Assessing the Positional Planimetric Accuracy of DBpedia Georeferenced Resources

Abdelfettah Feliachi, Nathalie Abadie, Fayçal Hamdi

Assessing the positional planimetric accuracy of DBpedia georeferenced resources

Abdelfettah Feliachi[1], Nathalie Abadie[1], and Fayçal Hamdi[2]

[1] LaSTIG-COGIT, Université Paris-Est, IGN / SRIG, Saint-Mandé, France
[2] CEDRIC - Conservatoire National des Arts et Métiers, Paris, France

Abstract. Assessing the quality of the main linked data sources on the Web like DBpedia or Yago is an important research topic. The existing approaches for quality assessment mostly focus on determining whether data sources are compliant with Web of data best practices or on their completeness, semantic accuracy, consistency, relevancy or trustworthiness. In this article, we aim at assessing the accuracy of a particular type of information often associated with Web of data resources: direct spatial references. We present the approaches currently used for assessing the planimetric accuracy of geographic databases. We explain why they cannot be directly applied to the resources of the Web of data. Eventually, we propose an approach for assessing the planimetric accuracy of DBpedia resources, adapted to the open nature of this knowledge base.

1 Context and objectives

Many data Web resources are associated with spatial location, directly (using coordinates or geometric primitives) or indirectly (with an address or a place name). When it is direct, this spatial reference is often used to produce spatial data analysis, cartographic visualization or georeference other resources. As any properties used to describe resources published on the Web of data, this spatial reference can be used to identify resources representing the same entity and interconnect them. In the latter case, its use is based on a simple hypothesis, in accordance with the first Tobler geography law of [13]: the closer two spatial references are, the bigger are the chances that their related resources represent the same real world entity.

In the Web of data, many resources are associated with some location on Earth, either directly (using coordinates or geometric primitives) or indirectly (with an address or a place name). Direct spatial references can be used for spatial data analysis, cartographic visualization or to georeference other resources. Like any other properties, spatial references can also be used to evaluate the similarity of the resources they describe for data linking purposes. Two resources of

This is the author-created version of the article. The final authenticated version is available online at https://doi.org/10.1007/978-3-319-70625-2_21

similar types described by similar spatial references may thus be considered as representing the same real world entity and therefore be linked to each other.

Each of the possible use cases involving the spatial references associated with the data requires taking into account their absolute positional planimetric accuracy, that is to say the difference between the locations provided by the spatial references and the locations considered as true [9]. [1] defines minimal positional planimetric accuracy values that Web gazetteer data should respect to provide accurate spatial references for resources that reuse their coordinates. However, information about the quality of these spatial references, often provided in the metadata of geographic datasets is, to the best of our knowledge, almost never available for resources published in the Web of data. Their evaluation seems little considered since it does not appear in the main state-of-the-art works dealing with the Web of data quality issues, such as [15].

This issue of assessing the quality, reliability and credibility of the Web of data resources is at the origin of the upper layers of the semantic Web "layer cake". Associating credibility ("Trust" layer) to data is based on provenance information and the facts inferred from it ("Proof" and "Logic" layers). Hence, W3C has published PROV [7], a set of recommendations for exchanging interoperable provenance data on the Web. They include a conceptual data model, the associated OWL2 ontology, a serialization text language, data integrity constraints, and so on.

In this article, we propose an approach to evaluate the positional planimetric accuracy of spatial references associated with the Web of data resources, adapted to the open nature of the Web data sources. We perform tests on resources extracted from the French DBpedia that describe the monuments of Paris.

2 Existing approaches for evaluating the absolute positional planimetric accuracy of georeferenced data

2.1 Approaches for geographic databases

Direct spatial references are used to provide a quantitative description of the characteristics of the real-world geographic entities, such as their location, shape, size or orientation . The representation of these characteristics through geometries depends on two main factors: the level of detail expected for the database, that is to say the level of geometric and semantic abstraction used in this database to represent real-world geographic entities [11], and the limitations dues to the resolution of the raw data sources used for geometry capture. Both may lead to simplified representations of the real world entities and thus to notable differences from one database to another [5].

`http://fr.dbpedia.org/`. Data extracted in December 2013 and containing 625 resources.

For the sake of brevity, we will use the term "geometries" instead of "direct spatial references" in the remainder of this article.

Rules, describing how geographic entities should be represented by geometries with boundaries captured along some given characteristic element of their shape, are provided to data-entry operators in order to guarantee a good homogeneity of the captured geometries. Besides, the absolute positional planimetric accuracy of the geometries depends strongly on the raw data used for their capture. In the case of geographic data provided by traditional data producers (e.g. national mapping agencies), the whole geometry acquisition process is designed to obtain data with a predefined absolute positional planimetric accuracy.

ISO 19157 standard on geographic data quality distinguishes two types of quality evaluation methods. The indirect methods are based on knowledge about data provided either by external sources or by the experience gained about data possibilities and limitations. The external knowledge sources may be qualitative metadata or genealogic information. This is closely related to the motivations of the PROV [7] recommendation: providing knowledge about the data lifecycle to assess the quality of a data source. Direct evaluation methods are based on the inspection of the data. Data may be analyzed on their own (absolute method) or they may be compared with other data sources (relative method). Relative methods require a high quality reference data source. These methods can be applied to the entire dataset or to a representative sample. The catalog of standardized quality measures of the ISO 19157 standard offers many numerical methods to evaluate the positional planimetric accuracy of the geometries. However, these methods are only applicable to data sources with a quite homogeneous data capture process. Otherwise, the standard recommends categorizing data by cause of heterogeneity and applying the chosen method independently to each subset.

2.2 Approaches for linked georeferenced data

Unlike in geographic databases, geometries associated with Web of data resources are not the main piece of information in their description. In recent years, many vocabularies have been proposed to represent geographic features geometries on the Web [2] and standardization work is currently under way [12]. Geometries may come from geographic databases converted to RDF and published by national mapping agencies. But most data sources on the Web include geometries of various provenances: Geonames (`http://www.geonames.org/`) gathers several traditional geographic datasets with crowdsourced data and the large georeferenced data sources DBpedia, Yago and LinkedGeoData are derived from the crowdsourcing projects Wikipedia and OpenStreetMap. When the data come from various sources, geometry acquisition processes are less controlled or less known. It may then be difficult to assess the positional planimetric accuracy of geometries, which can vary significantly from one geometry to another.

Works on linked data quality focus on their compliance with good practices in the Web of data. The survey on the quality assessment measures proposed by [15] gives priority to measures on dereferencing, licensing and interconnecting

This is the case for data published by the Ordnance Survey `http://data.ordnancesurvey.co.uk` or IGN Spain `http://geo.linkeddata.es/`

issues. Then come measures about the intrinsic quality of data, their fitness for user's needs, and their representation. However, none of the presented measures addresses the quality of the spatial references associated to resources. Nevertheless, the approach for the detection of aberrant numerical values in DBpedia proposed by [14] allows the identification of outliers in coordinates or altitudes values. In addition, [1] evaluates the positional planimetric accuracy of GeoNames coordinates with respect to their number of decimals.

Assessing the quality of the location information produced by crowdsourcing projects is a key issue addressed in many studies. [8] proposes an approach to evaluate the positional planimetric accuracy of geotags associated with FlickR images of remarkable buildings. The positional planimetric accuracy of the geotags is evaluated by calculating the average distance between their coordinates and those provided by the Wikipedia article describing the photographed buildings. The choice of Wikipedia as a reference data source for buildings location is, unfortunately, not discussed. Finally, OpenStreetMap (OSM) is probably the volunteered geographic data source whose quality has been the most extensively studied [3][6].[3] provides, for the OSM data, a set of indirect quality measures based on the available genealogy metadata. [6] evaluates the quality of OSM data on French territory using a set of standard direct measures. The positional planimetric accuracy of three OSM data samples (respectively with point, polyline and polygon geometries) is estimated by computing the mean distance between each geographic feature of these samples and their counterparts retrieved from a reference dataset produced by the French national mapping agency.

3 Genealogy of DBpedia Resources Geometries

DBpedia resources describing real-world geographic entities are georeferenced using points extracted from Wikipedia. Therefore, their positional planimetric accuracy depends directly on the coordinates provided by the Wikipedia contributors. Nearly 15% of Wikipedia articles are georeferenced and 16.25% derive their coordinates from Wikidata.

The "WikiProject Geographical Coordinates" aims to improve the quality of the Wikipedia articles coordinates by providing recommendations for their capture. First, they advocate the use of trusted sources, such as the geoportals of national mapping agencies, to find reliable coordinates. In addition, they indicate which characteristic shape element should be localized, depending on the type of geographic entity described by each article: the center of the inhabited area for municipalities or the main entrance for buildings. Recommendations also provide rules to round the coordinate values in order to have a coordinates precision consistent with the size of the geographic entities. For example, the coordinates

Source: https://fr.wikipedia.org/wiki/Projet:Géolocalisation

`https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geographical_coordinates`

Expressed in the WGS84 coordinates reference system

of a geographic entity located in France with a length between 50 and 100 meters should ideally have 4 decimals when they are expressed in decimal degrees.

These recommendations explain how to add these coordinates in the body of a Wikipedia article or in its infobox using the predefined template "Template:Coord". In addition to coordinates, it includes metadata. Most of them, such as "type", "dim" or "scale", are intended to define the most appropriate map scale to visualize the geographic entity. The metadata "source" indicates the provenance of the entered coordinates: coordinates obtained from the *Geographic Names Information System* should therefore mention "source:GNIS". The geographic coordinate extractor used by the French DBpedia searches for this template in Wikipedia articles and infoboxes and only keeps the coordinates to produce triplets based on geo:long, geo:lat and georss:point properties.

There is no guarantee that all Wikipedia contributors are aware of these recommendations and that they apply them correctly. Similarly, there is no evidence that they rely on data from national mapping agencies to determine the coordinates associated with articles or that they follow the recommendations about coordinates precision. Finally, nothing forces them to fill out the "Template: Coord" metadata, and even if they do, the DBpedia coordinates extractor does not keep them. Direct quality methods recommended by ISO 19157 were designed to evaluate the overall positional planimetric accuracy of homogeneous datasets. These methods are not directly applicable here. In addition, with no reliable genealogy metadata, indirect assessment methods cannot be used.

4 Direct evaluation of the absolute positional planimetric accuracy of DBpedia resources

In order to overcome the lack of genealogy metadata and to provide a single positional planimetric accuracy estimation for each resource, we propose to adapt the direct methods designed to assess the absolute positional planimetric accuracy of traditional geographic data to the specific case of Web of data resources. These require to be treated on a case-by-case basis.

4.1 The proposed approach

We propose a two-step approach. The first step aims to find, for each resource to be evaluated, which characteristic element of its shape was pointed out to define its coordinates. Then, the distance between each DBpedia point and its supposed counterpart within a reference geographic dataset is computed.

Using a point to locate a geographic entity provides a highly simplified geometric representation of that entity. Moreover, this requires deciding what characteristic element of its shape should be pointed out preferably. Therefore the

`https://en.wikipedia.org/wiki/Template:Coord`
`http://fr.dbpedia.org/doc/listeExtracteurs.html`
`http://www.w3.org/2003/01/geo/wgs84_pos#`

evaluation of the positional accuracy of such a point cannot be done without taking into account this representation choice: its coordinates must be compared with those of a point captured in the same way, but with a higher accuracy.

In the case of historical monuments, the recommendations made by the project "WikiProject Geographical Coordinates" seem to indicate to localize the entrance of each monument, preferably by picking its coordinates on the relevant national mapping agency geoportal. However, if we plot the coordinates of Paris DBpedia monuments on an IGN orthophotographic base map, we observe significant shifts with respect to the expected localization (see figure 1). It seems therefore impossible to rely on the recommendations of the project "WikiProject Geographical Coordinates" to determine which point of the shape of historical monuments is represented by DBpedia coordinates.

Fig. 1. DBpedia historical monuments, in yellow, on IGN orthophotographic base map.

We therefore propose to formulate hypotheses on the choices made by contributors when entering the coordinates of the monuments. Then, we compare monuments coordinates with points corresponding to each category of representation choice, selected from a reference geographic dataset. This provides us with indicators for classifying each monument coordinates by category of representation choice. This step can be carried out with common spatial analysis tools and a supervised classification method.

The second step of our approach consists in comparing the evaluated coordinates with those of the points identified as corresponding to the representation choice made by the contributors and selected within a reference geographic dataset having a better and well documented positional planimetric accuracy.

4.2 Implementation

When the points used for locating DBpedia historic monuments are plotted on an IGN orthophotographic base map, three types of representation choices can

be identified: close to the center of the building considered as a historical monument, close to its facade and finally near the road centerline in front of its facade. The two first representation choices correspond respectively to two types of recommendations from the "WikiProject Geographical Coordinates": for geographical entities with a broad spatial extent coordinates should be captured at their center, and for buildings at their main entrance. The third representation choice is a common practice for capturing addresses.

Relevant spatial indicators must then be defined to decide to which category of representation choice belongs each point [10]. We used two IGN geographic datasets, chosen for their consistency with the "WikiProject Geographical Coordinates" guidelines: the buildings from the BD PARCELLAIRE® and the roads from the BD TOPO®. Using the PostgreSQL/PostGIS database management system, we calculated three indicators based on the distance between the DBpedia points and the reference geometries: the distances to the barycenter and to the facade of the nearest building and the distance to the nearest road segment. Since the sizes of the buildings and the roads vary considerably according to the districts of Paris, we normalized these values (see figure 2).

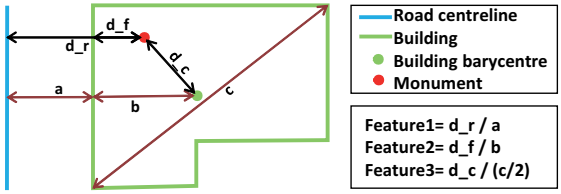

Fig. 2. Learning features for geometry capture rules.

We then manually prepared a learning sample of about 30 monuments of each type. Finally, we used Weka as it implements the most commonly used supervised classification algorithms. We applied several of the algorithms available in Weka to our data and we manually checked the results.

Finally, we computed the absolute positional planimetric accuracy of each DBpedia point by adding two values: its distance to the point in the geographic reference dataset identified as the representation choice made by its contributor - the building facade, the building barycenter or the road centerline - and the absolute positional planimetric accuracy of the reference geographic dataset.

Cadastral database produced by the IGN

Database describing the topography of the French territory and its infrastructures produced by the IGN

`http://www.cs.waikato.ac.nz/ml/weka/`

4.3 Results and discussion

The table 1 presents the results of the four classification algorithms tested in order to assign to each DBpedia point a category of representation choice (see [4] for details about the classifiers). These results are compared with a manual classification. The four tested algorithms provide good results, which tends to validate our choice of indicators.

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

Table 1. Learning results for some applied classification algorithms.

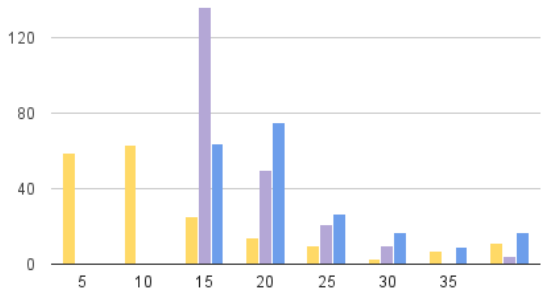

Fig. 3. Frequency of positional planimetric accuracy values for DBpedia's historical monuments in Paris according to the representation choice. The abscissa axis indicates the maximum accuracy values in meters and the ordinates axis the number of DBpedia resources. The yellow, mauve, and blue sticks represent the resources captured respectively at the road centerline, at the facade and at the building barycenter.

Figure 3 shows the distribution of the obtained positional planimetric accuracy values. The values of the estimated positional uncertainties are mostly low, with a strong predominance of values between 15 and 25 meters. The resources captured at the facade or at the center of buildings have planimetric accuracy values greater than 10 meters due to the relatively large average planimetric accuracy value given by the BD PARCELLAIRE® metadata . Those captured at the road centerline have relatively low values. This is probably due to spatial

`http://professionnels.ign.fr/sites/default/files/DC_BDPARCELLAIRE_1-2.pdf`

indicators used for the classification that require relatively small distances to the road centerline to assign a resource to this class, as well as the very low absolute planimetric accuracy values of the BD TOPO® for Paris road segments.

Figure 4 shows the results, represented by circles centered on the DBpedia monuments and of radius equal to their respective accuracy values. The monuments located by their barycenter have greater positional uncertainties than the others, which confirms the distribution of the figure 3.

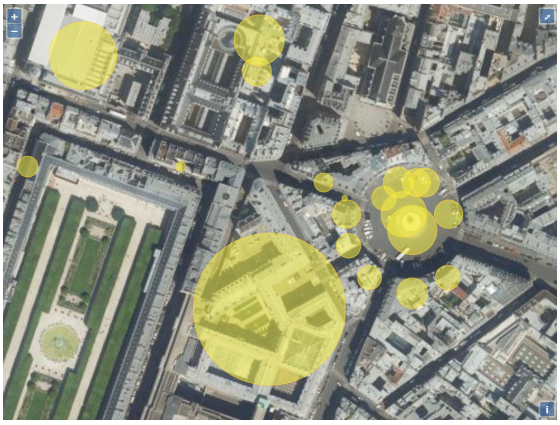

Fig. 4. Positional planimetric accuracy values of DBpedia monuments.

The results strongly depend on the initial assumptions underlying the overall evaluation. The categories of representation choices are defined by comparing the coordinates of DBpedia resources with geographic data recommended by the "WikiProject Geographical Coordinates" as reference sources. We thus assume that IGN databases have actually been used as coordinates sources by all Wikipedia contributors. Additionally, we assume that the coordinates of the DBpedia resources describing historical monuments are accurate enough so that their closest building in IGN databases can be considered as representing the same monument. This also assumes that DBpedia coordinates possess 4 decimals (or even 5 for the smallest buildings) as recommended by the "WikiProject Geographical Coordinates". From the 625 historical monuments analyzed, 606 have coordinates with at least 4 decimal places and 500 coordinates with at least 5 decimal places. This tends to confirm that the coordinates capture recommendations are rather respected on this point and that contributors were motivated to provide accurate spatial information. On the other hand, the distribution of resources in the three categories of representation choices tends to show that the capture recommendations about the characteristic element of the shape to be represented are not followed. In fact, resources are almost equally distributed between the three categories of choice in the manual classification: 33.4% for the buildings barycenter, 35.4% for their facade and 31.2% for the road centerline.

The first step of our approach, which aims to find for each resource what choice of representation has been made, is therefore essential.

Our approach is particularly suited to data sources that represent geographic entities, distinguishable as individual topographic objects, by points captured at the level of a characteristic element of their shape, a priori unknown and potentially different for each resource. It seems to be applicable for linear geographical entities as Wikipedia capture recommendations also encourage to represent them by points captured at the level of well defined shape characteristic elements. On the other hand, it is less applicable for geographic entities perceived by aggregation of individual objects, such as urban areas.

In order to implement our approach for each type of georeferenced resource, the possible categories of representation choices must be identified, spatial indicators must be defined and computed and learning samples must be created for each of category of representation choices to be considered. A first step towards its generalization to all the categories of DBPedia georeferenced resources could be to adapt it to different samples chosen for the variety of their types of representation choices and the required spatial indicators.

5 Conclusion and perspectives

The evaluation of the positional planimetric accuracy of the geometries associated with the DBpedia resources made us study their genealogy and compare the contributors practices with the capture recommendations formulated by the Wikipedia project. In the case of historical monuments in Paris, it appears that the recommendations on the number of decimals of the coordinates seem to be respected. Similarly, the predominantly low positional planimetric accuracy values suggest that contributors use accurate data sources. On the other hand, the capture recommendations concerning the choice of the shape characteristic element of the entity to be localized seem much less followed.

Providing georeferenced resources with genealogy information on the coordinates source and the choices of representation would make indirect evaluations of their positional planimetric accuracy possible or simplify the direct evaluation of this accuracy. In addition, such metadata would be useful to implement spatial analysis applications aware of the potential and limitations of the geometries associated with Web resources. For these purposes, extensions of vocabularies such as PROV-O or DQV using the ISO 19157 [9] standard could be considered.

References

1. Ahlers, D.: Assessment of the accuracy of geonames gazetteer data. In: Proceedings of the 7th Workshop on Geographic Information Retrieval. pp. 74–81. ACM (2013)

`https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Geographical_coordinates/Linear`
`https://www.w3.org/TR/vocab-dqv/`

2. Atemezing, G.A., Troncy, R.: Comparing vocabularies for representing geographical features and their geometry. In: Terra Cognita 2012 Workshop. vol. 3 (2012)
3. Barron, C., Neis, P., Zipf, A.: A comprehensive framework for intrinsic openstreetmap quality analysis. Transactions in GIS 18(6), 877–895 (2014)
4. George-Nektarios, T.: Weka classifiers summary. Athens University of Economics and Bussiness Intracom-Telecom, Athens (2013)
5. Girres, J.F.: Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques. Application aux mesures de longueur et de surface. Thèse de doctorat, spécialité sciences et technologies de l'information géographique, Université Paris-Est (2012)
6. Girres, J.F., Touya, G.: Quality assessment of the french openstreetmap dataset. Transactions in GIS 14(4), 435–459 (2010)
7. Groth, P., Moreau, L.: Prov-overview, an overview of the prov family of documents. W3c working group note 30 april 2013, W3C (https://www.w3.org/) (2013)
8. Hauff, C.: A study on the accuracy of flickr's geotag data. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. pp. 1037–1040. ACM (2013)
9. ISO: 19157: Geographic information – data quality. International standard, International Organization for Standardization (http://www.iso.org) (2013)
10. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press (2012)
11. Sarjakoski, L.: Conceptual models of generalisation and multiple representation. Generalisation of geographic information: Cartographic modelling and applications. Amsterdam, The Netherlands: Elsevier (2007)
12. Tandy, J., Barnaghi, P., van den Brink, L.: Spatial data on the web best practices. W3c working group note 25 october 2016, W3C and OGC (2016)
13. Tobler, W.R.: A computer movie simulating urban growth in the detroit region. Economic geography 46(sup1), 234–240 (1970)
14. Wienand, D., Paulheim, H.: Detecting incorrect numerical data in dbpedia. In: European Semantic Web Conference. pp. 504–518. Springer (2014)
15. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web 7(1), 63–93 (2016)