GeoBench : un outil d'alignement entre entités spatiales pour la construction d'un benchmark cartographique

Thomas Morel, Anthony Morana

28 février 2014

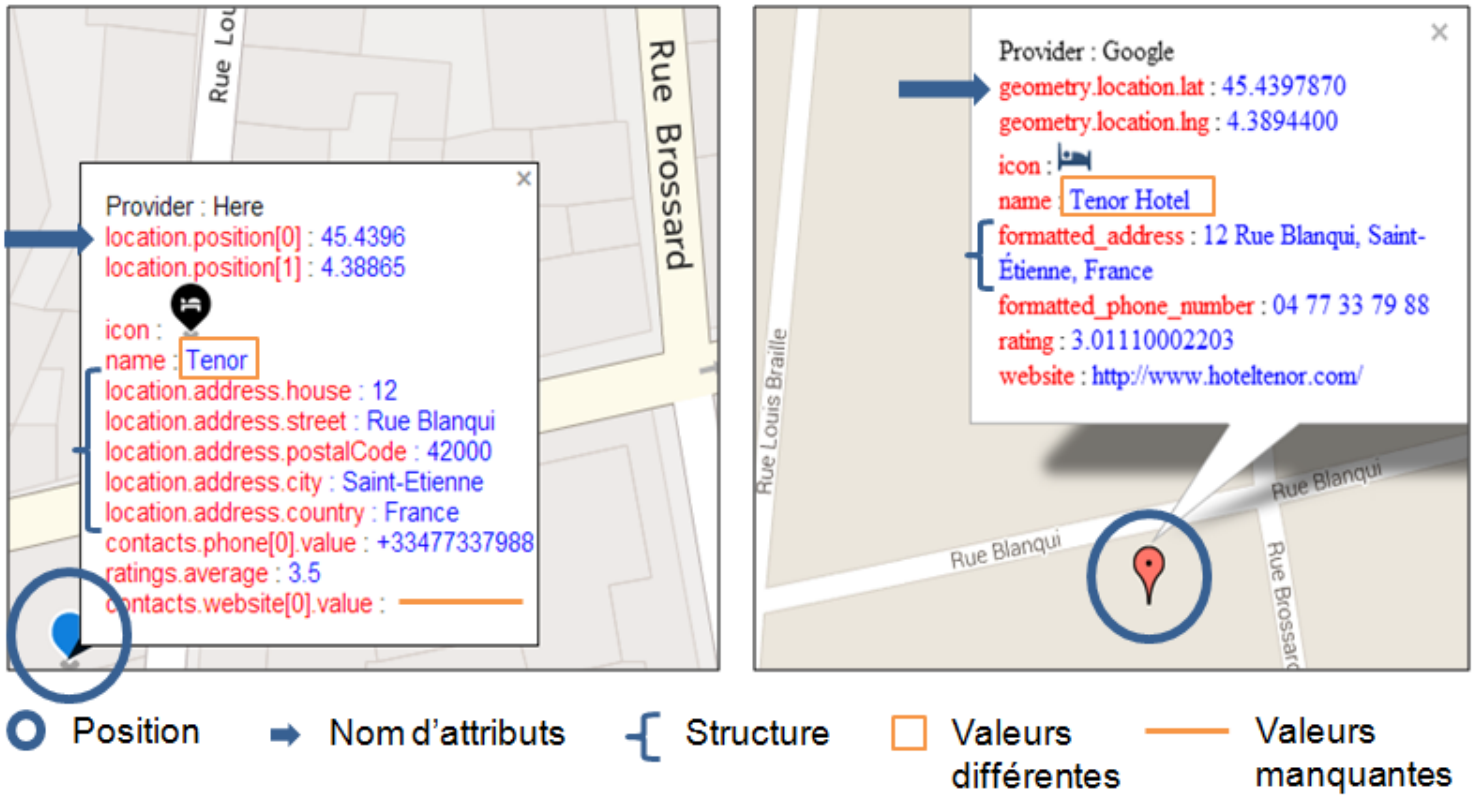
Encadrés par Fabien Duchateau et Bilal Berjawi

Contexte

- Applications mobiles géolocalisées en plein essor
- Un fournisseur cartographique possède ses points d'intérêts (POI), e.g., Basique de Fourvière
- Un POI est représenté par une ou plusieurs entités spatiales
- UNIMAP : projet IMU dont le but est d'intégrer les données qui proviennent d'entités spatiales de différents fournisseurs et représentant le même POI

Problématique

- Incohérence / incomplétude des données
- Problème de passage à l'échelle
- Problème d'évaluation des algorithmes



Travaux similaires

Approche d'alignement de Vivek et al. [SGV06] [1]

- Combinaisons d'attributs spatiaux et non spatiaux
- Jeu de données trop facile à aligner et non disponible

Approche de Safra et al. [SKS+10] [2]

- Seulement sur les attributs spatiaux
- Jeu de données indisponible

D'autres jeux de données sont disponibles [3] mais difficilement exploitables

1. V. Sehgal, L. Getoor, and P. Viechnicki. Entity resolution in geospatial data integration. ACM GIS, 2006
2. E. Safra, Y. Kanza, Y. Sagiv, Catriel Beeri, and Y. Doytsher. Location-based algorithms for finding sets of corresponding objects over several geo-spatial datasets. IJGIS, 2010
3. http ://www.cs.utexas.edu/users/ml/riddle/data.html

Contributions

- Développement de l'outil GeoBench
  - Permet de construire un jeu de données cartographique en proposant à l'expert des paires d'entités correspondantes
- Algorithme d'alignement d'entités provenant de fournisseurs hétérogènes
  - Bonne qualité d'alignement
  - Performant
- Génération et mise à disposition d'un jeu de données (benchmark)

Plan

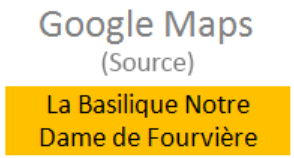
Blocking

Intérêt du blocking : regrouper un sous ensemble d'entités des fournisseurs à comparer avec une entité source (pour éviter le produit cartésien)

Notre proposition de blocking :

- Sélection d'un sous ensemble d'entités
  - Nom + coordonnées
  - Type + coordonnées
- Périmètre de recherche en fonction du type
- Retourne un ensemble d'entités à aligner avec l'entité source

Blocking (2)



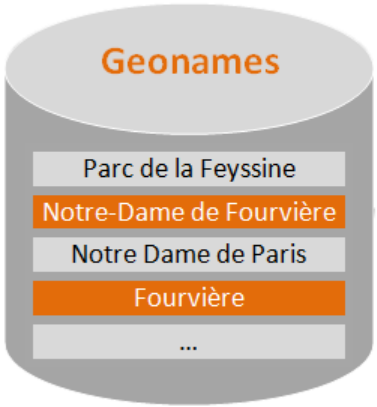


Figure: Illustration du blocking pour le POI « Basilique de Fourvière »

Blocking (2)

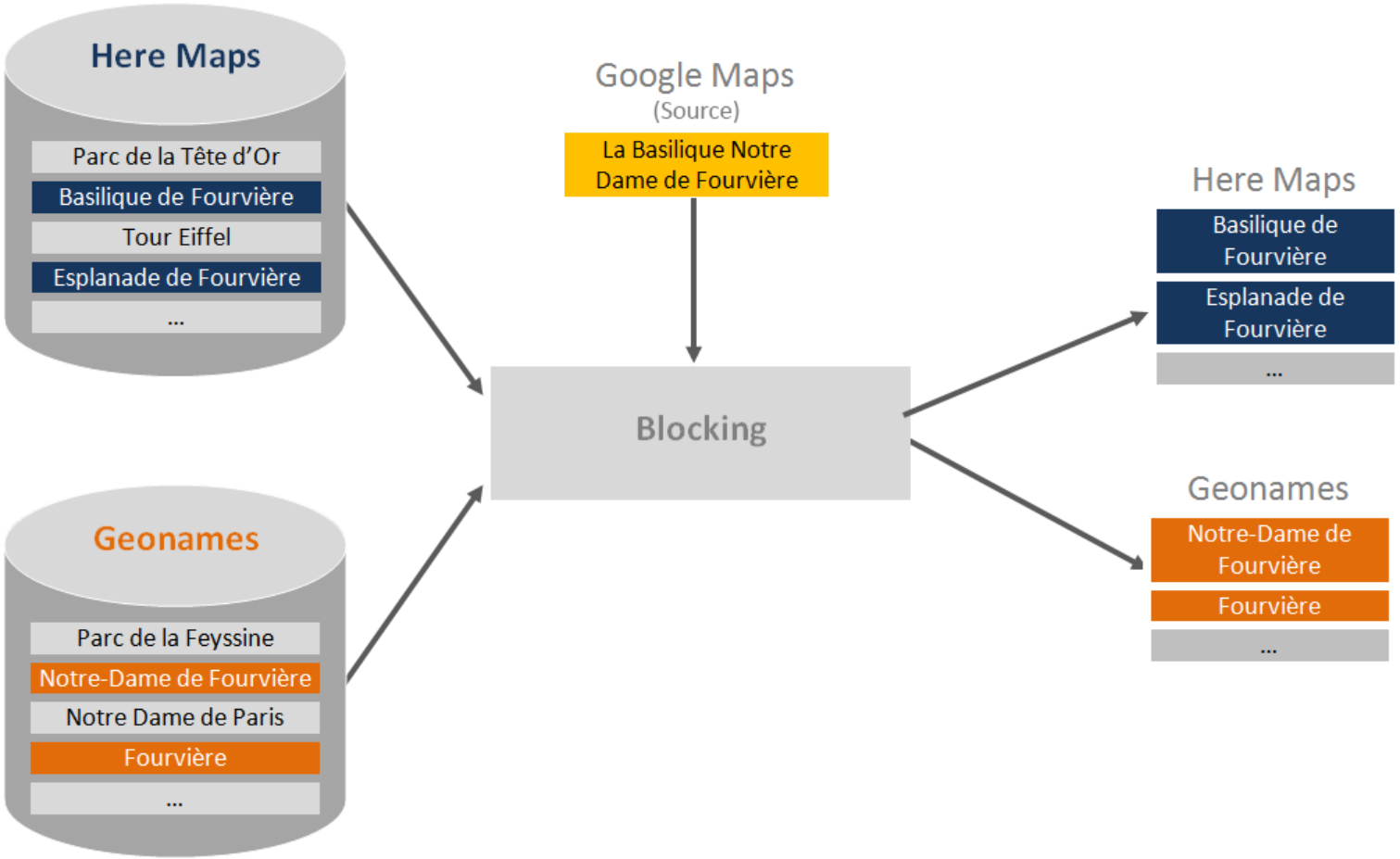


Figure: Illustration du blocking pour le POI « Basilique de Fourvière »

Alignement

Intérêt de l'alignement : trouver des correspondances entre entités spatiales en appliquant des mesures de similarité sur les sous ensembles résultant du blocking

Description d'une entité

- Attributs primaires : coordonnées, nom, type
- Attributs secondaires : adresse, téléphone, site web

Hypothèse : les correspondances entre attributs sont connues (schémas statiques de petite taille)

Alignement (2)

Notre proposition d'alignement

- Calcul de score de similarité entre attributs correspondants
  - Coordonnées : distance euclidienne (normalisée)
  - Types : comparaison booléenne à partir de la taxonomie des correspondances entre types de chaque fournisseur
  - Chaine de caractères : distance de Levenhstein (normalisée)
- Calcul d'un score de pertinence entre 0 et 1 pour chaque couple d'entités
  - Moyenne pondérée
  - 2/3 des scores sur les attributs primaires + 1/3 des scores sur les attributs secondaires

Alignement (3)

Figure: Illustration de l'alignement pour le POI « Basilique de Fourvière »

Alignement (3)

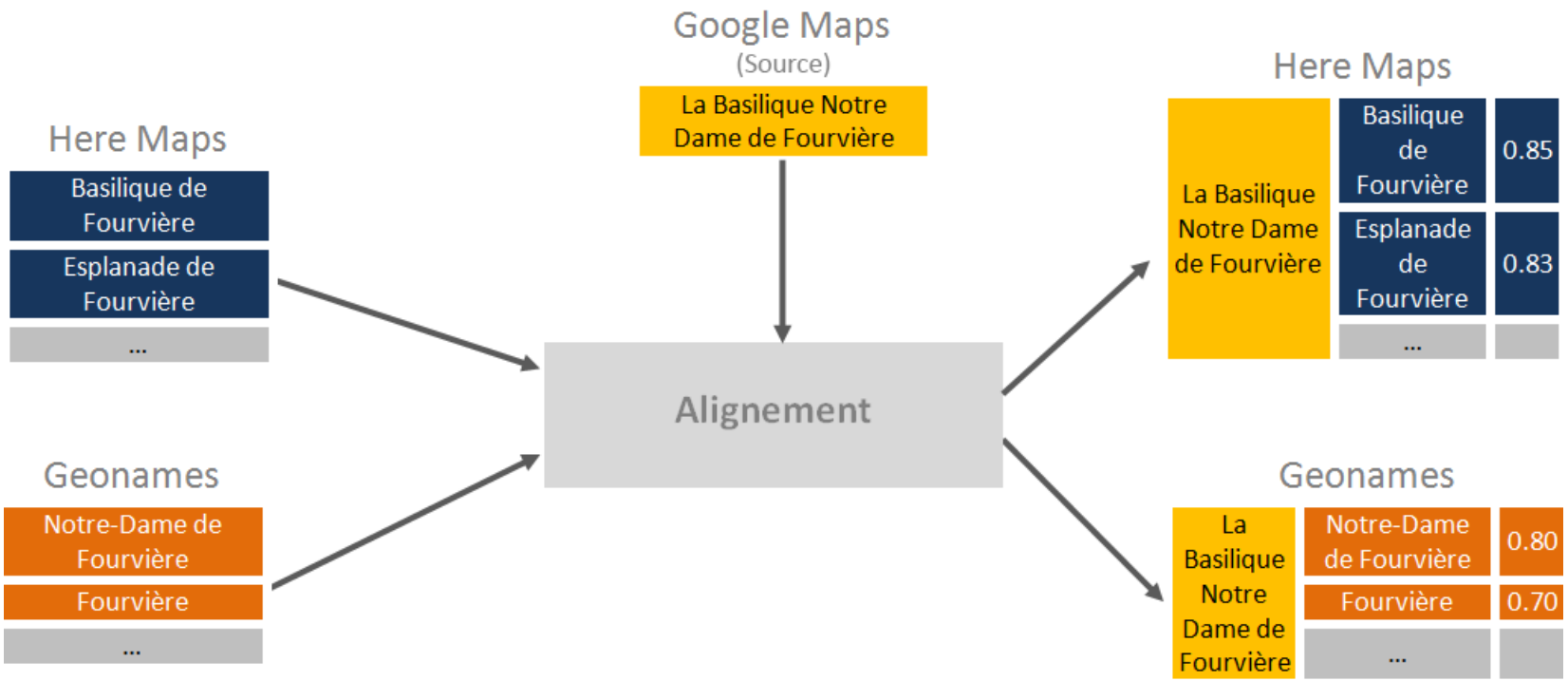
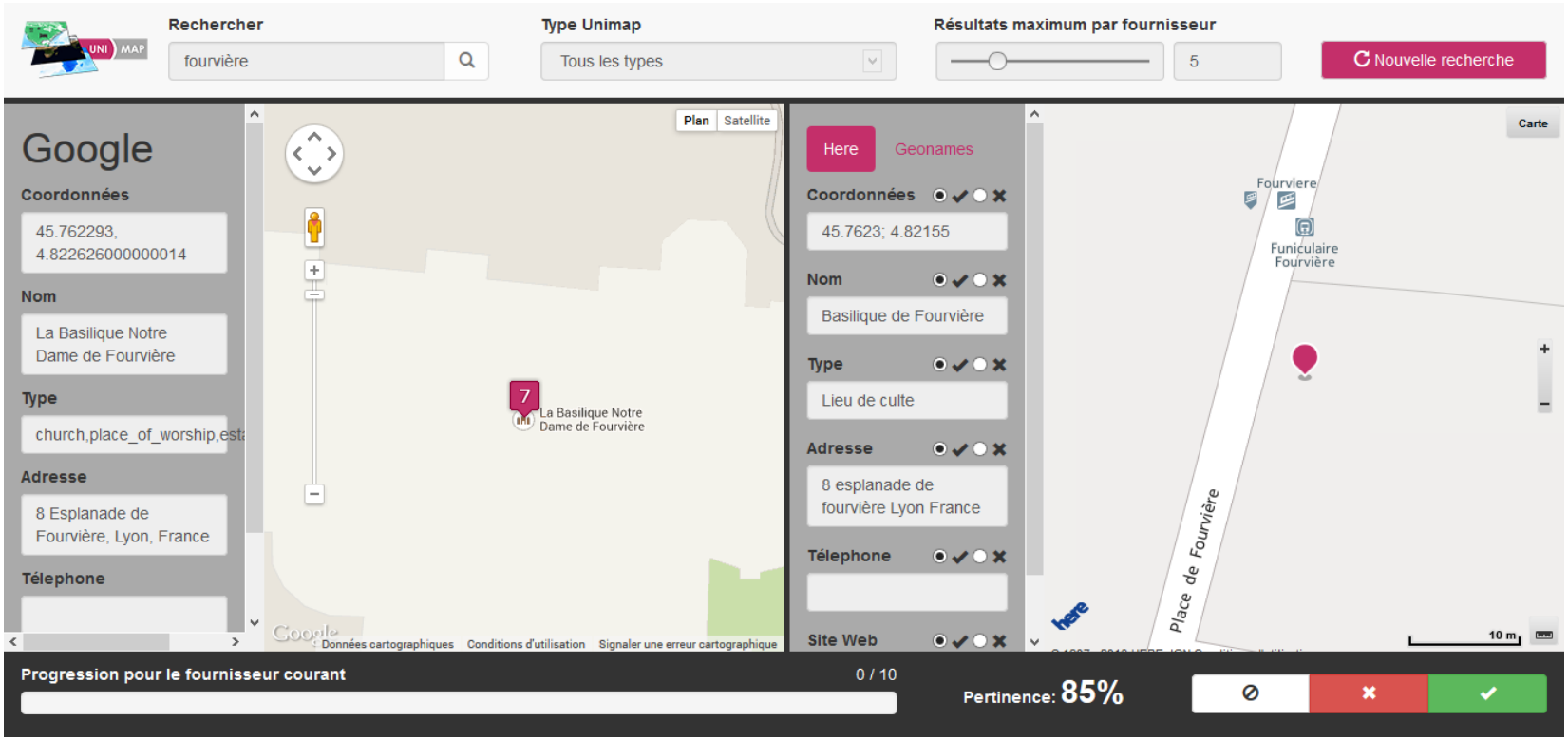


Figure: Illustration de l'alignement pour le POI « Basilique de Fourvière »

Plan

Aperçu de GeoBench



Alignement du POI Basilique de Fourvière avec Here à partir du fournisseur source Google Maps

Protocole d'évaluation

Utilisation de GeoBench par les experts Unimap pour construire un jeu de données

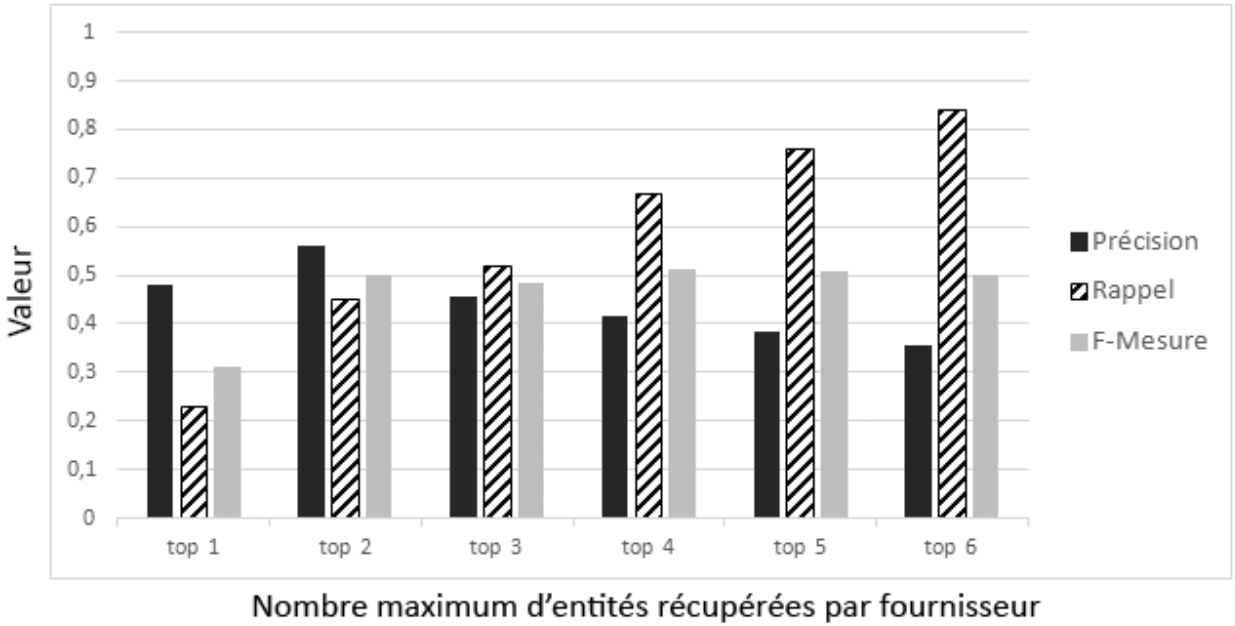
Fournisseur	Google Maps (source)	Here maps	Geonames
Nombre entités	100	310	157
Nombre entités en France	≈ 80	≈ 260	≈ 100
Nombre entités hors France	≈ 20	≈ 50	≈ 57
Nombre de correspondances correctes	-	68	31

Figure: Statistiques générales du jeu de données

Évaluation qualitative :

- Precision, rappel, F-Mesure
- Est-ce que GeoBench propose en haut du classement les entités correspondantes ?

Evaluation qualitative



- 85% des correspondances correctes
- L'augmentation du rappel compense la perte de précision (F-Mesure constante)

Conclusion

Développement de Geobench :

- Basé sur un algo d'alignement simple et performant (découverte de correspondances à la volée)
- 85% des entités correctes parmi les 6 premières propositions
- Algorithme de détection des différences entre les données d'entités correspondantes

Perspectives :

- Amélioration de la qualité (e.g., comparaison des types, poids du calcul de pertinence)
- Amélioration de l'ergonomie de GeoBench
- Ajout de nouveaux fournisseurs