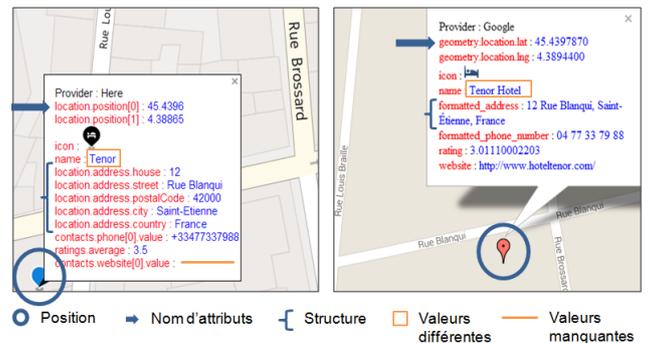


Analyse des données INSEE sur les quartiers

Fabien Duchateau, Franck Favetta (prénom.nom@liris.cnrs.fr)

Contexte. En France, l'INSEE propose un découpage du territoire en 50000 IRIS (Ilots Re-groupés pour l'Information Statistique), que l'on peut considérer comme des quartiers (e.g., environ 200 IRIS pour la ville de Lyon). Chacun de ces quartiers est décrit par 550 indicateurs en moyenne (e.g., nombre de restaurants, nombre de boulangeries, nombre de piscines). Comme les données fournies par l'INSEE sont éparpillées dans des fichiers Excel, nous les avons rassemblées dans une base de données MongoDB appelée [mongiris](#). Le principal problème vient du fait que ces données INSEE ne sont pas mises à jour chaque année (dans [mongiris](#), les données datent de 2014 et 2016), et que c'est à la charge des communes de renseigner ces informations.

Ce travail vise donc à vérifier l'exactitude des données INSEE en s'appuyant sur des applications cartographiques comme [Bing Maps](#), [Foursquare](#), [Here](#) ou [OpenStreetMap](#). Comme les fournisseurs retournent des points d'intérêt (POI) équivalents mais représentés différemment (cf figure ci-contre), une étape de comparaison (appariement spatial) de ces POI permet de détecter les doublons [1].



Objectifs. Ce projet cherche à atteindre les objectifs suivants :

- Étudier la littérature scientifique récente sur le *spatial entity matching* [1, 2, 3, 4];
- Sélectionner des indicateurs à vérifier, et développer un script Python permettant d'interroger au moins 2 fournisseurs cartographiques afin de dénombrer les POI d'une zone donnée (sans doublon);
- Analyser les résultats obtenus (e.g., catégorisation des quartiers contenant des données obsolètes/incomplètes).

Bibliographie.

- [1] Kai Sun, Yunqiang Zhu, and Jia Song. Progress and challenges on entity alignment of geographic knowledge bases. *ISPRS International Journal of Geo-Information*, 8(2):77, 2019. [Lien GI2019](#).
- [2] Emerson MA Xavier, Francisco J Ariza-López, and Manuel A Urena-Camara. A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys (CSUR)*, 49(2):1–34, 2016. [Lien CS2016](#).
- [3] Suela Isaj, Esteban Zimányi, and Torben Bach Pedersen. Multi-source spatial entity linkage. In *Symposium on Spatial and Temporal Databases*, pages 1–10, 2019. [Lien SSTD2019](#).
- [4] Setu Shah, Vamsi Meduri, and Mohamed Sarwat. Gem: An efficient entity matching framework for geospatial data. In *Geographic Information Systems*, pages 346–349, 2021. [Lien GIS2021](#).