GeoBench: a Geospatial Integration Tool for Building a Spatial Entity Matching Benchmark

A. Morana, T. Morel, B. Berjawi, F. Duchateau

Laboratoire d'InfoRmatique en Image et Systèmes d'information

UMR5205 CNRS / Université de Lyon / Université Claude Bernard Lyon 1 / INSA de Lyon

`http://geobench.liris.cnrs.fr`

Objectives

- For data integration practionners : facilitate the building of spatial entity matching datasets
- For end-users : build a map with complete information about their favourite places

Motivations

- Evaluate and compare spatial entity matching approaches
- Build a characterized spatial dataset for machine learning purposes

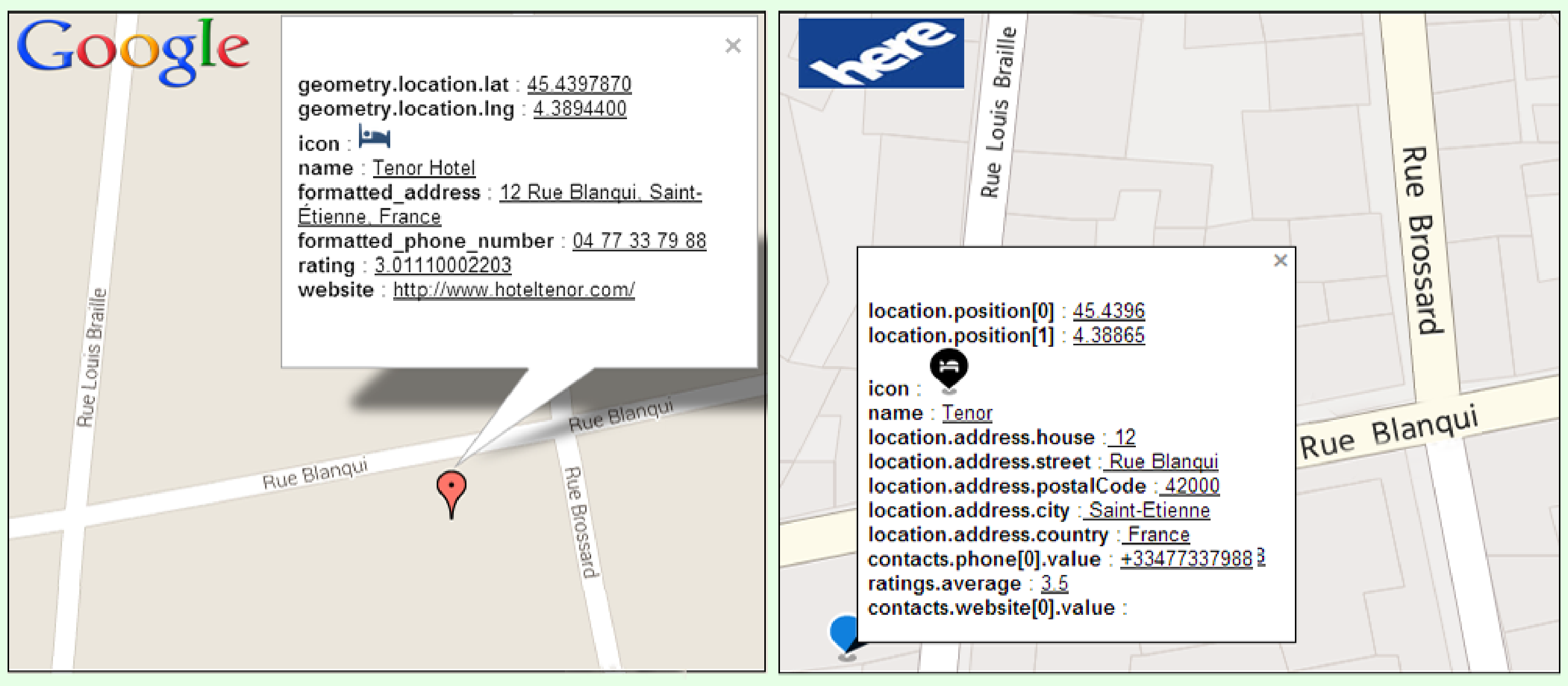
Issues and Contributions

- Location-based services providers offer incomplete and/or contradictory data about tourist places
- Recent works are proposed to discover spatial entities that refer to the same place
- These works have been evaluated using different test protocols
- Datasets used for evaluation are not made fully available

Example of heterogeneity between two LBS providers

Differences at :

- Positioning
- Attributes names
- Structures
- Symbols
- Values



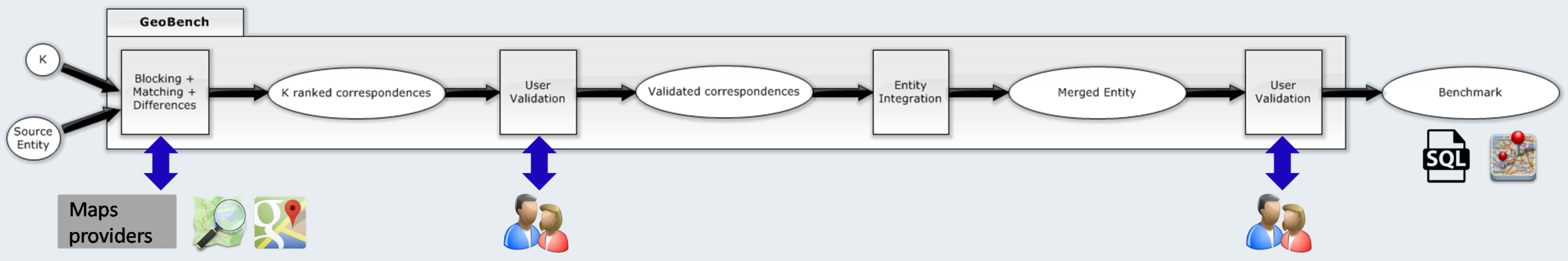
geometry.location.lat : 45.4397870
geometry.location.lng : 4.3894400
icon :
name : Tenor Hotel
formatted_address : 12 Rue Blanqui, Saint-Étienne, France
formatted_phone_number : 04 77 33 79 88
rating : 3.01110002203
website : http://www.hoteltenor.com/

location.position[0] : 45.4396
location.position[1] : 4.38865
icon :
name : Tenor
location.address.house : 12
location.address.street : Rue Blanqui
location.address.postalCode : 42000
location.address.city : Saint-Etienne
location.address.country : France
contacts.phone[0].value : +334773379883
ratings.average : 3.5
contacts.website[0].value :

Comparison of the same POI (Tenor Hotel) using two LBS providers

Overview of GeoBench

GeoBench is a tool which serves to build a benchmark for spatial entity matching by facilitating the discovery and the integration of corresponding spatial entities.



GeoBench phases

- **Blocking Algorithm** aims at quickly identifying a subset of entities among all those available which likely represent the source entity φ
- **Detecting differences** aims at classifying the terminological and spatial differences between the attributes of two entities
- **Matching Algorithm** aims at computing a confidence score between the entities of the blocking phase and the initial source entity φ
- **Integrating Corresponding Entities** offers the possibility to merge corresponding entities into a new integrated entity

Specifications

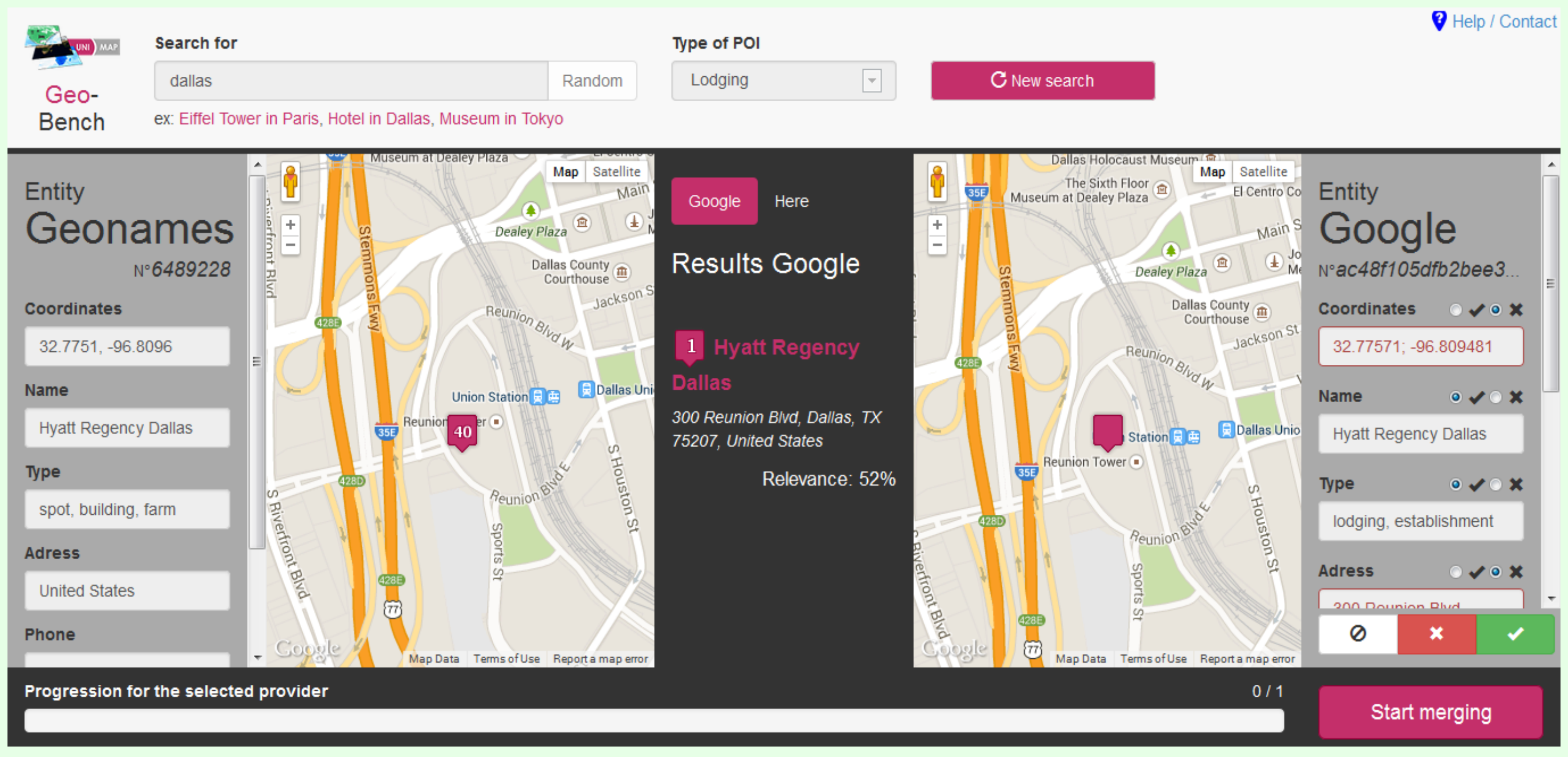
Blocking

- Based on the coordinates of the source entity φ and include all entities within a radius
- Entities of the blocking area whose name shares a token with φ's name and having the same type of φ

Matching

- Terminological : based on Levenhstein string similarity measure
- Spatial : based on Euclidean distance

Example of the matching process



Search for: dallas
Type of POI: Lodging
ex: Eiffel Tower in Paris, Hotel in Dallas, Museum in Tokyo

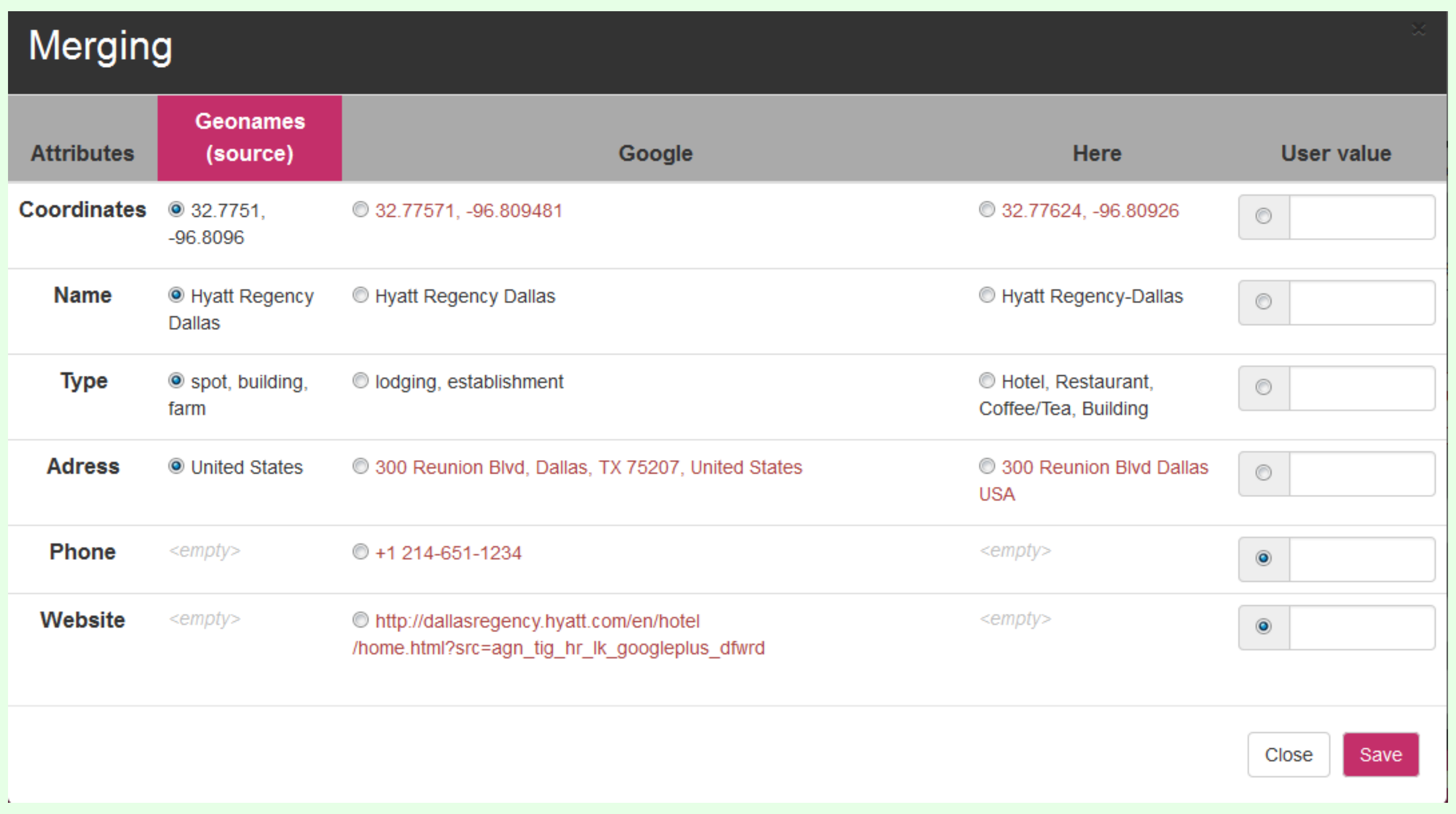
Entity
Geonames
N°6489228
Coordinates: 32.7751, -96.8096
Name: Hyatt Regency Dallas
Type: spot, building, farm
Adress: United States
Phone

Results Google
1 Hyatt Regency Dallas
300 Reunion Blvd, Dallas, TX 75207, United States
Relevance: 52%

Entity
Google
N°ac48f105dfb2bee3...
Coordinates: 32.77571; -96.809481
Name: Hyatt Regency Dallas
Type: lodging, establishment
Adress: 300 Reunion Blvd...

Progression for the selected provider 0 / 1

Example of the integration process



Merging

Attributes	Geonames (source)	Google	Here	User value
Coordinates	● 32.7751, -96.8096	○ 32.77571, -96.809481	○ 32.77624, -96.80926	○
Name	● Hyatt Regency Dallas	○ Hyatt Regency Dallas	○ Hyatt Regency-Dallas	○
Type	● spot, building, farm	○ lodging, establishment	○ Hotel, Restaurant, Coffee/Tea, Building	○
Adress	● United States	○ 300 Reunion Blvd, Dallas, TX 75207, United States	○ 300 Reunion Blvd Dallas USA	○
Phone	<empty>	○ +1 214-651-1234	<empty>	●
Website	<empty>	○ http://dallasregency.hyatt.com/en/hotel/home.html?src=agn_tig_hr_lk_googleplus_dfwrd	<empty>	●