PABench: Designing a Taxonomy and Implementing a Benchmark for Spatial Entity Matching

Bilal Berjawi, Fabien Duchateau, Franck Favetta, Maryvonne Miquel, Robert Laurini

PABench: Designing a Taxonomy and Implementing a Benchmark for Spatial Entity Matching

Bilal Berjawi*, Fabien Duchateau†, Franck Favetta‡, Maryvonne Miquel* and Robert Laurini*
*LIRIS, UMR 5205
INSA de Lyon, Lyon, France
†LIRIS, UMR 5205
Université Claude Bernard Lyon 1, Lyon, France
‡LIRIS, UMR 5205
Ecole Nationale Supérieure de la Nature et du Paysage, Blois, France
Email: firstname.lastname@liris.cnrs.fr

***Abstract*—The tremendous increase of data sources containing spatial information is bound up with the diversity of geospatial applications such as location-based services (LBS) and global positioning systems. LBS providers use maps to locate spatial entities referring to points of interest (POI), for instance restaurants or locations of events. In our study, we specifically focus on places that tourists can get through LBS. The multiplication of these providers has an impact on the quality of data: spatial entities referring to the same POI may include spatial and terminological properties with incomplete, inconsistent, inaccurate or even wrong data. Thus, entity matching approaches have been proposed to discover correspondences between spatial entities, and experimental validations are traditionally performed to demonstrate the performance of these approaches in terms of effectiveness and efficiency. However, the datasets used in these experiments are rarely available, thus limiting their reuse and a fair comparison between the proposed approaches. This paper introduces PABench, a benchmark for spatial entity matching, which is available for researchers to assess their algorithms. Our benchmark includes a taxonomy of observed differences, inconsistencies and errors, which enables the characterization of different LBS providers. PABench can provide a complete and accurate evaluation of the different aspects of spatial entity matching approaches, and also facilitate an understanding of their weaknesses and abilities with respect to spatial integration. This paper provides a conceptual platform to enable LBS interoperability.**

Keywords–Benchmark; Location-based services composition; Location-based services interoperability; Spatial entity matching; Taxonomy.

I. INTRODUCTION

With the proliferation of LBS and the increasing amount of geographic data, many issues arise related to the integration of several sources of spatial data. This integration is necessary in order to update information that changes daily [1] or to produce more complete and accurate information [2]. LBS are daily used in various applications, and cartographic providers (e.g., Google Maps, Bing Maps, MapQuest) play an essential role in offering POIs such as restaurants, hotels and tourist places. A POI can be defined as a geographic object that has a point geometric shape. A POI has spatial attributes, longitude and latitude, and terminological (non-spatial) attributes such as name and type (e.g., restaurant, hotel). Some providers may supply additional terminological attributes such as address, phone number, web site, customers' ratings, etc. Due to lack of completeness, noise, inaccurate and contradictory data, it is interesting to propose solutions for detecting corresponding entities (i.e., which refer to the same POI) as given from different providers. This challenge aims at improving the quality and the relevance of information, which has a significant impact in tourist applications.

Geospatial integration has been widely studied under the term "*map conflation*" where two whole maps are integrated. Integration of maps consists of identifying the corresponding entities and fusing them [3]. Ruiz et al. present a wide description of the art with respect to map conflation [4]. The authors describe existing works in map conflation regarding their formats (raster and vector) and their criteria such as spatial data, terminological data and topological relationships between entities. Some works have been proposed in map conflation using punctual entities [5][6][7], linear entities [8][9][10] and polygonal entities [11][12][13]. Thakkar et al. propose a tool to assess the quality of geospatial data sources [14]. It utilizes the information from geospatial data sources with known quality to estimate the quality of geospatial data sources with unknown quality. In the last decade, the integration problem mainly refers to the "*entity matching*" research domain, enhanced by a spatial aspect. The discovery of corresponding entities is performed either by exploiting only spatial information [15] or by computing and combining terminological similarities for selected attributes (e.g., name, type) [16][17]. Machine learning algorithms may be applied to tune the parameters (e.g., weights) of a matching process [18]. When corresponding entities have been detected, an interesting use case aims at displaying a merged entity, i.e., to use a crafted algorithm to fuse the attributes' values of these corresponding entities. Such merging algorithms are not 100% confident. For instance, two corresponding entities may have a different location and the algorithm needs to determine the correct position. Similarly, the names or the phone numbers of two corresponding entities may differ, and the choice of the correct values relies on the merging algorithm. A merged entity may therefore include at different levels some uncertainties, which have to be presented to end-users [19].

Unfortunately, it is impossible to have a fair evaluation and comparison between existing approaches due to the absence of benchmarks. Our objective is to design such a benchmarking system, called PABench (POI Alignment Benchmark) to compare the existing spatial entity matching approaches in order to improve them or to build a better one. First, it is important to find out how the corresponding entities differ between each

other in order to understand how they can be integrated. To do so, we formalize a taxonomy to distinguish the differences that may occur within the entity set of one single provider, denoted as Intra-Difference class, and the differences between the entity sets of several providers, denoted as Inter-Difference class. The former class helps evaluate the quality of the entity set of one provider (e.g., complete information, redundancies), while the latter helps evaluate the matching between the entity sets of several providers. This taxonomy is useful to produce statistics about the providers' datasets. The next step is to create a benchmark based on the characterization of the taxonomy to understand the weaknesses and strengths of spatial entity matching approaches.

Our contributions can be summarized as follows: (i) propose a model of LBS, (ii) construct the taxonomy and understand how it impacts the results' quality of a spatial entity matching system, (iii) create a benchmark that serves in evaluating and comparing the spatial entity matching approaches [20].

The paper is organized as follows. In Section II, related work is discussed. The taxonomy of differences is given in Section III. PABench is presented in Section IV. The construction of datasets is given in Section V and a use case is provided in Section VI. Finally, Section VII concludes the paper and outlines future work.

II. Related Work

In this section, we present existing benchmarks in data integration.

In **ontology matching**, the objective is to discover semantic correspondences between concepts and properties of different ontologies [21]. Ontology alignment researchers have designed Ontology Alignment Evaluation Initiative (OAEI) [22] to compare ontology alignment tools. The OAEI datasets fulfill various criteria. For instance, the *benchmark* dataset gathers many ontologies in which a specific type of information has been altered (modifications, deletions, etc.). Consequently, it aims at detecting the weaknesses of a tool according to available information. Other datasets might be very specific like the *Food and Agriculture Organization* ontologies, which usually require external resources such as dictionaries to obtain acceptable results. In 2013, the initial datasets have been extended with synthetic ones [23].

Schema matching and mapping can be defined as the discovery of correspondences between schema elements as well as the mapping functions to transform source instances into target instances [24]. The community has designed benchmarks for evaluating these two tasks. XBenchMatch enables the assessment of schema matching tools [25]. It includes a classification of task-oriented datasets and new metrics for computing the post-match effort. STBenchmark aims at evaluating the quality of the mapping functions and their execution time [26]. Datasets are gathered according to common transformations (e.g., copying, flattening) but they can be enriched using instance generators, which can be tuned with configuration parameters (e.g., kinds of join, nesting levels).

The **entity matching** task, which is directly related to spatial entity matching, consists of discovering correspondences between equivalent objects. It also benefits from two benchmarks. The former proposes a set of four datasets about e-commerce and scientific publications [27]. These static datasets were used to compare entity matching tools. On the other hand, EMBench is based on importing existing entities (e.g., from Linked Open Data) and applying modifiers to their features (e.g., abbreviation, synonyms) [28]. These changes generate a set of modified entities, which form an entity matching dataset when grouped with the original entities. Although these two entity matching benchmarks are useful in most contexts, they are insufficient when dealing with spatial matching.

To the best of our knowledge, there is no benchmark for evaluating spatial entity matching approaches. Kang et al. propose a tool to detect the corresponding spatial entities [29]. They take two sources of entities as input, the potentially corresponding entity pairs being automatically detected based on a specified similarity measure. Then, the user has to make a decision for each pair to be considered as corresponding or not. This tool may be interesting to create a training dataset but it is not enough for a benchmark. Beeri et al. implemented a random-dataset generator to evaluate their spatial entity matching approach [30]. They generate two datasets of spatial entities in which some entities are corresponding. Unfortunately, generated entities are described only by spatial information (longitude and latitude) because their proposed matching approach exploits only the spatial information to detect the correspondences. Otherwise, the datasets used in the spatial entity matching papers are not made fully available, for instance because of confidentiality issues. A few attempts are available, such as a dataset about restaurants [31]. Yet, they cannot be exploited for various reasons. Some of them are not challenging (e.g., a simple equality metric applied to the phone numbers in the restaurant dataset discovers all the correct corresponding entities). In addition, a specific dataset may be required, for instance to include all POI types (e.g., restaurants, museums, mountains) or all entities from a given area. This lack of benchmark does not facilitate a fair and accurate comparison of the different spatial matching approaches. We also argue that the properties of a dataset are useful, both for understanding why a spatial entity matching approach is (not) effective, and for using appropriate training data when needed.

III. Taxonomy

In this section, we present a taxonomy to characterize the differences between the LBS providers. We start by introducing preliminary definitions that describe a model of LBS providers.

A. Preliminary definitions

It is necessary to understand the context of the LBS in order to construct a process to integrate them. In this section we illustrate a model that describes the LBS context of multi-provider.

Definition 1: Point of Interest (POI)
A POI is a geographical object described by a set of properties. Among these properties, there is a name, a type (e.g., restaurant, castle), a location (positioning coordinates) and a geometric shape (e.g., point, line, polygon). It is defined by the tuple:

$$POI = (name, type, coordinates, shape)$$

For example, the tuple ET below represents the Eiffel Tower POI:

$$ET = (\text{Eiffel Tower, Tourist, (48.858439, 2.294474), Point})$$

We distinguish between *large POI* and *small POI* relative to the POI area in the real-world. For example, a pub is a *small POI* while a park is a *large POI*. Let us consider the set $\mathbb{P} = \{p_1, ..., p_q\}$ that contains all the POIs of the real world where q is the number of POIs. Each LBS provider offers a set of entities that refer to a subset of existing POIs. Currently, the entities are represented with a point geometrical shape. Regarding the entities that refer to POIs with large areas, they are approximated by points such as computing their center of gravity [30]. The entities offered by a provider are derived from a specific schema of that provider.

Definition 2: Schema of provider
The schema $\mathbb{S}_k$ describes the structure of entities offered by the provider k. It is defined by:

$$\mathbb{S}_k = \mathbb{I}_k \cup \mathbb{L}_k \cup \mathbb{A}_k \cup \mathbb{B}_k$$

where

- $\mathbb{I}_k = \{id_k\}$ is an internal identifier attribute that represents a given entity for the provider k.
- $\mathbb{L}_k = \{\text{LONGITUDE}_k.label, \text{LATITUDE}_k.label\}$ is a pair of spatial attributes standing for the spatial information.
- $\mathbb{A}_k = \{\text{NAME}_k.label, \text{POITYPE}_k.label\}$ is a pair of terminological attributes that are mandatory. We call them primary attributes because they exist in the schemas of all providers and always have values.
- $\mathbb{B}_k = \left\{att\mathbb{B}_k^1.label, ..., att\mathbb{B}_k^r.label\right\}$ is another set of terminological attributes that are optionally provided with $r = |\mathbb{B}_k|$. We call them secondary attributes because they may be either missed from some schemas or have null values.

Hypothetically, a schema of any provider k includes at least all attributes in $\mathbb{I}_k \cup \mathbb{L}_k \cup \mathbb{A}_k$. We note att_k^i any attribute of the schema $\mathbb{S}_k$. The abstract data type of att_k^i, denoted as $att_k^i.type$, may have one of the following data types: string, number, array or associative array. Note that a schema may be static or dynamic. A static schema has fixed labels and structures, while for a dynamic schema, labels and structures can be modified. As an example, the provider *OpenStreetMap* [32] has a dynamic schema in which the user can add new attributes for some entities. In contrast, *GoogleMap* [33] has a static schema, so that the number and the labels of the attributes are common for all the entities. The entity set of a provider k is denoted by $\mathbb{E}_k = \{e_1, ..., e_n\}$ where n is the number of entities.

Definition 3: Entity of POI
An entity of a POI of a provider k, denoted by $e \in \mathbb{E}_k$, is an instance of the schema $\mathbb{S}_k$ and refers to one real-world POI $p \in \mathbb{P}$.

$$\begin{aligned} e = \{&(id_k.label, id_k.val), (\text{LATITUDE}_k.label, \text{LATITUDE}_k.val), \\ &(\text{LONGITUDE}_k.label, \text{LONGITUDE}_k.val), \\ &(\text{NAME}_k.label, \text{NAME}_k.val), (\text{TYPE}_k.label, \text{TYPE}_k.val), \\ &\left(att\mathbb{B}_k^1.label, att\mathbb{B}_k^1.val\right), ..., (att\mathbb{B}_k^r.label, att\mathbb{B}_k^r.val)\} \end{aligned}$$

where r is the number of secondary attributes of the schema $\mathbb{S}_k$. Table I shows an example of two entities x and y offered by two different providers that represent the POI ET (*Eiffel Tower*) with two different schemas. We denote $\mathbb{E} = \bigcup_{k=1}^{m} \mathbb{E}_k$, the union set of m providers' entities sets.

Definition 4: Association function f
The association function f is defined by:

$$\begin{aligned} f{:}&\mathbb{E} \rightarrow \mathbb{P} \\ &e \rightarrow f(e) = p \end{aligned}$$

such that $e \in \mathbb{E}$ refers to p.

For example, the entity x of Table I refers to the POI ET (*Eiffel Tower*) and $f(x) = ET$.

Definition 5: Corresponding entities
Two entities $e_1 \in \mathbb{E}_1$ and $e_2 \in \mathbb{E}_2$ are corresponding entities, denoted $e_1 \equiv e_2$, iff

$$\exists\, p \in \mathbb{P} \backslash f(e_1) = f(e_2) = p$$

For example, the two entities x and y of Table I are corresponding entities ($x \equiv y$) because they refer to the same POI ET (*Eiffel Tower*).

Definition 6: Corresponding attributes
Two attributes $att_1^i \in \mathbb{S}_1$ and $att_2^j \in \mathbb{S}_2$ are two corresponding attributes, denoted $att_1^i \equiv att_2^j$, iff they represent the same concept.

In the literature of schema matching, the correspondences between attributes are represented by a relationship [24] such as equivalence, overlap, disjointness, exclusion. But in the context of LBS providers, we only consider the equivalence relationship.

In the next section, we use the above definitions to introduce the taxonomy of differences.

B. Taxonomy of differences

In this section, we propose a formalization of the various differences that may arise between the entities of two providers. To illustrate the comparison between providers, let us consider $\mathbb{E}_1$ and $\mathbb{E}_2$ as entity sets of two LBS providers. Let $\mathbb{S}_1 = \mathbb{I}_1 \cup \mathbb{L}_1 \cup \mathbb{A}_1 \cup \mathbb{B}_1$ and $\mathbb{S}_2 = \mathbb{I}_2 \cup \mathbb{L}_2 \cup \mathbb{A}_2 \cup \mathbb{B}_2$ be the schemas of $\mathbb{E}_1$ and $\mathbb{E}_2$ respectively. Also, consider the POI $p \in \mathbb{P}$ and two corresponding entities $e_1 \in \mathbb{E}_1$ and $e_2 \in \mathbb{E}_2$ that refer to p (i.e., $e_1 \equiv e_2$).

The potentially corresponding entities of several sets will be compared depending on four levels: 1) schema, 2) terminology, 3) spatial and 4) entities' availability.

1) Schema Differences:
This level explains the heterogeneity between distinct schemas where two differences are distinguished.
Generally, differences between schemas involve two providers, i.e., Inter-Difference. In the case of a provider with a dynamic schema, differences may be classified as Intra-Difference.

Attribute Heterogeneity:
The attribute heterogeneity consists of two corresponding attributes belonging to two distinct schemas and have different labels or different abstract data types. In Table I, the attribute

TABLE I. EXAMPLE OF TWO ENTITIES x AND y, OFFERED BY TWO DIFFERENT PROVIDERS, THAT REPRESENT THE POI ET (*Eiffel Tower*) WITH TWO DIFFERENT SCHEMAS

Model	Entity x (offered by provider 1)	Entity y (offered by provider 2)
$\mathbb{I}$	EntityID : *51190385*	id : *fd0cfb424bbd79bf28a832e1764f1c2aa5927714*
$\mathbb{L}$	Latitude : *48,858606* Longitude : *2,293971*	geometry : { location : { lat : *48.85837*, lng : *2.294481*}}
$\mathbb{A}$	DisplayName : *Tour Eiffel* EntityTypeID : *7999*	name : *Eiffel Tower* types : *establishment*
$\mathbb{B}$	Phone : *0892701239* CountryRegion : *FRA* Locality : *Paris* PostalCode : *75007* AddressLine : *Champ De Mars, Avenue Anatole France ...*	formatted_phone_number : *+33 892 70 12 39* website : *http://www.tour-eiffel.fr* formatted_address : *Champ de Mars,* *5 Avenue Anatole France, 75007 Paris, France* *...*

DisplayName in the schema of the provider 1 and the attribute *name* in the schema of the provider 2 represent the name of the POI but they have different labels. Consider two attributes $att_1^i \in \mathbb{S}_1$ and $att_2^j \in \mathbb{S}_2$, $\mathbb{S}_1$ and $\mathbb{S}_2$ have an attribute heterogeneity difference iff

$$\Big(att_1^i \equiv att_2^j\Big) \wedge \Big(att_1^i.label \neq att_2^j.label \vee att_1^i.type \neq att_2^j.type\Big)$$

Different Structures:
Schemas may have various structures. One attribute of one schema may correspond to two or more attributes of another schema. Returning to Table I, the address is represented by three attributes *Locality, PostalCode* and *AddressLine* in the schema of the provider 1 while the attribute *formatted_address* in the schema of the provider 2 represents the full address. That is, a concept is described by one attribute of the schema $\mathbb{S}_1$ and by two or more attributes of the schema $\mathbb{S}_2$, or vice versa.

$$att_1^i \equiv (att_2^1, att_2^2, \ldots) \vee (att_1^1, att_1^2, ...) \equiv att_2^j$$

There are complex correspondences between the structures of the schemas. For instance, more than one attribute of a schema may correspond to more than one attribute of another (i.e., [n:m] correspondences). We do not consider the complex correspondences in this paper since the schemas in the context of LBS are simple.

2) Terminological Differences:
This level is related to the heterogeneity of values for primary and secondary terminological attributes of two corresponding entities.

Different Data:
Two corresponding entities have different values for their corresponding terminological attributes (primary or secondary). e_1 and e_2 have different data iff

$$\exists\, att_1^x \in \mathbb{A}_1 \cup \mathbb{B}_1, \exists\, att_2^y \in \mathbb{A}_2 \cup \mathbb{B}_2 \setminus e_1 \equiv e_2 \wedge (e_1.att_1^x \equiv e_2.att_2^y) \wedge (e_1.att_1^x.val \neq e_2.att_2^y.val)$$

Note that the degree of difference between the data varies. This variation may be classified as semantic (SEM) or syntactic (SYN). The former consists of two corresponding attributes that have different values but are based on the same concept (e.g., *eat-drink* and *restaurant* are two POI types that have the same meaning). The latter is about the syntax of corresponding attributes' values. They are a consequence of the different ways that a value can be written in real life, without any alteration of its meaning, or a result of human errors (i.e., misspellings, word permutations, aliases, different standards, acronyms, abbreviations and multilingualism). In Table I, the type of the entity x is *7999* while it is *establishment* for the entity y. *Different Data* (SEM and SYN) is classified as Inter-Difference.

Missing Data (MD):
Two corresponding entities having a feature that is described by one entity and missed by the other. In Table I, the website is missing from the entity x while it is given by the entity y. This difference is classified as Inter-Difference. e_1 and e_2 have missing data iff

$$\Big(\exists\, att_1^x \in \mathbb{A}_1 \cup \mathbb{B}_1, \exists\, att_2^y \in \mathbb{A}_2 \cup \mathbb{B}_2 \setminus (att_1^x \equiv att_2^y \wedge (e_1.att_1^x.val = NULL \vee e_2.att_2^y.val = NULL))\Big) \vee \Big(\exists\, att_1^x \in \mathbb{A}_1 \cup \mathbb{B}_1, \forall\, att_2^y \in \mathbb{A}_2 \cup \mathbb{B}_2 \setminus (att_1^x \not\equiv att_2^y)\Big)$$

Similar Data:
It consists of two entities that have similar values for terminological attributes but refer to two distinct POIs of the same type. Consider two POIs $p' \in \mathbb{P}$ and $p'' \in \mathbb{P}$ of the same type, and two entities $e' \in \mathbb{E}$ and $e'' \in \mathbb{E}$ that refer to p' and p'' respectively. If e' and e'' have similar values for any terminological attributes, then the difference between e' and e'' is denoted as similar data.

$$\exists\, att' \in \mathbb{B}', \exists\, att'' \in \mathbb{B}'' \setminus (p' \neq p'') \wedge (p'.type = p''.type) \wedge (f(e') = p') \wedge (f(e'') = p'') \wedge ((e'.\text{NAME}.val \cong e''.\text{NAME}.val) \vee (e'.att'.val \cong e''.att''.val))$$

Usually, the *Similar Data* difference appears when we have two or more branches of the same organization. Entities that represent these branches are of the same type, have similar terminological values (e.g., place name), located in different

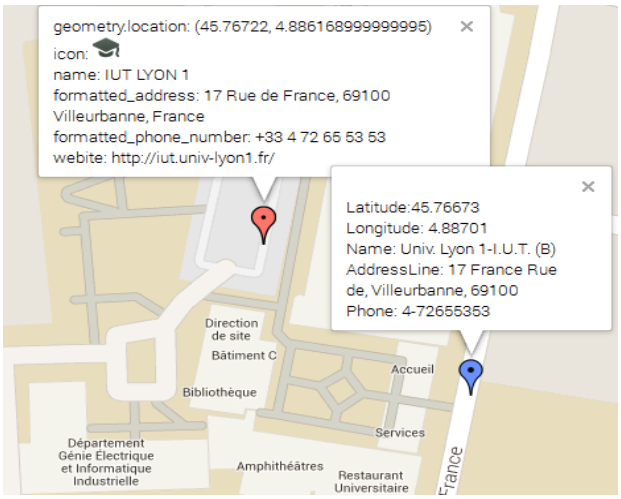


Figure 1. Example of the *Equipollent Positions* difference: Two corresponding entities refer to the IUT Lyon 1 University (large POI) having different locations but both are correct.

areas and not corresponding to each other because each branch is a distinct POI. This differences is classified as Similar Data is classified as Intra-Difference and Inter-Difference.

3) Spatial Differences:
At this level, we investigate the problem of positioning between the corresponding entities. Three differences can be distinguished.

Different Locations (DL):
Two corresponding entities have different values for their corresponding spatial attributes. The two entities of Table I refer to *Eiffel Tower*, but they have different longitude and latitude values. The distance between the two locations is approximately 226 meters. This difference is classified as Inter-Difference.

$$e_1 \equiv e_2 \wedge (e_1.\text{LATITUDE}.val \neq e_2.\text{LATITUDE}.val \vee e_1.\text{LONGITUDE}.val \neq e_2.\text{LONGITUDE}.val)$$

Equipollent Positions (EP):
This difference appears when the corresponding entities refer to a *large POI* and have different locations, but those locations are both correct with regard to the location of the POI that they represent. That is, corresponding entities' positions are equivalent in terms of concept but not in terms of values. This difference is classified as Inter-Difference.

$$(e_1 \equiv e_2) \wedge (e_1.\text{LONGITUDE}, e_1.\text{LATITUDE}) \subset p.coordinates \wedge (e_2.\text{LONGITUDE}, e_2.\text{LATITUDE}) \subset p.coordinates \wedge (e_1.\text{LONGITUDE}.val \neq e_2.\text{LONGITUDE}.val \vee e_1.\text{LATITUDE}.val \neq e_2.\text{LATITUDE}.val)$$

Figure 1 shows two corresponding entities refer to the *IUT Lyon 1 University* (large POI) and have different locations (center of gravity vs entrance gate) but both are correctly represented.

Superposition:
This consists of two entities that have the same locations but refer to two distinct POIs of the same type and it is classified as Intra-Difference and Inter-Difference. Usually, this case appears in shopping centers where two POIs of the same type are located one above the other on two different floors. Consider two POIs $p' \in \mathbb{P}$ and $p'' \in \mathbb{P}$ of the same type, and two entities $e' \in \mathbb{E}$ and $e'' \in \mathbb{E}$ that refer to p' and p'' respectively. If e' and e'' have the same location, then the difference between e' and e'' is denoted as superposition.

$$(p' \neq p'') \wedge (p'.type = p''.type) \wedge (f(e'') = p'') \wedge (f(e') = p') \wedge (e'.\text{LATITUDE}.val = e''.\text{LATITUDE}.val) \wedge (e'.\text{LONGITUDE}.val = e''.\text{LONGITUDE}.val)$$

4) Entity's Availability:
The entity's availability category takes into account errors that can be found in the entity set of a provider. Two differences can be distinguished at this level.

Not Found Entity:
This case, classified as Inter-Difference, consists of a POI that is given by one provider but not by the other. Considering the POI $p' \in \mathbb{P}$, p' is a Not found POI iff

$$\exists\, e_1 \in \mathbb{E}_1, \forall\, e_2 \in \mathbb{E}_2 \backslash f(e_1) = p' \wedge\ f(e_2) \neq p'$$

Duplicate Entities:
This case, classified as Intra-Difference, corresponds to two entities of the same provider that refer to the same POI. Consider two entities $e_1 \in \mathbb{E}_1$ and $e_1^{'} \in \mathbb{E}_1$, e_1 and $e_1^{'}$ are two duplicate entities iff

$$\exists\, p \in \mathbb{P} \backslash \left(f(e_1) = f\left(e_1^{'}\right) = p\right)$$

Although the differences described in the taxonomy are elementary, the detection of the corresponding entities requires some hard work because a combination of differences may occur when comparing two entities. For instance, the two entities x and y of Table I are two corresponding entities with a combination of four differences, namely 1) *Attribute Heterogeneity*, 2) *Different Structure*, 3) *Syntactic Different Data* (SYN) and 4) *Different Locations* (DL).

IV. Benchmark

The taxonomy of differences is useful to get statistics about LBS providers and to understand how they can be integrated. Also, it serves to create a benchmark that evaluates the performance of spatial entity matching approaches and to build a characterized dataset that serves for machine learning purposes. In this section we describe PABench [20] based on this taxonomy.

A. Overview of the matching process

In general, the data integration process consists of three consecutive phases namely 1) schema/ontology matching, 2) entity matching and 3) entity merging or fusion. The schema matching task helps finding corresponding attributes between two schemas in order to compare their values later in the entity matching task. It produces an alignment between the corresponding attributes accompanied by a transformation

function such as combination, split, etc. A schema matching approach must be able to handle the differences denoted in the schema category namely *Attribute Heterogeneity* and *Different structures*. In the context of LBS, the schema matching of providers can be done manually since their schemas are small and simple, so that there is no need for semi- or fully-automatic approaches to handle the schema matching task. Secondly, (spatial) entity matching approaches are used to find corresponding entities in several datasets to merge them together. It takes as input the datasets that need to be merged and the alignment of their schemas produced by the schema matching task. Entity matching can be done by computing a similarity score between each pair of entities. Then, a matching approach considers a pair of entities as corresponding if its similarity score is higher than a given threshold [34], or produces a list of pairs of entities ranked according to their similarity score. Concerning spatial entity matching systems, they measure the degree of similarity between entities using various techniques such as Euclidean distance between entities' locations, semantic equivalence and syntactic comparison of terminological information. These measures serve to compute a score that indicates their belief that two entities correspond. Finally, the entity merging phase takes as input the result of the entity matching task in order to fuse the corresponding entities. How the entities are merged depends on how these entities will be used, it may be done by a simple combination of values or based on specific rules in a given context. In our context, we intend to consider the tourist rules to merge the spatial entities. Note that the schema matching and the entity merging are no longer discussed in this paper; next we focus on the evaluation of spatial entity matching approaches.

To evaluate the performance and the results' quality of a spatial entity matching system, consider two datasets, namely source and target, for which a list of correct correspondences, called ground truth, is known in advance. For each entity in the source dataset, the matching system will try to find the corresponding entity from the target dataset. Thus, correspondences returned by the matching system are compared to the ground truth correspondences in order to measure how successfully the matching system detects the expected answer.

B. Benchmark construction

PABench has been constructed based on the differences defined in the taxonomy. Recall that a spatial entity consists of spatial and terminological (primary and secondary) information and refers to a real world POI. Deciding whether two spatial entities correspond is a challenging task due to the differences that occur between them. As previously mentioned, two corresponding entities being compared may have a combination of differences where each combination is a distinct situation of differences. To understand the weaknesses and strong points of an entity spatial matching system, the evaluation must be characterized according to the situations of differences that may occur between entities. In other words, it is required to evaluate a spatial entity matching system based on each situation of differences.

The possible situations of differences are computed based on the taxonomy of differences with respect to the entity matching task. Since the entity matching goal is to detect the corresponding entities, only the differences concerning corresponding entities are considered, namely *Different Locations* (DL) and *Equipollent Positions* (EP) from the spatial category and *Missing Data* (MD), *Semantic Different Data* (SEM) and *Syntactic Different Data* (SYN) from the terminological category. *Superposition*, *Similar Data* and *Not Found Entity* differences may be used to add noise entities (see below) when comparing the source and target datasets. Finally, *Duplicate Entities* must be pre-handled before the entity matching task using deduplication techniques to ensure the quality of used datasets.

Spatial information is only expressed by an entity's location, it may have zero (i.e., no difference) or only one difference in the spatial category differences. The set of spatial differences S_dif is given by:

$$S\_dif = \{\emptyset, DL, EP\}$$

Primary terminological information is expressed by an entity's name and type, it may have zero, one (i.e., at least one attribute has the difference) or two differences of the terminological category differences. The *Missing Data* (MD) difference cannot be considered because the primary terminological attributes are always provided and have values (see Section III-A). The set of primary terminological differences PT_dif is given by:

$$PT\_dif = \{\emptyset, SEM, SYN, (SEM, SYN)\}$$

Secondary terminological information varies from one provider to another, it may have zero, one (i.e., at least one attribute has the difference), two (i.e., each difference appears at least once) or three differences of the terminological category differences. The set of secondary terminological differences ST_dif is given by:

$$\begin{aligned} ST\_dif = \{&\emptyset, SEM, SYN, MD, (SEM, SYN, MD), \\ &(SEM, SYN), (SEM, MD), (SYN, MD)\} \end{aligned}$$

Let $Situations\_dif$ be the set of all possible combinations of differences that may occur between two corresponding entities at all levels (spatial, primary terminological and secondary terminological)

$$\begin{aligned} Situations\_dif = \{&a, b, c \setminus a \in S\_dif, b \in PT\_dif, \\ &c \in ST\_dif\} \end{aligned}$$

where

$$|Situations\_dif| = 3 \times 4 \times 8 = 96$$

Returning to Table I, the two corresponding entities x and y have a combination of differences. For spatial information they have *Different Locations* (spatial coordinates), for primary terminological information they have *Syntactic Different Data* (name, type) and for secondary terminological information they have *Syntactic Different Data* (phone, address) and *Missing Data* (website). The situation $s \in Situations\_dif$ between x and y is given by $s = \{DL, SYN, (SYN, MD)\}$.

To guarantee that the situations have no redundancy, each situation $s \in Situations\_dif$ must be unique and exclusive, in the sense that the situations do not share any relation between them such as intersection, inclusion, etc. Thus, our benchmark consists of comparing a source dataset with a target dataset in which the corresponding entities have a specific situation of differences. If the correct answer

TABLE II. CONTINGENCY TABLE OF EVALUATION MEASURES.

Matching approach \ Ground truth	Corresponding entities	Non-corresponding entities
Corresponding entities	True Positive (TP)	False Positive (FP)
Non-corresponding entities	False Negative(FN)	True Negative (TN)

(represented by the ground truth) is returned by a matching approach, it means that it is able to deal with the given situation.

Definition 7: **TestCase**
For each situation $s \in Situations\_dif$, we define a test case that consists of a source dataset $\mathbb{E}_S \subset \mathbb{E}$ ($\mathbb{E} = \bigcup_{k=1}^{m} \mathbb{E}_k$), a target dataset $\mathbb{E}_T \subset \mathbb{E}$ and a ground truth between both source and target datasets.

$$TestCase(s) = (\mathbb{E}_S,\, \mathbb{E}_T,\, groundTruth)$$

Noise entities may be added to $\mathbb{E}_S$ and $\mathbb{E}_T$. A noise entity is an entity that exists in one dataset and does not have any correspondence in the other dataset. The goal of adding noise entities is to explain whether a matching approach is able to avoid detecting two non-corresponding entities even if they have near locations or similar information. Noise entities should contain entities with *Not Found Entity*, *Similar Data* and *Superposition* differences. Concerning the *Not found Entity* difference, it can be easily detected from the real entities of LBS providers. But *Similar Data* and *Superposition* are hard to detect, in this case we intend to automatically generate entities with such differences. These test cases allow us to find the situations of differences that a matching approach is able to handle and to what degree this handling is possible in order to differentiate it from other similar approaches.

C. Quality measure and impact of differences

Results' quality of a matching system is measured by the standard performance measures that come from the information retrieval domain, precision, recall and F-measure [35]. These measures evaluate the performance of a matching system by comparing its results to ground truth's results. Also, they help to understand weaknesses and strengths of a matching approach for each test case. Table II classifies the contingency of evaluation measures' base. Precision calculates the proportion of correct correspondences detected by the matching system among all detected correspondences. Using the notations of Table II, the precision is given by formula (1). A 100% precision means that all correspondences detected by the matching system are true.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

Recall computes the proportion of correct correspondences detected by the matching system among all correct correspondences. The recall is given by formula (2). A 100% recall means that all correct correspondences have been found by the matching system.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

F-measure is a trade off between precision and recall and it is calculated with the formula (3). The β parameter of formula (3) regulates the respective influence of precision and recall ($\beta \in \mathbb{R}^+$). It is often set to 1 to give the same weight to these two evaluation measures.

$$F - measure\,(\beta) = \frac{\left(\beta^2 + 1\right) \times precision \times recall}{(\beta^2 \times precision) + recall} \tag{3}$$

It is important to analyze how the differences of the taxonomy impact these measures in order to discover the weaknesses and the strengths of an approach. Concerning the *Attribute Heterogeneity*, *Different Structure* and *Duplicate Entities* differences, they are pre-handled before launching the spatial entity matching process. *Different Data* (SEM and SYN), *Missing Data*, *Different Locations* and *Equipollent Positions* must be addressed through the matching system. They impose obstacles that may prevent a matching system from detecting the correct corresponding entities. Hence, if a matching approach fails to overcome those obstacles, True Positive (TP) decreases, False Negative (FN) increases and recall decreases. Concerning *Superposition* and *Similar Data*, two distinct entities with the same location or with similar data may be detected as corresponding. These differences increase False Positive (FP) and precision decreases. The *Not Found Entity* concerns a POI that is represented by one provider and not by the other, that means the entity of the first provider does not correspond to any of the entities of the second provider. But it risks a situation where a matching approach detects a correspondence for an entity of this difference, which increases FP and precision decreases. Also, this difference impacts the entity merging phase because as long as the number of available entities is small, we cannot ensure the correctness of information. In the case of *Duplicate Entities*, a matching approach may detect the same correspondence twice. This case will increase TP leading to a wrong precision value. That is why it is important to verify the quality of providers' datasets before starting the matching process using deduplication techniques. Table III summarizes the taxonomy of differences and their impacts on the quality measures.

V. DATASETS

A tool that consists of two modules has been implemented in order to generate the datasets of test cases. The first module, called GeoBench [36], is addressed to experts and it serves to build a characterized dataset in a semi-automatic process through the sets of three LBS providers namely Google Maps, Nokia Here Maps and Geonames. Let $\mathbb{E}_1$, $\mathbb{E}_2$ and $\mathbb{E}_3$ be the datasets of the three LBS providers respectively. Experts can search for a specific or random source entity from one provider, and then GeoBench searches for the nearby target entities from the two others separately. For each retrieved target entity, an expert has to decide whether it corresponds to the source entity and to select the differences that exist between the two entities at each level (spatial, primary and secondary terminological). Concerning the secondary terminological level, only the most common secondary information is considered namely phone number, website and address. These three secondary attributes may have *Missed Data* (MD) or *Syntactic Difference Data* (SYN) assuming that it is impossible to have a *Semantic Different Data* (SEM) according to the information that they represent (e.g., two corresponding phone numbers may be

TABLE III. TAXONOMY OF DIFFERENCES AND QUALITY MEASURES IMPACT.

Category	Difference	Intra-Diff	Inter-Diff	Impact
Schema	Attribute Heterogeneity	X(dynamic schema)	X	
	Different structure	X(dynamic schema)	X	
Terminology	Different Data (SEM)		X	TP ↘ FN ↗
	Different Data (SYN)		X	TP ↘ FN ↗
	Missing Data (MD)		X	TP ↘ FN ↗
	Similar Data (SD)	X	X	FP ↗
Spatial	Different locations (DL)		X	TP ↘ FN ↗
	Equipollent Positions (EP)		X	TP ↘ FN ↗
	Superposition (SUP)	X	X	FP ↗
Availability	Not found POI		X	FP ↗
	Duplicate Entities	X		TP (wrong value)

TABLE IV. NUMBER OF ENTITIES AND CORRESPONDENCES PRODUCED USING GEOBENCH (OCTOBER 2014).

Dataset	Number of entities
$\mathbb{E}_1$	715
$\mathbb{E}_2$	583
$\mathbb{E}_3$	282
Total	**1580**

Number of correspondences	
$\mathbb{E}_1$, $\mathbb{E}_2$	569
$\mathbb{E}_1$, $\mathbb{E}_3$	254
$\mathbb{E}_2$, $\mathbb{E}_3$	247
Total	**1070**

TABLE V. TOP FIVE TEST CASES ACCORDING TO THE NUMBER OF DETECTED CORRESPONDENCES (OCTOBER 2014).

Test Case #	Situation	Number of Correspondences
43	{EP, SYN, {SYN, MD}}	145
27	{DL, SYN, {SYN, MD}}	78
41	{EP, SYN, SYN}	69
35	{EP, SEM, {SYN, MD}}	62
11	{0, SYN, {SYN, MD}}	60

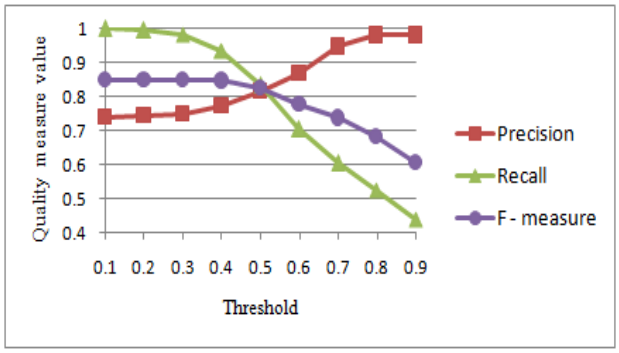
syntactically different but can never be semantically different). In this case, the number of the possible situations decreases from 96 to 48. In other contexts, the *Semantic different Data* (SEM) may be considered (e.g., when comparing the food type of two entities of restaurant's type, Pizza vs Italian food). GeoBench allows us to create a dataset $\mathbb{E} = \mathbb{E}_1 \cup \mathbb{E}_2 \cup \mathbb{E}_3$ in which, for each pair of entities, we know the relevance of correspondence and the situation of differences. Note that $\mathbb{E}_1$, $\mathbb{E}_2$ and $\mathbb{E}_3$ contain only the entities processed by GeoBench and not the whole set of entities of the three LBS providers.

The second module has been implemented to generate the test cases, it uses the dataset $\mathbb{E}$ created by GeoBench to generate source and target datasets for each situation $s \in Situations\_dif$. This module allows to configure the characteristics of a test case through a set of parameters in order to control aspects such as situation of differences, percentage of correspondences and percentage of noise. Once the characteristics of a test case are configured, source and target datasets are generated with a ground truth file so the evaluators can assess the results of their approaches. The module will search for all pairs of entities that match the requested situation of differences, then the entities of each pair will be distributed between the source and the target datasets and, the identifiers of corresponding entities will be listed together in the ground truth file. Finally, noise entities will be added to the target dataset or source dataset.

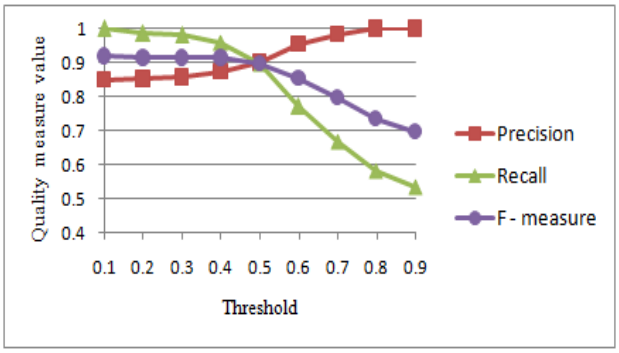
Statistics shown in Table IV represent the number of entities of each provider's dataset and the number of correspondences between them. Table V provides the top five test cases according to the number of detected correspondences. More statistics are available online along with PABench[20]. Retrieved datasets describe real world POIs where entities have been retrieved from several existing LBS providers using their web services. These datasets do not contain any redundancies or duplicated entities. The current version of the benchmark, as of October 2014, contains 1070 corresponding entities. All entities are available in CSV/SQL standard format to be easily parsed and used. To ensure that the test cases are rich enough, entities are distributed in several geographical zones/countries and refer to POIs of several types including *large POIs* and *small POIs*. Note that in the practice of the LBS context, some situations of differences rarely occur (e.g., the situation where two entities from different providers have no difference at all). In the future, we intend to develop an entity generator tool that takes a subset of source entities to modify the values of their attributes (spatial, primary and secondary terminological) in order to create a target dataset that expresses these rare situations.

VI. ANALYZE AND FIRST USE OF PROVIDERS' DATASETS

In this section, we demonstrate the resistance of the benchmark against frequently used basic measures. To reach this goal, a simple matching tool (one similarity measure and a threshold) is used to determine the difficulty of the spatial entity matching task and to show the heterogeneity of our collected dataset. Our basic matching tool consists in comparing the values of a single terminological attribute using a string similarity measure. Entity pairs that have the highest similarity score above a given threshold are considered as corresponding. This simple approach is used to match $\mathbb{E}_1$ with $\mathbb{E}_2$ and $\mathbb{E}_1$ with $\mathbb{E}_3$. Concerning the terminological information, we will compare the values of the NAME attribute using Levenshtein string similarity measure [37]. Experiments are repeated by varying the threshold value. The results are measured in terms of precision, recall and F-measure (see Section IV-C). Figures 2(a) and 2(b) show the matching quality of $\mathbb{E}_1$ with $\mathbb{E}_2$ and $\mathbb{E}_1$ with $\mathbb{E}_3$ respectively. The x-axis represents the values of threshold and the y-axis represents the values of precision, recall and F-measure. For a small value of threshold (0.1), the precision is low (75% for $\mathbb{E}_1$ vs $\mathbb{E}_2$ and 85% for $\mathbb{E}_1$ vs $\mathbb{E}_3$), which means that 25%-15% of detected corresponding entities are not correct according to the ground truth. However, the recall is high (99% in both cases), which means that the matching approach does not miss any of the ground truth correspondences. Increasing the threshold increases the precision and decreases the recall. For a high value of threshold (0.9), precision increases up to 98% in both cases, which indicates that most of the detected

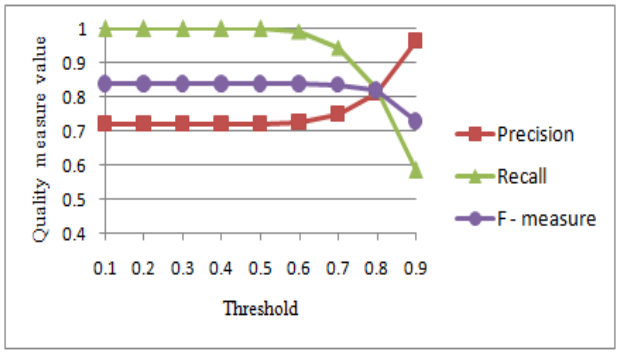


(a) Dataset E1 - E2

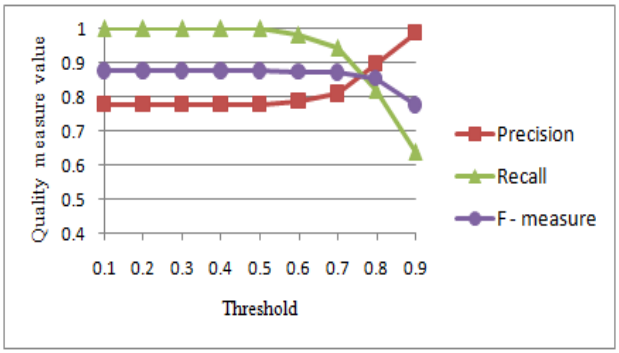


(b) Dataset E1 - E3

Figure 2. Results' quality in terms of precision, recall and F-measure using Levenshtein similarity measure.



(a) Dataset E1 - E2



(b) Dataset E1 - E3

Figure 3. Results' quality in terms of precision, recall and F-measure using JaroWinkler similarity measure.

TABLE VI. EXECUTION TIME FOR LEVENSHTEIN AND JAROWINKLER ACCORDING TO THE NUMBER OF ENTITIES IN MATCHED DATASETS.

Datasets	Number of comparisons	Execution time of Levenshtein (sec)	Execution time of JaroWinkler (sec)
$\mathbb{E}_1$ vs $\mathbb{E}_2$	715 × 583	11	840
$\mathbb{E}_1$ vs $\mathbb{E}_3$	715 × 282	5	400
$\mathbb{E}_2$ vs $\mathbb{E}_3$	583 × 282	4	320

correspondences are correct according to the ground truth. Conversely, the recall decreases to approximately 50%, which means that the half of the ground truth correspondences are missing. A trade-off is reached for a 0.5 threshold, in which case the *F-measure* is up to 80% for $\mathbb{E}_1$ vs $\mathbb{E}_2$ and 90% for $\mathbb{E}_1$ vs $\mathbb{E}_3$, which indicates that using a simple matching method is not enough to resolve the matching problem. Similar results are obtained when the Levenshtein measure is replaced by the JaroWinkler measure [38][39][40] (see Figures 3(a) and 3(b)). A trade off is achieved with a 0.8 threshold and the *F-measure* is up to 80% for $\mathbb{E}_1$ vs $\mathbb{E}_2$ and 87% for $\mathbb{E}_1$ vs $\mathbb{E}_3$. Except that JaroWinkler keeps the same quality with threshold up to 0.6 while it decreases with Levenshtein from 0.4, this means that the scores calculated by Levenshtein for corresponding entities is lower than the scores calculated by JaroWinkler. The two similarity measures are approximately equivalent in terms of results' quality, but the Levenshtein metric is more efficient in terms of execution time. Table VI compares their execution times according to the number of entities in matched datasets.

These experiments show that basic similarity measures are not enough to match the real and heterogeneous data of PABench. Note that matching $\mathbb{E}_2$ with $\mathbb{E}_3$ have the same trend as matching $\mathbb{E}_1$ with $\mathbb{E}_2$ and $\mathbb{E}_1$ with $\mathbb{E}_3$.

VII. CONCLUSION

Spatial entity matching has become a basic problem in many application domains such as heterogeneous location-based services. In this paper, we highlighted the absence of a benchmark to compare and evaluate spatial entity matching approaches. We proposed a taxonomy that characterizes differences, heterogeneities and errors between LBS providers at four levels: schema, terminology, spatial and availability. We studied the impact of the identified differences on the results' quality of a matching approach and we proposed the necessary specifications to design a benchmark, called PABench, that serves to evaluate and compare spatial entity matching approaches. We believe that our proposition will allow researchers to better evaluate their matching approaches, identify the capabilities of their approaches, and also guide performance improvements in existing spatial entity matching approaches. In the future, PABench may be extended by 1) adding more entities and 2) automatically generating entities to cover the situations of differences that occur only rarely in the LBS context. Also, we intend to create a survey that compares and evaluates existing approaches in terms of results' quality and execution time using our benchmark. This evaluation will explain the weaknesses and the strengths of current works,

which will help to propose a better matching approach. On the other hand, the proposed taxonomy is limited to punctual geographical objects, but it may be extended to cover complex objects (e.g., polygons and lines) in order to be used for complex geographical data.

ACKNOWLEDGMENT

This work was supported by the LABEX IMU (ANR-10-LABX-0088) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). Authors would like to thank Google Maps, Geonames, Open Street Map and Nokia Here Maps for providing POIs' entities.

REFERENCES

[1] I. N. Gregory, "Time-variant gis databases of changing historical administrative boundaries: A european comparison," Transactions in GIS, vol. 6, no. 2, 2002, pp. 161–178.

[2] M. A. Cobb, F. E. Petry, and K. B. Shaw, "Fuzzy spatial relationship refinements based on minimum bounding rectangle variations," Fuzzy Sets and Systems, vol. 113, no. 1, 2000, pp. 111–120.

[3] M. L. Casado, "Some basic mathematical constraints for the geometric conflation problem," in Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Lisboa, Instituto Geogrfico Portugus, 2006, pp. 264–274.

[4] J. J. Ruiz, F. J. Ariza, M. A. Ureña, and E. B. Blázquez, "Digital map conflation: a review of the process and a proposal for classification," International Journal of Geographical Information Science, vol. 25, no. 9, 2011, pp. 1439–1466.

[5] A. Saalfeld, "A fast rubber-sheeting transformation using simplicial coordinates," The American Cartographer, vol. 12, no. 2, 1985, pp. 169–173.

[6] C.-C. Chen, S. Thakkar, C. A. Knoblock, and C. Shahabi, "Automatically annotating and integrating spatial datasets," in Advances in Spatial and Temporal Databases, 2003, pp. 469–488.

[7] S. Volz, "An iterative approach for matching multiple representations of street data," in Proceedings of the JOINT ISPRS Workshop on Multiple Representations and Interoperability of Spatial Data, Hannover, 2006, pp. 101–110.

[8] A. Saalfeld, "Conflation automated map compilation," International Journal of Geographical Information System, vol. 2, no. 3, 1988, pp. 217–228.

[9] Y. Doytsher, "A rubber sheeting algorithm for non-rectangular maps," Computers & Geosciences, vol. 26, no. 9, 2000, pp. 1001–1010.

[10] M. Zhang, W. Shi, and L. Meng, "A generic matching algorithm for line networks of different resolutions," in Workshop of ICA Commission on Generalization and Multiple Representation Computering Faculty of A Coruña University-Campus de Elviña, Spain, 2005.

[11] E. M. Arkin, L. P. Chew, D. P. Huttenlocher, K. Kedem, and J. S. B. Mitchell, "An efficiently computable metric for comparing polygonal shapes," IEEE Trans. Pattern Anal. Mach. Intell., vol. 13, no. 3, 1991, pp. 209–216.

[12] M. Gombosoi, B. Zalik, and S. Krivograd, "Comparing two sets of polygons," International Journal of Geographical Information Science, vol. 17, no. 5, 2003, pp. 431–443.

[13] A. Masuyama, "Methods for detecting apparent differences between spatial tessellations at different time points," International Journal of Geographical Information Science, vol. 20, no. 6, 2006, pp. 633–648.

[14] S. Thakkar, C. A. Knoblock, and J. L. Ambite, "Quality-driven geospatial data integration," in ACM International Symposium on Geographic Information Systems, Washington, USA, 7-9 November, 2007, p. 16.

[15] E. Safra, Y. Kanza, Y. Sagiv, C. Beeri, and Y. Doytsher, "Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets," International Journal of Geographical Information Science, vol. 24, no. 1, 2010, pp. 69–106.

[16] A.-M. O. Raimond and S. Mustire, "Data matching - a matter of belief," in International Symposium on Spatial Data Handling (SDH), 2008, pp. 501–519.

[17] R. Karam, F. Favetta, R. Kilany, and R. Laurini, "Integration of similar location based services proposed by several providers," in Networked Digital Technologies, 2010, pp. 136–144.

[18] V. Sehgal, L. Getoor, and P. Viechnicki, "Entity resolution in geospatial data integration." in ACM International Symposium on Geographic Information Systems, 2006, pp. 83–90.

[19] B. Berjawi, E. Chesneau, F. Duchateau, F. Favetta, C. Cunty, M. Miquel, and R. Laurini, "Representing uncertainty in visual integration," in Proceedings of the 20th International Conference on Distributed Multimedia Systems, Pittsburgh, USA, 27-29 August, 2014, pp. 365–372.

[20] "PABench," URL: http://liris-unimap01.insa-lyon.fr/benchmark [accessed: 2014-12-03].

[21] J. Euzenat and P. Shvaiko, Ontology matching. Heidelberg, Germany: Springer-Verlag, 2007, ISBN: 3-540-49611-4.

[22] "Ontology Alignment Evaluation Initiative," URL: http://oaei.ontologymatching.org [accessed: 2014-12-03].

[23] J. Euzenat, M.-E. Rosoiu, and C. Trojahn, "Ontology matching benchmarks: generation, stability, and discriminability," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 21, 2013.

[24] Z. Bellahsene, A. Bonifati, and E. Rahm, Schema Matching and Mapping. Heidelberg, Germany: Springer-Verlag, 2011, ISBN: 978-3-642-16517-7.

[25] F. Duchateau and Z. Bellahsene, "Designing a benchmark for the assessment of schema matching tools," in Open Journal of Databases (OJDB), vol. 1, no. 1. RonPub, Germany, 2014, pp. 3–25.

[26] B. Alexe, W. C. Tan, and Y. Velegrakis, "Stbenchmark: towards a benchmark for mapping systems," Proceedings of the VLDB, vol. 1, no. 1, 2008, pp. 230–244.

[27] H. Köpcke, A. Thor, and E. Rahm, "Evaluation of entity resolution approaches on real-world match problems," PVLDB, vol. 3, no. 1, 2010, pp. 484–493.

[28] E. Ioannou, N. Rassadko, and Y. Velegrakis, "On generating benchmark data for entity matching," Journal on Data Semantics, vol. 2, no. 1, 2013, pp. 37–56.

[29] H. Kang, V. Sehgal, and L. Getoor, "Geoddupe: A novel interface for interactive entity resolution in geospatial data," in International Conference on Information Visualisation, 2007, pp. 489–496.

[30] C. Beeri, Y. Doytsher, Y. Kanza, E. Safra, and Y. Sagiv, "Finding corresponding objects when integrating several geo-spatial datasets," in ACM International Workshop on Geographic Information Systems, 2005, pp. 87–96.

[31] "Restaurants' dataset," URL: http://cs.utexas.edu/users/ml/riddle/data.html [accessed: 2014-12-03].

[32] "OpenStreetMap," URL: http://www.openstreetmap.org [accessed: 2014-12-03].

[33] "Google Maps," URL: http://maps.google.com [accessed: 2014-12-03].

[34] J. Madhavan, P. A. Bernstein, and E. Rahm, "Generic schema matching with cupid," in VLDB, 2001, pp. 49–58.

[35] C. J. V. Rijsbergen, Information Retrieval, 2nd ed. Newton, Massachusetts, USA: Butterworth-Heinemann, 1979, ISBN: 0408709294.

[36] A. Morana, T. Morel, B. Berjawi, and F. Duchateau, "Geobench: a geospatial integration tool for building a spatial entity matching benchmark," in ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas, USA, 2014, in press.

[37] V. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," Soviet Physics Doklady, vol. 10, no. 8, 1966, pp. 707–710.

[38] M. A. Jaro, "Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida," Journal of the American Statistical Association, vol. 84, no. 406, 1989, pp. 414–420.

[39] W. E. Winkler, "The state of record linkage and current research problems," in Statistical Research Division, US Census Bureau. Citeseer, 1999. [Online]. Available: http://www.census.gov/srd/www/byname.html

[40] M. A. Jaro, "Probabilistic linkage of large public health data files," Statistics in medicine, vol. 14, no. 5-7, 1995, pp. 491–498.