Matching and Visualizing Thematic Linked Data: An Approach Based on Geographic Reference Data

Editor(s): Name Surname, University, Country
Solicited review(s): Name Surname, University, Country
Open review(s): Name Surname, University, Country

Abdelfettah Feliachi [a], Nathalie Abadie [b] and Fayçal Hamdi [c]
[a] *COGIT laboratory, Université Paris-Est, IGN / SRIG, Saint-Mandé, France.*
E-mail: Abdelfettah.Feliachi@ign.fr
[b] *COGIT laboratory, Université Paris-Est, IGN / SRIG, Saint-Mandé, France.*
E-mail: Nathalie-F.Abadie@ign.fr
[c] *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 beneficially for data matching or cartographic visualization purposes. Indeed, they may be used as instance matching criteria: 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 criteria 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 the mean of background reference geodataset. We also propose to take advantage of links created between thematic resources and geographic resources for designing better maps for data visualization at different scales.

Keywords: data matching with background knowledge, geographic reference data, data matching for multiscale data visualization

1. Context and objectives

Among the data published on the Web of Data, many resources are associated to a position in the geographic space, 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. For example, in the LOD cloud[1], each of the the properties geo[2]:long and geo:lat of the W3C vocabulary Geo is 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.

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 dataset to another. This

[1] LOD cloud statistics, consulted on 28/07/2014.
[2] `http://www.w3.org/2003/01/geo/wgs84_pos#`

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, different information sources or different levels of accuracy 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 geographic reference databases for matching and visualizing thematic data described by heterogeneous spatial references. We follow the intuition that anchoring thematic resources on geographic reference data should help to identify relationships between these resources. A common way to help users discover data is to provide them with data visualization applications. For georeferenced data, maps are the most intuitive visualization approach. Therefore, we also propose a cartographic application that visualize thematic data together with geographic reference data at different scales 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 fields of georeferenced data matching and cartographic visualization of Linked Data. In section 3, we describe the data matching approach based on a background reference geodataset 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 4. The multiscale cartographic visualization application based on anchor links and generalization techniques is described in section 5. Finally, we conclude and give some perspectives in section 6.

2. Related works

Several research efforts have dealt with the matching and the cartographic visualization of data by means of their spatial references. In the following we describe works that are most closely related to our research.

2.1. *Thematic data matching by means of to 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 [6]. 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 [8].

Many approaches and tools have been proposed for automatically identifying relationships between instances 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 instances 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 instance. Many approaches have already been proposed for comparing direct spatial references (ie. geometries). [9], [3], and [16] have proposed geographic data matching approaches based on point comparisons. [20] and [10] 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 [11].

At the crossroads of these two 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.

2.2. 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. 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[3]. On the visualization portal proposed by the GeoLinkedData.es initiative[4], Linked Data with geometries of a "point" type are displayed on a GoogleMaps base map. Semmap[5] 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[6] or Palladio[7] provide an easy way to visualize and explore the data. Optionally, they include a point visualization over a base map when the data contains spatial location coordinates. Geographic data with different kinds of geometries, and published according to the RDF model, can be visualized with data from LinkedGeoData[8] 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. [14] 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 convex hull is created from the venues location points of a cluster. This layer is added over a base map to provide a user friendly interface for summaries exploring. In the approach proposed by (Owusu-Banahene and Coetzee, 2013)[12], 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 countries names from both tables.

3. A data matching approach based on a background reference geodataset

Instance matching approaches for linking georeferenced data are usually based on a set of measures that could be used to compare spatial references of the same type. These measures include geometric distances between geographic features or string-based measures used for comparing character strings that represent addresses or geographical names. However, different datasets may use different types of spatial references. In such cases, a first step of geocoding must be performed in order to translate available indirect spatial references into direct spatial references: resources described by postal addresses or geonames are matched to geographic reference resources, described by both indirect and direct spatial references. Besides, differences of geometrical accuracy or differences in geometry capture rules can occur even between datasets that use the same type of spatial references. They may cause difficulties for the instance matching process: e.g. 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 completely distinct entities. In order to overcome these heterogeneities, we propose to take advantage of geographic reference databases by using them as background knowledge resources in order to match heterogeneously georeferenced thematic data. Analogously to the approach proposed by [1] in

[3] `http://data.ordnancesurvey.co.uk/`

[4] `http://geo.linkeddata.es/browser`

[5] `http://open-data.europa.eu/semmap/`

[6] `http://en.lodlive.it/`

[7] `http://palladio.designhumanities.org/`

[8] `http://linkedgeodata.org/`: RDF data represenatation of information collected by the OpenStreetMap project.

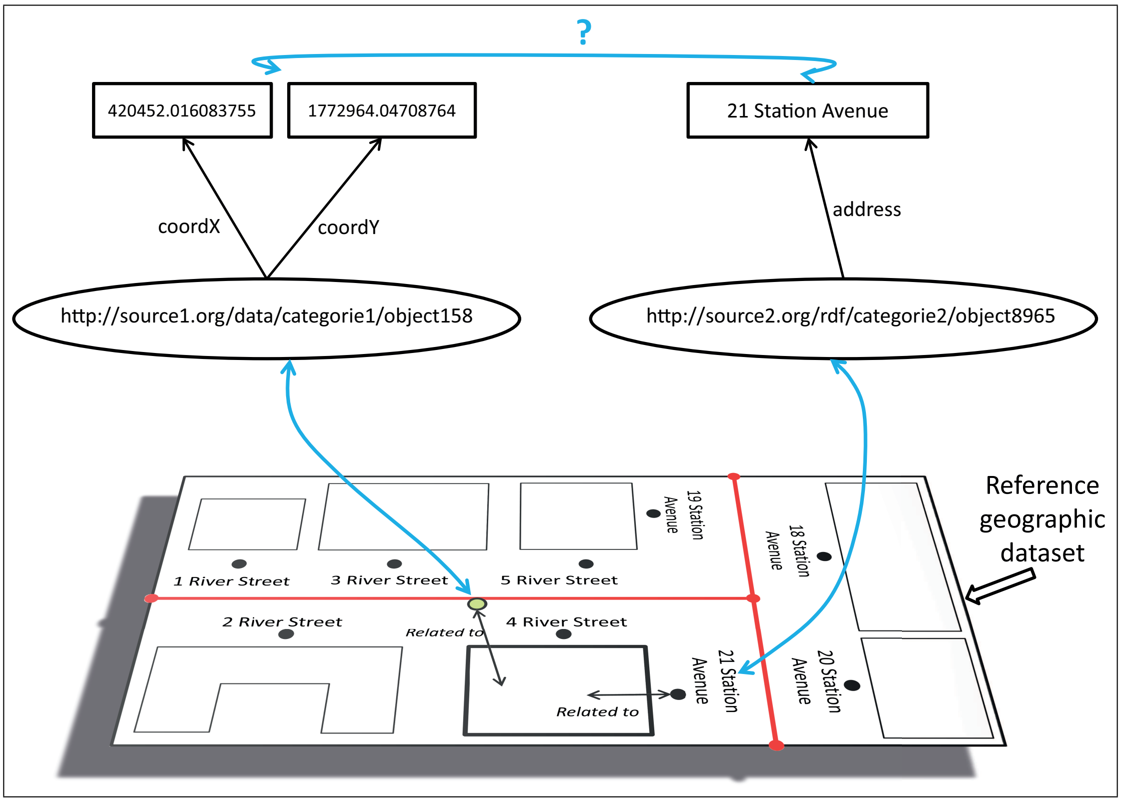


Fig. 1. Using geographic reference data for data linking purposes

the ontology matching field, we propose to first **anchor** the different thematic resources (matching candidates) to the same reference geodataset, and then to **derive** equivalence - or other - relationships between these thematic resources from the anchoring relationships between thematic and geographic resources. Figure 1 presents how an equivalence relationship can be derived between a resource located by coordinates and a resource located by an address. The latter can be geocoded by matching the character string describing its address with a geographic reference database that represents addresses and their coordinates. However the coordinates linked to the second resource from the address geographic reference database are still too far from the coordinates used in the description of the first resource to enable both resources to be matched by direct comparison of their locations.But when the first resource is first linked to the geographic reference database that represents buildings, a relationship between both resources can be derived from the fact that they are anchored to geographic reference resources (namely a building and an addresse) that are close to each other.

Anchoring thematic resources to a reference geodataset requires executing a data linking process that considers the spatial references as the main criteria. 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 also be used for 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 linkage 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 approach.

3.1. Problem definition

We define a dataset S is to be a tuple
$(C, E, I_C, R_T, R_I, R_D, I_R)$ where:
C is a set of concepts c_n with $n \in \{1, 2, .., |C|\}$ (*e.g.* topo[9]:Place or dbpedia-owl[10]:Monument).
E is a set of entities e^X with $i \in \{1, 2, .., |E|\}$ (*e.g.* dbpedia-en[11]:Eiffel_Tower).
I_C is a set of instantiation tuples (e, c) with $e \in E\, and\, c \in C$ (*e.g.* (dbpedia-en:Eiffel_Tower, dbpedia-owl:Monument)).
I_R is a set of relation tuples (e_i, r, e_j) where $(e_i, c_m) \in I_C\ and\ r \in R_I \cup R_D \cup R_T$. e_j may be an entity (in this case $(e_j, c_n) \in I_C$), or a literal.
R_T is set of non-spatial (thematic) properties (*e.g.* prop-fr[12]:style).
R_I is set of indirect spatial referencing properties (*e.g.* prop-fr:adresse).
R_D is a set of direct spatial referencing properties (*e.g.* geometrie[13]:geometry).
$R_I \cap R_D = \emptyset$.

We would like to discover a matching link set M between two thematic datasets $S^A(C^A, E^A, I_C^A, R_T^A, R_I^A, R_D^A, I_R^A)$ and $S^B(C^B, E^B, I_C^B, R_T^B, R_I^B, R_D^B, I_R^B)$. This link set is defined as: $M(S^A, S^B) = \{(e^A, e^B, \alpha) \mid (e^A, c^A) \in I_C^A\ and\ (e^B, c^B) \in I_C^B\ and\ c^A \alpha c^B\ and\ \alpha = \{\equiv\}\}$. This means we match only the entities that have semantically equivalent classes.

3.2. Approach formulation

Let S^A, S^B be two thematic datasets where. We note S^X for $X = A, B$. To be matched by ou approach, S^X must fulfill the condition $R_I^X \cup R_D^X \neq \emptyset$, i.e the data must be spatially referenced.

3.2.1. Geographic reference dataset definition

A reference geodataset S^{GRD} is defined as a dataset $(C^{GRD}, E^{GRD}, I_C^{GRD}, R_T^{GRD}, R_I^{GRD}, R_D^{GRD}, I_R^{GRD})$. A geodataset describes geographic features and their geometries. Depending on the type of the spatial references existing in S^X, S^{GRD} must fullfill the following conditions in order to satisfy the goal of our approach:

- If $R_I^X \neq \emptyset$ then we must have:
  $\exists r^X \in R_I^X, \exists r^{GRD} \in R_I^{GRD} | r^X \equiv r^{GRD}$. i.e. in this case, there must be at least two indirect spatial referencing properties in S^X and S^{GRD} that are semantically equivalent.
- If $R_D^X \neq \emptyset$ then we must have:
  $R_D^{GRD} \neq \emptyset$.

3.2.2. Anchor link set defintion

An anchor link set is computed between S^X and S^{GRD}. It is defined as $AL^X(S^X, S^{GRD}) = \{(e^X, e^{GRD}, \beta) \mid (e^X, c^X) \in I_C^X\ and\,(e^{GRD}, c^{GRD}) \in I_C^{GRD}\ and\ c^X\ \beta\ c^{GRD}\}$.

β is a common sense relationship (in a broader sense) that semantically associates the thematic class $c^X \in C^X$ with the geographic class $c^{GRD} \in C^{GRD}$. In our approach, β is used to anchor resources of type C^X to geographic resource of type C^{GRD}. e.g: e^X = dbpedia-en:Eiffel_Tower, c^X = dbpedia-owl:Monument. e^{GRD} = the building bdtopo:xxx, c^{GRD} =topo:Bati. The coherence between c^X and c^{GRD} is the reason why can seek for an anchor link between e^X and e^{GRD}.

An anchor link set between S^X and S^{GRD} is computed by comparing direct or indirect spatial references. In the indirect spatial reference case, anchoring a thematic resource e^X to a geographic resource e^{GRD} is like performing a geocoding operation on e^X. In both direct and indirect spatial reference cases, e^X and e^{GRD} are compared by the means of distances (metrics). we define a distance as a function

$$\begin{aligned} d : L^X \times L^{GRD} &\to \quad \mathbb{R} \\ (l^X, l^{GRD}) &\mapsto f(l^X, l^{GRD}). \end{aligned}$$

Here, we distinguish two cases:

- *Comparing indirect spatial references:* (*e.g.* addresses, geonames). In this case L^X (resp. L^{GRD}) is a subset of E^X (resp. E^{GRD}) that contains indirect spatial references, they are defined as so:
  $L^X = \{l^X | (e^X, r^X, l^X) \in I_R^X\ and\, r^X \in R_I^X\}$.
  $L^{GRD} = \{l_1^{GRD} \mid (e^{GRD}, r_1^{GRD}, l_1^{GRD}) \in I_R^{GRD}\ and\, r_1^{GRD} \in R_I^{GRD}\ and$
  $(\exists (e^{GRD}, r_2^{GRD}, l_2^{GRD}) \in I_R^{GRD}$ such that
  $r_2^{GRD} \in R_D^{GRD})\}$ (i.e. geographic resources must have both direct and indirect spatial references).
  In this case d is a distance measure for syntaxique similarity (*e.g. Levenstein, Jaro-Winkler, Jaccard...*).

[9] http://data.ign.fr/def/topo#
[10] http://dbpedia.org/ontology/
[11] http://dbpedia.org/resource/
[12] http://fr.dbpedia.org/property/
[13] http://data.ign.fr/def/geometrie#

- *Comparing direct spatial references:* (*e.g.* geometries). Let S^X be a thematic data set with direct spatial references and S^{GRD} the reference geodataset.
  In this case L^X (resp. L^{GRD}) is a subset of E^X (resp. E^{GRD}) that contains direct spatial references, they are defined as so:
  $L^X = \{l^X | (e^X, r^X, l^X) \in I_R^X \; and \; r^X \in R_D^A\}$.
  $L^{GRD} = \{l^{GRD} | (e^{GRD}, r^{GRD}, l^{GRD}) \in I_R^{GRD} \; and \; r^{GRD} \in R_D^{GRD}\}$.
  In this case, d is a geographic or a geometrique distance measure (*e.g. Orthodromic distance, Euclidian distance, Hausdorf distance...*).

3.2.3. Derivation definition

Let $LA^A(S^A, S^{GRD})$ be the computed anchor link set between S^A and S^{GRD} and $LA^B(S^B, S^{GRD})$ the computed anchor link set between S^B and S^{GRD}. A matching link set $M(S^A, S^B)$ can be derived from LA^A and LA^B.

We define the derivation function A from the two anchor link sets LA^A and LA^B as:

$$\begin{aligned} A : LA^A \times LA^B \times SR \rightarrow &\quad M \\ (la^A, la^B, sr) &\mapsto (e^A, e^B, \sigma). \end{aligned}$$

with $la^A = (e^A, e_1^{GDR}, a)$, $la^B = (e^B, e_2^{GDR}, b)$, $sr = (e_1^{GDR}, e_2^{GDR}, s)$ and $\sigma = a \circ s \circ b$.

SR is a set of triples containing couples of geographic resources, and the spatial relationship between them (e.g. a point addresse is "near of" a building). The tuples $(e_i^{GDR}, e_i^{GDR}, =)$ are included in SR. The nature of the σ link between two thematic objects is defined by the composition of a, s and b. Thus, the choise of the spatial relation triples sr and the anchor links a and b, depends on the nature of the wanted matching link σ.

4. 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 accurate than those used for describing geographic features in the geographic reference dataset, such as thematic data described by points instead of polylines or polygons.

To implement and test our approach, we chose two thematic and open datasets on historical monuments in the city of Paris. The first one was retrieved from French DBpedia[14]. The second one, the Mérimée[15] database, is produced and maintained by the French Ministry of Culture and Communication. To select the DBpedia monuments we explored the entities that have dbpedia-fr[16]: Catégorie: Monument_historique_de_Paris as an object for the dcterms[17]: subject property, and the entities related to dbpedia-fr:Catégorie: Monument_parisien by a skos[18]: broader|dcterms: subject path. We selected only the georeferenced entities with prop-fr: longitude and prop-fr: longitude properties or geo[19]: 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 PARCELLAIRE®[20] and the BD ADRESSE®[21] databases produced by the French national mapping agency (IGN[22]). These databases are provided as ESRI shapefiles[23] and freely available for education and research purposes. The BD PARCELLAIRE® represents real-world buildings by geometries of a "polygon" type. It has been chosen instead of other topographic databases such as BD TOPO®[24], because 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 geometry of a "point" type.

4.1. Architecture and tools

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

Reference geodata were translated into RDF using the Datalift[25] platform. This platform provides a

[14] `http://fr.dbpedia.org/`

[15] `https://www.data.gouv.fr/fr/datasets/liste-des-immeubles-proteges-au-titre-des-monuments-historiques/`, downloaded in july 2013.

[16] `http://fr.dbpedia.org/resource/`

[17] `http://purl.org/dc/terms/`

[18] `http://www.w3.org/2004/02/skos/core#`

[19] `http://www.w3.org/2003/01/geo/wgs84_pos#`

[20] IGN's land parcel database.

[21] IGN's adress database.

[22] The National Institute of Geographic and Forest Information.

[23] A geospatial vector data format.

[24] IGN's topographic database.

[25] `http://datalift.org/`

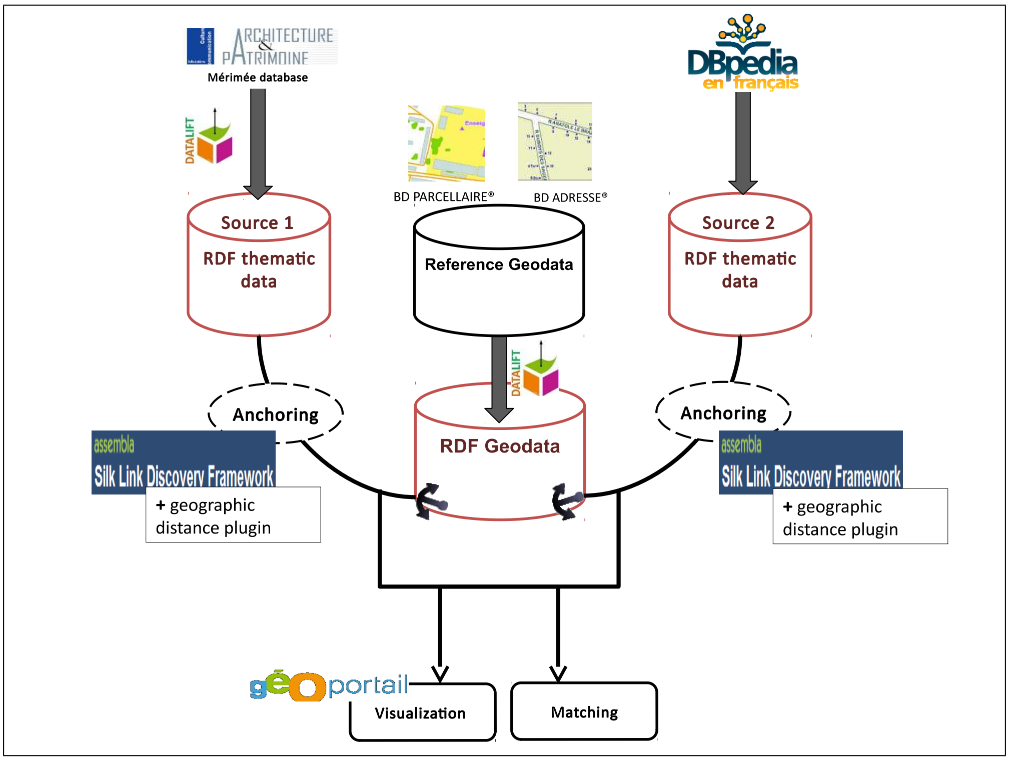


Fig. 2. Implementation of the background reference geodataset matching approach

way to transform data from many models and formats, including geodata (GML, SHP, etc), to 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 CONSTRUCT 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[26] which has been extended with a spatial plugin that can be used 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.

4.2. 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

[26] http://www.assembla.com/spaces/silk/

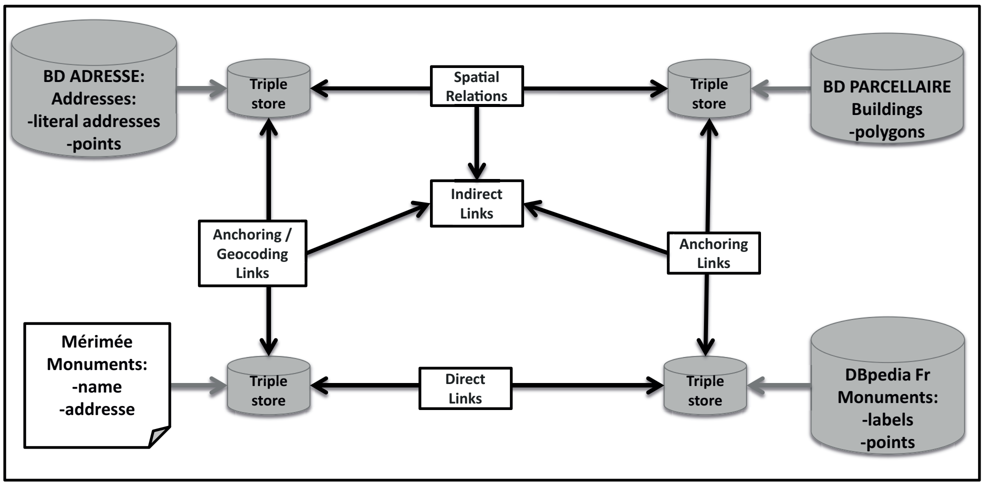


Fig. 3. The different performed linking tasks.

properly[27], we start by projecting all the coordinates to the same appropriate CRS given as a parameter.

This distance measure is implemented and integrated to Silk as a distance plugin using Geotools[28] java library.

4.3. Data matching configuration

Let Mérimée database be the first thematic dataset $S^A(C^A, E^A, I_C^A, R_T^A, R_I^A, R_D^A, I_R^A)$. In this case, $R_D^A = \emptyset$ and $R_I^A =$\{mérimée:adresse, mérimée: codeInsee\}.

Let DBpedia Paris historical monuments be the second thematic dataset $S^B(C^B, E^B, I_C^B, R_T^B, R_I^B, R_D^B,$ $I_R^B)$. $R_I^B = \emptyset$ and $R_D^B =$\{prop-fr:long, prop-fr:lat, geo:long, geo:lat\}.

Let the union of BD PARCELLAIRE® and BD ADRESSE® be the geodataset $S^{GRD} = (C^{GRD},$ $E^{GRD}, I_C^{GRD}, R_T^{GRD}, R_I^{GRD}, R_D^{GRD}, I_R^{GRD})$. In this case $R_D^{GRD} =$ \{geometrie:geometry\} and $R_I^{GRD} =$ \{bdadresse:nomVoie, bdadresse: numero, bdadresse: repetionIndex, bdadresse:codeInsee, bdadresse: codePostal, bdadresse:complement \}[29].

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^A) are georeferenced by indirect spatial references. Therefore, a first data linking process is performed to geocode these addresses, since it was necessary for the two applied approaches. 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^{GRD}) (Figure 3). Addresses elements in each database are combined before being compared using a token wise variant of the Levenshtein measure. Let $LA_1(S^A, S^{GRD})$ be the resulting anchor link set of this first anchoring task.

Once Mérimée instances have been geocoded, they can be matched with DBpedia resources on Paris historical monuments. This data matching task is performed according to two distinct approaches. The first one 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 geometry of the instance of the BD ADRESSE® database to which it has been anchored with the links of $LA_1(S^A, S^{GRD})$. This operation was necessary in order to homogenize the types of spatial references used by the two thematic sources. The second data matching approach is the one that we proposed in section 3, and which involves anchoring thematic data sources on a reference geodataset and then deriving links between thematic resources from the anchor links created between thematic and geographic resources. This matching process is made up of three steps.

DBpedia resources on Paris historical monuments (S^B) are anchored to the instances of (S^{GRD}) that rep-

[27] The Geotools distance that we reuse is based on an Euclidian distance and thus requires projected coordinates.

[28] http://geotools.org/

[29] Data was lifted to RDF locally, thus, "mérimée" and "bdadresse" are local base URIs, and are not accessible over the Web.

resents the BD PARCELLAIRE® buildings (see Figure 3). For this step, the only useable matching criterion is the location. The shortest distance between each point describing a DBpedia resource and the polygons representing buildings in the BD PARCELLAIRE® 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 PARCELLAIRE® database (10 m) and the empiric estimation of DBpedia monuments location accuracy. Let $LA_2(S^B, S^{GRD})$ be the resulting anchor link set of this matching task.

Then, a spatial relationship set SR was computed to create explicit links between the addresses and the buildings contained in S^{GRD}. In the same way as in the previous step, SR was created by matching the addresse points of BD ADRESSE® with the nearest building in BD PARCELLAIRE®. We considerated only the addresses that geocode the Mérimée resources with the anchor links contained in $LA_1(S^A, S^{GRD})$.

Finally, between Mérimée (S^A)resources and DBpedia (S^B) resources, the matching set ($M(S^A, S^B)$) is derived from the anchor links LA_1, LA_2 and the sptial relationship set SR created in the previous steps, as defined in section 3. To simplify the process, Anchor links, spatial relation links and links between thematic resources are all of type owl:sameAs. Thus, the derived link set was computed from by applying a SPARQL query to look for transitivity between LA_1, SR and LA_2: if a DBpedia monument is anchored to building bg and a Mérimée monument is anchored to an addresse ad and bg is spatially related to ad , then an equivalence link was created between the two monument resources .

4.4. Results and evaluation

The first data matching task between the Mérimée database and the BD ADRESSE® database is performed for geocoding 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, a reference mapping was manually created 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.

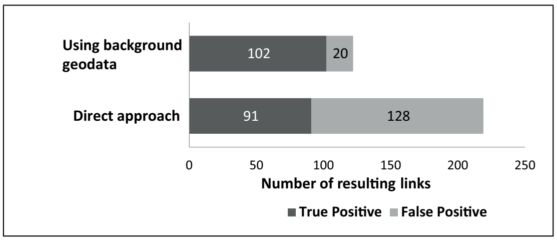


Fig. 4. Linking tasks results from both approaches

The second data matching task is performed directly between the 1347 geocoded instances of Mérimée and the 369 resources retrieved from DBpedia. Running the Silk script with 40 meters as a threshold distance between these two datasets results in 217 links to be compared to those produced by our approach based on a background reference geodataset.

The last data matching task was based on a background reference geodataset. The first data matching subtask 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 second data matching task, between the addresses resources that geocode Mérimée monuments and the BD PARCELLAIRE ® buildings, matched (with a spatial relationship) the totality of the 1964 addresses to a building, due to the geometrical consistency between the BD ADRESSE® and the BD PARCELLAIRE®. The derivation step based on previous anchoring and spatial relationship results generated 122 links between Mérimée resources and DBpedia resources.

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. Figure 4 compares the results produced from both direct and background reference geodataset-based approaches. Every single link created by each approach was checked manually.

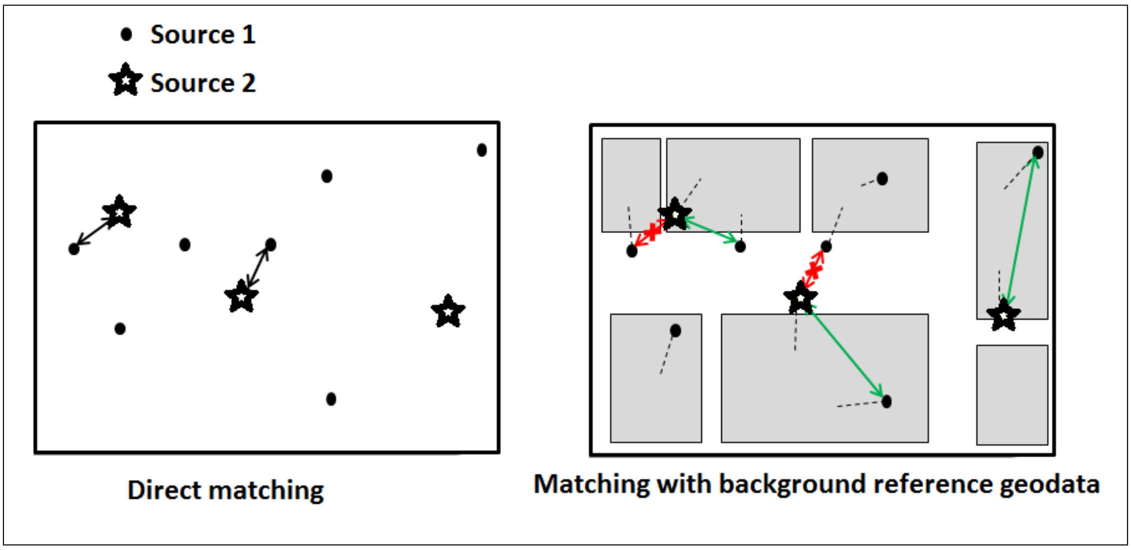


Fig. 5. An example of avoided false links and new gained links using the background reference geodataset approach

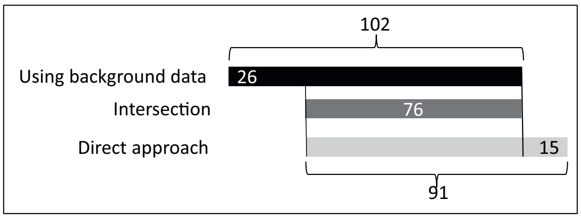
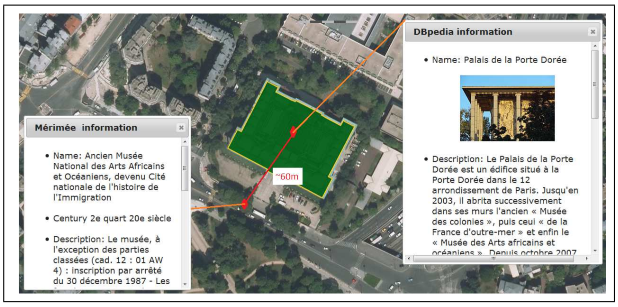


Fig. 6. Comparison of true postive links links from both approaches

For the direct linking approach only 91 out of 217 links (41.93%) were correct. For our approach 102 out of 122 links (83.60%) were correct.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. This result tendes to show that our approach has a better performance than the direct matching approach (better precision and recall scores).



Mérimée information

- Name: Ancien Musée National des Arts Africains et Océaniens, devenu Cité nationale de l'histoire de l'Immigration
- Century 2e quart 20e siècle
- Description: Le musée, à l'exception des parties classées (cad. 12 : 01 AW 4) : inscription par arrêté du 30 décembre 1987 - Les

DBpedia information

- Name: Palais de la Porte Dorée
- Description: Le Palais de la Porte Dorée est un édifice situé à la Porte Dorée dans le 12 arrondissement de Paris. Jusqu'en 2003, il abrita successivement dans ses murs l'ancien « Musée des colonies », puis ceui « de la France d'outre-mer » et enfin le « Musée des Arts africains et océaniens ». Depuis octobre 2007

Fig. 7. An example of a link derived only by the background-geodataset based approach

Besides, there were only 76 common links in the intersection of the true positive parts of both links sets, as shown in Figure 5. 26 true positive links were detected only by the background reference geodataset-based approach. In most cases, these links relate 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. Contrarily to that, 15 links were detected only by the direct matching approach. They correspond to cases where the spatial references associated to DBpedia resources are so inaccurate that they refer to positions in space which are far from the buildings they are intended to locate and close to other buildings that do not correspond to these historical monuments.

This tends to prove that using background reference geodata for thematic-data matching could compensate for the heterogenity of spatial references between different thematic data sources in the data matching process. So far, this approach based on geographic reference data showed two positive behaviors (Figure 7). The first is matching distant but equivalent resources. The second involves avoiding a lot of false links by using the reference geodataset as a disambiguation information source, particularly when there is a high density of spatial references in the same geographic area.

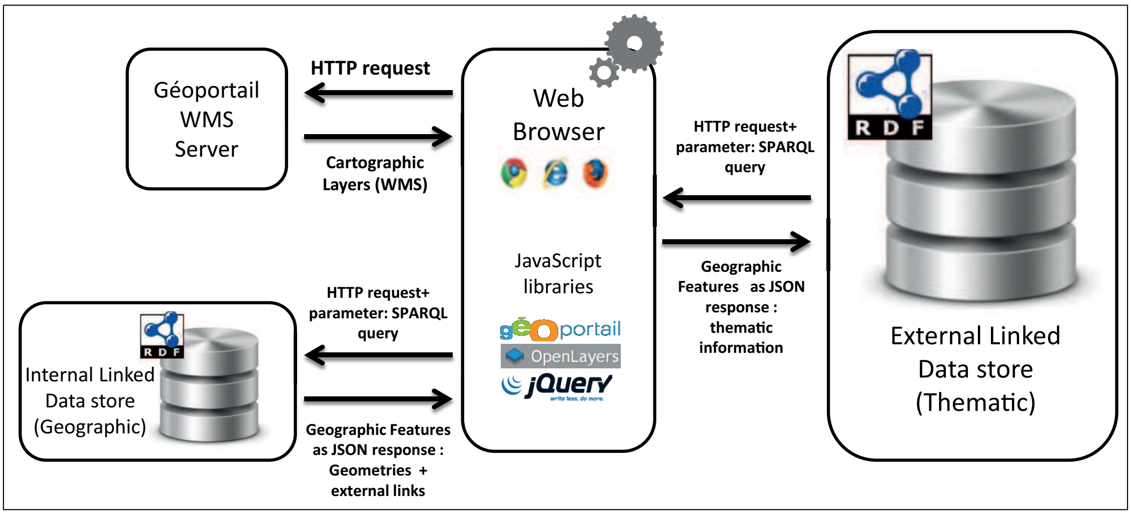

Géoportail WMS Server
HTTP request
Cartographic Layers (WMS)
Web Browser
JavaScript libraries
géoportail
OpenLayers
jQuery
write less, do more.
HTTP request+ parameter: SPARQL query
Geographic Features as JSON response : thematic information
RDF
External Linked Data store (Thematic)
HTTP request+ parameter: SPARQL query
Geographic Features as JSON response : Geometries + external links
RDF
Internal Linked Data store (Geographic)

Fig. 8. Web mapping application architecture for visualizing thematic data with reference geodataset

5. Visualizing georeferenced Linked Data with a background reference geodataset

In the case of georeferenced Linked Data, cartographic visualization is a reliable manner to explore the data and enhance their usability. In our case of study, it also has the advantage to offer an efficient way to visualize the results of the data linking task.

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 geographic reference 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, co-visualizing data from multiple thematic sources and using amalgamation techniques to produce a better multiscale visualization.

5.1. Architecture and tools

Figure 8 summarizes the architecture of our cartographic application. It is in fact an interactive Web client interface, implemented using OpenLayers[30] and the API of Géoportail[31] .

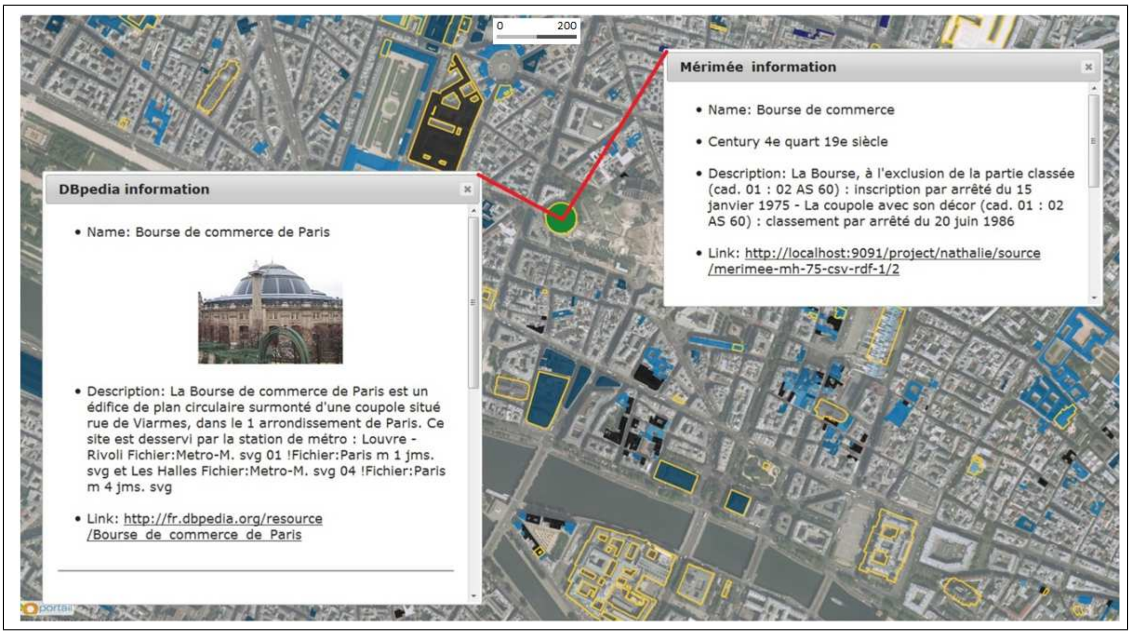
Openlayers is a JavaScript library that provides a set of functions to display layers of data stored in geoservers and retrieved respecting the protocols described by OGC (Open Geospatial Consortium). Géoportail API provides several cartographic and orthophotographic base maps produced by the French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

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 9 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 used an orthophotographic base map.

As shown in Figure 9, this application has an interactive side too; by clicking on the geometry of a linked building, some information is retrieved on the fly from the corresponding source through an HTTP request and is displayed to the user.

[30] http://openlayers.org/
[31] http://www.geoportail.gouv.fr/



Mérimée information

- Name: Bourse de commerce
- Century 4e quart 19e siècle
- Description: La Bourse, à l'exclusion de la partie classée (cad. 01 : 02 AS 60) : inscription par arrêté du 15 janvier 1975 - La coupole avec son décor (cad. 01 : 02 AS 60) : classement par arrêté du 20 juin 1986
- Link: http://localhost:9091/project/nathalie/source /merimee-mh-75-csv-rdf-1/2

DBpedia information

- Name: Bourse de commerce de Paris
- Description: La Bourse de commerce de Paris est un édifice de plan circulaire surmonté d'une coupole situé rue de Viarmes, dans le 1 arrondissement de Paris. Ce site est desservi par la station de métro : Louvre - Rivoli Fichier:Metro-M. svg 01 !Fichier:Paris m 1 jms. svg et Les Halles Fichier:Metro-M. svg 04 !Fichier:Paris m 4 jms. svg
- Link: http://fr.dbpedia.org/resource /Bourse_de_commerce_de_Paris

Fig. 9. An example of co-visualizing thematic data from two sources and displaying on the fly retrieved information

5.2. *Multiscale visualization*

At a large scale, building geometries can be visualized directly as shown in Figure 9. Indeed, at this level of detail, the geometries remain clearly distinguishable. Problems occur when zooming out to smaller scales. Geometries become too small for the users' visual acuity. In order to provide an easily readable and clear cartography over different scales, some approaches, called generalization approaches, are generally used. Generalization is a synthesis process that can be compared to a text summarizing process, where we try to reduce the number of words while keeping the main ideas [15]. The cartographic generalization aims at simplifying the information while changing the viewing scale, on condition of keeping the same meaning of the map.

We want to take advantage of these approaches to create a multiscale cartographic visualization application relying on both reference geodata and information retrieved from the linked thematic sources. Geometries of reference geodata are used, not only for supporting co-visualization of thematic information, but they are also grouped based on their spatial proximity and on the thematic information held by their corresponding thematic resources for creating new amalgamated geometries which are visible at smaller scales.

We adapt the cluster amalgamation approach described by [13]. For our purposes, the geometries of buildings are amalgamated depending on their spatial proximity and the century of construction of their related historical monument as described in the Mérimée database.

The implemented algorithm (Algorithm 1) is based on three main steps:

- *Grouping*: The first step consists in grouping the geographic features on the basis of their spatial proximity and a thematic similarity criterion. Spatial proximity is assessed by creating a buffer of a radius d around the geometry of each geographic feature and looking for any eventual overlapping between resulting buffers. Thematic similarity is assessed by comparing the values of some thematic properties. In our case, this task is performed by simply checking the correspondance of the construction century between the monuments. The output of this step is a list of sets of geographic features and their corresponding thematic information.
- *Amalgamating*: The second step consists in replacing every set of geographic features by a single geographic feature. This implies creating a new single geometry for this resulting feature which provides the same visual effect as the group of geometries of the initial features. For that purpose, we have applied an algorithm based on a Delaunay triangulation and developed by [5] for

```
Data: A set of thematic features $S$. for $s \in S$, $s.th$ is the thematic atribute and $s.geom$ is the geometrique
    attribute
Result: A set of agregated features for small scale visualization $S_a$.
Initialise a grouping distance threshold $d$;
Initialise a feature area filter $f$;
Initialise a distance $t$ for concave hull computation;
Duplicate $S$ on set $S^B$;
//Grouping.
foreach $s$ in $S$ do
    foreach $s_2$ in $S^B$ do
        if $s \neq s_2$ AND $s.th = s_2.th$ AND distance( $s.geom$ , $s_2.geom$) $\leq d$ then
            $s.geom$ gets the union of $s.geom$ and $s_2.geom$;
            The other attributes of $s$ are concatenated with $s_2$ attributes;
            $s_2$ is deleted from both $S$ and $S^B$;
            Break the current loop;
        end
    end
end
//Amalgamating and first step of filtering.
foreach $s$ in $S$ do
    Densify $s.geom$ with a distance stricly smaller than $t$;
    $s.geom$ gets the computed concave hull of $s.geom$ with the deletion parameter $t$;
    if the area of $s.geom$ is $\leq f$ then
        Remove $s$ from $S$;
    end
end
//Second step of filtering.
Sort $S$ in a descendant order according to the geometries areas;
foreach $s$ in $S$ do
    forall the $s_2$ in $S$ with $s_2.geom$ smaller than $s.geom$ do
        if $s.geom$ overlaps $s_2.geom$ AND the differentiation between $s_2.geom$ and $s.geom$ is smaller than $f$
        then
            Remove $s_2$ from $S$;
        end
    end
end
```

Algorithm 1: Features amalgamation for smaller scale visualising

creating a concave hull from a set of points. An already existing java implementation[32] of this algorithm has been reused. It takes any type of geometric primitive as an input. The principle of this algorithm is to use a chosen length threshold t to remove edges from the Delaunay triangulation. This triangulation being created from a set of points , the algorithm has a better behavior when it is applied to a set of points, than when it is applied to a set of polygons: the result of the deletion keeps all the points (the vertexes in the case of polygons) while it is not the case for the polygons' original segments (segments longer than t are deleted). The solution applied to overcome this issue is to start by a densification of the geometries with a distance smaller than t before applying this algorithm. The rest of the attributes of each amalgamated object are formed from the concatenation of the attributes of its components.

[32] www.rotefabrik.free.fr/concave_hull/

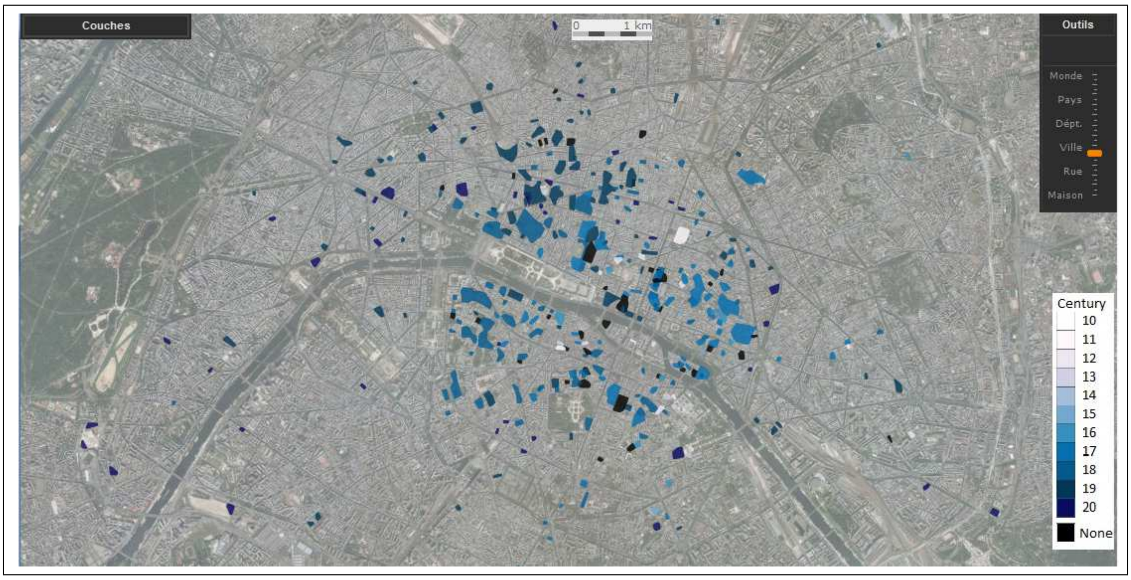


Fig. 10. Monuments visualization at quarter level.

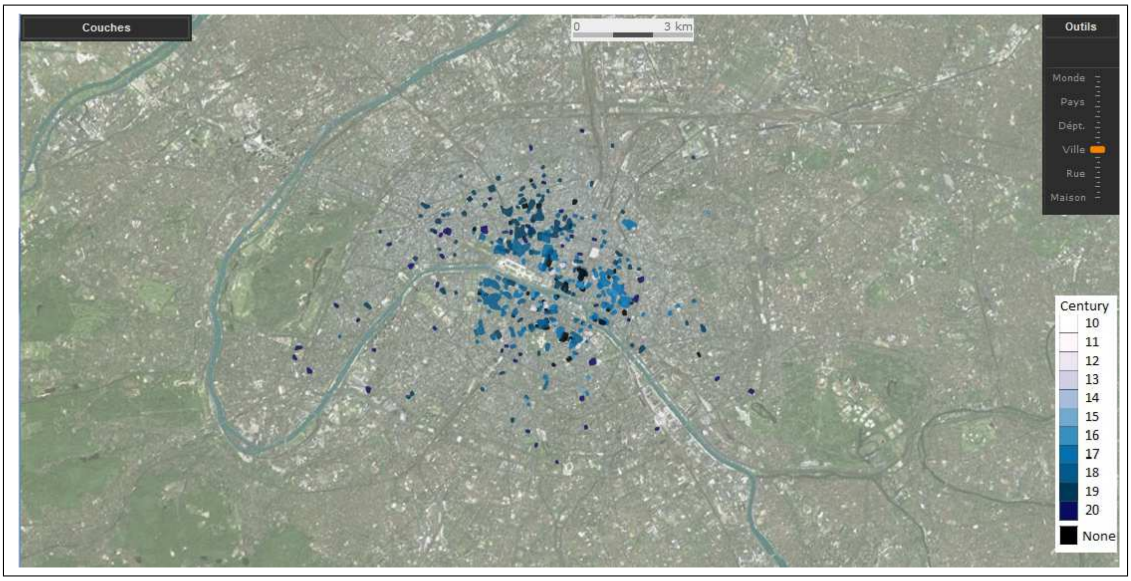


Fig. 11. Monuments visualization at city level.

- *Filtering*: The last step consists in filtering all the objects we consider irrelevant for the cartographic visualization application. The purpose of our amalgamation approach is to create geographic features represented by geometries which are large enough to be displayed at smaller scales than the geometries of the original features. Therefore, resulting geographic features are filtered by comparing their areas to a chosen filter threshold f. This step also handles the issue of overlapping amalgamated objects so that:

  ∗ The wider geometries are put forward.
  ∗ If two geometries are overlapping and if their differentiation is smaller than the filter threshold f then the smaller feature is excluded from the visualization.

The results are shown in Figure 10. d and f were adapted to a quarter level (between road level and city level of Geoporail) . At a much smaller scale (city level or smaller), other modifications could be used, such as enlarging the amalgamated geometries and increasing the grouping distance d and the filtering threshold f (Figure 11).

6. Conclusion

In this paper, we proposed an indirect matching approach for georeferenced objects over different thematic Linked Data sources. Differently from the direct matching approach, we used a reference geodataset as a background source in order to overcome heterogeneities and imprecisions related to location representation. We compared the results of the two approaches for a same pair of datasets and showed how using a background geodataset enhances the performances of the matching process. We showed in a last part an example of how matching using a reference geodataset can be beneficial for some applications. We took, for instance, cartographic visualizing and implemented a multiscale Web mapping application. For this application, we suggested a method for grouping and amalgamating the geometries of geographic reference resources.

In this paper, we used tentatively owl:sameAs links to anchor thematic resources on geographic resources. A perspective for our work is to consider the real link natures between these two kinds of resources. We believe that this will improve both the matching and the visualization of resources. Another perspective is to take into consideration the topological relations between geographical resources when looking for indirect links to decrease the effects of location imprecision; we have seen, for instance that some resources were matched with the direct approach and were not matched with our approach because one resource was not located near of its real corresponding building, but near of a neighboring building.

7. 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

[1] Z. Aleksovski,W. ten Kate and F. van Harmelen, *Exploiting the structure of background knowledge used in ontology matching*, in Ontology Matching Workshop, 5th International Semantic Web Conference, Athens, Georgia, USA, 2006, pp. 13–24.

[2] G. A. Atemezing and R. Troncy, *Towards Interoperable Visualization Applications Over Linked Data*, in Talk Given at the 2nd European Data Forum (EDF), Dublin, Ireland, 2013.

[3] C. Beeri,Y. Kanza, E. Safra and Y. Sagiv , *Object fusion in geographic information systems*, in VLDB, (2006), pp. 816–827.

[4] A. Bel Hadj Ali, *Qualité géométrique des entités géographiques surfaciques - Application à l'appariement et définition d'une typologie des écarts géométriques*, Ph. D. thesis, Université de Marne-la-Vallée, 2001.

[5] M. Duckham, L. Kulik, M. Worboys and A. Galton, *Efficient generation of simple polygons for characterizing the shape of a set of points in the plane*, Pattern Recognition, 41(**10**), (2008), 3224–3236.

[6] A. Ferraram, A. Nikolov, and F. Scharffe , *Data linking for the semantic Web*, Semantic Web : Ontology and Knowledge Base Enabled Tools, Services, and Applications,(2013),169.

[7] S. Hahmann and D. Burghardt, *Connecting linkedgeodata and geonames in the spatial semantic Web*, in Proc. of GIScience 2010 Extended Abstracts, Purves, R. and Weibel, R.(eds.), (010)pp. 28–34.

[8] T. Heath and C. Bizer, *Linked data : Evolving the Web into a global data space*, Synthesis lectures on the semantic Web : theory and technology 1(1), (2011),1–136.

[9] M. Minami, *Using ArcMap*, ESRI Press,(2000).

[10] S. Mustière and T. Devogele, *text Matching networks with different levels of detail*, Geoinformatica 12(**4**), (2008), 435–453.

[11] A.-M. Olteanu, *Appariement de données spatiales par prise en compte de connaissances imprécises*, Ph. D. thesis, Université de Marne-la-Vallée, 2008.

[12] W. Owusu-Banahene and S. Coetzee, *Integrating linked open data into open source Web mapping*, in International Cartographic Conference. International Cartographic Association, 2013.

[13] N. Regnauld, *Algorithms for the amalgamation of topographic data*, in Proceedings of the 21st International Cartographic Conference, Durban, South Africa, 2003.

[14] G. Rizzo, G. Falcone, R. Meo, R. G. Pensa, R.Troncy, and V. Milicic, *Geographic Summaries from Crowdsourced Data*, in European Semantic Web Conference , Crete, 2014.

[15] A. Ruas,(Ed.), *Généralisation et représentation multiple*, Information Géographique et Aménagement du Territoire, Hermes Lavoisier, 2002.

[16] E. Safra, Y. Kanza, Y. Sagiv and Y. Doytsher, *Efficient integration of road maps*, in Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems, GIS '06, New York, NY, USA, (2006),pp. 59–66. ACM.

[17] J. Salas and A. Harth, *Finding spatial equivalences accross multiple rdf datasets*, in Terra Cognita Workshop, 2011, pp. 114–126.

[18] A. Samal, S. Seth and K. Cueto1, *A feature-based approach to conflation of geospatial sources*, International Journal of Geographical Information Science 18(**5**), (2004),459–489.

[19] P. Shvaiko, F. Farazi, V. Maltese, A. Ivanyukovich, V. Rizzi, D. Ferrari and G. Ucelli, *Trentino government linked open geo-data*

: a case study, in The Semantic Web-ISWC 2012, (2012), pp. 196–211, Springer.

[20] V. Walter and D. Fritsch, *matching spatial data sets : a statistical approach*, volume 13, 1999,pp. 445–473.