LIRIS

UMR 5205 CNRS

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

B. Berjawi, F. Duchateau, F. Favetta, M. Miquel, R. Laurini

Laboratoire d'InfoRmatique en Image et Systèmes d'information

GEOProcessing 2015

Lisbon, Portugal

cnrs INSA Lyon 1 UNIVERSITÉ LUMIÈRE LYON 2 ÉCOLE CENTRALELYON

Motivation

- **Multiplication of Points of Interest (POI) and data sources**
  - Several Location-Based Services (LBS) providers
  - Incomplete, inconsistent, inaccurate, wrong information
- **Integration of multiple sources**
  - Similarity measures
  - Probability measures
  - Learning-based methods
- **How to evaluate and compare spatial integration methods?**

Related Work

- **Ontology matching:**
  - Ontology Alignment Evaluation Initiative (OAEI) [1]
- **Schema matching:**
  - XBenchMatch [2]
  - STBenchmark [3]
- **Entity matching:**
  - EMBench [4]

[1]: Ontology Alignment Evaluation Initiative," URL: http://oaei.ontologymatching.org
[2]: 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.
[3]: 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.
[4]: 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.

Related Work

- **Spatial Entity matching:**
  - Geoddupe [5]
  - Random-spatial-dataset generator [6]

Need for a spatial entity matching benchmark

PABench: Point of Interest Alignment Benchmark

[5]: 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.
[6]: 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.

Contributions

- **Taxonomy of LBS**
  - Describe the context of LBS providers
  - Compare the LBS providers
  - Characterize the differences that occur between LBS providers
- **Benchmark**
  - Construct PABench based on the taxonomy characterization
  - Generate a characterized training dataset using real data

Outline

- Introduction
- Related Work
- **Taxonomy of LBS**
  - Preliminary definitions
  - Differences
- **Benchmark**
  - Benchmark construction
  - Datasets
- **Conclusion and Future Work**

Taxonomy - Preliminary definitions

- **POI: geographical object described by a set of properties**

  POI = (name, type, coordinates, shape)

- **Schema of provider: structure of entities offered by the provider**

  I: Internal identifier

  L: Spatial attributes

  A: Primary terminological

  B: Secondary terminological

- **Entity of POI: instance of a schema and refers to one real-world POI**

  $e = \{(id_k:label, id_k:val), (LATITUDE_k:label, LATITUDE_k:val), \ldots\}$

Taxonomy - Preliminary definitions

- **Association function f: returns the POI described by a given entity**

$$f : E \rightarrow P$$
$$e \rightarrow f(e) = p$$

- **Corresponding entities: two entities from two distinct providers refer to the same POI** ($e_1 \equiv e_2$)

$$\exists\, p \in P \setminus f(e_1) = f(e_2) = p$$

- **Corresponding attributes: two attributes from two distinct schemas represent the same concept** ($att_1 \equiv att_2$)

Taxonomy - Differences

Category	Difference
Schema	Attribute Heterogeneity
	Different structure
Terminology	Semantic Different Data (SEM)
	Syntactic Different Data (SYN)
	Missing Data (MD)
	Similar Data (SD)
Spatial	Different locations (DL)
	Equipollent Positions (EP)
	Superposition (SUP)
Availability	Not found POI
	Duplicate Entities

Differences of corresponding entities

Differences of non-corresponding entities

Taxonomy - Example

Entity x (offered by provider 1)	Entity y (offered by provider 2)
EntityID: 51190385	id: fd0cfb424bbd79bf28a832e1764f1c2
Latitude: 48,858606 Longitude: 2,293971	geometry: { location : { lat : 48.85837, lng: 2.294481}}
DisplayName: Tour Eiffel EntityTypeID: 7999	name: Eiffel Tower types: establishment
Phone: 0892701239 CountryRegion: FRA Locality: Paris PostalCode: 75007 AddressLine: Champ De Mars, Avenue Anatole France …	formatted phone number: +33892701239 website: http://www.tour-eiffel.fr formatted address: Champ de Mars, 5 Avenue Anatole France, 75007 Paris, France

Attribute Heterogeneity

$(att_i \equiv attj) \wedge (atti.label \neq attj.label \vee atti.type \neq attj.type)$

Taxonomy - Differences

Category	Difference
Schema	Attribute Heterogeneity
	Different structure
Terminology	Semantic Different Data (SEM)
	Syntactic Different Data (SYN)
	Missing Data (MD)
	Similar Data (SD)
Spatial	Different locations (DL)
	Equipollent Positions (EP)
	Superposition (SUP)
Availability	Not found POI
	Duplicate Entities

Differences of corresponding entities

Differences of non-corresponding entities

Taxonomy - Example

Entity x (offered by provider 1)	Entity y (offered by provider 2)
EntityID: 51190385	id: fd0cfb424bbd79bf28a832e1764f1c2
Latitude: 48,858606 Longitude: 2,293971	geometry: { location : { lat : 48.85837, lng: 2.294481}}
DisplayName: Tour Eiffel EntityTypeID: 7999	name: Eiffel Tower types: establishment
Phone: 0892701239 CountryRegion: FRA Locality: Paris PostalCode: 75007 AddressLine: Champ De Mars, Avenue Anatole France …	formatted phone number: +33892701239 website: http://www.tour-eiffel.fr formatted address: Champ de Mars, 5 Avenue Anatole France, 75007 Paris, France

Different Structure

$att_i \equiv (att1, att2, \ldots) \vee (att1, att2, \ldots) \equiv att_j$

Taxonomy - Differences

Category	Difference
Schema	Attribute Heterogeneity
	Different structure
Terminology	Semantic Different Data (SEM)
	Syntactic Different Data (SYN)
	Missing Data (MD)
	Similar Data (SD)
Spatial	Different locations (DL)
	Equipollent Positions (EP)
	Superposition (SUP)
Availability	Not found POI
	Duplicate Entities

Differences of corresponding entities

Differences of non-corresponding entities

Taxonomy - Example

Entity x (offered by provider 1)	Entity y (offered by provider 2)
EntityID: 51190385	id: fd0cfb424bbd79bf28a832e1764f1c2
Latitude: 48,858606 Longitude: 2,293971	geometry: { location : { lat : 48.85837, lng: 2.294481}}
DisplayName: Tour Eiffel EntityTypeID: Touristic place	name: Eiffel Tower types: Landmark - attraction
Phone: 0892701239 CountryRegion: FRA Locality: Paris PostalCode: 75007 AddressLine: Champ De Mars, Avenue Anatole France ...	formatted phone number: +33892701239 website: http://www.tour-eiffel.fr formatted address: Champ de Mars, 5 Avenue Anatole France, 75007 Paris, France

Semantic and Syntactic Different Data

$\exists\ att_i \in A1 \cup B1, \exists\ attj \in A2 \cup B2 \setminus$
$e1 \equiv e2 \wedge (e1.atti \equiv e2.attj) \wedge (e1.atti.val \equiv e2.attj.val)$

Taxonomy - Differences

Category	Difference
Schema	Attribute Heterogeneity
	Different structure
Terminology	Semantic Different Data (SEM)
	Syntactic Different Data (SYN)
	Missing Data (MD)
	Similar Data (SD)
Spatial	Different locations (DL)
	Equipollent Positions (EP)
	Superposition (SUP)
Availability	Not found POI
	Duplicate Entities

Differences of corresponding entities

Differences of non-corresponding entities

Taxonomy - Example

Entity x (offered by provider 1)	Entity y (offered by provider 2)
EntityID: 51190385	id: fd0cfb424bbd79bf28a832e1764f1c2
Latitude: 48,858606 Longitude: 2,293971	geometry: { location : { lat : 48.85837, lng: 2.294481}}
DisplayName: Tour Eiffel EntityTypeID: Touristic place	name: Eiffel Tower types: Landmark - attraction
Phone: 0892701239 CountryRegion: FRA Locality: Paris PostalCode: 75007 AddressLine: Champ De Mars, Avenue Anatole France	formatted phone number: +33892701239 website: http://www.tour-eiffel.fr formatted address: Champ de Mars, 5 Avenue Anatole France, 75007 Paris, France

Missing Data

$\exists\ att_i \in A1 \cup B1, \exists\ att_j \in A2 \cup B2 \setminus$
$(att_i \equiv att_j) \wedge (e1.att_i.val = NULL \vee e2.\ att_j.val = NULL)$

Taxonomy - Differences

Category	Difference
Schema	Attribute Heterogeneity
	Different structure
Terminology	Semantic Different Data (SEM)
	Syntactic Different Data (SYN)
	Missing Data (MD)
	Similar Data (SD)
Spatial	Different locations (DL)
	Equipollent Positions (EP)
	Superposition (SUP)
Availability	Not found POI
	Duplicate Entities

Differences of corresponding entities

Differences of non-corresponding entities

Taxonomy - Example

Entity x (offered by provider 1)	Entity y (offered by provider 2)
EntityID: 51190385	id: fd0cfb424bbd79bf28a832e1764f1c2
Latitude: 48,858606 Longitude: 2,293971	geometry: { location : { lat : 48.85837, lng: 2.294481}}
DisplayName: Tour Eiffel EntityTypeID: Touristic place	name: Eiffel Tower types: Landmark - attraction
Phone: 0892701239 CountryRegion: FRA Locality: Paris PostalCode: 75007 AddressLine: Champ De Mars, Avenue Anatole France	formatted phone number: +33892701239 website: http://www.tour-eiffel.fr formatted address: Champ de Mars, 5 Avenue Anatole France, 75007 Paris, France

Different Location

$e1 \equiv e2 \wedge (e1.LATITUDE.val \neq e2.LATITUDE.val \vee e1.LONGITUDE.val \neq e2.LONGITUDE.val)$

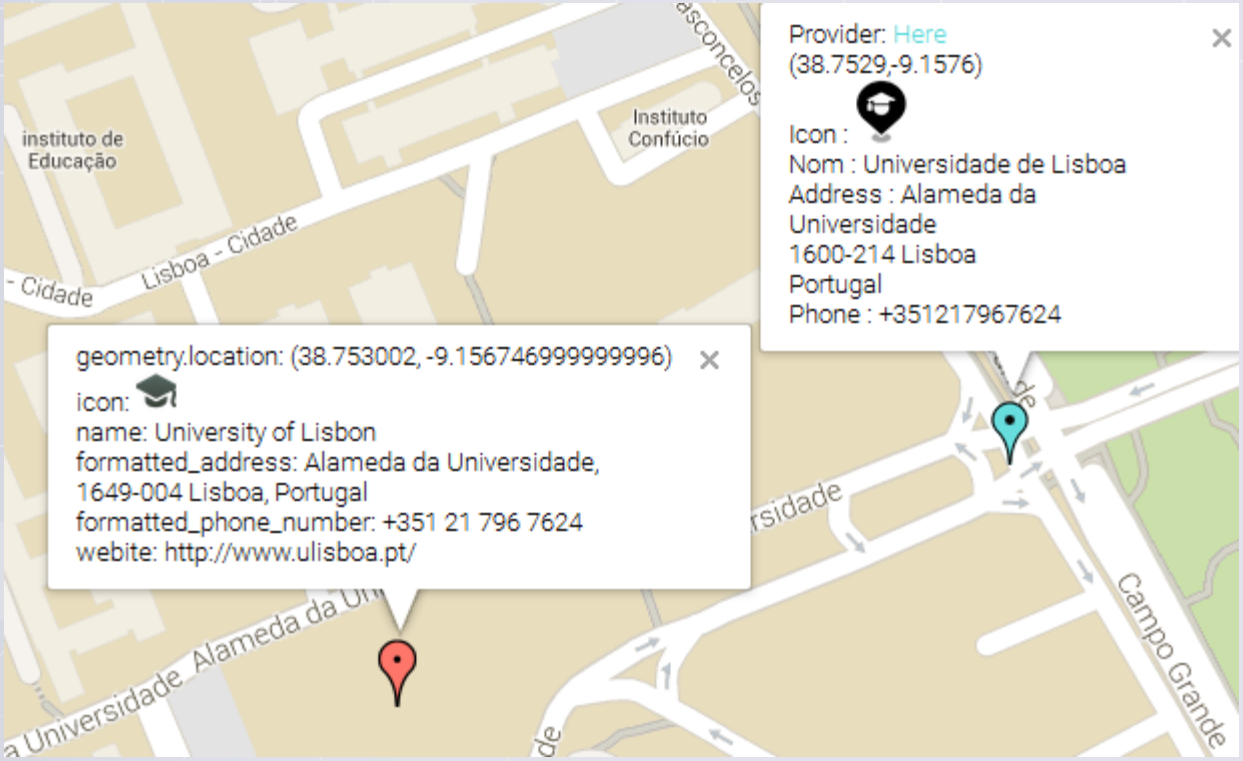
Taxonomy - Differences

Category	Difference
Schema	Attribute Heterogeneity
	Different structure
Terminology	Semantic Different Data (SEM)
	Syntactic Different Data (SYN)
	Missing Data (MD)
	Similar Data (SD)
Spatial	Different locations (DL)
	Equipollent Positions (EP)
	Superposition (SUP)
Availability	Not found POI
	Duplicate Entities

Differences of corresponding entities

Differences of non-corresponding entities

Taxonomy - Example



Equipollent Positions

$(e_1 \equiv e2) \wedge (e1.LATITUDE, e1.LONGITUDE) \subset p.coordinates \wedge (e2.LATITUDE, e2.LONGITUDE) \subset p.coordinates \wedge (e1.LONGITUDE.val \neq e2.LONGITUDE.val) \wedge (e1.LATITUDE.val \neq e2.LATITUDE.val)$

Taxonomy - Differences

Category	Difference
Schema	Attribute Heterogeneity
	Different structure
Terminology	Semantic Different Data (SEM)
	Syntactic Different Data (SYN)
	Missing Data (MD)
	Similar Data (SD)
Spatial	Different locations (DL)
	Equipollent Positions (EP)
	Superposition (SUP)
Availability	Not found POI
	Duplicate Entities

Differences of corresponding entities

Differences of non-corresponding entities

Benchmark - Construction

Differences concerning corresponding entities:

Level	Attributes	Set of possible differences
Spatial	Location	∅,DL, EP
Primary Terminological	Name and Type	∅, SEM, SYN, {SEM , SYN}
Secondary Terminological	Phone, Address, Site, etc.	∅, MD, SEM, SYN, {SEM, SYN, MD}, {SEM, SYN}, {SEM, MD}, {SYN, MD}

96 (3x4x8) distinct situations of differences

Generate a test case for each of the 96 situations

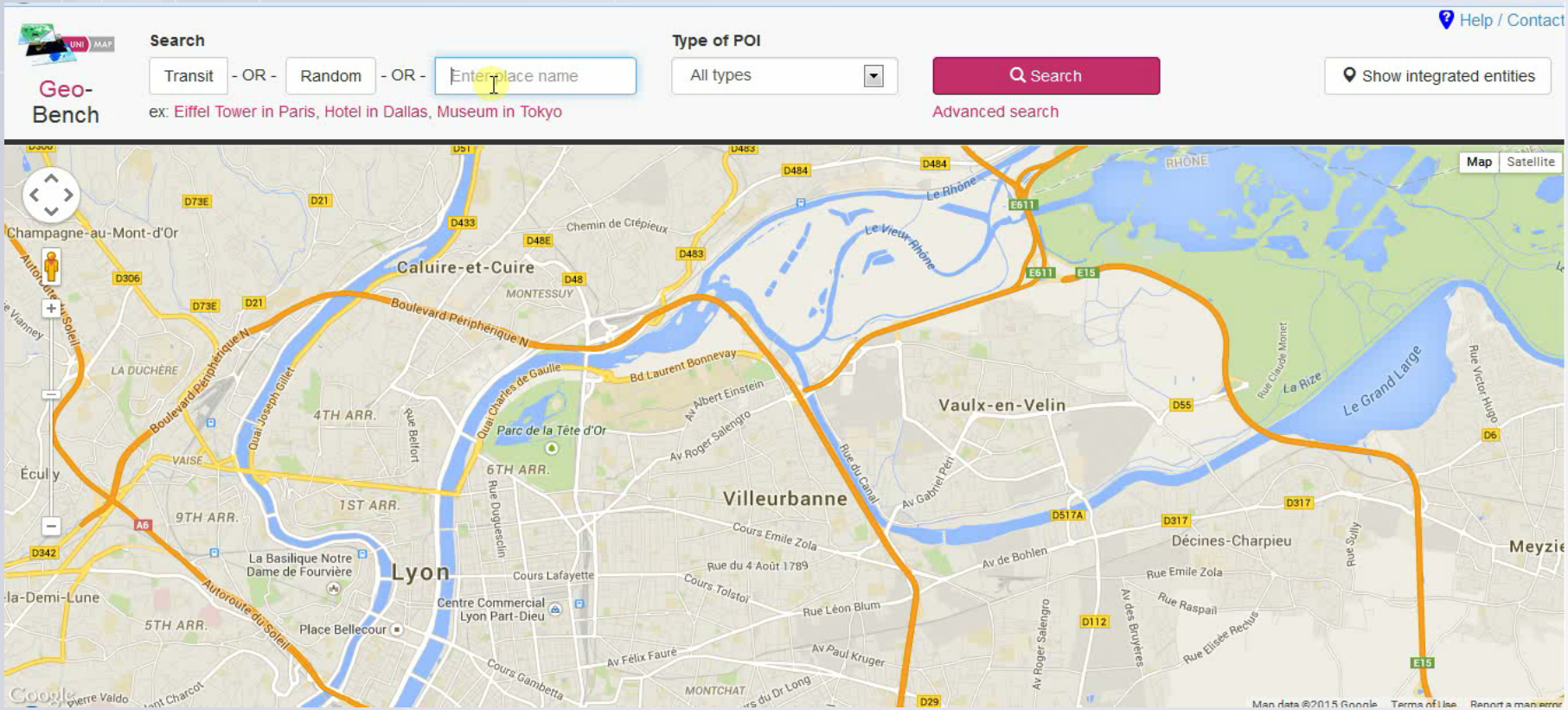
Example of a situation: s= {DL, {SEM,SYN}, MD}
Test_case(s)= (Source dataset, Target dataset, Ground_truth)

Remaining differences will be used to add noise

- Superposition, Similar Data, Not found POI

Benchmark - Datasets

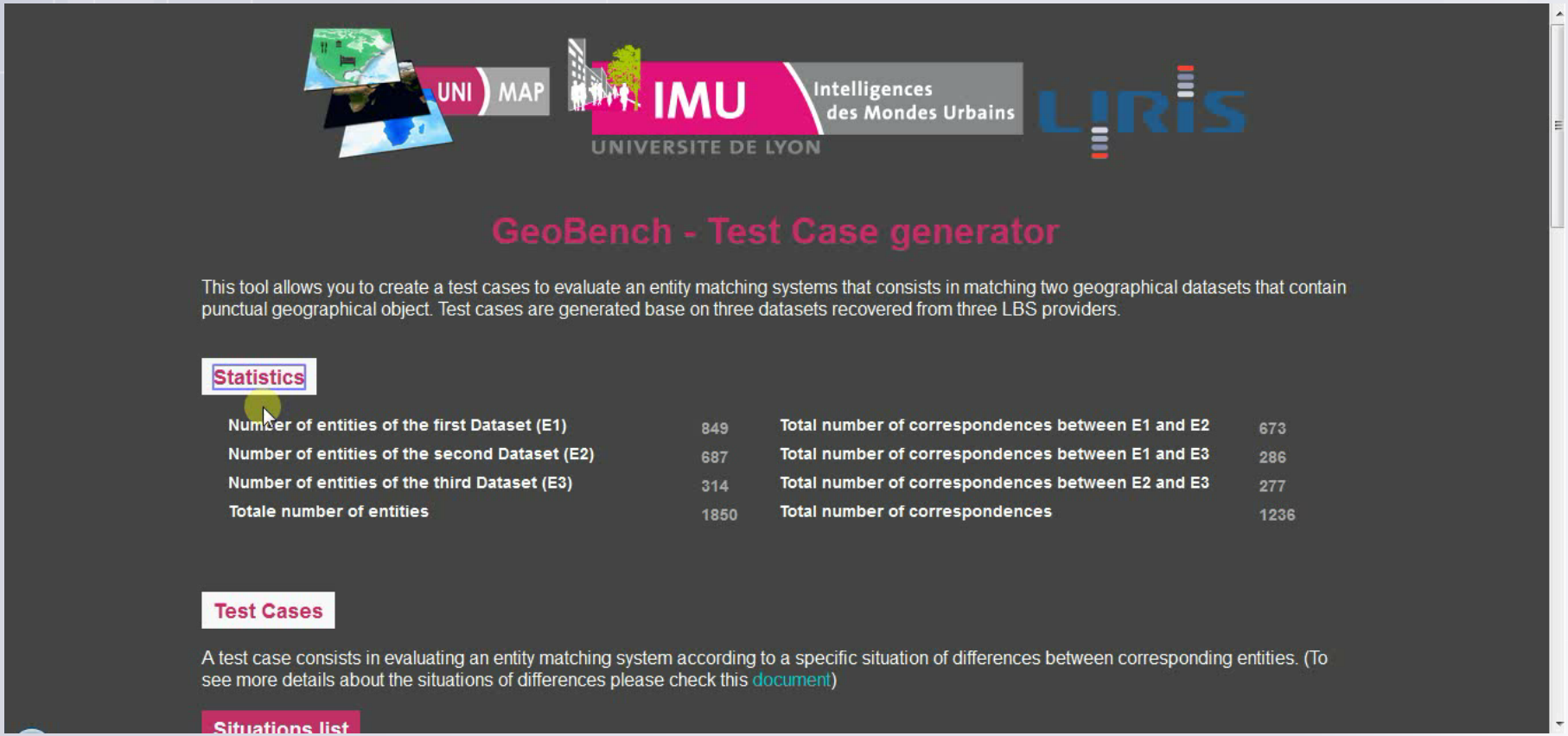
Create a characterized dataset – GeoBench tool [7]



[7]: G. Morana, T. Morel, B. Berjawi, and F. Duchateau, "GeoBench: a Geospatial Integration Tool for Building a Spatial Entity Matching Benchmark (Demo), " in ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas, Texas, USA, 4-7 November, 2014, pp. 533-536. **http://tinyurl.com/p3dbmpj**

Benchmark - Datasets

Test cases generator



http://tinyurl.com/nc4rurr

Conclusion and Future Work

- **Contribution:**
  - Taxonomy that describes LBS context
  - Necessary specifications to design PABench
  - GeoBench tool to create a characterized dataset and a test case generator
- **Future Work:**
  - Extend PABench by adding more entities
  - Create a survey that compares and evaluates existing approaches using PABench
  - Extend the taxonomy to cover complex objects

Thank you for your attention

Bilal Berjawi
LIRIS, INSA de Lyon, France
bberjawi@liris.cnrs.fr
http://unimap.liris.cnrs.fr

Benchmark – Datasets statistics

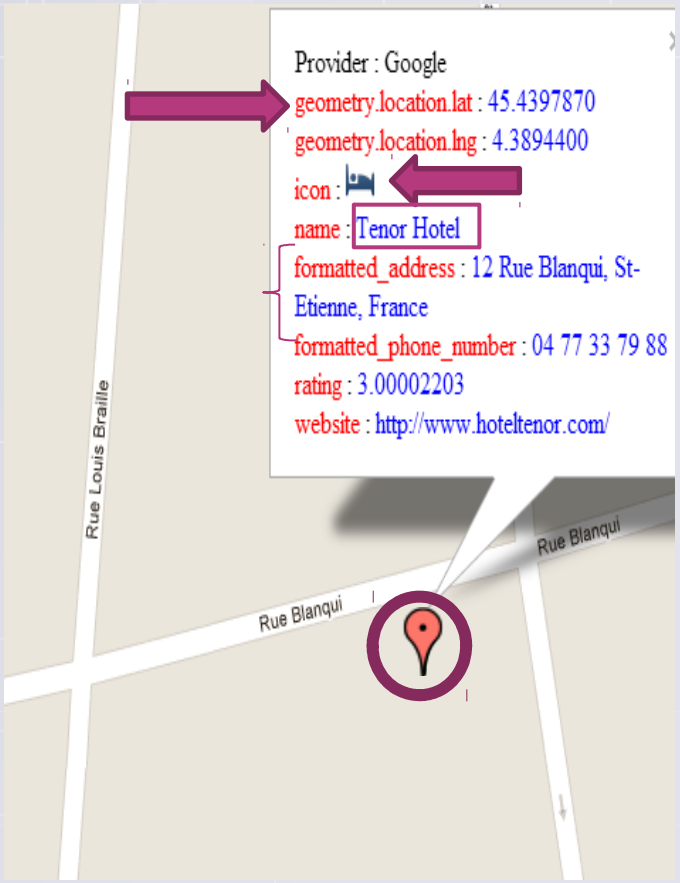
Dataset	Number of Entities
E1	846
E2	685
E3	314
Total	**1845**

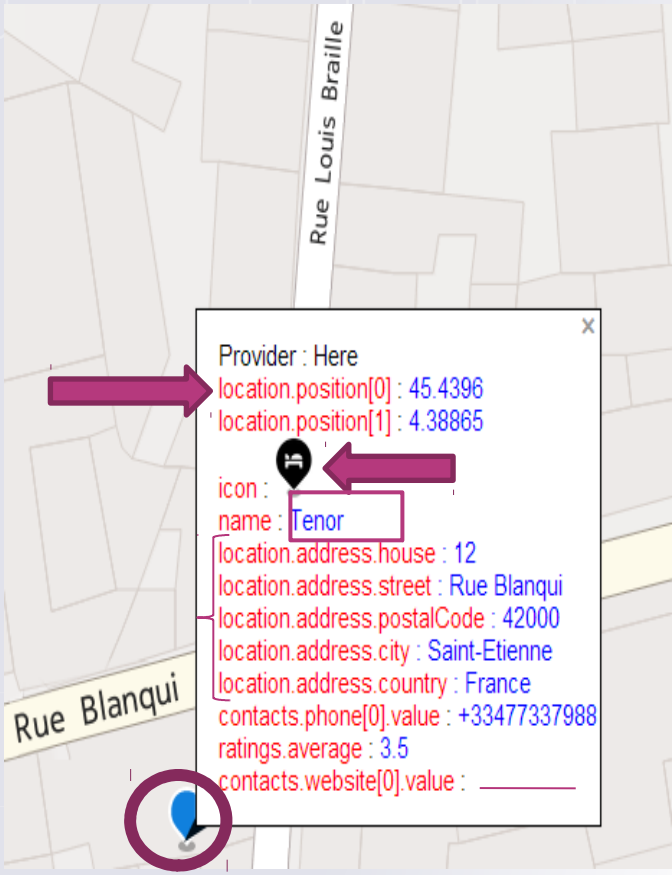
Number of correspondences	
E1, E2	671
E1, E3	286
E2, E3	277
Total	**1234**

Situations of differences	Number of correspondences
{EP, SYN, {SYN, MD}}	147
{DL, SYN, {SYN, MD}}	93
{EP, SYN, SYN}	71
{∅, SYN, {SYN, MD}}	70
{EP, ∅, {SYN, MD}}	63

Example of differences

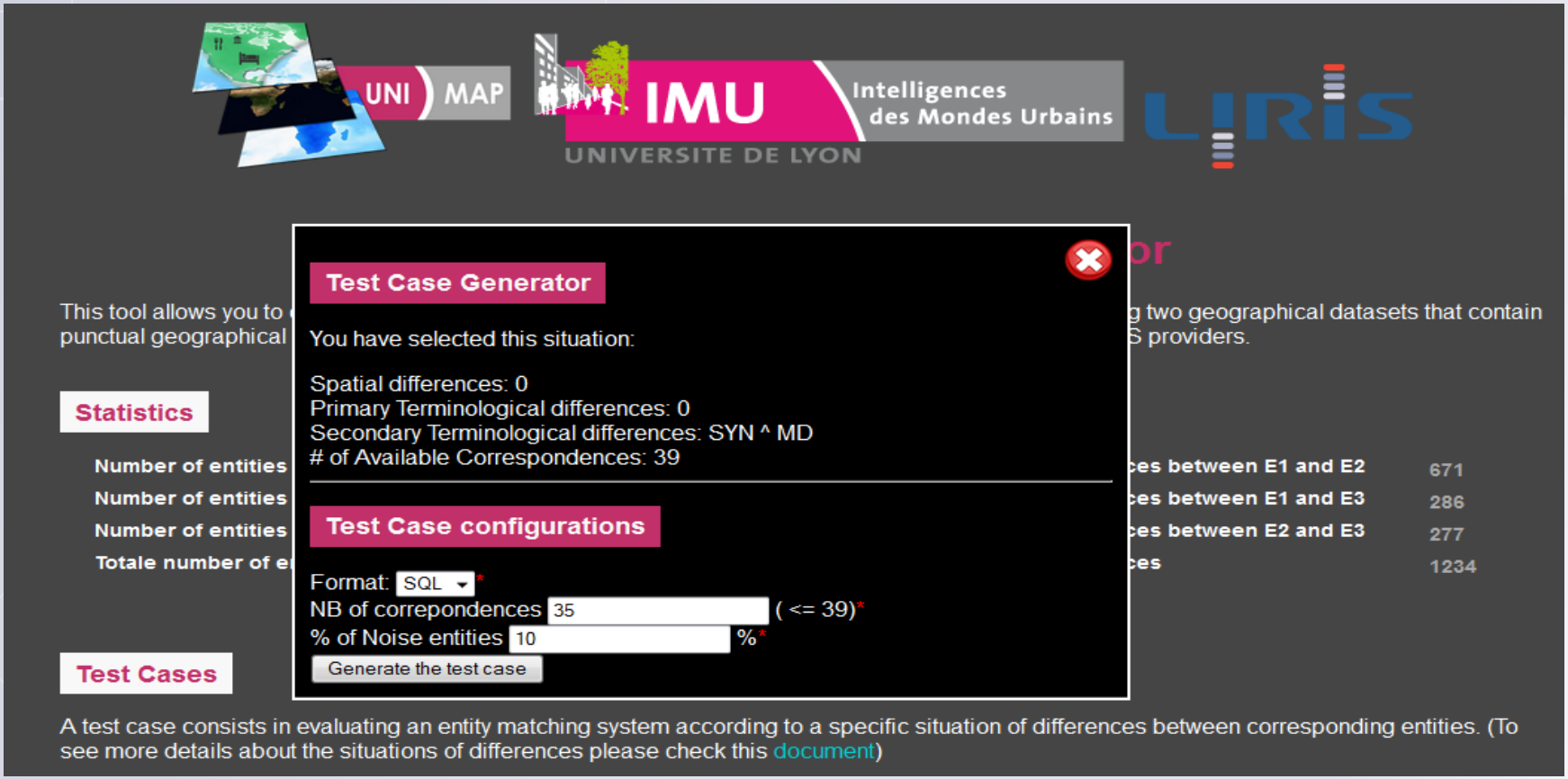
- Positioning
- Attributes names
- Structure
- Legends
- Different values
- Missing values





Benchmark - Datasets

Statistics and Test Cases generator



http://tinyurl.com/nc4rurr