatial data. It is not a comprehensive vocabulary for representing spatial information but it defines upper-level classes, properties and datatypes that offer a flexible way to describe spatial entities and geometries. *GeoSPARQL* is based on the *Simple Feature Access*[7] which is a set of standards that define (1) a common architecture of geometry and a representation in text (WKT), and (2) a spatial extension of SQL functions.

In summary, the overall scope of this work is to gather the representation of real estate information from advertisements as well as uncertain spatial information in one knowledge graph. To the best of our knowledge, this is the first one, in the real estate domain, combining real estate attributes, location and up-to-date data to search properties or analyze the market. Moreover, we identified a lack of formalization to represent and query vernacular or cognitive places that we adress in this work.

3 From Real Estate advertisements to Knowledge Graph

This section describes (Fig. 1) how we processed a corpus of real estate advertisements in order to extract a knowledge graph of meaningful information about real estate and its vague location, while respecting the Semantic Web standards. The result of this work is referred to as the SURE-KG dataset.

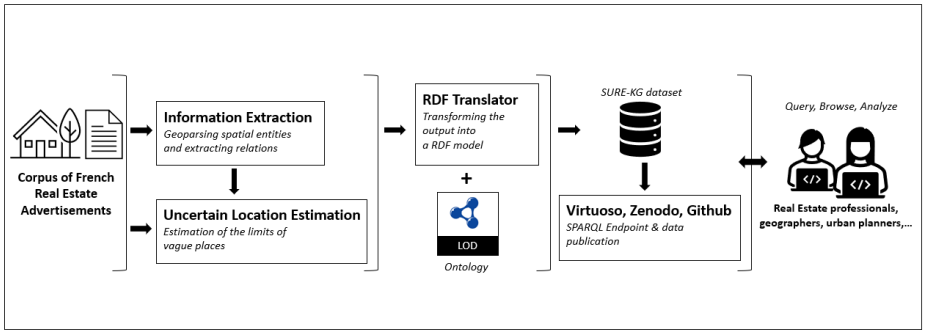

Fig. 1. SURE-KG overview: pipeline, resources and services

3.1 Initial Dataset

The initial dataset[8] is a corpus gathering real estate advertisements written in French and located in the French Riviera from various online advertisers and

[6] http://www.opengeospatial.org/standards/geosparql
[7] `https://www.ogc.org/standard/sfa/`
[8] `http://github.com/Wimmics/sure/tree/main/dataset`

provided by Septeo Proptech, our industrial partner. A real estate advertisement is mainly compound of a text describing the property and its location and is surrounding by pictures and metadata such as the price, the floor area, the city and the coordinates (latitude and longitude). We focused on the text and metadata to extract Real Estate attributes and location information (proximity to amenities, neighborhood, scenic view, etc.). We processed 100,000+ ads located in the Alpes-Maritimes, France.

3.2 Ontological Formalization: SURE

The SURE[9] (Spatial Uncertainty and Real Estate) ontology has been developed to formalize the metadata, and follows the standards and best practices of the linked open data. This ontology serves to represent the real estate, its attributes and its location as well as uncertain spatial entities and their boundaries. Listing 1 gives an example of this representation. In the following, we use the prefix *sure:* to refer to our ontology.

Real Estate: The Real Estate is the accommodation described in an advertisement through the text and metadata. We identified three kinds of information to describe a real estate :

1. Type (house, apartment, etc.);
2. Features (price, floor size, floor level, number of rooms, etc.) ;
3. Location (coordinates, city, neighborhood, proximity to the amenities, etc.).

We used the *GeoSPARQL* and *schema.org* vocabularies to formalize the real estate information. First, the classes *schema:Apartment* and *schema:House* define the type of accommodation since we only have extracted houses and apartments. Also, properties underlying these classes describe some features (*schema:floorLevel*, *schema:numberOfRooms*, etc.). Furthermore, the real estate is also a spatial object represented as a point (latitude/longitude) or located through other places (i.e., qualitative spatial relations). Thus, we have created the class *sure:RealEstate* that is a sub-class of *geosparql:Feature*. An instance of *sure:RealEstate* might have a geometry (*sf:Point*) if the coordinates are given, or be linked to other places (*sure:locatedIn*).

Place: A place is a spatial object that is often described by its name and its feature (e.g., Nice, Masséna Square, etc.). For instance, digital gazetteers mostly refer to a place thanks to its names and its attributes. However, a significant number of places are vernacular and might be compound of a spatial relation (e.g., "city center", "the old downtown", "West of Nice", "Nearby the Promenade des Anglais"). In [7] and [8], the authors define two type of places : absolute vs relative places. An absolute place is a named place (e.g., Promenade des

[9] `http://ns.inria.fr/sure`

Anglais) while a relative place is a place that needs a linguistic or spatial reasoning processes (e.g, "West of Nice", "Downtown Nice"). We proposed to follow this definition by creating two classes, *sure:AbsolutePlace* and *sure:RelativePlace*, to represent a place in the ontology. The first one represents all places (named or not) where the real estate is located "in" while the latter only describes the place compound of proximity-related relations (e.g., "nearby", "5 kilometers", "10 minutes", etc.). We added two properties to this class to define the spatial relation and its object: *sure:hasSpatialRelation* and *sure:hasAnchor*. We used *rdfs:label* to refer to the name of a place and *rdf:type* to specify its class (e.g., "neighborhood", "street", "school", etc.). In the literature, many vocabularies have been developed to represent geographic features. *GeoNames*[10] uses the *SKOS* concepts to describe upper-level classes. *GeoLinkedData*[11] defines three ontologies depending on the use (administrative, transport, hydrography) and uses existing vocabularies. Finally, the National Institute of Geographic and Forest Information (*IGN*[12]) developed its own ontology using the dataset *BDTOPO* to describe topographic and administrative entities (buildings, road network, green area, etc.). Since our application focuses on a representation of a place according to its use and its perception, while the existing vocabularies describe the spatial entity according to their nature and their topography, their use is limited. Thus, we defined two upper-classes, *sure:LocativeArea* and *sure:Amenity*, to distinguish the amenities from the place of living. Then, we chose to extract and generate classes from the ads despite they might be noisy.

Uncertain Location: A spatial object is often linked to a geometry to represent its boundary. However, the places extracted in the ads, are described in the view of the real estate agent. Thus, the agent might not have the same limits as the official ones, or exaggerate them in order to sell [5]. The location is vague and can not be represented as a single point or polygon. A classic way to overcome the vagueness is the use of the fuzzy set theory ([32], [33]) that we used to approximate a vague place.

In the fuzzy set theory, a fuzzy subset A of a set E is defined by a function called membership function μ_A. The function gives the degree of membership to the set A for each element x of E. The degree is often ranged between 0 and 1. If $\mu_A(x) = 1$, then x completely belongs to A while $\mu_A(x) = 0$, then x does not belong to A. We applied this theory to the places to capture the uncertainty of its location by computing the membership degree for each point in the space. Also, we could retrieve crisp sets using alpha-cuts. An alpha-cut $\tilde{A}_\alpha$ is a crisp subset where each element has a membership degree greater than α.

$$\tilde{A}_\alpha = \left\{x \in A; \mu_{\tilde{A}}(x) \geq \alpha\right\}.$$

[10] http://geonames.org/ontology/ontology_v3.0.rdf
[11] http://geo.linkeddata.es
[12] http://data.ign.fr/def/topo/20190212.htm

The core and the support are specific α-cuts where α is respectively equal to 1 and 0 :

$$cor(A) = \{x \in A; \mu_A(x) = 1\},$$
$$supp(A) = \{x \in A; \mu_A(x) > 0\}.$$

In the ontology, we have represented the geometry of a place as a collection of alpha-cut. We defined *sure:AlphaCut* as a subclass of *geosparql:Geometry*, and its property *sure:hasAlpha* to set the membership degree. Then, *GeoSPARQL* allows to associate a collection of geometries to the same object. Thus, a place could have several alpha-cuts in order to represent as reliable as possible its uncertain boundaries.

3.3 Generation Pipeline

The ontological formalization helped us to decide what kind of information should be extracted and processed from the advertisements. Our pipeline involves two main steps: (1) process each document of the corpus to extract information and, in particular, spatial information ; (2) estimate the uncertain boundary of each place extracted. Finally, we have translated the output of both treatments into a unified and consistent RDF dataset.

Spatial Information Extraction: Geoparsing is the task to detect geographic terms from text and has been widely used in various types of texts such as travel blogs [25], social media in emergencies [26,27], housing advertisements [21], or fictional novels [28]. This method is often a subtask of Named Entity Recognition (NER) applied to geographic entities. However, most of the geoparsers are designed to only detect Toponyms in the English language. Hence, we specifically designed a model to extract spatial information in real estate advertisements that has been described and evaluated in our previous work [30]. This method is a two-stage pipeline involving Named Entity Recognition and Relationship Extraction. The Named Entity Recognition model architecture is a *BiLSTM+CRF model* combined with a text embedding. The Relationship Extraction is based on Dependency Parsing methods. Both have been trained from scratch and evaluated on a corpus of French Real Estate advertisements written in French and located in the French Riviera. The NER method detects 4 type of entities to better capture the spatial information in a Real Estate advertisement:

- **Feature**: entity representing the type of a place (e.g., natural features, constructions and subdivisions of land).
- **Toponym**: entity referring to the proper name of a place (also known as place name or geographic name)
- **Spatial Relation entity**: entity describing a proximity-related spatial relation between two places (e.g., "nearby", "5 minutes away")

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix geosparql: <http://www.opengis.net/ont/geosparql#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix schema: <http://schema.org/> .
@prefix : <http://ns.inria.fr/sure#> .

############# Classes and Properties
:RealEstate rdfs:subClassOf geosparql:Feature.
:Amenity rdfs:subClassOf geosparql:Feature.
:TrainStation rdfs:subClassOf :Amenity.
:AbsolutePlace rdfs:subClassOf geosparql:Feature.
:RelativePlace rdfs:subClassOf geosparql:Feature.
:AlphaCut rdfs:subClassOf geosparql:Geometry.

:hasAlpha a rdfs:Property ;
   rdfs:domain :AlphaCut ;
   rdfs:range xsd:double .

:hasAnchor a rdfs:Property ;
   rdfs:domain :RelativePlace ;
   rdfs:range geo:Feature.

:hasSpatialRelation a rdfs:Property ;
   rdfs:domain :RelativePlace ;
   rdfs:range xsd:string.

############# Instances
:RealEstate1 a schema:House, :RealEstate ;
   schema:floorSize "200"^^xsd:double;
   schema:hasPrice "975000"^^xsd:double;
   schema:numberOfRooms "6"^^xsd:double;
   sure:locatedIn :Proche_Gare_Riquier ;
   :hasDescription :Text1 .

:Gare_Riquier a :TrainStation, :Place;
  rdfs:label "riquier"@fr.

:Proche_Gare_Riquier a :RelativePlace;
  :hasAnchor :Gare_Riquier;
  :hasSpatialRelation "proche";
  geosparql:hasGeometry :AlphaCut1, :AlphaCut2.

:AlphaCut1 a :AlphaCut ;:hasAlpha "0.5"^^xsd:double;
    geosparql:asWKT "MULTIPOLYGON (((43.6957 7.280889,..., 43.69578 7.280882)))"^^geosparql:
     wktLiteral.

:AlphaCut2 a :AlphaCut ;:hasAlpha "0.8"^^xsd:double;
    geosparql:asWKT "MULTIPOLYGON (((43.695 7.2808,..., 43.6955 7.280892)))"^^geo:wktLiteral
     .
```

Listing 1: Example of the RDF representation of real estate information and vague places.

– **Mode of transportation**: entity referring to the travel mode between two places (e.g., "walk")

Also, the model gives a structured knowledge by reconstructing relations such as the spatial relation (i.e., relations between a Spatial Relation entity and a Toponym or Feature) or the nature of a place (i.e., relation between a feature and

a Toponym). Afterwards, we post-processed the output to clean data from misspelling, plural and abbreviations. We replaced the well-known abbreviations by its correct terms (e.g., "min" for "minutes", "m" for "meters"). We also applied the Jaro-Winkler distance to retrieve very similar terms and correct misspeling (threshold 0.9 and for toponym 0.95 because more possibilities). Lastly, we retrieved the most obvious hyponyms for the Feature entities by applying a simple term inclusion heuristic. Table 1 shows shows general statistics about the spatial information extraction and post-processing output. We can see that the post-processing part significantly reduced the number of unique entities and increased the number of times each entity is mentioned in the advertisements.

Nb of ads processed	102,335
Nb of ads with spatial information extracted ≥ 1	80,200
Median of nb of spatial information extracted by ad	3
Maximum of nb of spatial information extracted by ad	53

Type of entity	Nb of entity before post-processing	Nb of entity after post-processing	Median of the count of each entity before post-processing	Median of the count of each entity after post-processing
Feature	4,212	486	2	35
Toponym	11,084	5,501	2	2
Spatial Relation	491	50	2	113
Mode of Transportation	14	10	8	4

Table 1. Statistics about (1) the spatial information extraction and (2) post processing

Uncertain Location Estimation The second stage of the pipeline consists in the estimation of the limits of each place. Since we have extracted a significant number of vernacular places that could not be retrieved in official database, and the real estate agents might exaggerate the limit of a place, we decided to create our own knowledge from the geolocated advertisements. We used the Kernel Density Estimation method, which is a non-parametric estimation method that infers the shape of a variable from a sample, and gives a probability (density) for each point of the support. In our study, we chose Gaussian kernels to approximate the boundary of a spatial object, mainly because they are well-supported by existing libraries and Gaussian membership functions are a popular choice for fuzzy sets. For each extracted place, we selected all geolocated ads mentioning it, removed outliers and estimated its footprint based on the advertisements' coordinates. In order to get a reliable estimation, we only applied the method for each place with a minimum of 10 ads mentioning it. Finally, the method returns a density that could be easily transformed into a membership function of a fuzzy set. This method has been implemented and evaluated in our previous work [35] where we used it to retrieve a location estimation of real estate advertisements that are not or wrongly geocoded.

Fig. 2 shows an example of the fuzzy representation of the uncertain place *Downtown* in 4 large cities in the Alpes-Maritimes. *Downtown* is a very vague place since it does not exist any official limits, and everyone has is own opinion on where it starts and where it ends. A high number of studies in GIS have been conducted to automatically estimate the position of *Downtown* in different cities ([34]). Our method and dataset give a partial answer of what and where is *Downtown* in Antibes, Cannes, Cagnes-sur-mer and Nice according to the real estate agents. We can see that the downtown in Nice and Cagnes-sur-mer is pretty well defined around a main avenue or square. On the other hand, Antibes and Cannes seem to have two downtowns since both have two neighborhoods far from the center that might be considered as smaller cities (Juan-les-Pins and Cannes La Bocca). Overall, our method seems to give an accurate representation for these 4 cities.

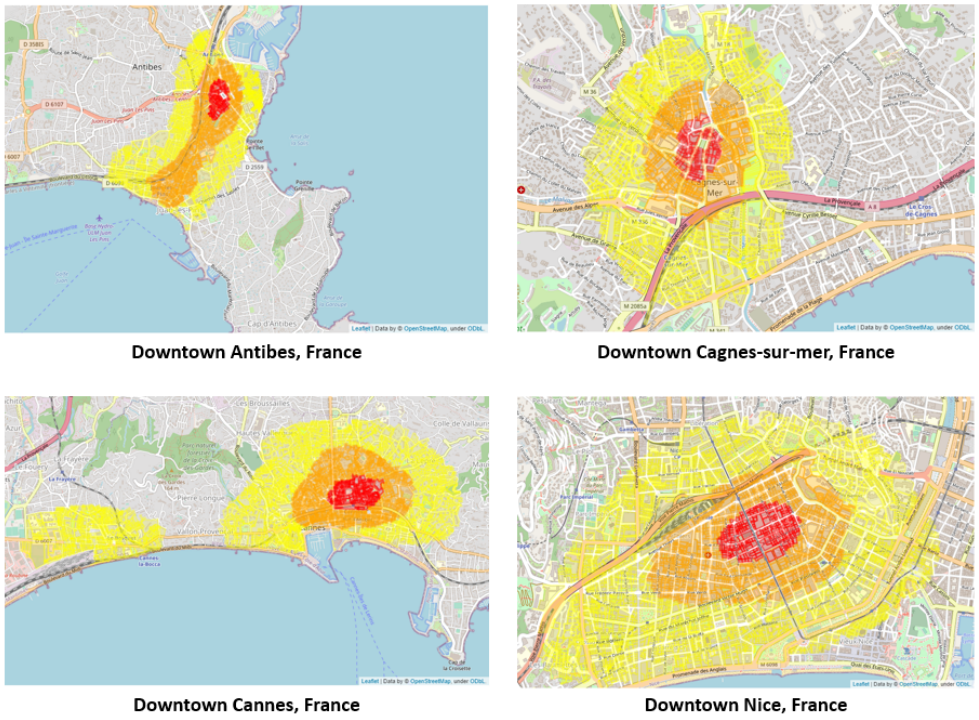

Fig. 2. Examples of the estimation of three alpha-cuts (yellow : $\alpha > 0$, orange: $\alpha > 0.5$, red $\alpha = 1$) of *Downtown* in Antibes, Cagnes-sur-mer, Cannes and Nice, France.

RDF Generation The final stage of the pipeline is the translation of both treatments into a RDF model. We created a script[13], available on our Github repository, using the python library RDFLib[14]. We defined a template of triples about the real estate and spatial information. We also linked the cities mentionned in the metadata of the real estate ads to Geonames using the library Geocoder[15] and Geonames' attributes (e.g., feature class). Among 281 unique cities, we found 182 Geonames related entities that we linked with the predicate *owl:sameAs*. Finally, we added the named entity annotations given by the NER model to the knowledge graph to propose a reusable dataset. We chose the Web Annotation Data Model vocabulary to annotate the text as presented in Listing 2). The annotation points to the annotated ad and the text position (the target), and the named entity category (the body). We also give a confidence score of the extraction (sure:confidence) provided by the Named Entity Recognition model.

```
@prefix dc: <http://purl.org/dc/elements/1.1> .
@prefix oa: <http://www.w3.org/ns/oa#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema: <http://schema.org/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix : <https://ns.inria.fr/sure#> .

############# Instances
:RealEstate1 :hasDescription :Text1

:Text1 dc:description "Biot : dans domaine ferme a proximite du village [...].";
   dc:language "fr" .

:Entity0 a oa:Annotation ;
   oa:hasBody :Toponym ;
   oa:hasTarget [oa:hasSource sure:Text1],
       [a oa:TextPositionSelector ;
          oa:end "0"^^xsd:double ;
          oa:start "0"^^xsd:double] ;
   oa:motivatedBy oa:classifying ;
   :confidence "0.99"^^xsd:double.

:Entity1 a oa:Annotation ;
   oa:hasBody :SpatialRelation ;
   oa:hasTarget [oa:hasSource sure:Text1],
       [a oa:TextPositionSelector ;
          oa:end "6"^^xsd:double ;
          oa:start "6"^^xsd:double ];
   oa:motivatedBy oa:classifying ;
   :confidence "0.99"^^xsd:double.
```

Listing 2: Example of RDF representation of annotations of Toponym and Spatial Relation.

[13] https://github.com/Wimmics/sure/tree/main/src/GraphGeneration
[14] http://rdflib.readthedocs.io
[15] http://geocoder.readthedocs.io/

4 Knowledge graph and resulting linked dataset

The SURE-KG dataset is a RDF graph that provides an RDF representation of Real Estate ads and the spatial information automatically derived from the textual data and metadata. It contains more than 7M triples, 100K Real Estate ads and 6K places. Table 2 reports some statistics about class and property instances.

Dataset Description. In line with best practices [30], the dataset comes with a thorough self-description, comprising licensing, authorship and provenance information, used vocabularies, interlinking and access information, described with Dublin Core Metadata Information, DCAT and VOID.

Dataset Accessibility. The dataset is made available by means of a DOI identified RDF dump downloadable from Zenodo, and a public SPARQL endpoint. A Github repository provides a comprehensive documentation, source codes and query templates. The ontology has been published following the standards and best practices of the linked open data. This information is summarized in Table 3.

Reproducibility In compliance with the open science principles, all the scripts and files involved in the pipeline are provided in the project's Github repository under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, so that anyone may rerun the whole processing pipeline.

Dataset Licensing The SURE-KG RDF Knowledge Graph is published under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License[16]. In particular, this license allows anyone to use the dataset for a non-commercial purpose since the data come from an industrial project.

Sustainability Plan. In the short term, we plan to apply the pipeline to all the regions of France. This will be the opportunity to assess the quality of the method on other data and areas. Furthermore, we would like to carry on an evaluation of the ontology with experts of the real estate domain by answering competency questions. In the middle and long term, we intend to improve the pipeline (e.g., linking spatial information to other gazetteers) and to fit it to other language (e.g., English). Furthermore, we have deployed a server to host the SPARQL endpoint that benefits from a high availability infrastructure and 24/7 support.

5 Potential Impact and Reusability

Target audiences and expected uses. We adopted a user-centered approach to design this project. The industrial partner (Septeo Proptech) and geographers were closely involved to help us to identify motivating scenarios and potential users.

Scenario 1: Helping potential buyers to find a Real Estate property. In this scenario, we aim to retrieve information about a property (e.g., price, number of

[16] `https://creativecommons.org/licenses/by-nc-sa/4.0/`

Class URI	nb of instances
http://ns.inria.fr/sure#RealEstate	102,335
http://ns.inria.fr/sure#AbsolutePlace	3,760
http://ns.inria.fr/sure#RelativePlace	2,273
http://ns.inria.fr/sure#AlphaCut	20,530
http://ns.inria.fr/sure#LocativeArea	1,904
http://ns.inria.fr/sure#Amenity	77
http://ns.inria.fr/sure#Quartier	193

Property URI	nb of instances
http://www.w3.org/2000/01/rdf-schema#label	1,736
http://www.opengis.net/ont/geosparql#hasGeometry	122,865
http://ns.inria.fr/sure#hasAnchor	2,273
http://ns.inria.fr/sure#hasSpatialRelation	2,273
http://ns.inria.fr/sure#hasDescription	81,911
http://ns.inria.fr/sure#locatedIn	330,654

Table 2. Selected statistics on typical properties and classes.

Dataset DOI	10.5281/zenodo.7885757
Downloadable RDF dump	`https://doi.org/10.5281/zenodo.7885757`
Public SPARQL endpoint	`http://erebe-vm2.i3s.unice.fr:5000/sparql/`
Source Code and Documentation	`http://github.com/Wimmics/sure`
URIs Namespace	`http://ns.inria.fr/sure#`
Dataset URI	`http://ns.inria.fr/sure/data/`
Citation	Lucie CADOREL, Fabien GANDON, Andrea G. B. TETTAMANZI, 2023. SURE-KG dataset. `https://doi.org/10.5281/zenodo.7885757`

Table 3. Dataset availability.

rooms, floor size, etc.) and its location (e.g., neighborhood, proximity to the amenities, quiet environment, etc.) in order to help a potential buyer to take a decision.

Scenario 2: Helping real estate professional to analyse the real estate market A real estate agent might need to understand the market and to know the real properties sold/for sale to align the real property he has to sell. He needs to know the similar properties, the mean price, the number of sells in the neighborhood, etc.

Scenario 3: Helping geographers and urban planners to analyse the territory. The real estate agents have a good knowledge of the territory and give its description through the advertisements that could be used by geographers or urban planners. For instance, they could study the social representations to understand which part of a territory better suits to one type of population than another. Urban planners could also analyzed how real estate agents mention the amenities (e.g., transports, schools, shops, etc.) to highlight a lack of services in a neighboorhood.

Scenario 4: Helping NLP researchers to test their model. In this scenario, we aim to give annotated textual data to help NLP researchers to test models to retrieve geographic named entities.

Current Use. The processing pipeline and the dataset are used in the context of the industrial need of the company. The *scenario 2* is one of the use case that the company faces on to provide a tool to the real estate agents. On the oher hand, ongoing research works are conducted by geographers from the Unversity Nice Côte d'Azur about the social representation of the city of Nice according to the real estate agents [31]. Finally, in [35], we implemented a method to retrieve the location estimation of real estate advertisements, since around 80% of the advertisements are not or wrongly geocoded (i.e., the given coordinates correspond to the city center). We performed information fusion using the fuzzy boundaries of the spatial information found in the text (e.g., "Promenade des Anglais", "near the beach"). Fig.3 provides an example of the result of the information fusion of the three spatial information extracted from the text. We can see that the exact coordinates of the property match with a parcel with a high degree of membership.

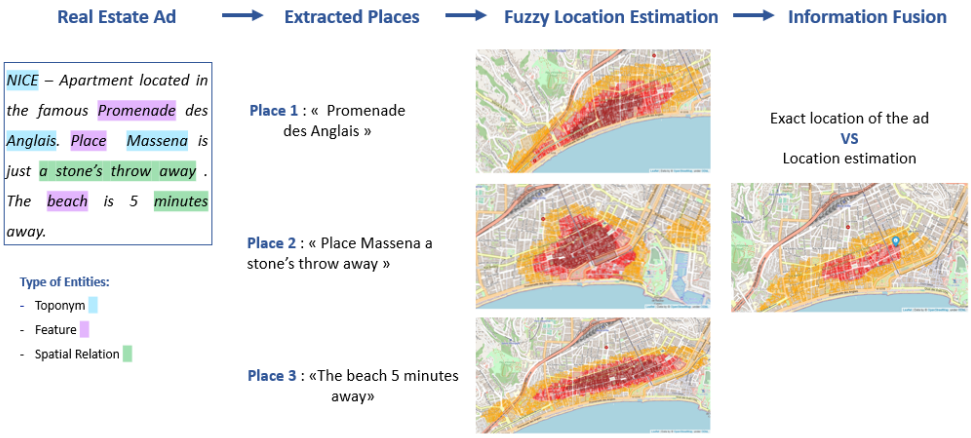

Fig. 3. Example of information fusion to retrieve the location of a real property

Interest of communities in using the dataset. Beyond the real estate domain, the dataset could be used by the Semantic Web community in works and experimentation with uncertain and spatial data. For instance, GeoSPARQL implements topological relations (e.g., union, intersection, overlaps, etc.) for crisp geometries, but it could be extended to fuzzy geometries [36]. Furthermore, the GIS community could be interested in this dataset to enrich gazetteers with vernacular and cognitive places, or to study qualitative spatial relations, such as "near", according to their context [37].

Potential for reuse. To the best of our knowledge, the SURE-KG dataset is the first one integrating uncertain and vague spatial data in a knowledge graph that could serve as benchmarking spatial algorithms. We provide a documentation on the Github repository including a demonstration notebook to help users to query the graph. For a potential wider application, the processing pipeline do not require any adaptation for the extraction of spatial information from French texts, but might need small adjustments to extract metadata (e.g., price, floor size, etc.). For the adaptation to the other languages, a more substantial work may be required to train a model on the data.

6 Conclusion and Perspectives

In this paper, we described the SURE-KG dataset: a new knowledge graph to represent and query real estate and uncertain and vague spatial information from advertisements, and used in an industrial application with our partner, Septeo Proptech. To the best of our knowledge, this is the first application that integrates uncertain and vague spatial data in a knowledge graph. We first proposed the SURE ontology to design and formalize the need to represent real estate and uncertain spatial data. Then, we presented the generation pipeline to (1) extract spatial information, (2) approximate location of vague places and (3) translate the output to a RDF dataset. We published and made available the knowledge graph by means of a DOI identified RDF dump downloadable from Zenodo and a SPARQL endpoint in order to easily query the dataset. Through interactions with our industrial partner but also geographic researchers, we are ensuring that our approach is guided by and aligned with the actual needs of potential users from different domains. Moreover, we have shown that our dataset is already involved in ongoing works conducted by geographers from the Unversity Nice Côte d'Azur. Finally, great care has been taken to produce a dataset and software that meet the open and reproducible science goals and the FAIR principles.

Several directions could be considered to expand this work. First, we only have applied our pipeline on advertisements located in the Alpes-Maritimes. We would like to apply it to all the regions of France. It will give us the opportunity to evaluate the method's ability to adapt to new data. Second, we would like to carry on an evaluation of the ontology. We plan to design competency questions to test with experts of the real estate domain. Last, we have not yet linked place-names extracted from the advertisements, except the cities, to entities from other datasets (e.g., Geonames, DBPedia, etc.). Linking our place-names to other datasets could enrich gazetteers and be used to compare official boundaries to cognitive ones.

References

1. Uschold, Mike, and Michael Gruninger. "Ontologies: Principles, methods and applications." The knowledge engineering review 11.2 (1996): 93-136.

2. Bosvieux, Jean. "L'immobilier, poids lourd de l'économie." Constructif 1 (2018): 10-14.
3. Bekoulis, Ioannis, et al. "Reconstructing the house from the ad: Structured prediction on real estate classifieds." EACL2017, the 15th Conference on the European Chapter of the Association for Computational Linguistics. 2017.
4. Zadeh, Lotfi A. "Fuzzy sets." Inform Control 8 (1965): 338-353.
5. McKenzie, Grant, and Yingjie Hu. "The "Nearby" exaggeration in real estate." Proceedings of the Cognitive Scales of Spatial Information Workshop (CoSSI 2017), L'Aquila, Italy. 2017.
6. Bennett, Brandon, and Pragya Agarwal. "Semantic categories underlying the meaning of 'place'." Spatial Information Theory: 8th International Conference, COSIT 2007, Melbourne, Australiia, September 19-23, 2007. Proceedings 8. Springer Berlin Heidelberg, 2007.
7. Lesbegueries, Julien, Christian Sallaberry, and Mauro Gaio. "Associating spatial patterns to text-units for summarizing geographic information." ACM SIGIR 2006. GIR, Geographic Information Retrieval, Workshop. 2006.
8. Syed, Mehtab Alam, et al. "GeoXTag: Relative spatial information extraction and tagging of unstructured text." AGILE: GIScience Series 3 (2022): 16.
9. Shi, Ling, and Dumitru Roman. "Ontologies for the real property domain." Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics. 2018.
10. Sladić, Dubravka, et al. "Ontology for real estate cadastre." Survey Review 45.332 (2013): 357-371.
11. Stubkjaer, Erik. The ontology and modelling of real estate transactions. Routledge, 2017.
12. Paasch, Jesper M. "Legal Cadastral Domain Model: An Object-orientated Approach." Nordic Journal of Surveying and Real Estate Research 2.1 (2005): 117-136.
13. Shi, Ling, et al. "The prodatamarket ontology for publishing and integrating cross-domain real property data." journal" Territorio Italia. Land Administration, Cadastre and Real Estate 2 (2017).
14. Laddada, Wissame, et al. "Ontology-based approach for neighborhood and real estate recommendations." Proceedings of the 4th ACM SIGSPATIAL Workshop on Location-Based Recommendations, Geosocial Networks, and Geoadvertising. 2020.
15. Janowicz, Krzysztof, et al. "Know, Know Where, KnowWhereGraph: A densely connected, cross-domain knowledge graph and geo-enrichment service stack for applications in environmental intelligence." AI Magazine 43.1 (2022): 30-39.
16. Dsouza, Alishiba, et al. "Worldkg: A world-scale geographic knowledge graph." Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021.
17. Karalis, Nikolaos, Georgios Mandilaras, and Manolis Koubarakis. "Extending the YAGO2 knowledge graph with precise geospatial knowledge." The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part II 18. Springer International Publishing, 2019.
18. Auer, Sören, et al. "Dbpedia: A nucleus for a web of open data." The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings. Springer Berlin Heidelberg, 2007.
19. Jones, Christopher B., et al. "Modelling vague places with knowledge from the Web." International Journal of Geographical Information Science 22.10 (2008): 1045-1065.

20. Grothe, Christian, and Jochen Schaab. "An evaluation of kernel density estimation and support vector machines for automated generation of footprints for imprecise regions from geotags." International Workshop of Computational Models of Place (PLACE'08). Park City, Utah, USA. 2008.
21. Hu, Yingjie, Huina Mao, and Grant McKenzie. "A natural language processing and geospatial clustering framework for harvesting local place names from geotagged housing advertisements." International Journal of Geographical Information Science 33.4 (2019): 714-738.
22. Keßler, Carsten, Krzysztof Janowicz, and Mohamed Bishr. "An agenda for the next generation gazetteer: Geographic information contribution and retrieval." Proceedings of the 17th ACM SIGSPATIAL international conference on advances in Geographic Information Systems. 2009.
23. Grossner, Karl, and Ruth Mostern. "Linked places in world historical gazetteer." Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities. 2021.
24. Ballatore, Andrea. "Prolegomena for an ontology of place." Advancing geographic information science (2016): 91-103.
25. Adams, Benjamin, and Krzysztof Janowicz. "On the geo-indicativeness of non-georeferenced text." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 6. No. 1. 2012.
26. Grace, Rob. "Toponym usage in social media in emergencies." International Journal of Disaster Risk Reduction 52 (2021): 101923.
27. Hu, Yingjie and Wang, Jimin. "How Do People Describe Locations During a Natural Disaster: An Analysis of Tweets from Hurricane Harvey." In K. Janowicz and J. A. Verstegen, editors, the 11th International Conference on Geographic Information Science (GIScience 2021) - Part I, volume 177, pages 6:1-6:16, Dagstuhl, Germany. 2021.
28. Moncla, Ludovic, et al. "Automated geoparsing of paris street names in 19th century novels." Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities. 2017.
29. Cadorel, Lucie, Alicia Blanchi, and Andrea GB Tettamanzi. "Geospatial Knowledge in Housing Advertisements: Capturing and Extracting Spatial Information from Text." Proceedings of the 11th on Knowledge Capture Conference. 2021.
30. Lóscio, B. F., C. Burle, and N. Calegari. "Data on the Web Best Practices, W3C recommendation, 2012." (2018).
31. Blanchi, Alicia, et al. "Studying Urban Space from Textual Data: Toward a Methodological Protocol to Extract Geographic Knowledge from Real Estate Ads." Computational Science and Its Applications–ICCSA 2022 Workshops: Malaga, Spain, July 4–7, 2022, Proceedings, Part II. Cham: Springer International Publishing, 2022.
32. Schneider, Markus. "Uncertainty management for spatial datain databases: Fuzzy spatial data types." Advances in Spatial Databases: 6th International Symposium, SSD'99 Hong Kong, China, July 20—23, 1999 Proceedings 6. Springer Berlin Heidelberg, 1999.
33. Bunel, Mattia, Ana-Maria Olteanu-Raimond, and Cécile Duchêne. "Référencement spatial indirect: modélisation à base de relations et d'objets spatiaux vagues." Sageo 2018. 2018.
34. Montello, Daniel R., et al. "Where's downtown?: Behavioral methods for determining referents of vague spatial queries." Spatial cognition and computation. Psychology Press, 2017. 185-204.

35. Cadorel, Lucie, Denis Overal, and Andrea GB Tettamanzi. "Fuzzy representation of vague spatial descriptions in real estate advertisements." Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising. 2022.
36. Schockaert, Steven, Martine De Cock, and Etienne E. Kerre. Reasoning about fuzzy temporal and spatial information from the web. Vol. 3. World Scientific, 2011.
37. Aflaki, Niloofar, et al. "What Do You Mean You're in Trafalgar Square? Comparing Distance Thresholds for Geospatial Prepositions." 15th International Conference on Spatial Information Theory (COSIT 2022). Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.