D6 – SPATIAL ENTITY MATCHING WITH GEOALIGN

N. Barret*, F. Duchateau*, F. Favetta*, L. Moncla**
*Université de Lyon, LIRIS UMR 5205, **INSA Lyon, LIRIS UMR 5205

CONTEXT

- Several cartographic providers for POI (restaurant, hotels…)
- Several spatial entities may refer to the same POI: accuracy, completeness

How to get complete, minimal and up-to-date information for a POI?

→ Matching and merging of the corresponding entities

→ Evaluation of matching algorithms (with or without ground truth)

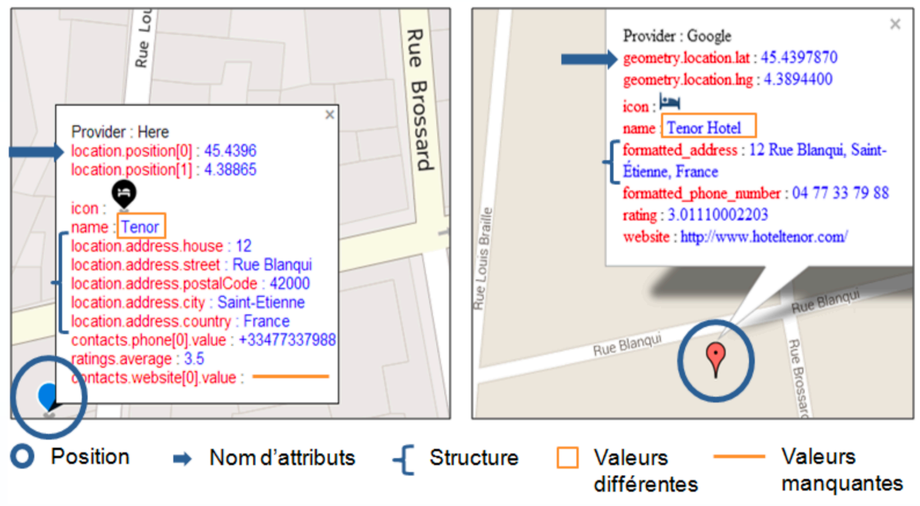
Related works

Spatial entity matching: GeoDDupe [1, 2]

→ Similarity score fixed and not customizable (rule-based or learning-based approaches)

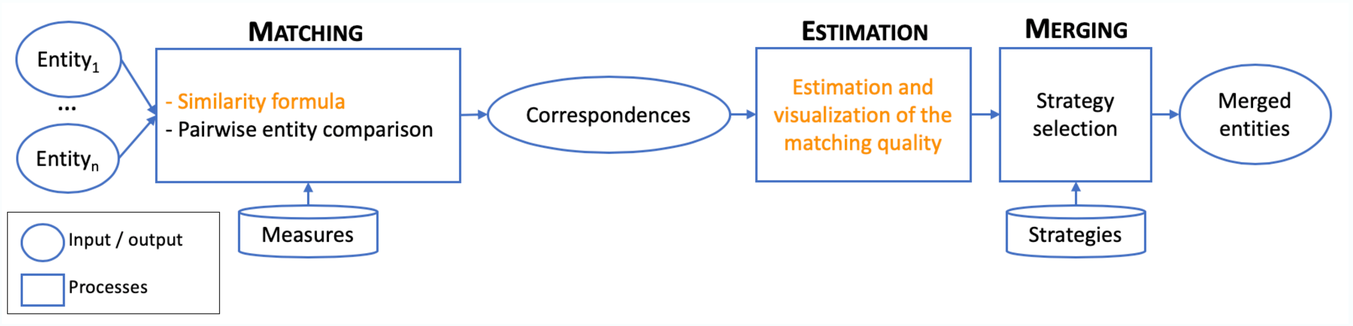
Building a benchmark: GeoBench [2]

→ Static datasets with small coverage and manual operations



[1] Hyunmo Kang, et al. Geoddupe: a novel interface for interactive entity resolution in geospatial data. In IV 2007.
[2] Vivek Sehgal, et al. Entity resolution in geospatial data integration. In GIS 2006.
[3] Anthony Morana, et al. GeoBench: a geospatial integration tool for building a spatial entity matching benchmark. In SIGSPATIAL 2014.

OVERVIEW OF GEOALIGN



CUSTOMIZING SIMILARITY FUNCTION

- Combination of *tokens* : weighted sum of the similarity values computed between attributes
- Threshold for decision making

$$\sum_{i=1}^{n} weight_i * sim_i(attribut_i) > threshold$$

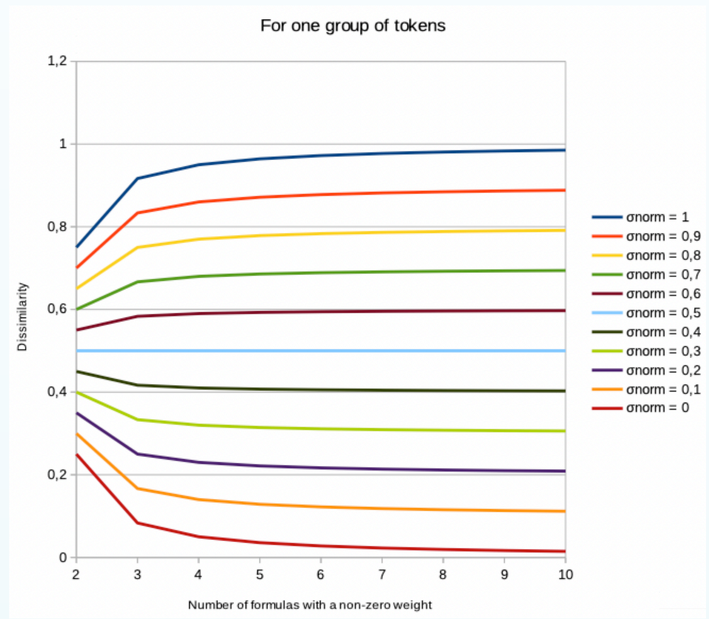
ESTIMATION OF THE MATCHING QUALITY (WITHOUT GROUND TRUTH)

Assumption: a candidate correspondence has more chances to be correct if it has been detected by several different functions

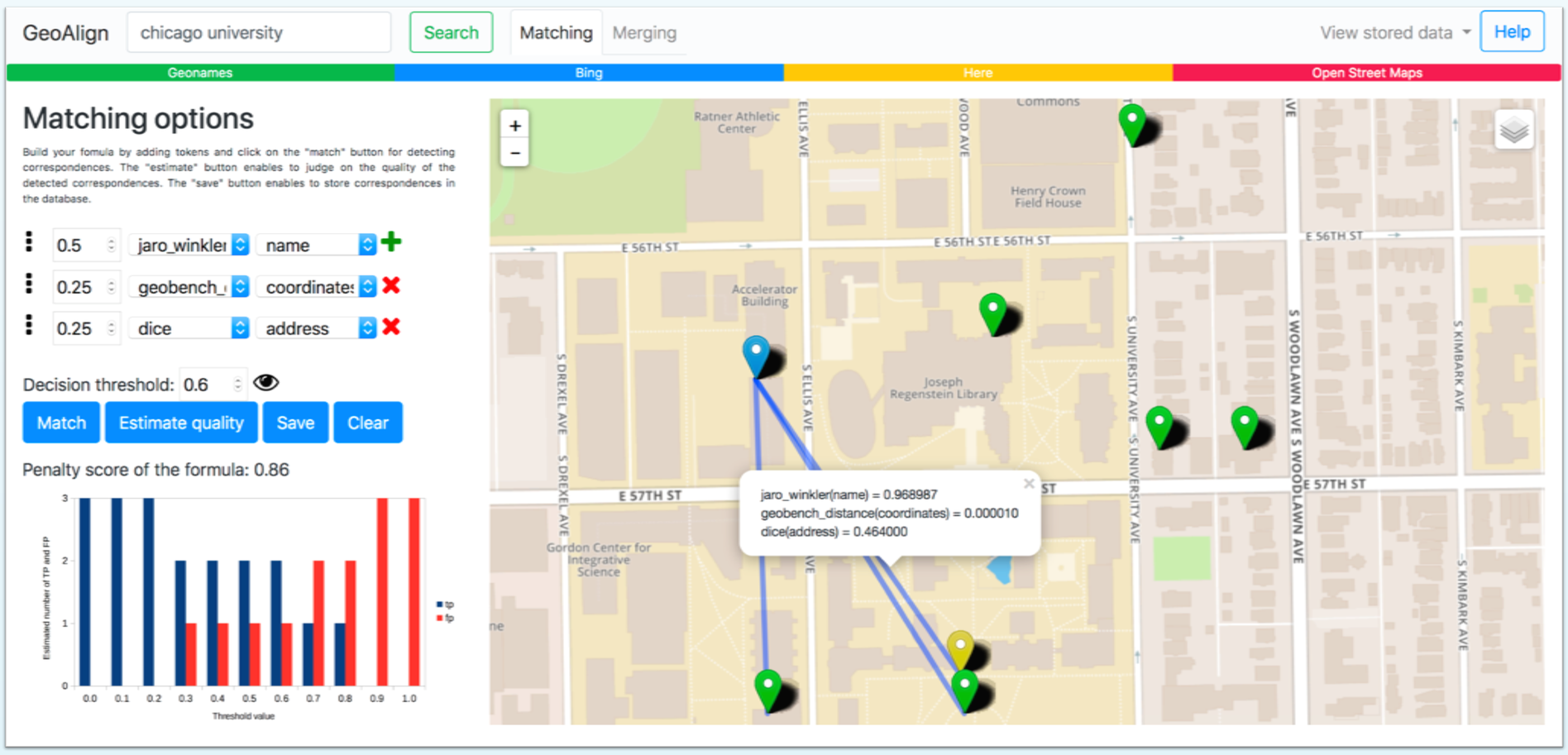
Computing the dissimilarity score between similarity functions

- Two tokens belong to the same group if they deal with the same attribute and if their measure belong to the same category (e.g. string-based, spatial…)
- We compute the dissimilarity Δj from a group of tokens
  - Computing the standard deviation of the weights
  - Weighting the standard deviation using an hyperbola function
  - Computing the average dissimilarity Δj:

$$\Delta j = f(x) = \begin{cases} 1, & n_j = 1 \\ \left(\frac{\sigma_j}{\sigma max_j} - 0.5\right) * c_j + 0.5, with\ c_j = 1 - \frac{0.25}{n_j - 1.5} \end{cases}$$



GEOALIGN PROTOTYPE



PERSPECTIVES

- More experiments for evaluating the estimation of the matching quality of detected correspondences
- Generalization with a rule-based combination (need assistance or learning to build the formula)