An Adaptive Approach for Interlinking Georeferenced Data

Abdelfettah Feliachi, Nathalie Abadie, Fayçal Hamdi

An Adaptive Approach for Interlinking Georeferenced Data

Abdelfettah Feliachi, Nathalie Abadie, Fayçal Hamdi

An Adaptive Approach for Interlinking Georeferenced Data

Abdelfettah Feliachi
Univ. Paris-Est, LASTIG COGIT, IGN, ENSG
Saint-Mande, France
Abdelfettah.Feliachi@ign.fr

Nathalie Abadie
Univ. Paris-Est, LASTIG COGIT, IGN, ENSG
Saint-Mande, France
Nathalie-F.Abadie@ign.fr

Fayçal Hamdi
CEDRIC, CNAM
Paris, France
Faycal.Hamdi@cnam.fr

ABSTRACT

The resources published on the Web of data are often described by spatial references such as coordinates. The common data linking approaches are mainly based on the hypothesis that spatially close resources are more likely to represent the same thing. However, this assumption is valid only when the spatial references that are compared have been produced with the same positional accuracy, and when they actually represent the same spatial characteristic of the resources captured in an unambiguous way. Otherwise, spatial distance-based matching algorithms may produce erroneous links. In this article, we first suggest to formalize and acquire the knowledge about the spatial references, namely their positional accuracy, their geometric modeling, their level of detail, and the vagueness of the spatial entities they represent. We then propose an interlinking approach that dynamically adapts the way spatial references are compared, based on this knowledge.

CCS CONCEPTS

• **Information systems → Semantic web description languages**; *Geographic information systems*; *Data extraction and integration*; *Web data description languages*;

KEYWORDS

Spatial references, instances matching, linked data

ACM Reference Format:
Abdelfettah Feliachi, Nathalie Abadie, and Fayçal Hamdi. 2017. An Adaptive Approach for Interlinking Georeferenced Data. In *K-CAP 2017: K-CAP 2017: Knowledge Capture Conference, December 4–6, 2017, Austin, TX, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3148011.3148025

1 INTRODUCTION

Interlinking data on the Web is a crucial step in the data publication process. This step aims to identify and create links between resources which represent the same real-world entity, or are related to each other by some kind of relationship. Many of the existing approaches, used for data linking on the Web, are mainly inspired by previous works on data integration and entity resolution. The data matching subtask is thus generally performed by comparing the values of similar properties used by resources from heterogeneous data sources for describing real-world entities in order to estimate the degree of similarity between these resources. The higher the similarity score is between two resources, the more they are likely to represent the same real-world entity [12]. Tools like Silk[1] [16] or LIMES[2] [22] implement such approaches. They allow one to compare the property values that describe the resources, by means of various distance measures. For each property comparison, a confidence value is computed based on the distance value and parameters defined by a data linking expert. Then, a function is applied to aggregate these confidence values into one value, used to decide whether to create a link or not for this pair of resources. This is a typical multicriteria decision problem.

In the Web of data, many resources are associated, via a spatial reference, to a location in the geographic space. Spatial references may be direct, such as geographic coordinates or geometric primitives (points, linestrings or polygons), or indirect such as postal addresses or names of administrative units. Like any other property, spatial references can be used to evaluate the similarity of resources in a data matching process. In the field of geographic data matching, many measures have been proposed to evaluate the similarity of geometries[3]. They are progressively implemented in data linking tools. However, they have been designed for traditional geographic databases matching and may not provide good results when directly reused for georeferenced resources of the Web. The open nature of Web of data sources, mainly produced by crowdsourcing, poses indeed new challenges for geometric similarity evaluation due to the heterogeneity of the geometric quality within and between data sources .

In this work we follow the intuition that improving the spatial data matching results requires one to adapt each pair of geometries similarity evaluation to the characteristics of the tested geometries. The remainder of this paper is organized as follows: section 2 presents related works about geometry similarity evaluation on the Web of data. In section 3, we present a vocabulary to describe geometry characteristics that must be taken into account for geometry comparisons and in section 4 we detail our adaptive geometry similarity evaluation approach. Section 5 details the experiments that we carried out to validate our approach.

2 COMPARING GEOMETRIES ON THE WEB OF DATA

Previous works in the field of geographic data matching have focused on approaches for evaluating the similarity of geographic features, mainly based on the evaluation of their geometries similarity. As two geographic feature located far from each other in space

[1] http://silkframework.org/
[2] http://aksw.org/Projects/LIMES.html
[3] For the sake of brevity, we will use the term "geometries" instead of "direct spatial references" throughout this article.

are not likely to represent the same real world phenomenon, geometry similarity prevails indeed over any other properties similarity to decide whether two geographic features should be matched or not [2]. In this section, we present the approaches proposed in the field of geographic data matching for geometric similarity evaluation in order to adapt them for georeferenced resources linking.

2.1 Geometry similarity measures for geographic data matching

Geographic databases are created through a process of abstraction of real world phenomena. Geometries are used to provide a quantitative description of spatial characteristics of real world entities, such as their dimension, position, size, shape, and orientation [25]. Due to the quality of raw data sources and the discrete nature of the geometrical primitives in spatial databases, the spatial characteristics of real world entities can be captured only in a simplified way. This has an impact on the quality and the data capture rules of the resulting geometries [13].

Geometry similarity measures designed for geographic databases matching are all based on one or more geometry similarity function(s), that evaluate(s) geometry similarity with respect to some particular descriptor [18]. These functions are chosen depending on the types of the geometries and the criteria with respect to which they are compared: distance functions based on euclidean, orthodromic or elliptic curve spatial distances, like the min distance function, deal with location of points sets [30]; boolean functions based on the inclusion of the evaluated geometry in some buffer built around a given geometry are used in [34], [14] and [32] for comparing geometry locations; the surface distance is used by [5] to compute the similarity between polygons with regards to their location and the area of their overlapping surfaces; [6] and [3] propose two measures for comparing polygon shapes, respectively based on distances and angles values; the Hausdorff and Fréchet distances, which deal with both location and shape of linestrings, are widely used for linestring and polygon similarity evaluation (see for example [26], [7]); [24] uses a function for comparing linestrings orientations and [35] functions for comparing polygons orientations, area and length; [7], [24] also use functions for comparing geometries neighborhoods (i.e. their topologically related or spatially close geometries).

All geometry similarity measures use at least one location-based similarity function. This function can be combined with functions based on other descriptors. To that end, they are standardized to values between 0 and 1 by means of various normalization functions and (eventually weighted) aggregation methods are used to compute an overall standardized geometry similarity value [18].

Parameters such as buffer size or normalization function thresholds are used to define to what extent differences between geometries with regards to some given descriptor are considered acceptable. Setting such parameters is usually assigned to experts, who define their values based on their knowledge about the databases to be matched and the functions behaviors. Parameters related to the evaluation of location-based geometry similarity are the most intuitive. Most of the time, they represent the maximum acceptable spatial distance between two geometries for them being considered as potentially representing the same real world entity; above this value, geometries are considered too far from other to represent the same thing. This parameter is thus closely related to the absolute positional planimetric accuracy of the databases, defined by the ISO 19157 standard as the "closeness of reported coordinate values to values accepted as or being true". It may also be affected by geometry capture rules or geographic feature boundary vagueness. For example, a postal address represented by a point might be captured in various ways: within the extent of the building located by the address, at the entrance of the building, on the centerline of the street in front of the building, etc. [24] details what information is needed to configure confidence functions used for geometry similarity evaluation based on location, neighborhood and orientation.

Most of geometry similarity measures designed for complex geometries such as linestrings and polygons combine several geometric similarity functions [24], [35]. This is usually done to overcome some data integration conflicts due to differences in the levels of detail of the datasets, i.e. the degree of geometric and semantic abstraction used for representing real world entities in these geographic datasets [27]. Geographic databases integration conflicts caused by differences of level of detail have been thoroughly described by [8]. In [33] and [21], the road networks to be matched have different levels of granularity, i.e. the road segments are more detailed in one of the databases than in the other. This conflict is solved by the similarity functions based on the spatial neighborhood of road edges and nodes.

Some approaches also apply geometry transformation operations before computing geometry similarity in order to reduce the gap between the geometries to be compared. For example, [34] performs conflation to lower the location differences between the road segments to be matched due to each database positional planimetric accuracy. [36] uses generalisation algorithms to harmonise the levels of detail of the compared databases.

Many approaches have thus been propose to compute geometry similarity for geographic databases matching. Some of them are progressively introduced for georeferenced resource linking.

2.2 The heterogeneity of geometries on the Web of data

Unlike geometries of geographic databases, geometries used on the Web of data are not the main piece of information of the resources they describe. They are not necessarily available in the description of the resources. Besides, spatial data on the Web may have various origins: they may be extracted or transformed from geographic databases provided by traditional data producers such as national mapping agencies (e.g. Ordnance Survey data, geo.linkeddata.es), but they may also be a fusion of many data sources as they can be produced by crowdsourcing (e.g. DBpedia[4], Geonames[5]). In such cases, inside a same data source, spatial references may have been captured differently which leads to what we call "internal heterogeneity".

Standardization efforts, specifically the recommendation of the OGC/W3C working group on the best practices when publishing spatial data[6] tend to solve syntactic heterogeneity of geometries

[4] http://www.dbpedia.org/
[5] http://www.geonames.org/
[6] https://www.w3.org/TR/sdw-bp/

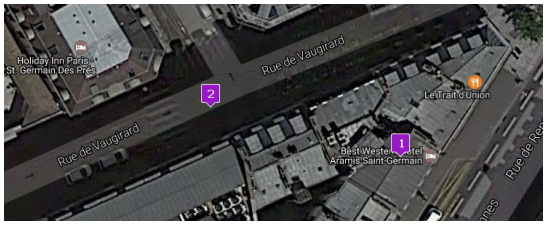

Figure 1: Internal geometric heterogeneity in Geonames.

on the Web of data. However, crowdsourced geometries may have been produced with different capture rules and at different levels of details. Fig. 1 shows an example of internal heterogeneity of geometries. In Geonames, hotels are located by points. One was captured on the building while the other was captured on the roadway. As Geonames is an open voluntary data source, this discrepancy could be due to some difference of positional accuracy or geometry capture choices.

Classical geometry similarity measures presented in 2.1 are designed to compare homogeneous geometry sets, each set being produced at the same level of detail, with the same positional planimetric accuracy and the same capture rules. They may thus be inadequate for geometries used to georeferenced linked data.

2.3 Geometry similarity functions selection, combination and tuning

Choosing automatically the adequate setting of a matching process has been the subject of many works in both ontology matching and data linking fields, and still poses challenges [31]. Approaches such as [10] and [23] propose to take advantage of the alignment of the vocabularies that structure the data to select the properties that should be compared and the distance measures used for that purpose. Other approaches addressed the question of tuning the parameters of the matching process such as the approach proposed in [29] to automatically compute the comparison criteria weights. The challenge of tuning the ontology matching process has been addressed by various approaches. [9] propose an approach for choosing the comparison criteria and computing a decision tree to aggregate them. [15] propose a classification of matching algorithms and use a set of decision rules to assess to each ontology context an adequate matching algorithm. [20] propose also an approach based on decision rules to tune the matching process using the metadata describing the ontologies and those describing the matching algorithms

The self-tuning of the matching process provide an adaptation of its settings to the context of matching while reducing the intervention of the expert. In fact, techniques such as using decision rules provide a materialization of the experts knowledge about the best setting of a matching process according to different contexts. In this work, our intuition is to use decision rules to automatically select and tune the adequate similarity measures between geometries while taking into account the heterogeneities that may exist between every pair of geometries.

3 THE XY SEMANTICS ONTOLOGY

We have seen in the previous section that geometries used for representing real world geographic features may be produced by different capture processes and may therefore be different from one data source to another. In addition, human and material input errors and the collaborative open nature of some data sources on the Web may accentuate geometric heterogeneities within a single dataset. A geometric level of detail and well-defined data specifications allow us to understand the meaning of each geometry: what it represents, how it was captured, how it is modeled, how accurate it is, and so on. In other words, what is the semantics carried by this geometry? The heterogeneities between the geometries are therefore nothing but differences in their semantics. We thus define **"the semantics of the XY"** as *"the set of geometry characteristics related to the geometric level of detail and to the capture process"*. Naming our ontology as so is motivated by the definition of geographic data semantics given by [19] as the relationship between the data and the real world phenomenon they represent.

From the heterogeneities faced in the geographic databases matching approaches, to the challenges faced in the context of the Web of data presented in section 2, we have identified the following characteristics that are more likely to affect the setting of a spatial data matching process:

- the absolute positional accuracy of geometries,
- the geometry capture rules (geometric modeling),
- the vagueness of the spatial characteristics of the geographic entities represented by the geometries,
- the level of detail of the data sources.

3.1 Vocabulary description

We propose an ontology[7], called XY semantics, that describes these characteristics, and thus, enables using them as knowledge through an interconnection process. We have chosen only these four characteristics since we assume that, although they are the most important for understanding heterogeneities, the only way to take advantage of them is to make them explicit. Indeed, other geometric features such as orientation, elongation, area, etc. are implicitly present in geometry, and therefore they are not difficult to extract on the fly.

The XY semantics ontology is based on the ISO standards on geographic databases metadata ISO 19115 and geographic data quality ISO 19157. It also includes works related to spatial entities vagueness [28] and geometry capture rules [1]. Fig. 2 shows an excerpt of XY semantics ontology.

The XY semantics ontology enables to associate to each geometry elements describing each of the four characteristics of geometry semantics listed above. As an example, the positional planimetric accuracy is described by a method and an evaluation result. The evaluation method specifies whether the evaluation result is derived from another quality element or assessed from the data. This latter is the most often employed. A data-based evaluation can be carried out by looking to a sample of the data or to their genealogy. For this reason we used the *Entity* class from the PROV-O[8] ontology.

[7] http://data.ign.fr/def/xysemantics
[8] https://www.w3.org/TR/prov-o/

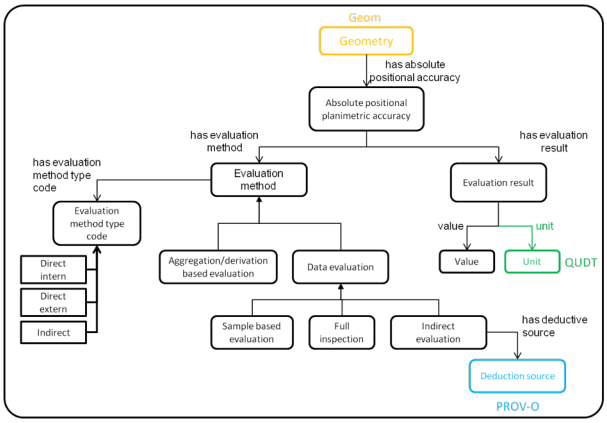

Figure 2: Excerpt of the XY semantics ontology describing the planimetric accuracy of geometries.

The evaluation result is described by its numerical value as well as its unit of measure defined from the QUDT[9] ontology.

Populating the XY semantics ontology is a task that can be very complicated. In the case of geographic data produced by mapping agencies, descriptive records and metadata, including geometric metadata, are often provided with datasets. Moreover, these datasets guarantee an internal homogeneity of the geometrical characteristics: for the same class, geometries often have often the same geometrical modeling, the same planimetric accuracy, etc. Even when the geometries of the same class have different characteristics, additional indications are often provided to explain the different situations (e.g. the different possible planimetric accuracy of addresses). In contrast, in the context of the Web of data, metadata about the methods used for georeferencing or its quality are rarely provided. In addition, datasets built collaboratively do not necessarily provide guidance on how to represent spatial references, and even if they do, they do not necessarily guarantee that contributors comply with these guidelines.

We describe in the following how to populate the XY semantics ontology in the presence and in the absence of geometric metadata.

3.2 Populating the ontology when geometric metadata are provided with datasets

The metadata of the geographical data are often provided in descriptive files. Those provided with the authoritative data of mapping agencies make populating our ontology much easier. According to the metadata of the IGN[10] address database, address points are captured in various locations: at the address sign, at the entrance of the building, 4.5m from the axis of the street (by projections from centroids of plots, by interpolations or arbitrarily), in an addressing area or in the center of the city. Moreover, the different planimetric accuracy values of the geometries are provided. These metadata can be easily translated into RDF data structures according to our XY semantics ontology and associated them with each geometry through SPARQL data insertion queries.

[9]http://qudt.org/schema/qudt
[10]The French national mapping agency

3.3 Populating the ontology when geometric metadata are not provided

Identifying geometric characteristics when they are not described in metadata is a laborious task if performed manually for each geometry in a data source. Thus, populating the XY semantics ontology becomes a complicated task. To deal with this issue, we propose a two-steps approach that automatically identify the geometry capture rules. First, it finds for each resource within the same dataset, which characteristic element of its form was chosen, when the coordinates used to locate it were entered. Identifying this characteristic allows then the evaluation of the planimetric accuracy of the spatial references. This latter is carried out by adapting the direct estimation methods for the absolute planimetric accuracy of geographical data [17] to the collaborative data.

We propose to use a "reverse-engineering" mode to identify the different geometric modelings of the spatial references. We start from the main assumption that a geometry results from an intentional choice of geometric modeling by the contributor. We propose to formulate the different hypotheses on the choices made by the contributors when entering geometries of resources. These hypothetical choices may be determined by visually comparing the spatial references of the resources to the geometries used to represent geographical entities of semantically similar or equivalent types within a geographic dataset. This empirical visual analysis allows mainly to identify the various hypothetical patterns (trends) of the geometric modeling choices that emerge from the data.

The next step is to associate each spatial reference to one of the identified geometric modeling patterns. We formalize our problem as follows: we have a geometry population G and a set of geometric modeling classes $\{C_1, ..., C_n\}$. We must therefore define a set of relevant descriptors D and select a set of learning $S \times \{C_1, ..., C_n\}$ (with $S \subset G$) in order to define the classification function C. Relevant descriptors D must be descriptive indicators whose values combination allow to discriminate the different classes. To define them, we propose to analyze the geometries of resources in comparison with geometries which represent geographical entities of semantically similar or equivalent types within a reference geographic dataset. The descriptors can therefore be a distance or a relationship between the analyzed geometries and the characteristic elements of the shape of the geographical features represented by reference geometries. We can for example consider the distance between each analyzed geometry and the closest linestring used for representing a road centerline in the reference dataset. The selection of a learning set consists in finding for each class of geometric modeling a representative sample of easily recognizable geometries in the analyzed dataset. Then, we apply a learning algorithm to assign each geometry to a geometric modeling class.

In order to evaluate our approach, we applied it to 625 resources from the French DBpedia[11] that describe historical monuments of Paris. The monuments in DBpedia are spatially referenced by prop-fr[12]:longitude and prop-fr:latitude properties. DBpedia is extracted from Wikipedia, a volunteered encyclopedia, where the location of the resources are provided without any metadata about their geometric modeling or their positional accuracy. Though, by

[11]http://fr.dbpedia.org/. Version of December 2013.
[12]http://fr.dbpedia.org/property/

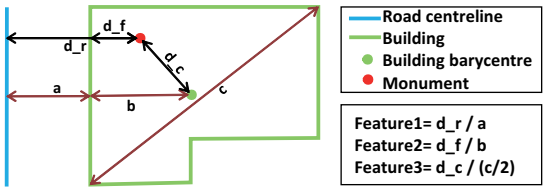

Figure 3: Learning features for geometry capture rules

Table 1: Learning results for each classifying algorithms

Method	Precision	Recall	F-measure
Bayes Network	91,6%	91,3%	91,3%
JRIP	96,3%	96,3%	96,3%
Decision Table	96,4%	96,3%	96,2%
Random Forest	98,8%	98,8%	98,7%

plotting the data on a base map we can intuitively distinguish three principal representations of the monuments locations: near the building center, near the building facade or near the road centerline. The assumption that these are the actual intended geometric representation cannot be verified because of the open volunteered nature of this source. However, we take it as hypothesis in order to automatically learn the geometry capture rules and estimate the positional accuracy of the geometries.

We used two reference geographic data sources about buildings[13] and road network[14] and we investigated some possible learning features computed with state of the art GIS tools and presented in Fig.3. We prepared a training set by manually labeling ~30 monuments of each class. Then, we applied some of Weka[15] learning algorithms and validated the results manually (see results in table 1). The resulting classification is interpreted on the level of the geometries by adding metadata about their capture rules.

Finally, we estimated the absolute positional accuracy of each point by summing two values: the distance between the point and its intended location (the building facade, the building barycenter or the road centerline) and the planimetric accuracy of the geographic feature that represents this intended location. More details on the population approach described in this section are presented in [11]

4 AN ADAPTIVE APPROACH FOR GEOMETRY SIMILARITY EVALUATION

In this section, we present how to use parameters about thresholds or the confidence value function, inferred using the XY semantics ontology, to automatically adapt the in-progress geometry similarity evaluation process.

4.1 General Description of the Approach

We have seen in section 2.1 that geometry similarity measures are usually based on some main choices: one or more similarity functions, the behavior of their normalization function, their parameters such as thresholds, the way they are combined through aggregation operators and weights. All these choices can be made based on the knowledge provided by the metadata about geometries capture process represented consistently with the XY semantics vocabulary.

As shown in Fig. 4, we suggest choosing the geometry distance function and customizing the confidence function for every comparison of geometries, depending on their metadata. For example, the threshold, the confidence function and the weight of the spatial criterion can be adapted. This can be decided through some decision rules that take as input the metadata of two geometries and give as output the parameters for their comparison. The decision rules must be defined by a data matching expert. The decisions concern different cases and can impact distinct parameters:

- The value of the distance threshold.
- The behavior of the confidence function.
- The weight of the spatial criterion.
- The potential neutrality of the spatial criterion.

For example, when two geometries have different capture rules, we can expect a considerable gap between them. In this case we can be less strict with distance values (i.e. define a higher distance threshold), and thus we can provide higher confidence values for the same distance. Moreover, when two geometries have a bad absolute positional accuracy, or when they are captured based on some vague geographic entity, we can increase or decrease the weight of the spatial criterion or even removing it from the aggregation, depending on its estimated reliability.

4.2 Approach Implementation

Silk is a very well known and maintained data linking tool with many interesting features. In order to capitalize on the assets of Silk, we implemented our approach as described in Fig. 5 to make Silk compatible with our adaptive approach for geometry similarity evaluation.

In the case of Silk, the confidence value[16] is obtained as described in equation 1, where d is the distance computed between property values and θ is the threshold chosen by the user.

$$confidence = \begin{cases} 1 - (d/\theta) & if \quad d \in [0, 2 \times \theta] \\ -1 & else \end{cases} \tag{1}$$

As explained in 4.1, decision rules based on the metadata that describe the geometries may impact the confidence function behavior. We thus suggest changing Silk's default confidence function and replacing it by the following function:

$$confidence = \begin{cases} 1 - (d/\theta)^{\alpha} & if \quad d \in [0, \theta] \\ -((d-\theta)/\theta)^{\beta} & if \quad d \in [\theta, 2 \times \theta] \\ -1 & else \end{cases} \tag{2}$$

Where d is the distance computed between geometries and θ, α and β are the parameters affected by the decision rules. θ represents the threshold. It can be computed by summing the absolute positional accuracy values of the geometries. α and β are the convexity/concavity factors for respectively the positive and the negative parts of the confidence function. Both α and β should be positive values (c.f. Fig. 6). When the decision rules affect one or more of these three parameters, the confidence function becomes more or less strict

[13] From the BD PARCELLAIRE®, IGN's land parcels database.
[14] From the BD TOPO®, IGN's topographic database.
[15] http://www.cs.waikato.ac.nz/ml/weka/
[16] https://github.com/silk-framework/silk/blob/master/doc/LinkageRules.md

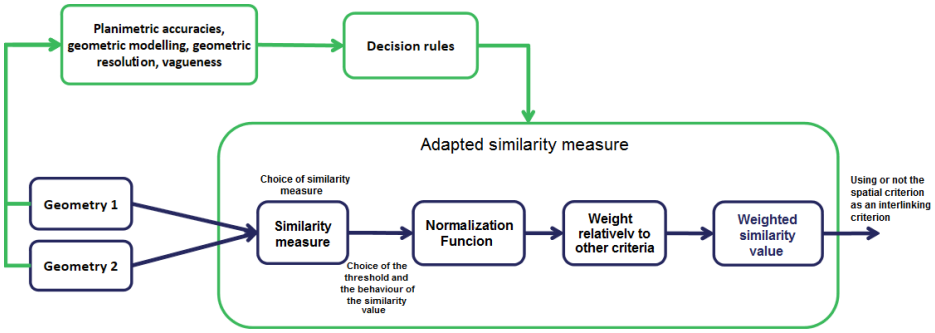

Figure 4: The global approach for the self-adaptive comparisons of geometries

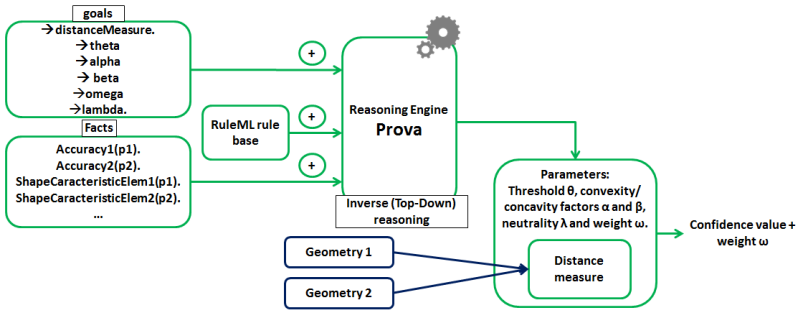

Figure 5: Implementation of the self-adaptive linking approach

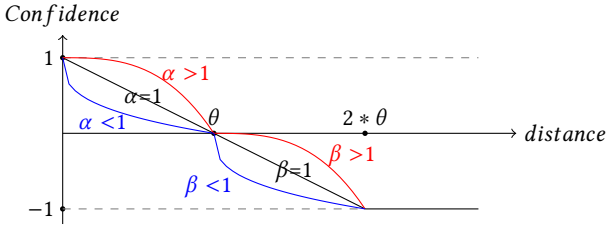

Figure 6: Variations of confidence function behavior

while keeping its monotony. Two other parameters are eventually necessary: λ the neutrality of the spatial criterion and ω its weight in the case of a weighted aggregation.

All the parameters used by the geometry comparison process are defined by means of a decision rule base defined in advance by a data linking expert. The rule base is written in RuleML[17].The pieces of knowledge about geometries capture process and planimetric accuracy are formatted as RuleML facts. The resulting values of the requested parameters are represented as RuleML goals.

Then, a backward reasoning is executed by PROVA[18] rule engine on the whole rule base to infer the comparison parameters. θ, α, β, ω and λ are thus given as an output from the reasoning engine. They are then used together with the spatial distance to compute the confidence value of the spatial criterion respecting the formula described by the equation 2. A backward (Top-Down) reasoning works similarly to most Prolog systems[4]. It gives answers to a set of goals by reasoning on a base of rules and facts. In order to make sure that our approach works safely, the rule base must be decidable. Since we have a fixed number of goals, the complexity of the reasoner is linear with the number of the rules declared in the rules base. Using PROVA in Silk for every comparison results in a total quadratic complexity. This implementation is scalable though, because it remains compatible with the MapReduce version of Silk.

5 INTERLINKING THE MONUMENTS OF PARIS

To evaluate our adaptive geometry similarity measure, we applied it on datasets on Paris historical monuments. The first one is the Mérimée[19]database, which is a national monuments registry produced and maintained by the French Ministry of Culture and Communication. Mérimée database contains 1582 instances, it is provided as a CSV file and its monuments are located by textual addresses. We transformed the data to RDF with the Datalift[20] platform. Then we selected only the monuments located in Paris and we geocoded these monuments by linking their addresses to their corresponding BD ADRESSE® [21] features. The second comes from DBpedia and is presented in 3.3

[17] Rule Markup Language, http://wiki.ruleml.org/
[18] https://prova.ws/
[19] https://www.data.gouv.fr/fr/datasets/immeubles-proteges-au-titre-des-monuments-historiques/
[20] http://datalift.org/
[21] Address database of IGN (French national mapping agency).

Extracting metadata about the absolute positional accuracy and the geometry capture rules of Mérimée resources is quite straightforward. Indeed, their geometries come from the BD ADRESSE® database, which is a well-documented traditional geographic database. Its implementing rules provide information about the four different geometry capture rules applied for addresses: at the address sign, by projection on the corresponding street centerline, by interpolation on the corresponding section of the street centerline, and more rarely in the center of the addressing zone. Depending on these geometry capture rules, three different positional accuracy values are possible: 12, 18 and 30 meters. We structured these metadata according to the vocabulary presented in 3 and we added them to Mérimée geometries with Sparql update queries. In the case of DBpedia monuments, we used the approach presented in section 3.3 to learn their metadata.

5.1 Comparative Matching Tests

In order to evaluate our approach, we have performed two spatial matching tasks on the datasets described above: one with Silk's default spatial distance operator and one with our adaptive geometry similarity measure.

For Silk's default spatial distance operator, the confidence value of each comparison is computed according to the formula 1. We performed several runs with different distance thresholds θ to find out which value gives the best results. The top table of Table.2 outlines these results. The best f-measure is obtained for θ=40 meters. The runtime of this approach is around 2 seconds.

For our approach, we used two rule bases to define θ, α, and λ parameters (see section 4.1). Since it is a mono-criterion matching task, the parameters β and ω are not needed. Since the geometries in the two datasets are points only, we chose to use a euclidean distance measure between them. With respect to what is usually done in geographic data matching, we define the threshold θ as the sum of the positional accuracies of the two compared geometries (here named *accu*1 and *accu*2). When the geometries capture rules target different spatial characteristic (*characelem*1 and *characelem*2) of the real world entities they intend to represent (e.g. the first geometry is the barycenter of the building while the second is a point of the facade of a building), we add a bias named $delta_c$. When they target different types of real world entities (*host*1 and *host*2), we add another bias named $delta_h$. For our use case, we set $delta_c$ at 10 m and $delta_h$ at 15 m. The rule base $rb1$ is summarized below and the results we get with it are shown on Table.2:

```
theta(X):-host1=host2,characelem1=characelem2,X=accu1+accu2.
theta(X):-host1=host2,characelem1!=characelem2,X=accu1+accu2+delta_c
theta(X):-host1!=host2,X=accu1+accu2+delta_h.
distance("euclidian").
alpha(1). lambda(false).
```

Based on the experience of the test performed with the rule base $rb1$, we defined a second rule base by adding more fine-grained rules, namely $rb2$. In this rule base, we set the values of $delta_c$ at 20 m and $delta_h$ at 30 m[22]. We also change the convexity of the confidence function when geometry capture rules are different and when the targeted real world geographic entities are too vague or too wide, we neutralize the spatial criterion. These additional rules are described below and the results are also presented in Table.2. These rules replace the last line of $rb1$:

```
alpha(1):- host1=host2,characelem1=characelem2.
alpha(2):- host1=host2,characelem1!=characelem2.
alpha(3):- host1!=host2.
lambda(true):-host1= addressZone. lambda(true):-host2= addressZone.
lambda(true):-host1= commune. lambda(true):-host2= commune.
lambda(false):-host1!=addressZone,host2!=addressZone,host2!=commune,
              host2!=commune.
```

The runtime of the matching task using these rule bases is 9∼14s.

[22] $delta_c$ and $delta_h$ were estimated by investigating the bias in some cases where two geometries with different geometric modelings locate two equivalent resources

Table 2: Instance matching results compared to a reference links set produced with Wikidata information and completed manually

θ	Precision	Recall	F-measure
10	84,55%	20,90%	33,51%
20	74,15%	39,33%	51,40%
30	64,15%	51,46%	57,11%
40	57,96%	58,88%	**58,42%**
50	50,09%	63,15%	55,86%
60	44,75%	65,17%	53,06%
70	40,43%	66,97%	50,42%
80	36,76%	68,31%	47,80%
90	33,99%	69,44%	45,64%
100	31,68%	70,34%	43,68%
Using Bayes Network learning results			
Rule base	Precision	Recall	F-measure
rb1	70,43%	58,88%	64,14%
rb2	71,43%	62,92%	**66,91%**
Using learning results after correction			
Rule base	Precision	Recall	F-measure
rb1	73,46%	59,10%	65,50%
rb2	70,99%	62,70%	**66,59%**

5.2 Discussion

Extracting knowledge about the causes of heterogeneity between geometries by using supervised learning method shows promising results. Nonetheless, the choice of adequate learning features is conditioned by the context of the data and has an important impact on the results. For instance, we ran another learning test on the address data in the city of Lyon in France. In this case the geometries had also three possible geometry capture rules: in the center of the building, at the entrance of the building and on the street centerline. In this city, the buildings sizes and their distances to the road are more homogeneous than in Paris. In this case, simple learning features such as d_f, d_r and d_c (Fig.3) were sufficient to obtain very good results. The choice of the training set is also crucial: the entities must be clearly representative of the different learning classes. The classification errors induced by the learning step show a low effect on the final linking results compared to the improvement brought by the linking approach.

The matching results of our approach show clearly better f-measure scores than the default approach. Adapting the parameters of the geometry comparison measure ensures some of the benefits

of both small and big distance thresholds. Compared to the best result of the classical approach (θ=40), we avoid 50% of the false positive links using $rb1$ and 40% using $rb2$. We do not significantly add new true positive links using $rb1$ but we increase their number by 6% using $rb2$.

Unsurprisingly, our approach has a clearly higher runtime. The complexity of the implementation depends on the size of the rule base. This is why we have tried to define the minimum number of decidable rules that can sufficiently adapt the parameters of the confidence function. A more detailed rule base could have provided better results but it would have been much less efficient in runtime. The user has to find the best trade-off between efficiency and performances. As a matter of fact, our approach is better suited to instance matching tasks of data sources which have a high spatial density and instances described by geometries with a lot of internal geometric heterogeneities.

6 CONCLUSION AND FUTURE WORKS

In this work we tackled the problem of the geometry similarity evaluation for georeferenced resources linking. We proposed an ontology to represent knowledge about geometry positional accuracy and capture rules and an approach to extract it from the considered spatial data and geographic reference data by using automatic supervised learning. We also defined a data matching approach that relies on this knowledge to adapt the comparison of geometries during its runtime. The matching results show better performances than the classical non-adaptive approach.

Yet, the main downside of our approach is the time complexity of the current implementation that should be improved. This could be done by adding a cache system for the reasoning results in order to reduce the workload of the reasoning engine. Further tests, with bigger and more heterogeneous datasets, especially datasets with different types of geometry, could also bring new insights to this proposal. Future tests should also include the two remaining aspects of the XY semantics, namely the geometry resolution and its vagueness, in both populating and interlinking approaches.

REFERENCES

[1] Nathalie Abadie. 2012. *Formalisation, acquisition et mise en œuvre de connaissances pour l'intégration virtuelle de bases de données géographiques: les spécifications au cœur du processus d'intégration.* Ph.D. Dissertation. Université Paris-Est.

[2] Benjamin Adams, Linna Li, Martin Raubal, and Michael F Goodchild. 2010. A general framework for conflation. *Extended Abstracts Volume, GIScience* (2010).

[3] Esther M Arkin, L Paul Chew, Daniel P Huttenlocher, Klara Kedem, and Joseph S Mitchell. 1991. *An efficiently computable metric for comparing polygonal shapes.* Technical Report. CORNELL UNIV ITHACA NY.

[4] Marcel Ball, Harold Boley, David Hirtle, Jing Mei, and Bruce Spencer. 2005. Implementing RuleML Using Schemas, Translators, and Bidirectional Interpreters. (2005). https://www.w3.org/2004/12/rules-ws/paper/49/

[5] Atef Bel Hadj Ali. 1999. Geometrical Matching of Polygons in GISs and Assessment of Geometrical Quality of Polygons. (1999).

[6] Scott D Cohen and Leonidas J Guibas. 1997. Partial matching of planar polylines under similarity transformations. In *8th Annual ACMSIAM Symposium on Discrete Algorithms*. 777–786.

[7] Benoît Costes. 2014. Matching old hydrographic vector data from Cassini's maps. (2014), 51–65 pages.

[8] Thomas Devogele, Christine Parent, and Stefano Spaccapietra. 1998. On spatial database integration. *International Journal of Geographical Information Science* 12, 4 (1998), 335–352.

[9] Fabien Duchateau, Zohra Bellahsene, and Remi Coletta. 2008. A flexible approach for planning schema matching algorithms. *On the Move to Meaningful Internet Systems: OTM 2008* (2008), 249–264.

[10] Zhengjie Fan, Jérôme Euzenat, and François Scharffe. 2014. Learning concise pattern for interlinking with extended version space. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, Vol. 1. IEEE, 70–77.

[11] Abdelfettah Feliachi, Nathalie Abadie, and Fayçal Hamdi. 2017. Assessing the planimetric accuracy of georeferenced data on the Web: A case study on DBpedia.. In *QMMQ 2017 workshop, in conjunction with ER2017 conference.*

[12] Alfio Ferraram, Andriy Nikolov, and François Scharffe. 2013. Data linking for the semantic web. *Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications* 169 (2013), 326.

[13] JF Girres. 2012. Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques. Application aux mesures de longueur et de surface. *PhD, Université Paris–Est, France* (2012).

[14] S Hahmann and D Burghardt. 2010. Connecting linkedgeodata and geonames in the spatial semantic web. In *6th International GIScience Conference.*

[15] Mirella Huza, Mounira Harzallah, and Francky Trichet. 2007. OntoMas: a tutoring system dedicated to ontology matching. In *Enterprise Interoperability II.* Springer, 377–388.

[16] Robert Isele, Anja Jentzsch, and Christian Bizer. 2011. Efficient Multidimensional Blocking for Link Discovery without losing Recall. In *WebDB.*

[17] ISO. 2013. *19157: Geographic information – Data quality.* International Standard. International Organization for Standardization (http://www.iso.org).

[18] Krzysztof Janowicz, Martin Raubal, and Werner Kuhn. 2011. The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science* 2011, 2 (2011), 29–57.

[19] Marinos Kavouras and Margarita Kokla. 2007. *Theories of geographic concepts: ontological approaches to semantic integration.* CRC Press.

[20] Malgorzata Mochol and Anja Jentzsch. 2008. Towards a rule-based matcher selection. In *International Conference on Knowledge Engineering and Knowledge Management.* Springer, 109–119.

[21] Sébastien Mustière and Thomas Devogele. 2008. Matching networks with different levels of detail. *GeoInformatica* 12, 4 (2008), 435–453.

[22] Axel-Cyrille Ngonga Ngomo. 2013. Orchid–reduction-ratio-optimal computation of geo-spatial distances for link discovery. In *International Semantic Web Conference.* Springer, 395–410.

[23] Andriy Nikolov, Victoria Uren, Enrico Motta, and Anne N De Roeck. 2008. Integration of semantically annotated data by the KnoFuss architecture. In *EKAW.* Springer, 265–274.

[24] Ana-Maria Olteanu-Raimond, Sébastien Mustière, and Anne Ruas. 2015. Knowledge formalization for vector data matching using Belief Theory. (2015), 21-46 pages.

[25] George Percivall, Carl Reed, Lew Leinenweber, Chris Tucker, and Tina Cary. 2003. OGC reference model. (2003). http://rap.opengeospatial.org/orm.php

[26] J Salas and Andreas Harth. 2011. Finding spatial equivalences accross multiple RDF datasets. In *Proceedings of the Terra Cognita Workshop on Foundations, Technologies and Applications of the Geospatial Web.* 114–126.

[27] LT Sarjakoski. 2007. Conceptual models of generalisation and multiple representation. *Generalisation of geographic information: cartographic modelling and applications* (2007), 11–35.

[28] S Schade. 2010. Computer-tractable translation of geospatial data. *International Journal of Spatial Data Infrastructures Research, Revue en ligne publiée par le Joint Research Centre (European Commission)* 5 (2010).

[29] Md Seddiqui, Rudra Pratap Deb Nath, Masaki Aono, et al. 2015. An efficient metric of automatic weight generation for properties in instance matching technique. *arXiv preprint arXiv:1502.03556* (2015).

[30] Mohamed Ahmed Sherif and Axel-Cyrille Ngonga Ngomo. 2015. A systematic survey of point set distance measures for link discovery. *Semantic Web Journal.(Cited on page 18.)* (2015).

[31] Pavel Shvaiko and Jérôme Euzenat. 2013. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering* 25, 1 (2013), 158–176.

[32] Luis M Vilches-Blázquez, Víctor Saquicela, and Oscar Corcho. 2012. Interlinking geospatial information in the web of data. *Bridging the Geographic Information Sciences* (2012), 119–139.

[33] Steffen Volz. 2006. An iterative approach for matching multiple representations of street data. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 36, Part 2/W40 (2006), 101–110.

[34] Volker Walter and Dieter Fritsch. 1999. Matching spatial data sets: a statistical approach. *International Journal of geographical information science* 13, 5 (1999), 445–473.

[35] Yanxia Wang, Deng Chen, Zhiyuan Zhao, Fu Ren, and Qingyun Du. 2015. A Back-Propagation Neural Network-Based Approach for Multi-Represented Feature Matching in Update Propagation. *Transactions in GIS* 19, 6 (2015), 964–993.

[36] Bisheng Yang, Xuechen Luan, and Yunfei Zhang. 2014. A Pattern-Based Approach for Matching Nodes in Heterogeneous Urban Road Networks. *Transactions in GIS* 18, 5 (2014), 718–739.