Thematic Linked Data Matching: an Approach based on Geographic Reference Data

Abdelfettah Feliachi\textsuperscript{a}, Nathalie Abadie\textsuperscript{a} and Fayçal Hamdi\textsuperscript{b}

\textsuperscript{a} COGIT laboratory, Université Paris-Est, IGN / SRIG, Saint-Mandé, France. E-mails: Abdelfettah.Feliachi@ign.fr, Nathalie-Abadie@ign.fr

\textsuperscript{b} CEDRIC Laboratory, Conservatoire National Des Arts Et Métiers (CNAM), Paris, France. E-mail: faycal.hamdi@cnam.fr

Abstract. Many resources published on the Web of Data are described by either direct or indirect spatial references. These spatial references can be used for data matching or cartographic visualization purposes. Indeed, they may be used as instance matching criterion: two resources that are very close in space may represent the same thing, or at least they may have some semantic relationship. However, heterogeneities between spatial references may make their use as instance matching criterion not very reliable or even impossible. In this article, we propose to reduce the data matching difficulties caused by the heterogeneity of spatial references by means of a geographic reference dataset. We also propose to take advantage of links created between thematic resources and geographic resources for designing better maps for data visualization.

Keywords: data matching with background knowledge, spatial references, geographic reference data, data and links visualization

1. Context and objectives

Among the data published on the Web of Data, many resources are associated either directly through geographic coordinates or geometric primitives such as points, polylines or polygons, or indirectly through postal addresses, names of administrative units or points of interest, to a position in the geographic space. For example, the properties geo\textsubscript{1}:long and geo:lat of the W3C vocabulary Geo are respectively used in more than 100 000 triples, around 300 000 triples reuse classes that describe postal addresses, and more than 60 properties with semantics close to "locatedIn" or "hasLocation" are currently used by datasets\textsuperscript{2}.

Describing a resource with some spatial reference implies that this resource is somehow related to some real world geographic entity. Two resources which are described by identical or spatially close spatial references are therefore very likely to have some semantic relation. Thus, using spatial references is generally beneficial for data matching purposes. Spatial references are most often taken into account in the data matching process by computing geographic distance measures between resources that are compared. However, spatial references associated to resources may be very heterogeneous from one resource to another. This heterogeneity can be caused by the use of different types of spatial references (direct or indirect), the use of different vocabularies for describing spatial references, spatial reference sources with different levels of accuracy, different geometric modelling
choices or different level of details from one dataset to another. This heterogeneity between spatial references may make the use of geographic distance measures for data matching purposes not very reliable, and sometimes even impossible.

In this article, we propose to take advantage of reference geographic databases for matching and visualizing thematic data described by heterogeneous spatial references. We follow the intuition that anchoring thematic resources on reference geographic data should help to identify relationships between these resources. Data visualization applications are a common way to help the users discover and understand the data. For georeferenced data, maps are the most intuitive visualization approach. Therefore, we also propose a cartographic application that visualize thematic data together with reference geographic data by taking advantage of the anchor links created at the data matching step.

The paper is organized as follows. In the next section, we present some related research in the field of georeferenced data matching. Section 3 exposes the matching challenges related to spatial references in the Web of Data. We describe in section 4 the data matching approach based on a geographic reference dataset that we propose. This approach was implemented on datasets describing French historical monuments. This use case is presented and the results are evaluated in section 5. The cartographic visualization application based on anchor links is described in section 6. Section 7 is a discussion about the results and some outlooks for our work. Finally, we give a conclusion in section 8.

2. Related works on data matching by means of their spatial references

Data linking is the step in the data publication process aiming 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. Its data matching subtask is 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 between two resources is, the more they are likely to represent the same real-world entity [5]. This similarity evaluation task is practically based on approaches and measures proposed in different communities, which also need to identify relationships between resources such as reference reconciliation or ontology matching [9].

Many approaches and tools have been proposed for automatically identifying relationships between objects from heterogeneous geographic databases which represent the same real-world entity. Like data linking, geographic data matching is practically performed by comparing properties of geographic databases elements in order to evaluate their degree of similarity. In the case of geographic data, the main matching criterion is generally their spatial reference, which represents the position of the real-world geographic entity represented by each database element. Many approaches have already been proposed for comparing direct spatial references (i.e. geometries). [10], [3], and [16] have proposed geographic data matching approaches based on point comparisons. [20] and [11] have proposed approaches for matching linear data used for representing networks respectively at the same level of detail and at different levels of detail. Finally, [4] and [18] have proposed approaches for comparing polygons. In all cases, a specific measure is used for quantifying the level of similarity between geometries. Multicriteria approaches based on the comparisons of geometries, attributes and topological properties have also been proposed, such as [12].

At the crossroads of the data linking and the geographic data matching domains, approaches which aim at finding equivalence relationships between Linked Data described by spatial references have also been proposed. They are based on similarity measures between geometries directly inspired by those used in the field of geographic data matching [17]. Most of the time, they are applied on data from traditional geographic databases, represented according to the RDF model and published on the Web of Data. In the approach proposed by [7], resources described by geometries of the “point” type are compared on the basis of two criteria: their geometries and their names. The similarity measure between geometries used in that approach depends on the bilateral inclusion of a given candidate point into a bounding box built around a potentially matching point. In the approach proposed by [17], overlapping between polygonal geometries is evaluated by means of the Hausdorff distance.
3. Matching spatial references from Web of Data sources

As we have seen in the previous section, matching approaches based on the spatial references criterion take advantage of a set of measures. These measures are generally used to compare spatial references of the same type. They include geometric distance measures between geographic features or string-based distance measures used for comparing strings that represent addresses or geographical names. Furthermore, different datasets may use different types of spatial references. In such cases, a first step of geocoding must be performed. This means that descriptive locational data, such as postal addresses or named places, must be turned into direct spatial references [6]. Besides, differences in geometrical accuracy or in geometry capture rules can occur even between datasets that use the same type of spatial references. These differences may cause difficulties for the data matching process; for instance, a large distance between the geometries of two features representing the same real world geographic entity, or a very short distance between features representing two distinct entities. These heterogeneities are usually dealt with by means of a geometric distance threshold chosen by considering each dataset mean positional accuracy and multicriteria matching approaches for disambiguating the results.

However, in the context of the Web of Data, spatial references are often used for locating the resources they describe and not for representing them in a detailed way for cartographic purposes. Resources are mostly located by postal addresses, longitude and latitude coordinates or geometries of a "point" type. Detailed geometries such as linestrings or polygons are more common in datasets derived from traditional geographic data sources. Therefore, in addition to the problems usually faced with geographic data, matching georeferenced resources in the context of the Web of data raises some additional challenges.

First, as we have seen, spatial references are very likely to have a low level of detail, which implies that matching criteria based on spatial properties like location, shape or orientation may be either less reliable or completely inoperable.

Furthermore, unlike geometries in traditional geographic datasets, spatial references are not the main pieces of information in the description Web of Data resources. Therefore, they may not always be captured, and thus not always be available as matching criterion. In a matching process where the spatial criterion is required, this may lead to a difference of completeness from one source to another.

Besides, the granularity of the data sources contents may differ: one resource in a first dataset may be detailed by many resources in a second dataset. This raises the issue of proposing a matching approach that could generate n:m links instead of 1:1 links.

Contrary to geographic datasets provided by traditional producers such as national mapping agencies, Web of Data spatial references may be produced with different levels of accuracy and different geometric modelling choices within the same dataset. In such cases, defining a value for the geometric distance threshold, which takes into account positional differences due to both positional accuracy and geometric modelling of each resource of both data sources, but which remains low enough to avoid false positive results, is at least challenging and most of the time impossible.

Lastly, all these causes of heterogeneities are most of the time not explicitly documented, which makes them even more difficult to handle.

Problem definition: Let $S_1$ and $S_2$ be two thematic datasets containing a set of resources (e.g. dbpedia-en³:Eiffel_Tower) belonging to a same category (e.g. dbpedia-owl⁴:Monument), and having at least a direct or indirect spatial reference property (e.g. prop-fr⁵:adresse, geometrie⁶:geometry, etc.).

Due to heterogeneities between spatial references, matching $S_1$ and $S_2$ by comparing the geometric distance values between resources of both datasets with a threshold would provide results with low precision and recall. We want to propose an approach based on the spatial reference criterion that overcomes the heterogeneities of the Web of Data resources described above, in order to generate a link set $M$ with better Precision and recall than the link set obtained by the classical direct approach. Therefore, we introduce a geographic reference dataset as background knowledge resource.

³http://dbpedia.org/resource/
⁴http://dbpedia.org/ontology/
⁵http://fr.dbpedia.org:8890/property/
⁶http://data.ign.fr/def/geometrie
4. A data matching approach based on geographic reference data

Geographic reference data are defined by the The French National Geographic Information Council (CNIG) as "a minimum set of complementary and consistent data that can be used for georeferencing data from any organism and producing domain specific reference geodata. Therefore, it is useful for everybody. The data in a geographic reference dataset are clearly identified and are under the responsibility of a public structure, itself clearly identified as responsible for them...". To fulfill their intended use, geographic reference data must be consistent. They generally describe the real world phenomena with well detailed geometries, and provide a set of metadata about their content, their quality, their capture methods, etc. In this paper, we call thematic data any domain specific data that are not specifically designed for representing real world topographic entities. Therefore, even if they have spatial references, thematic data are not part of a geographic reference dataset.

In order to overcome the heterogeneities of georeferenced thematic data mentioned in section 4, we propose to take advantage of geographic reference data by using them as background knowledge resources. Analogously to Aleksovski et al. [1] approach, in the ontology matching field, we propose to first anchor different resources to the same reference geodataset and then to derive equivalence - or other - relationships between thematic resources from anchoring relationships between thematic and geographic resources.

Figure 1 is an example taken from the use case of the section 5, that presents how an equivalence relationship can be derived between a first resource located by coordinates and a second resource located by an address. The latter can be geocoded by matching the string describing its address with a geographic reference dataset that represents addresses and their coordinates. When the coordinates attributed to the second resource are directly compared to those of the first resource, it is not possible to establish a relationship between them because the coordinates are too far from each other (regarding a threshold value defined for mean distance between building locations in urban area). However, when both resources are first matched to the geographic reference dataset that represents buildings, a relationship between them can be derived from the fact that they are both anchored to the same building.

Anchoring thematic resources to a reference geodataset requires executing a data linking process that considers the spatial references as the main criterion. But even if thematic resources and geographic data do not represent the same things, the spatial references
used to describe thematic resources refer to points of the geographical realm that are close to the geographic features to which they should be anchored. Therefore, approaches traditionally used for linking georeferenced data can be used in this step. We thus propose to reuse existing data linking tools and to add plugins dedicated to geographic distance measures to them. This implies converting geographic data into RDF, and reusing a data linking tool, which is extensible, so that it can implement geographic distance measures adapted to the different geometric primitives and coordinate reference systems.

Beyond using them for data linking between different data sources, the links between geographic and thematic data could be beneficial for many other purposes. For instance, they can be reused for spatial data analysis or thematic mapping. In this article we propose to take advantage of these geographic-thematic links in a cartographic visualization application.

**Approach formulation**

Let \( S_1, S_2 \) be two (directly or indirectly) spatially referenced thematic datasets. Let \( S_{GR} \) be a geographic reference dataset describing geographic features and their geometries. \( S_{GR} \) must include the same types of spatial references as \( S_1 \) and \( S_2 \).

**Definition 1 (Anchor Link Set)** An anchor link set \( A \) is computed between a thematic dataset \( (S_1 \) or \( S_2) \) and \( S_{GR} \). An anchor link is a relationship (in a broader sense) that joins semantically a resource of a thematic class with a resource of the suitable geographic class.

**Example.** Thematic resources of category dbpedia-owl:Monument should be anchored to geographic resources of class topo8:Bati.

An anchor link set is computed by comparing direct or indirect spatial references. In the indirect spatial reference case, anchoring a thematic resource to a geographic resource is a manner of geocoding its location. In both direct and indirect spatial reference cases, the entities are compared by means of a distance measure \( d \). Here, we distinguish the two cases:

- **comparing direct spatial references:** In this case, \( d \) is a geographic or geometric distance measure (e.g. Orthodromic, Euclidian distance, Hausdorff, etc.).

**Definition 2 (Derivation)** Let \( A_1 \) be the computed anchor link set between \( S_1 \) and \( S_{GR} \) and \( A_2 \) the computed anchor link set between \( S_2 \) and \( S_{GR} \). The matching link set \( M \) can be derived from \( A_1 \) and \( A_2 \). We define the derivation function \( L \) from the two anchor link sets \( A_1 \) and \( A_2 \) as:

\[
L : A_1 \times A_2 \rightarrow M
\]

\[
(a_1, a_2) \mapsto m,
\]

\( a_1 \) is an anchoring link between the thematic entity \( e_1 \in S_1 \) and the geographic reference entity \( e_g \in S_{GR} \), and \( a_2 \) is an anchoring link between the thematic entity \( e_2 \in S_2 \) and the same geographic entity \( e_g \). The nature of the link \( m \) between thematic objects depends on the nature of the entities \( e_1, e_2 \).

5. **Linking datasets on French historical monuments by means of address and building reference data**

The proposed approach is inherently more adapted to thematic data described by spatial references less detailed than those used for describing geographic features in the reference geographic dataset, such as thematic data described by points instead of polylines or polygons.

To be representative of the issues presented in section 3, the use case must involve datasets with the following properties:

- They should be both georeferenced, possibly with different types of spatial references.
- They should describe resources with little detailed spatial references.
- They should describe resources with different levels of granularity.
- They should present a high density of spatial references in the covered area.
- They should describe resources with spatial references captured respecting different geometric modelling choices.
- They should describe resources with spatial references captured with different planimetric accuracies.

To implement and test our approach, we chose two thematic open datasets on historical monuments in
the city of Paris. The first one was retrieved from the French DBpedia\(^9\). The second one, the Mérimée\(^10\) database, is produced and maintained by the French Ministry of Culture and Communication. A first exploration of the data shows that these datasets fit with the requirements listed above. The spatial references are of different types, and are few detailed (points and addresses only). The amount of monuments in Paris causes high density of spatial references in some areas. Also, one monument in DBpedia could refer too many monuments in Mérimée and vice versa. Positional accuracy and spatial references capture rules are different from a resource to another.

To select the DBpedia monuments, we explored the entities that have "dbpedia-fr": Catégorie: Monument_historique_de_Paris" as an object for the "dc-terms: subject" property, and the entities related to "dbpedia-fr:Catégorie: Monument_parisien" by a "skos: broader | dcterms: subject" path. We selected only the georeferenced entities with prop-fr: longitude and prop-fr: longitude properties or geo\(^14\): long and geo: lat properties. Mérimée monuments are provided as a CSV file and are georeferenced with literal addresses. From this dataset, we extracted only the monuments that are located in the city of Paris by means of their postal code.

The reference geodataset used for describing the topography of the city of Paris is composed of data from the BD PARCELNAIRE\(^15\) and the BD ADRESSE\(^16\) databases produced by the French national mapping agency (IGN\(^17\)). These databases are provided as ESRI shapefiles\(^18\) and are freely available for education and research purposes. The BD PARCELNAIRE\(^\circ\) represents real-world buildings by geometries of a "polygon" type. It has been chosen instead of other topographic databases such as BD TOPO\(^\circ\), because

---

\(^9\)http://fr.dbpedia.org/
\(^11\)http://fr.dbpedia.org/resource/
\(^12\)http://purl.org/dc/terms/
\(^13\)http://www.w3.org/2004/02/skos/core#
\(^14\)http://www.w3.org/2003/01/geo/wgs84_pos#
\(^15\)IGN’s cadastral database.
\(^16\)IGN’s address database.
\(^17\)The National Institute of Geographic and Forest Information.
it provides a more fine-grained and less aggregated representation of buildings. The BD ADRESSE® is a database that represents addresses in a well-structured way, and that associates to each one of them a "point" geometry. These geographic reference data are provided with metadata documents which describe their characteristics, the quality of the data and the geometric representation. BD ADRESSE® is of a known planimetric accuracy that does not exceed 10m in the city of Paris, while the planimetric accuracy of BD PARCELLAIRE® does not exceed 2m. Both databases are geometrically consistent. This means that every address is located near its corresponding building. Even thought there is no explicit relation between BD ADRESSE® addresses and BD PARCELLAIRE® buildings, it is quite straightforward to retrieve them by means of a spatial matching step.

5.1. Architecture and tools

Figure 2 illustrates the global architecture we chose to implement our approach.

Geographic reference data were translated into RDF using the Datalift\(^{19}\) platform. This platform provides a way to transform data from many models and formats, including geodata (GML, SHP, etc)\(^{[8]}\), into RDF model. Thus, we have used this platform to convert BD PARCELLAIRE® buildings and BD ADRESSE® addresses from shape files to RDF datasets.

Mérimée data initially presented as a CSV file were also converted into RDF by using Datalift. Historical monuments described in this database are georeferenced by multiple literal postal addresses separated from each other by semicolons and commas. Thus, these data have been cleaned using SPARQL UPDATE queries in order to split these multiple addresses into simple addresses so that they could be processed individually.

After their conversion or their extraction, all the datasets were hosted in local triple stores (OpenRDF Sesame and Openlink Virtuoso) to ease the data linking process.

The data linking process is performed using Silk Link Discovery Framework\(^{20}\) which has been extended with a spatial plugin to compute geographic distances between geometries. Thereby, we can still use all the measures already implemented in this framework and benefit from all the optimizations and the multicriteria aggregation approaches proposed by Silk.

5.2. Distance measure between geometries

The most commonly used spatial references on the Web of Data and in open data thematic datasets are postal addresses, longitude and latitude coordinates, and more rarely geometries of a "point" type. Inversely, spatial references used in geographic databases are generally more detailed. Besides points, geometries of "linestring" or "polygon" types are often used to provide a more realistic representation of the location and shape of geographic entities. The distance measure that we chose takes into consideration these differences and returns the minimal distance between a point and any other type of geometry, or returns 0 if the point is included in or is equal to the geometry.

From a data source to another, geometries may be described by coordinates defined within different coordinate reference systems (CRS). Therefore, in order to unify the coordinates reference systems of two different sources and make sure that the distance measure is applied properly\(^{21}\), we start by projecting all the coordinates to the same appropriate CRS given as a parameter\(^{22}\).

We chose to use a distance rather than an inclusion in a buffer because it provides a way to sort matching candidates from nearest to furthest. In fact, buffers around polygons may overlap. Therefore a point may be included simultaneously in the buffers around two polygons. In this case we could not decide to which polygon the point should be anchored. By computing the distance between the point and each polygon, we can tell which one is closer, and thus anchor the point to it.

This distance measure is implemented using Geotools java library\(^{23}\), and integrated to Silk as a distance plugin.

5.3. Matching by means of geographic reference data

Let Mérimée database be the first thematic dataset \(S_1\). In this dataset, monuments are georeferenced

\(^{19}\)http://datalift.org/  
\(^{20}\)http://www.assembla.com/spaces/silk/  
\(^{21}\)The Geotools distance that we reuse is based on an Euclidian distance and thus requires projected coordinates.  
\(^{22}\)In our case, Lambert 93, the French legal projection CRS  
\(^{23}\)http://geotools.org/
by the properties mérimée:adresse and mérimée:codeInsee.

Let DBpedia Paris historical monuments be the second thematic dataset \( S_2 \). In this dataset, monuments are georeferenced by the properties prop-fr:long, prop-fr:lat, geo:long, geo:lat.

Let the union of BD PARCELLEAIRE® and BD ADRESSE® be the geodataset \( S_{GR} \). The available direct spatial references are of geometric:Point type for addresses and of geometric:Multipolygon type for buildings. The indirect spatial reference of the addresses is the combination of the properties badresse:numero, badresse:indiceRepetition, badresse:nomVoie, badresse:codeInsee, badresse:codePostal and badresse:complement. We created explicit links between each address and its corresponding building by taking advantage of their geometrical consistency.

Figure 3 illustrates the data matching operations performed between the two thematic datasets and the reference geodataset.

The monuments described in the Mérimée database \( (S_1) \) are georeferenced by indirect spatial references. Therefore, a first data linking process is performed to geocode these addresses (Figure 3 (1)). This geocoding process is performed by matching the addresses used in Mérimée for identifying the positions of monuments with addresses provided in the BD ADRESSE® database \( (S_{GR}) \). Addresses elements in each database are combined before being compared using a token wise variant of the Levenshtein measure. Anchor links between monuments and buildings are inferred from the links created between monuments and addresses and the explicit relations between addresses and buildings. Let \( A_1 \) be the resulting anchor link set of this first anchoring task.

DBpedia resources on Paris historical monuments \( (S_2) \) are anchored to the instances of \( (S_{GR}) \) that represents the BD PARCELLEAIRE® buildings (Figure 3 (2)). For this step, the only available matching criterion is the location. The shortest distance between each point describing a DBpedia resource and the polygons representing buildings in the BD PARCELLEAIRE® database was computed. The threshold distance for creating links is fixed to 40 m. This choice was made based on the known planimetric accuracy of the BD PARCELLEAIRE® database and the empiric estimation of DBpedia monuments location accuracy. Let \( A_2 \) be the resulting anchor link set of this matching task.

Finally, between Mérimée \( (S_1) \) resources and DBpedia \( (S_2) \) resources, the matching set \( (M) \) is derived from the anchor links \( A_1 \) and \( A_2 \) created in the pre-

---

24 Data was lifted to RDF locally, thus, “mérimée” and “badresse” are local base URIs, and are not accessible over the Web.

25 Number, repetition index, street name, INSEE code, postal code and address addition.
vious steps (Figure 3 (3)), as defined in our approach in section 4. To simplify the process, anchor links and links between thematic resources are all of type owl:sameAs. Thus, the derived link set was computed by applying a SPARQL query to look for transitivity between $A_1$ and $A_2$: if a DBpedia monument is anchored to some building $b$ and a Mérimée monument is anchored to the same building $b$, then an equivalence link is created between the two monument resources ("indirect links" in Figure 3).

**Results:** The first data matching task between the Mérimée database and the BD ADRESSE® database is performed for both geocoding and anchoring purposes. Among the 3024 monuments initially listed in the Mérimée database, only 2122 have an address. We managed to link 1347 of them to 1964 addresses from the BD ADRESSE® database (one monument can have several addresses). To evaluate the results of this data linking task, we manually created a reference mapping for the monuments of the first arrondissement of the city of Paris. The comparison between the results obtained in this arrondissement and this reference mapping had a precision of 100% and a recall of 90.35%. This precision score can be explained by the parameters chosen for matching Mérimée addresses and BD ADRESSE® resources: high matching scores are preferred and less sure results are eliminated. The loss in recall is partly due to some street naming differences between the two datasets. There are also some character strings in the "address" property of Mérimée database that are not consistent with postal addresses, such as "in front of the church" or "on the main square"., and that could not be handled by our geocoding approach. All geocoded Mérimée monuments are anchored to one or many buildings thanks to the explicit spatial relationships between addresses and buildings.

The second data matching task between DBpedia resources describing Paris historical monuments and the BD PARCELLAIRE® buildings, results in 319 anchor links. The unmatched DBpedia monuments are too far (with distances greater than 40 meters) from any building. The last step of our approach is the derivation step based on previous anchoring results. This step generated 122 links between Mérimée resources and DBpedia resources.

### 5.4. Direct matching

As we did not have any full reference link set between DBpedia and Mérimée, we cannot evaluate the results produced by our approach in absolute terms, but just relatively to the results produced by the direct matching approach. This approach is the classical geometrical approach, that matches resources in the first dataset with the spatially closest resources in the second dataset. To do so, a SPARQL construct query is processed in order to assign to each monument of the Mérimée database the point geometry of the instance of the BD ADRESSE® database to which it has been anchored in the first step. This operation was necessary in order to homogenize the types of spatial references used by the two thematic sources. Then, a matching task is created (Figure 3 (4)) to compare the direct spatial references using the same geometric distance measure with the same threshold as for the second anchoring task. The execution of this task results in a set of "direct links" (Figure 3 (4)).

**Results:** The data matching task is performed directly between the 1347 geocoded instances of Mérimée and the 369 resources retrieved from DBpedia. Running the Silk script on these two datasets results in 217 links to be compared to those produced by our approach. This comparison will be detailed and discussed in section 7.

### 6. Visualizing georeferenced Linked Data

In the case of georeferenced Linked Data, cartographic visualization is a reliable manner to explore the data and enhance their usability. In our use case, it also has the advantage to offer an efficient way to visualize the results of the data linking task. We present here some related works on visualizing georeferenced Linked Data, then we propose an architecture for visualizing them by taking advantage of geographic reference data.

#### 6.1. Related works on cartographic visualization of Linked Data

One of the main goals of the Linked Data architecture is to increase the interoperability and the usability of the data over the Web. To achieve this aim, data visualization represents an important asset to understand the structure and the content of the data. In [14] for ex-
ample, it provides an interactive way for understanding and monitoring the schema harmonization of geodata. Many Linked Data visualization solutions exist currently and vary between general (exploring) applications and data-specific applications (mashups) [2].

Most of the approaches proposed for cartographic visualization of Linked Data aim at exploring and displaying geographic data published according to the RDF model using traditional Web mapping tools. For instance, it is the case of the data visualization application implemented by the Ordnance Survey [26]. On the visualization portal proposed by the GeoLinkedData.es initiative [27], Linked Data with geometries of a "point" type are displayed on a Google map. Semmap [28] is a more recent Web application that can be used to explore SPARQL endpoints in a geographic manner. The application browses a specified graph of data retrieved from an endpoint, in order to suggest a list of facets. Contents related to the chosen facet(s) are then displayed on a base map whose extent is defined by the user. Other Linked Data browsing applications like Lodlive [29] or Palladio [30] provide an easy way to visualize and explore the data. Optionally, they include a point visualization over a base map when the data contains coordinates. Geographic data with different kinds of geometries, and published according to the RDF model, can be visualized with data from LinkedGeoData [31] that have geometries of a "point" type and georeferenced pictures from Flickr, on a GoogleMaps base map within the mashup application presented by [19].

In the approaches listed above, Linked Data with spatial references are just plotted on top of a base map. [15] propose an application prototype that creates geographic summaries from crowdsourced data. They suggest an approach for clustering Foursquare knowledge-base venues. The output summaries are presented as a layer that contains convex hulls: each cluster is represented by a convex hull created from its venues location points. This layer is added over a base map to provide a user friendly interface for summaries exploring. In the approach proposed by [13], thematic Linked Data from DBpedia are firstly linked to geographic data. Then, the geographic data and their related thematic information are used together for creating thematic maps. In this case, the instance matching step is performed within the spatial DBMS PostGIS. DBpedia data about several countries are imported into a PostGIS database and joined with a table of geographic data about countries and their boundaries by comparing country names from both tables.

---

**Note:**

26Example: http://data.ordnancesurvey.co.uk/doc/7000000000018821
27http://geo.linkeddata.es/browser
28http://open-data.europa.es/semmap/
29http://en.lodlive.it/
30http://palladio.designhumanities.org/
31http://linkedgeodata.org/: RDF data representation of information collected by the OpenStreetMap project.
6.2. Visualizing georeferenced Linked Data with geographic reference data

We propose an approach for visualizing Mérimée and DBpedia resources on Paris historical monuments based on the anchor links between these thematic data and the reference geographic data used in the data linking approach. Rather than plotting thematic resources spatial reference points on the top of a base map, we propose to take advantage of the geometry of their corresponding geographic objects. This solution provides us many opportunities for creating thematic maps and co-visualizing data from multiple thematic sources.

Figure 4 summarizes the architecture of our cartographic application. It is in fact an interactive Web client interface, implemented using OpenLayers\(^3\)\(^2\) and the API of Géoportail\(^3\)\(^3\).

Openlayers is a JavaScript library that provides a set of functions to display layers of data stored in geoservers and retrieved respecting the protocols described by OGC (Open Geospatial Consortium). Géoportail API provides several cartographic and orthophotographic base maps produced by the French national mapping agency.

Openlayers also provides a way to create vector layers from data retrieved using other protocols, particularly the HTTP protocol. Therefore it provides an easy way for querying SPARQL endpoints and formatting the retrieved resources before visualizing them on maps. Thus, from geographic data transformed and stored as RDF data for the needs of our data linking approach, we can create vector layers.

The anchor links are used as a way of referring to the thematic information stored in external triple stores. Figure 5 offers an example of co-visualizing geodata with thematic information from corresponding resources in DBpedia and Mérimée. Buildings linked with a Mérimée monument are colored with a shade of blue. The shade represents the construction century of the monument, if indicated in the monument description; otherwise, it is replaced by a black color. The buildings with a thick orange boundary are those linked with a DBpedia monument. Having a blue shade color and a thick orange boundary means that the building is linked to monument descriptions from both sources. For this example we have used an orthophotographic base map.

As shown in Figure 5, this application has an interactive side too; by clicking on the geometry of a linked building, some information are retrieved on the fly from the corresponding source through an HTTP request and is displayed to the user.

This provides us with a user-friendly interface for navigating in data linking results and visually evaluating them.

---

\(^3\)\(^2\)http://openlayers.org/
\(^3\)\(^3\)http://api.ign.fr
7. Discussion

In this section we present a further discussion about the outcomes and the outlooks of our approach.

7.1. Improving the reliability of the spatial matching criterion

Table 1 compares the results produced by both direct and geographic reference dataset based approaches. Every single link created by each approach was checked manually.

<table>
<thead>
<tr>
<th></th>
<th>True Positive links</th>
<th>False Positive links</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using geographic reference data</td>
<td>102</td>
<td>20</td>
<td>122</td>
</tr>
<tr>
<td>Direct linking</td>
<td>91</td>
<td>128</td>
<td>219</td>
</tr>
</tbody>
</table>

With the direct linking approach only 91 out of 217 links were correct (41.93% of precision). With our approach 102 out of 122 links were correct (83.60% of precision). With our approach, a lot of false positive links were avoided, and some more true positive links were discovered, than with the direct matching approach. In the absence of reference link set, we could not quantify the false negative part of the results, and thus recall can not be computed in both cases. However, the fact that our approach resulted on more true positive links implies that it has a better recall than the direct linking approach. With a better precision and a better recall our approach has thus a better F-measure than the direct linking approach.

A deeper analysis of the matching results shows that there were only 76 common links in the intersection of the true positive parts of both links sets, as shown in Table 2. 26 true positive links were detected only by the geographic reference dataset based approach. These links connect resources of DBpedia and Mérimée that actually represent the same real-world historical monument but whose spatial references refer to positions located quite far from each other. Figure 6 illustrates this case. This tends to show that using geographic reference data for thematic data matching could compensate for the heterogeneity of spatial references between different thematic data sources in the data matching process.

Contrary to that, 15 true positive links were detected only by the direct matching approach. They reflect cases where the spatial references associated to DBpedia resources are so inaccurate that they refer to positions in space which are far from the building they are intended to locate and close to other buildings that do not correspond to these historical monuments (e.g. Figure 7). This shows that in some cases the inaccuracy of spatial references can cause some limitations for our approach. In the same time, these cases are good opportunities to detect spatial references errors in DBpedia.

So far, this approach based on reference geographic data showed two positive behaviors. The first is matching distant but equivalent resources (as seen in Figure 6). The second involves avoiding a lot of false links by using the reference geodataset as a disambiguation information source (e.g. Figure 8), particularly when there is a high density of spatial references in the same geographic area.

Our aim is not to set up an approach against the existing multicriteria approaches, but to strengthen the reliability of the spatial references criterion. We believe that one outlook of our approach is to be combined with other criteria: names can be checked in the derivation step, for example.

7.2. Issues related to n:m relations cardinality

Under the hypothesis of using only the spatial references criterion, matching data with a n:m cardinality using the direct approach may be very hazardous. Using a fixed distance threshold \( t \) in this case means that a monument of the first dataset will be linked to the \( n \) closest monuments of the second datasets that are located in a radius \( s \). However, by using geographic reference data, a monument in the first dataset will be linked only to the monuments of the second dataset that share the same anchoring building. An example of n:m links is shown in Figure 9: the description of the monument in Mérimée includes La Sorbonne building and its chapel. These two components are described separately in DBpedia.
Fig. 6. An example of a true positive link derived only by our approach based on geographic reference data.

Fig. 7. An example of a true positive link derived only by the direct approach (false negative in our approach despite their short distance of 6m). In this case, the DBpedia monument is located on the wrong building.
Fig. 8. An example of a true negative link detected with our approach (detected false positive with the direct approach). These are two distinct but very close monuments facing each other.

Fig. 9. An example of n:m link detected by our approach.
7.3. A generic approach?

The gain in precision and recall of our approach is mainly due to some characteristics of the geographic reference data chosen for the process. Well detailed geometries, positional accuracy, consistency and homogeneity of spatial references help offset the lack of geometric detail, positional accuracy and homogeneous modelling of spatial references of thematic data.

The quality of the anchoring results is crucial for the quality of the overall matching process results: a wrong anchoring link may result in wrong derivation and thus in a wrong matching link. This is why we focused more on the precision during the geocoding step. This should be taken into account while configuring anchoring tasks: what distance measure, geocoding method or even multicriteria matching approach would give the best anchoring results? What would be the impact of an increase of the anchoring step recall on the final matching results quality?

In this article, we have used a very basic approach for deriving links from anchoring link sets: a link was created between thematic resources anchored to the same building. This approach could be extended by the use of complementary criteria: thematic resources are linked if they are anchored to the same geographic resource and if they have similar labels. Analogously with approach of [1], which takes advantage of the background ontology structure to infer relationships between source and target ontology concepts, we could also take advantage of spatial relationships between geographic resources to infer links that could not be found by our approach. For example, in area with a high density of adjacent and small buildings (like in the example presented in figure 7), when the positional accuracy of thematic spatial references is low, anchoring thematic resources anchored to adjacent geographic resources may produce some interesting results.

In our approach, the type of geographic reference data used in the process is highly dependent on the domain covered by the thematic data. In the use case, we chose to use geographic reference data about buildings because of the adequacy of this geographic feature type with the theme of monuments. By choosing the suitable reference data, we believe that this approach can be applicable with other use cases, e.g.: public services georeferenced by points could be anchored to buildings, animals GPS tracks could be anchored to preserved natural areas, traffic accidents could be anchored to road sections, etc. However, some adaptations should be taken into account:

- The chosen geographic reference dataset should be adapted by selecting the feature types that are appropriate for the theme and the level of detail of the thematic sources. The level of detail of this reference geographic dataset must be sufficient, i.e. it should have sufficient planimetric precision, resolution and granularity. This means that spatial references in the chosen geographic reference data should be geometrically more detailed and more accurate than those used in thematic data. In addition, the geographic reference dataset should be exhaustive enough on the area covered by the use case.

- Depending on the geometric primitives used for describing the use case resources, other distance measures used in the field of geodata matching could be implemented (e.g. Hausdorff, Fréchet, etc.).

- The links type between geographic and thematic resources should be more relevant than owl:sameAs. In this paper, we used tentatively owl:sameAs links to anchor thematic resources on geographic resources in order to illustrate the approach, but they are not to be published in this state. An outlook for our work is to consider the real link types between these two kinds of resources. This link type is strongly related to the nature of both thematic and geographic resources. For example public services will be linked to buildings with a link hostedIn, data about fauna and flora will be linked to forest area with links of types hasHabitat and growsIn.

8. Conclusion and future works

In this paper, we tackled the problem of matching georeferenced resources over different thematic sources by means of their spatial references. In the context of open and Linked Data, spatial references are likely to have a low level of detail, different granularity or completeness from one source to another, or different positional accuracies or geometric modelling rules from one resource to another. We suggested to use geographic reference data as background knowledge resources in an indirect matching approach, in order to overcome these heterogeneities. We illustrated with a use case how our approach may enhance the matching results, by comparing them with the outcomes of a direct matching approach. We showed how links between thematic and geographic resources may be ben-
icial for thematic data visualization, by taking advantage of the geographic reference data geometries.

The results of our approach are promising, but need to be confirmed by testing the approach against other use cases involving larger datasets or geographic reference data with other types of geometric primitives.

The outcomes of the matching approach bring some opportunities for the visualization of thematic data and their links. Even thought in the case of Paris monuments the Web mapping application remains quite fluid, we think that a bigger amount of data can be technically heavy for the light-client based architecture. We are working on improving this architecture and thinking about better ways to visualize links between data. Multi-scale visualization also should be explored. The links between thematic and geographic resources should be invested more in cartographic generalization. In fact, thematic information can be used as criteria for basic actions of generalization (such as reclassification, amalgamation, deformation, deletion, etc.). With anchor links between thematic and geographic data, thematic information can be retrieved on the fly while processing geographic resources. This can be an interesting outlook for designing on-the-fly generalization applications.

9. Acknowledgments

This work has been partially supported by the French National Research Agency (ANR) within the Datalift Project, under the grant number ANR-10-CORD-009.

References


