Alignement d'entités spatiales avec GeoAlign

Nelly Barret

30 mai 2019

Résumé

Ce projet a pour but de créer un outil permettant la détection et la fusion d'entités géographiques. Il reprend les bases de l'outil GeoBench et apporte trois principales améliorations, que sont la personnalisation de la formule de similarité, l'estimation de la qualité des correspondances détectées et la fusion automatique de ces correspondances.

Mots-clés – appariement d'entités spatiales, qualité de l'appariement, fusion d'entités, intégration de données

1 Introduction

Dans le cadre du second semestre du Master 1 Informatique, j'ai travaillé sur le sujet POM (Projet d'Orientation en Master) intitulé « GeoBench v2 » sous l'encadrement de Messieurs Fabien Duchateau et Franck Favetta, maîtres de conférence et membres de l'équipe Bases de Données au LIRIS (Laboratoire d'InfoRmatique en Image et Système d'information). L'objectif de ce POM est de mettre à jour et de faire évoluer l'outil GeoBench, proposé en 2014, en un nouvel outil : GeoAlign. Dans ce rapport, nous aborderons dans un premier temps le contexte de ce projet ainsi que ses problématiques. Dans un second temps, nous donnerons une vue d'ensemble du travail réalisé en abordant la détection, l'estimation de la qualité et la fusion des correspondances. Enfin, nous conclurons et terminerons sur les perspectives de ce projet.

2 Contexte scientifique

De nos jours, les fournisseurs de données cartographiques sont au centre de nombreuses applications et sites web, e.g. pour la recherche d'itinéraire ou les objets connectés. Ces fournisseurs proposent des fonctionnalités variées telles que la recherche et l'affichage de POI (« Point Of Interest », comme des restaurants, des musées, etc.). Un POI est représenté par une ou plusieurs entités spatiales. Il est composé de deux types d'attributs : des attributs primaires, i.e. le nom, les coordonnées, le type de POI et des attributs secondaires, i.e. l'adresse, le numéro de téléphone et le site web. Les attributs primaires sont obligatoires pour chaque POI tandis que les secondaires, optionnels, permettent d'avoir plus de détails. Cependant, cette représentation est différente d'un fournisseur à un autre au niveau de l'exactitude, de la précision ou de l'exhaustivité des données. Par exemple, la figure ci-dessous illustre ces différences avec l'hôtel Ténor. Le fournisseur Here propose une localisation très détaillée tandis que Google Maps en propose une formatée. En revanche, Google Maps a des coordonnées géographiques (latitude et longitude) plus précises que Here. Enfin, le nom diffère entre les deux fournisseurs, ce qui constitue un challenge pour l'appariement d'entités.

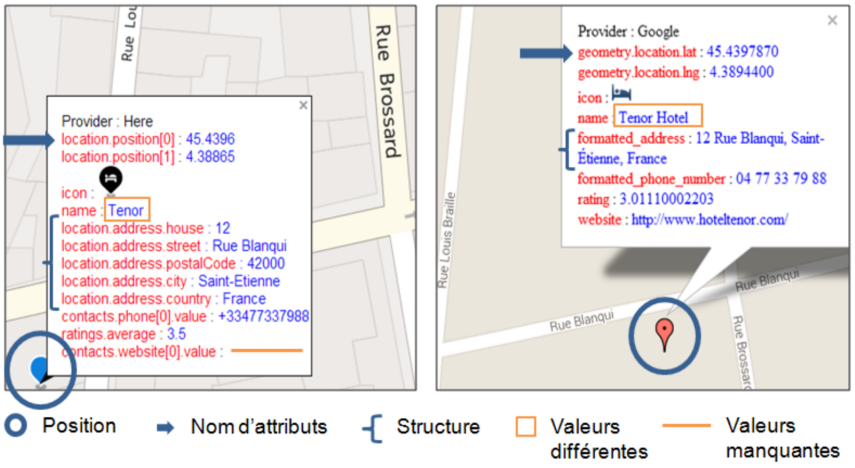

FIGURE 1 – POI de *l'hôtel Ténor à Saint-Etienne*, représenté par une entité Here (gauche) et une entité Google Maps (droite). Les catégories de différences entre les deux entités sont également spécifiées (e.g., position, structure)

Afin de réduire ces différences, il est possible de détecter, grâce à l'appariement d'entités (« *entity-matching* »), des correspondances entre les entités qui référencent le même POI. Ces entités peuvent ensuite être comparées et fusionnées pour améliorer la qualité des données relatives aux POI. Afin de détecter les entités similaires, il est d'abord nécessaire de connaître les attributs à comparer et donc d'apparier (aligner) les schémas des différents fournisseurs. L'appariement de schémas consiste à détecter les correspondances entre les attributs définis par des schémas hétérogènes, i.e. trouver les attributs équivalents des schémas proposés par les fournisseurs. Une fois les schémas appariés, il est possible de comparer les entités (appariement d'entités spa-

tiales). En général, les correspondances sont de type 1:1 (une entité d'un fournisseur ne correspond qu'à une seule entité d'un autre fournisseur) mais l'on trouve parfois des correspondances complexes (1:N ou N:M). Par exemple, une grande surface qui propose un service de station essence peut être représentée par une seule entité (la grande surface contient la station service) ou deux entités (la grande surface et la station service). L'état de l'art sur l'appariement d'entités décrit des approches qui exploitent les données descriptives (e.g. nom, adresse, type) et spatiales (e.g. coordonnées géographiques). Les mesures de similarité permettent de calculer un score de similarité entre deux attributs comparables. Pour l'étape d'appariement, plusieurs scores de similarité sont calculés, et il est ensuite nécessaire de les combiner pour décider si la paire d'entités est une correspondance ou non. L'une des techniques les plus utilisées consiste à calculer une moyenne pondérée d'un ensemble de scores de similarité, puis de décider selon une valeur seuil.

C'est le cas des outils GeoDDupe [3], Olteanu [7] et GeoBench [6] qui utilisent une fonction numérique pour la combinaison. Dans Sehgal [8] et MacKenzie [5], un apprentissage est mis en place pour le seuil de décision. Malgré la diversité des différentes approches d'appariement, la personnalisation de la formule de similarité reste limitée (principalement les poids associés à chaque mesure de similarité). Bien que le seuil de décision soit en général paramétrable, le fait que les mesures de similarité ainsi que les attributs sur lesquels elles s'appliquent soient « fixés » restreint aussi la personnalisation de ce seuil.
Lorsqu'une approche d'appariement d'entités est utilisée, il est important d'évaluer la qualité des correspondances obtenues. C'est l'objectif de GeoBench, qui facilite la construction d'un benchmark : la recherche d'entités est imposée chez un fournisseur spécifique (Google Maps) puis des suggestions d'entités correspondantes chez les autres fournisseurs sont proposées [6]. La validation des correspondances et des entités fusionnées est réalisée manuellement, bien que cela soit un travail fastidieux et qui semble difficilement possible dans le cas d'un passage à une échelle supérieure (e.g. nationale). Le benchmark construit est utile pour évaluer la qualité d'un algorithme, mais reste limité à quelques milliers d'entités réparties dans le monde. L'évaluation reste un problème ouvert car il n'est pas possible de connaître la « réalité terrain » et donc d'évaluer la qualité des correspondances pour l'ensemble de la planète.
Enfin, certaines approches permettent la fusion des entités correspondantes. L'objectif est d'obtenir une seule entité fusionnée avec, si possible, des données complètes, à jour et pertinentes. Différentes stratégies de fusion sont détaillées dans la littérature [2], comme le vote majoritaire, la moyenne ou la fraîcheur des données. Dans GeoBench, la fusion est également manuelle, bien qu'une assistance sélectionne par défaut, pour un attribut, la valeur la moins différente des autres. Par exemple, le fournisseur Google Maps propose l'entité nommée « Tour Eiffel, Paris, France », le fournisseur Here nomme l'entité correspondante « Eiffel Tower, Paris, France » et Geonames indique la valeur « Tour Eiffel, France ». L'assistance va présélectionner le nom issu de l'entité Geonames, qui est la moins différente selon la distance de Levenshtein.
Pour répondre à ces problèmes, nous présentons notre approche GeoAlign pour l'appa-

riement d'entités spatiales. Tout d'abord, GeoAlign inclut la construction personnalisée d'une formule de similarité pour la détection des correspondances. Il est possible de choisir les poids et le seuil, mais également les mesures de similarité et les attributs de la formule (cf section 3.1.1). Dans notre approche, l'appariement est automatique (au lieu de l'appariement manuel comme dans GeoBench). L'automatisation ainsi que la visualisation sur une carte facilitent les tests de différentes formules, mais cela ne permet pas de juger rapidement de la qualité obtenue par une formule. C'est pourquoi GeoAlign inclut un module d'estimation de la qualité des correspondances, basée sur le nombre et l'hétérogénéité des formules qui ont détecté une correspondance (cf section 3.1.2). Cette estimation, bien que préliminaire, offre une aide à l'utilisateur pour modifier sa formule de similarité par exemple. Enfin, concernant la fusion des entités correspondantes, plusieurs stratégies existantes ont été implémentées. Le choix de cette stratégie est important puisqu'il déterminera la qualité et la quantité de données pour les entités fusionnées. Ce processus de fusion est automatisé (cf section 3.2).

3 Travail réalisé

La figure 2 illustre le processus général de l'approche GeoAlign. Nos contributions principales apparaissent en orange. L'étape d'appariement se base sur les entités présentes dans l'emprise de la carte (partie visible de la carte) [1]. Cette emprise évite l'utilisation d'un algorithme de blocking (qui restreint le nombre d'entités à comparer). À partir de cet ensemble d'entités, l'étape d'appariement utilise une formule de similarité (somme pondérée de mesures appliquées à des attributs) et un seuil de décision pour faire correspondre (ou non) deux entités. Une estimation de la qualité des correspondances générées est calculée et présentée à l'utilisateur, qui peut alors décider de relancer l'appariement avec de nouveaux paramètres ou de continuer vers la fusion des entités correspondantes.

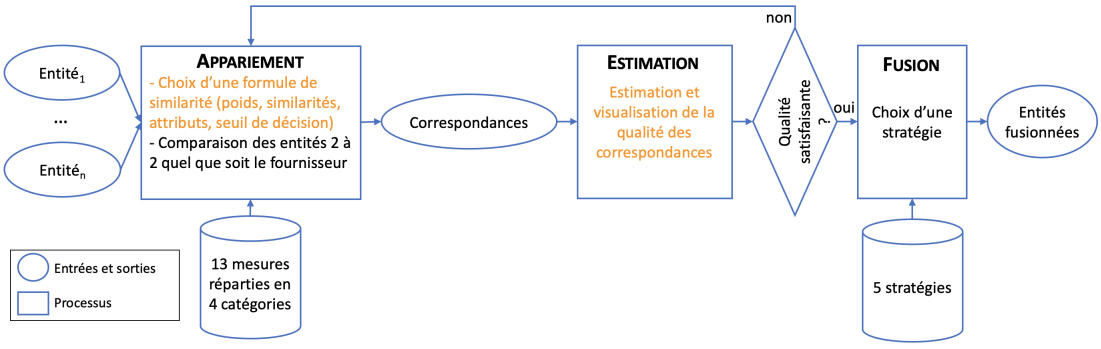

FIGURE 2 – schématisation des étapes pour l'appariement et la fusion des données.

1. À cause des limitations imposées par les API des fournisseurs, l'emprise contient généralement quelques centaines d'entités. Dans les cas où elle en contient davantage, l'utilisateur doit zoomer.

3.1 Détection des correspondances

La première étape est la détection des correspondances entre les entités. Cette détection se base sur la similarité de celles-ci grâce à une formule de similarité puis estime la qualité des correspondances détectées.

3.1.1 Calcul de similarité entre entités

Les quatre fournisseurs (Geonames, Bing, Here et Open Street Maps) ayant chacun leur propre hiérarchie de types de POI (e.g., la catégorie « restaurant » peut inclure les bars, les restaurants et les cafés selon le fournisseur), il a été nécessaire de construire notre propre hiérarchie. Celle-ci permet donc de mettre en relation les types de POI proposés par les différents fournisseurs. Le type d'un fournisseur donné n'est associé qu'à un seul type de notre hiérarchie comme le montre l'annexe 5.1. Cette relation exclusive permet notamment de comparer les entités qui ont des types similaires (e.g., un type « restaurant » et un type « tourisme »), comme illustré dans l'annexe 5.2.

Dans un premier temps, il est nécessaire de détecter les entités représentant le même POI. Pour cela, il faut modéliser une formule capable de déterminer le degré de correspondance entre deux entités et ainsi déterminer si celles-ci sont assez similaires pour correspondre au même POI. La formule de similarité s'exprime par la combinaison des scores de similarité obtenus pour différents attributs. Ces scores sont calculés par des mesures appartenant à différentes catégories, e.g. terminologique (Levenshtein, Jaro-Winkler) ou spatiale (distance euclidienne). Au total, GeoAlign implémente treize mesures réparties en quatre catégories, comme présenté dans l'annexe 5.3. À partir de ces différents scores, une combinaison qui a fait ses preuves [4] est la moyenne pondérée (elle permet en plus de représenter d'autres fonctions de combinaison). La moyenne pondérée produit un score de similarité global entre 2 entités. Enfin, il faut déterminer si le score obtenu par la formule de similarité permet de considérer une paire d'entités comme une correspondance ou pas. Un seuil permet cette étape de décision.

$$\text{Calcul des similarités : } f(e_1, e_2) = \sum_{i=1}^{n} poids_i * sim_i(attribut_i)$$

$$\text{Décision : } f(e_1, e_2) > seuil$$

GeoAlign permet de construire cette formule de similarité en choisissant les attributs, les mesures de similarité et les poids via une interface graphique. Il est aussi possible de choisir le seuil de décision. Cette formule personnalisée est ensuite appliquée sur les entités dans l'emprise courante de la carte. Pour chaque paire d'entités, si $f(e_1, e_2) > seuil$ où f est la formule de similarité et e_1, e_2 sont des entités, alors la

paire est considérée comme correspondante et elle est affichée sur la carte au moyen d'un trait entre les deux entités.

La possibilité de construire sa propre formule nécessite de vérifier que la formule soit correcte (somme des poids égale à 1) mais surtout qu'elle soit pertinente. En effet, le choix des poids et du seuil de décision a une influence non négligeable sur le nombre de correspondances trouvées ainsi que sur la qualité de celles-ci. Une formule trop permissive crée des correspondances qui n'existent pas dans la réalité (faux-positif) et une formule trop stricte ne détecte pas toutes les correspondances qui existent réellement (faux-négatif). Le choix des différents poids ainsi que celui du seuil global n'est donc pas trivial. Il est possible d'appliquer des restrictions afin de limiter certaines formules (surtout celles qui sont trop permissives), e.g. en limitant le nombre de correspondances calculé. En effet, nous faisons l'hypothèse que la plupart des correspondances sont de type 1:1 (ce qui se vérifie dans la majorité des cas, comme le montre les données du benchmark construit avec GeoBench [6]). Cela signifie qu'en moyenne, on détecte pour une entité donnée une seule entité correspondante par fournisseur. Ainsi, il est possible de calculer le nombre moyen de correspondances par fournisseur noté λ et un score de pénalité ϵ tel que $\epsilon = \frac{1}{\lambda}$. Par exemple, si une fonction de similarité détecte 1,2 correspondance en moyenne, son score de pénalité sera de 0,83. Ce score rentre en compte dans l'estimation de la qualité des correspondances détectées.

3.1.2 Estimation de la qualité

Après avoir détecté des correspondances, et n'ayant pas de réalité-terrain, il semble important d'estimer leur qualité pour fournir une indication à l'utilisateur. L'intuition sur laquelle repose cette estimation est qu'une correspondance a plus de chances d'être correcte si elle est détectée par plusieurs formules et si ces formules sont très différentes (c'est-à-dire qu'elles utilisent des mesures de similarité différentes appliquées à des attributs différents). Cette estimation de la qualité se découpe en plusieurs grandes parties. La première est la dissimilarité entre les fonctions de similarité. En effet, chaque correspondance est détectée par minimum 1 formule de similarité et k formules maximum. Chaque formule de similarité est une somme de tokens : un token est l'application d'un poids sur une mesure portant sur un attribut, e.g. $f(e_1, e_2) = token_1 + token_2$ où $token_1$ = *0.4*levenshtein(nom)* et $token_2$ = *0.6*distance(coordonnées)*.
Afin de comparer deux formules, il est nécessaire, pour la seconde étape, d'analyser quels sont leurs tokens similaires. Deux tokens sont similaires s'ils portent sur le même attribut et si leurs mesures appartiennent à la même catégorie (c.f. annexe 5.3). Par exemple, *0.6*levenshtein(nom)* et *0.7*jaro(nom)* sont deux tokens similaires selon cette définition. Pour chaque formule ayant détecté une correspondance donnée, les tokens similaires sont rassemblés pour former un groupe de tokens similaires. Puis pour chaque groupe de tokens similaires, on calcule l'écart-type de leurs poids puis on normalise cette valeur (par la valeur d'écart-type maximale qui dépend du nombre

de formules dans le groupe). Ce calcul d'écart-type permet de prendre en compte la répartition des tokens sur l'ensemble des formules dans leurs groupes (e.g., si l'écart-type d'un groupe est faible, cela signifie que les tokens ont des poids plutôt similaires dans ce groupe, sinon c'est que leurs poids sont plutôt différents).
La troisième étape correspond au calcul de la dissimilarité d'un groupe de tokens en modifiant l'écart-type normalisé selon une courbe basée sur une hyperbole (qui dépend du nombre de formules dans le groupe (c.f. annexe 5.4). Cette hyperbole varie rapidement pour un nombre bas de formules puis plus lentement au-delà (à partir d'environ 4 à 5 formules). L'idée est de modifier fortement la dissimilarité que représente l'écart-type normalisé quand le nombre de formules est faible, et de le modifier plus faiblement quand ce nombre croît. Les groupes de tokens contenant un seul token ont une dissimilarité fixée à 1 (car ces tokens « uniques » contribuent grandement à rendre un ensemble de formules dissimilaires). La dissimilarité Δj d'un groupe de tokens j est caractérisée par la formule suivante :

$$\Delta j = \begin{cases} j = 1 \text{ si taille(groupe)} = 1 \\ (\frac{\sigma_j}{\sigma_{max_j}} - 0.5).c_j + 0.5 \text{ sinon, où } c_j = 1 - \frac{0.25}{n_j - 1.5} \end{cases}$$

Enfin, la dissimilarité globale pour l'ensemble des formules associées à une correspondance est la moyenne des dissimilarités de l'ensemble des groupes de tokens :

$$\Delta GT = \frac{\sum_{j=1}^{n_{GT}} \Delta j}{n_{GT}}$$

où n_{GT} est le nombre de groupes de tokens pour la correspondance

Ce score de dissimilarité renvoie un nombre appartenant à l'intervalle [0, 1], sachant que plus ce score est proche de 1, plus les fonctions de similarité de la correspondance sont différentes (au pire aucune d'entre elles ne partage un token avec une autre). Une fois la dissimilarité d'une correspondance calculée, il faut estimer la pertinence de cette correspondance, i.e. s'il s'agit d'un vrai positif (TP) ou d'un faux positif (FP). Un vrai positif est une correspondance détectée qui devait l'être tandis qu'un faux positif est une correspondance détectée qui ne devrait pas l'être.

Une correspondance c est détectée par une ou plusieurs formules de similarité, i.e. un ensemble de groupes de tokens. Le score de pertinence ϕ de cette correspondance, estimé entre 0 et 1, est la moyenne pondérée de tous ses scores de similarité par le score de dissimilarité de ses fonctions, comme le montre la formule :

$$\phi(c) = \frac{\sum_{i=1}^{k} f_i(c)}{k} \Delta GT$$

Lorsque ce score de pertinence a été calculé pour toutes les correspondances, il est possible d'estimer la qualité globale. En appliquant un seuil ou un top-K, certaines correspondances seront considérées comme correctes (TP), les autres seront classées comme incorrectes (FP). À partir de ces chiffres, il est même possible d'estimer la précision en tant que qualité globale avec la formule :

$$\text{Précision} = \frac{TP}{TP+FP}$$

Dans le prototype GeoAlign, nous avons décidé de ne pas choisir arbitrairement une valeur seuil ou top-K. Un graphique est affiché avec le nombre estimé de TP et de FP pour des seuils variant de 0.1 dans l'intervalle [0, 1]. Aussi, l'utilisateur peut visualiser l'estimation de la qualité et son évolution (c.f. annexe 5.5).

L'un des inconvénients de cette approche vient du fait que les formules, saisies par les différents utilisateurs, peuvent être plus ou moins adaptées. En particulier, certaines formules peuvent être très laxistes (e.g., avec un seuil très faible), et donc produire une quantité importante de correspondances (dont de nombreux faux positifs). Pour pallier ce problème, nous utilisons le score de pénalité ϵ (présenté en fin de section 3.1.1) afin de pénaliser les scores de similarité de ces formules trop permissives. Les correspondances peuvent être utilisés par la suite pour la fusion.

3.2 Fusion des correspondances

Après avoir détecté les correspondances, il est possible de fusionner les données afin de produire une entité unifiée représentant le POI [1, 2]. Lorsque plusieurs entités correspondent à un POI, il faut choisir quelles sont les données les plus pertinentes. Le choix entre les différentes valeurs peut se faire selon plusieurs stratégies [2] :

- Aléatoire : pour chaque attribut, une valeur est sélectionnée au hasard.
- Basé sur un fournisseur : les données conservées lors de la fusion sont celles mentionnées par le fournisseur choisi.
- Vote majoritaire : cette méthode consiste à choisir la décision prise par le maximum de sources (i.e. de valeurs) mais nécessite dans notre contexte d'autoriser une faible imprécision (car deux valeurs sont rarement strictement identiques). Elle est particulièrement adaptée pour la prise de décisions.

Après avoir présenté la détection et la fusion des correspondances, deux scénarios sont présentés pour illustrer l'utilisation de l'outil GeoAlign.

3.3 Scénarios

Dans cette section, un premier scénario s'attache à démontrer l'utilisation de GeoAlign pour l'appariement et la fusion automatiques. Un second scénario illustre l'intérêt de la personnalisation des formules ainsi que l'estimation de la qualité des correspondances produites. Des captures d'écran du prototype figurent dans l'annexe 5.5.

3.3.1 Scénario 1 : automatisation de l'appariement et de la fusion

Alice vit à Lyon depuis quelques mois et elle souhaite faire un restaurant avec ses amis. Elle décide d'utiliser GeoAlign afin de trouver des établissements avec leurs jours d'ouverture et leur numéro de téléphone pour réserver. Elle saisit « restaurant Lyon » dans la barre de recherche. GeoAlign place sur la carte, à l'aide de marqueurs, les différents établissements qui correspondent à sa requête. En cliquant sur ces marqueurs, les informations secondaires sont affichées dans une infobulle. Alice remarque le restaurant « Aromatic », mais celui-ci n'a pas de numéro de téléphone. Elle décide alors d'apparier les restaurants puis de les fusionner afin d'obtenir toutes les informations nécessaires sur ce restaurant. Grâce à l'appariement et la fusion automatique, Alice peut ainsi rapidement vérifier que l'établissement est bien ouvert puis téléphoner pour réserver une table.

3.3.2 Scénario 2 : personnalisation de la formule de similarité

Bob est sociologue et étudie l'animation des quartiers en fonction de facteurs sociaux. Il travaille sur les données que fournit l'INSEE (Institut National de la Statistique et des Etudes Economiques). Seulement, ces données datent de 2015. Dans le quartier Croix Rousse, l'INSEE indique un total de 144 restaurants. Bob veut vérifier que ce nombre reflète toujours la réalité. Il décide alors d'utiliser GeoAlign. Il personnalise la formule de similarité afin d'adapter au mieux les correspondances détectées. Après avoir construit, testé et adapté des formules de similarité variées (i.e., exploitant différentes caractéristiques), Bob peut estimer et visualiser le nombre de correspondances correctes, ce qui le conforte dans l'idée que de nouveaux restaurants ont ouvert depuis 2015 dans ce quartier.

4 Conclusion et perspectives

L'objectif initial de ce POM était de reprendre et améliorer GeoBench, mais cette mission a évolué avec la proposition de deux contributions importantes : la personnalisation d'une formule de similarité et l'estimation de la qualité des correspondances. Ces dernières sont implémentées dans le prototype GeoAlign destiné à l'appariement et la fusion d'entités spatiales. Durant ces cinq mois de projet, je me suis d'abord penchée sur l'intégration des différents fournisseurs dans l'application. Je me suis ensuite concentrée sur la détection des correspondances : construction dynamique de la formule de similarité, application de cette formule sur les entités présentes dans l'emprise courante de la carte, stockage des entités et des correspondances détectées dans une base de données MySQL. Enfin, j'ai travaillé sur l'étape de fusion des entités en implémentant quelques stratégies existantes, puis ajouté l'estimation de la qualité des correspondances. En parallèle de l'implémentation de ces différents modules, j'ai travaillé sur l'interface du prototype : ajout dynamique d'éléments dans la formule de similarité, affichage des données relatives aux entités et affichage des correspondances. Ce prototype fait l'objet d'un article qui sera soumis mi-juin à la conférence ACM SIGSPATIAL.

Ce projet a plusieurs perspectives. La première serait sur le plan technique. En effet, GeoAlign pourrait proposer des fonctionnalités de gestion des données, e.g. la mise à jour des valeurs des entités déjà fusionnées. Une autre perspective serait de mesurer si l'estimation de la qualité est effectivement pertinente, et à partir de combien de formules. Pour cela, il est possible d'utiliser les benchmarks existants (OAEI pour l'alignement d'ontologies, Leipzig datasets pour l'appariement d'entités). Bien que GeoAlign utilise l'emprise de la carte afin de limiter le nombre d'entités utilisées dans la détection des correspondances, il serait possible de mettre en place un blocking adaptatif. Pour la formule de similarité, il serait possible de mettre en place des formules plus complexes qu'une somme pondérée, e.g. avec un arbre de poids et de mesures, ou de la modéliser selon une approche *rule-based*, i.e. une conjonction de tokens. La démarche serait d'assimiler les tokens (calculés pour l'estimation de la qualité des correspondances) à des règles et d'assimiler une formule de similarité à une conjonction de règles (i.e. de tokens).

Références

[1] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Kevin Murphy, Shaohua Sun, and Wei Zhang. From data fusion to knowledge fusion. *PVLDB*, 7(10) :881–892, 2014.

[2] Xin Luna Dong and Felix Naumann. Data fusion - resolving data conflicts for integration. *PVLDB*, 2(2) :1654–1655, 2009.

[3] Hyunmo Kang, Vivek Sehgal, and Lise Getoor. Geoddupe : A novel interface for interactive entity resolution in geospatial data. In *International Conference on Information Visualisation*, pages 489–496, 2007.

[4] Hanna Köpcke and Erhard Rahm. Frameworks for entity matching : A comparison. *Data & Knowledge Engineering*, 69(2) :197–210, 2010.

[5] Grant McKenzie, Krzysztof Janowicz, and Benjamin Adams. Weighted multi-attribute matching of user-generated points of interest. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 440–443, New York, NY, USA, 2013. ACM.

[6] Anthony Morana, Thomas Morel, Bilal Berjawi, and Fabien Duchateau. Geobench : a geospatial integration tool for building a spatial entity matching benchmark. In *International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, pages 533–536. ACM, 2014.

[7] AM Olteanu. A multi-criteria fusion approach for geographical data matching. *International Symposium in Spatial Data Quality*, 2007.

[8] Vivek Sehgal, Lise Getoor, and Peter Viechnicki. Entity resolution in geospatial data integration. In Rolf A. de By and Silvia Nittel, editors, *GIS*, pages 83–90. ACM, 2006.

5 Annexes

5.1 Annexe 1 : hiérarchie commune et appariement des types

L'objectif principal de la hiérarchisation des types est que tous les fournisseurs aient les mêmes catégories de types afin de pouvoir comparer les entités uniformément. Pour cela, nous avons construit manuellement une hiérarchie de types (c.f. Figure 3). Elle s'inspire des regroupements de types que l'on trouve chez certains fournisseurs (e.g Geonames).

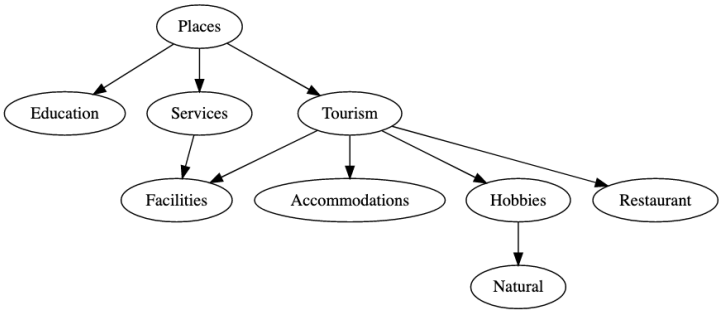

FIGURE 3 – hiérarchie des types.

Comme chaque fournisseur possède une liste de types plus ou moins fournie et détaillée (e.g., une dizaine de types chez Bing contre plus de 600 types chez Geonames), nous avons manuellement apparié les types de chaque fournisseur avec notre hiérarchie de types. La figure ci-dessous montre cet appariement pour le fournisseur Here (format JSON). À chaque type de notre hiérarchie (i.e., clé du dictionnaire, comme « accommodations » ou « places »), on associe les types correspondants équivalents ou inclus du fournisseur sous forme d'une liste de valeurs, comme « accommodation », « hotel », « camping ».

```
{
  "accommodations":["accomodation","hotel","camping",...],
  "places":["administrative-region","city-town-village",...],
  "education":["education-facility"],
  "facilities":["railway-station","parking-facility",...],
  "hobbies":["cinema","theatre-music-culture", ...],
  "natural":["forest-health-vegetation",...],
  "restaurant":["restaurant","snacks-fast-food",...],
  "services":["shopping","pharmacy","fire-department",...],
  "tourism":["sights-museums","landmark-attraction",...]
}
```

Grâce à notre hiérarchie commune de types et l'appariement des types de chaque fournisseur, nous pouvons comparer facilement les types, que ce soit lors d'une recherche ou pour calculer un score de similarité entre deux types (cf Annexe 5.2). Cette solution permet également de prendre en compte l'évolution des types.

5.2 Annexe 2 : similarités dans la hiérarchisation des types

Les types de notre hiérarchie ne sont pas exclusifs, et certains sont sémantiquement liés (e.g., « tourisme » et « restaurant »). Pour comparer les types de deux entités, nous nous inspirons de la mesure de Resnik qui calcule une similarité selon le nombre d'arêtes entre deux types dans la hiérarchie. Comme notre arbre est de petite taille, nous avons directement calculé le coefficient de similarité entre chaque paire de types. La figure 4 donne ces coefficients entre types (dans l'intervalle [0, 1]). Un coefficient à 0 indique que les deux types n'ont aucune caractéristique en commun, et un coefficient situé entre 0 et 1 indique le pourcentage de ressemblance entre les deux types. Par exemple, le type générique « places » a un coefficient de 1 avec lui-même et un coefficient de 0.8 avec les autres types.

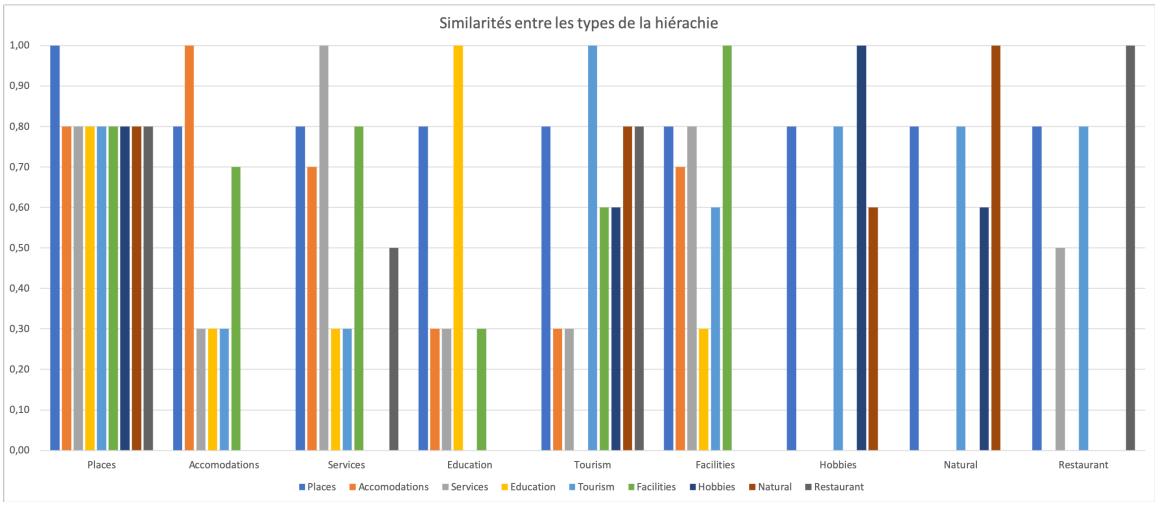

FIGURE 4 – similarités entre les types de la hiérarchie.

5.3 Annexe 3 : catégories de mesures

Le tableau ci-dessous présente les mesures de similarité implémentées dans GeoAlign, classées par catégorie.

String-based	Language-based	Phonétique	Spatiale	Sémantique
Jaro Jaro-Winkler Levenshtein Dice Jaccard	Trigrams	MRA Caverphone Metaphone	GeoBench's distance HSpSim 1 HSpSim 2 Euclide	Similarité des types

Concernant les mesures spatiales, la « GeoBench's distance » provient du travail sur GeoBench [6] tandis que les mesures « HSpSim1 » et « HSpSim2 » sont des propositions que nous avons définies. Des détails sur ces trois mesures sont donnés ci-dessous.

Distance de GeoBench

$$distance([lat1, lng1], [lat2, lng2]) = \frac{1}{(\arccos(\sin(lat1) * \sin(lat2) + \cos(lat1) * \cos(lat2) * \cos(lng2 - lng1)) * 6378137)/10}$$

Cette mesure calcule la distance angulaire entre deux points [lat1, lng1] et [lat2, lng2]. En multipliant cette distance par le rayon de la Terre (6 378 137 mètres) puis en divisant le résultat par 10, on obtient la distance séparant les deux points en décamètre.

HSpSim 1 et HSpSim 2

L'idée de ces mesures est de calculer une distance moins « abrupte » que l'inverse de la distance euclidienne ou celle proposée par GeoBench. HSpSim1 et HSpSim2 se basent sur respectivement une hyperbole et une double hyperbole. Chaque hyperbole est paramétrable en passant par trois points. Les deux courbes passent par un point de distance proche où la similarité est estimée maximale (1) et un point de distance éloignée où la similarité est estimée minimale (0). HSpSim1 passe par un point de distance intermédiaire où la similarité est considérée comme moyenne (0.5). HSpSim2 passe par trois points intermédiaires dont l'un où les hyperboles se rejoignent. Ces mesures nécessitent de futures expérimentations afin d'évaluer leur pertinence.

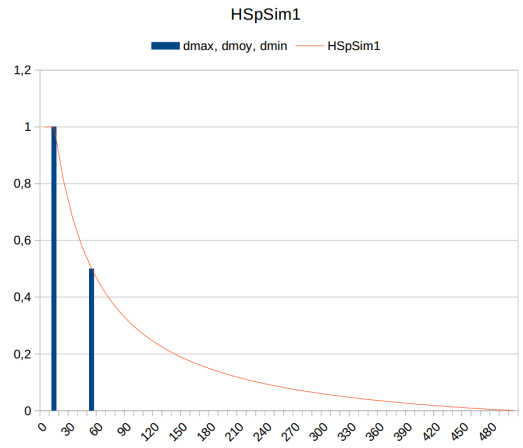

FIGURE 5 – courbe selon HSpSim1.

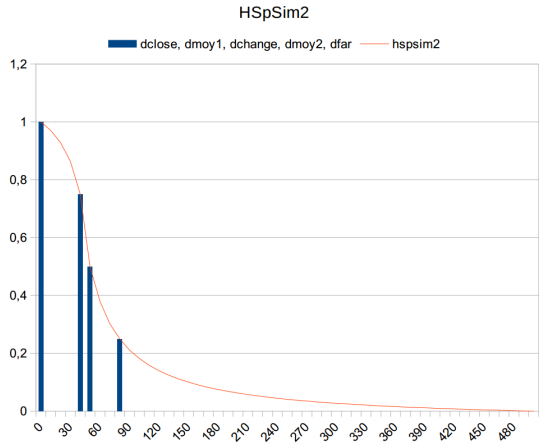

FIGURE 6 – courbe selon HSpSim2.

5.4 Annexe 4 : courbe d'ajustement de la dissimilarité

Cette annexe décrit la courbe basée sur une hyperbole utilisée dans le calcul de la dissimilarité. Elle permet d'ajuster le score de dissimilarité en fonction du nombre de tokens (i.e. de formules) dans un groupe de tokens. Par exemple, si un groupe de tokens contient deux tokens issus de deux formules, sa dissimilarité (exprimée par son écart-type normalisé) est plus atténuée que celle d'un groupe contenant 7 tokens issus de 7 formules.

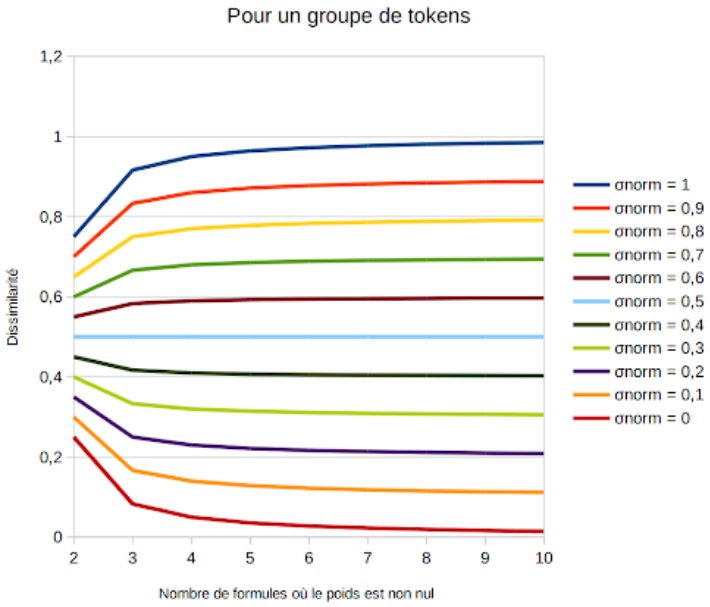

FIGURE 7 – hyperbole d'ajustement pour le score de dissimilarité.

5.5 Annexe 5 : capture d'écran de GeoAlign

Cette annexe présente une capture d'écran de la fenêtre principale de GeoAlign. On y voit notamment le formulaire pour construire la formule personnalisée de similarité (en haut à gauche), les correspondances découvertes entre deux entités sur la carte et l'estimation de la qualité de ces correspondances au moyen d'un graphique (en haut à droite).

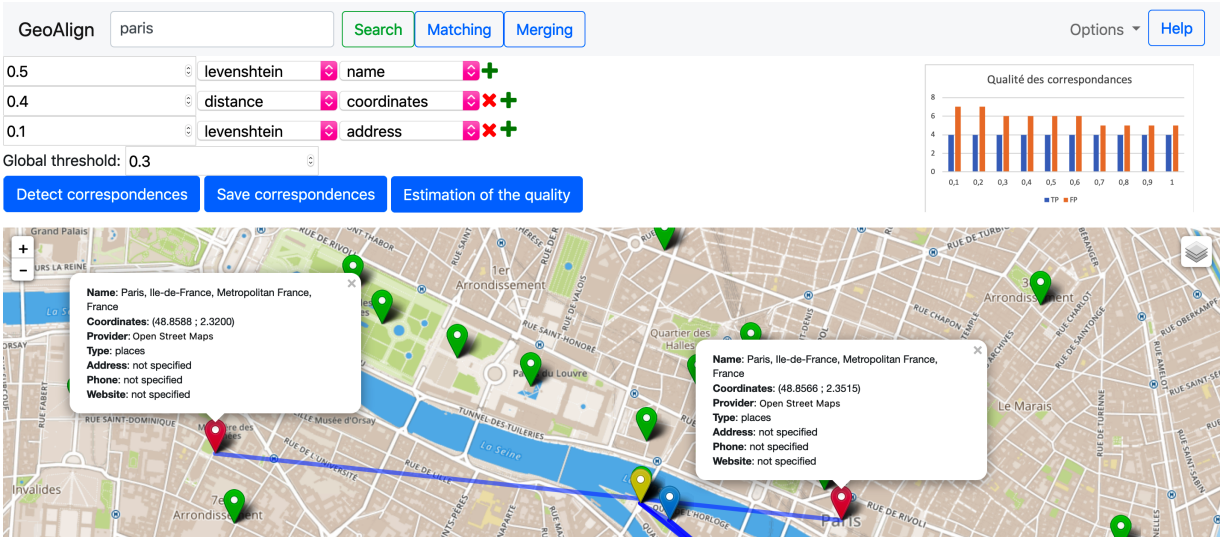

FIGURE 8 – capture d'écran du prototype GeoAlign.