Résolution d'entités nommées dans des corpus de documents historiques de grande taille et partiellement redondants

Le cas des annuaires commerciaux de Paris du XIX^{ème} siècle.

Mots-clés : résolution d'entités nommées, graphe géohistorique, données massives et passage à l'échelle, approche générique et reproductible.

A. Contexte : le projet de recherche SoDUCo pour explorer les transformations de Paris avec des données géohistoriques fines, nombreuses et ouvertes

Le projet SODUCO (2019-2023), financé par l'agence nationale de la recherche, construit des bases de données géographiques et des outils pour étudier les interactions entre la morphogénèse urbaine et les dynamiques sociales de Paris de la Révolution jusqu'au XX^{e} siècle.

Ces deux dimensions, spatiales et sociales, sont approchées à l'aide de deux corpus de sources historiques :

- les plans et cadastres, qui représentent les structures urbaines de la ville et de ses environs : rues, ilôts, bâtiments, etc. ;
- les annuaires commerciaux, sorte de "pages jaunes" avant l'heure, qui contiennent les noms, statuts sociaux, activités professionelles et adresses des commerçants et marchands parisiens.

Le corpus des annuaires est composé de 213 volumes numérisés issues de 20 collections différentes, Les efforts sont concentrés sur 3 collections principales : l'*Almanach du Commerce* (1797-1856), l'*Annuaire Général du Commerce* (1838-1857) et l'*Annuaire-Almanach du Commerce Bottin-Didot* (1857+). Malgré des différences de mise en page, tous les annuaires contiennent des listes d'individus avec des informations similaires (voir figure 1).

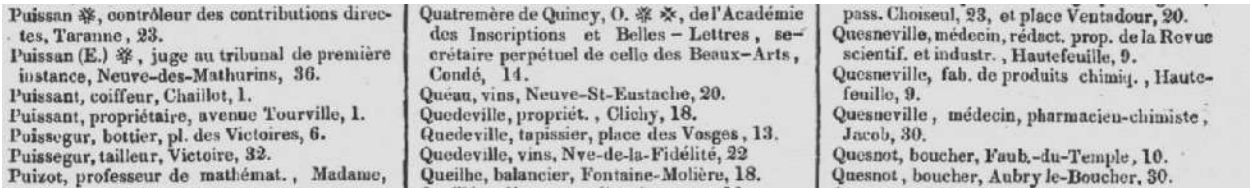
Puissan ✻, contrôleur des contributions directes, Taranne, 23.
Puissan (E.) ✻, juge au tribunal de première instance, Neuve-des-Mathurins, 36.
Puissant, coiffeur, Chaillot, 1.
Puissant, propriétaire, avenue Tourville, 1.
Puissegur, bottier, pl. des Victoires, 6.
Puissegur, tailleur, Victoire, 32.
Puizot, professeur de mathémat., Madame,
Quatremère de Quincy, O. ✻ ✻, de l'Académie des Inscriptions et Belles-Lettres, secrétaire perpétuel de celle des Beaux-Arts, Condé, 14.
Queau, vins, Neuve-St-Eustache, 20.
Quedeville, propriét., Clichy, 18.
Quedeville, tapissier, place des Vosges, 13.
Quedeville, vins, Nve-de-la-Fidélité, 22
Queilhe, balancier, Fontaine-Molière, 18.
pass. Choiseul, 23, et place Ventadour, 20.
Quesneville, médecin, rédact. prop. de la Revue scientif. et industr., Hautefeuille, 9.
Quesneville, fab. de produits chimiq., Hautefeuille, 9.
Quesneville, médecin, pharmacien-chimiste, Jacob, 30.
Quesnot, boucher, Faub.-du-Temple, 10.
Quesnot, boucher, Aubry le-Boucher, 30.

Fig 1 : Entrées de la liste alphabétique de l'*Annuaire Général du Commerce* pour l'année 1850, p.350.

Une chaîne d'extraction du contenu de ces annuaires numérisés a été mise en place [1], en vue de constituer une base de données spatio-temporelle permettant de suivre, au niveau le plus fin, les transformations de l'espace socio-professionnel de Paris tel qu'il nous est donné à voir par les annuaires.

Cette chaîne contient actuellement quatre maillons :

1. **[Layout analysis]** : les pages sont segmentées automatiquement pour en extraire la structure de chacune et isoler chaque entrée de l'annuaire;
2. **[OCR]** : le texte des entrées est ensuite reconnu automatiquement et extrait;
3. **[NER]** : les entités nommées constituant chaque entrée (nom, titres, activité, adresse, etc.) sont identifiées grâce à un modèle de NER (*named entity recognition*) spécialement entraîné pour ce type de corpus;
4. **[Geocoding]** : chaque entrée est géocodée en utilisant les points adresses extraits des plans et cartes historiques de Paris traités dans le projet.

En résulte une base de données de grande taille (~10 000 000 d'entrées) contenant les informations extraites de chaque annuaire. La figure 2 donne un exemple d'entrée extraite.

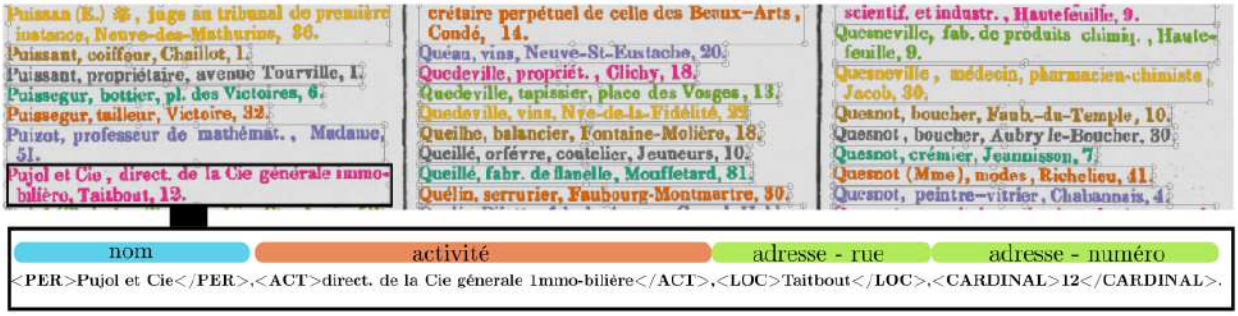

Fig 2 : en haut, les entrées segmentées avec leur texte extrait par OCR; en bas les entités nommées détectées dans l'entrée "Pujol et Cie, [...]".

Un premier stage [2] a permis de proposer une approche d'appariement de données pour identifier les entrées représentant un même commerce d'une année à l'autre. Appliquée avec succès sur un corpus restreint (~34 000 entrées relatives à la photographie), elle ne passe cependant pas à l'échelle et n'est donc pas généralisable au corpus entier.

Par ailleurs, cette approche ne traite que l'appariement des entrées d'annuaires représentant un même commerce. Or, le corpus permet également d'envisager l'appariement des adresses reconnues dans les entrées pour produire un référentiel spatio-temporel des adresses de Paris depuis la fin du XVIII^{e} siècle.

Références:

[1] N. Abadie, E. Carlinet, J. Chazalon, B. Duménieu. A Benchmark of Named Entity Recognition Approaches in Historical Documents: Application to 19th Century French Directories. *Document Analysis Systems. DAS 2022.* Mai 2022, La Rochelle, France.

[2] S. Tual. Construction de données spatio-temporelles à partir de sources historiques sérielles: Représenter les transformations du tissu professionnel parisien à l'échelle individuelle à partir d'annuaires du commerce du XIXe siècle. 26 septembre 2022. *Rapport de stage de Master 2 IGAST (Université Gustave Eiffel, ENSG).*

B. Objectifs du stage

Ce sujet de stage comporte un objectif méthodologique principal et trois objectifs applicatifs qui pourront être plus ou moins approfondis selon le profil du candidat ou de la candidate.

Objectif méthodologique : Proposer et mettre en œuvre une approche reproductible et qui **passe à l'échelle** pour **identifier des relations d'équivalence entre les principales entités nommées** extraites dans les annuaires du commerce parisien du XIX^{e} siècle et qui représentent un même élément du monde réel.

Objectifs applicatifs:

1) Produire une **graphe spatio-temporel** permettant le **suivi des commerces** parisiens sur la période étudiée. Il s'agira d'expliciter les apparitions, les déménagements, les transmissions et les disparitions de commerces tels qu'ils nous sont donnés à voir par les annuaires.
2) Produire une **graphe spatio-temporel** permettant le **suivi des adresses** à Paris sur la période étudiée. Il s'agira d'expliciter les adresses successives d'un même lieu (renommage de rues, changement de système d'adressage, etc.) telles qu'elles nous sont données à voir par les annuaires.
3) Produire un **graphe spatio-temporel** permettant le **suivi des descriptions d'activités** et de commerces fournies dans les annuaires sur la période étudiée. Il s'agira d'expliciter les différentes descriptions associées à un type de commerce ou d'activité au cours du temps, telles qu'elles nous sont données à voir par les annuaires. Celles-ci présentent en effet l'intérêt de permettre d'identifier les activités similaires mais décrites dans des termes différents selon les entrées. En outre, elles peuvent refléter l'évolution des activités au cours du temps, ou s'apparenter à de la publicité, etc.

C. Verrous scientifiques

La réalisation de ces objectifs suppose de proposer des solutions pour :

- ➔ Caractériser l'identité d'un commerce, d'une adresse, et d'un type d'activité pour identifier les entités nommées équivalentes,

- ➔ Proposer une approche d'appariement multisource générique, reproductible, et qui passe à l'échelle pour traiter les 10 millions d'entrées extraites,

- ➔ Exploiter la redondance des informations d'un annuaire à l'autre ou d'un index à l'autre pour compenser les erreurs et les lacunes liées au processus d'extraction d'informations,

- ➔ Évaluer la qualité des liens d'équivalence créés dans un contexte où une vérification manuelle n'est pas possible.

D. Compétences et formation requises

Formation : Master 2 ou troisième année d'école d'ingénieur en informatique, ou en géomatique.

Compétences et connaissances :

- Données géographiques structurées, données spatio-temporelles,
- Résolution d'entités nommées, liage, appariement,
- Graphes de connaissances géohistoriques,
- Développement Python,
- Un intérêt pour l'histoire sociale est un plus.

E. Informations pratiques

Modalités de candidature : envoyer CV et lettre de motivation adaptée au sujet **par email au format PDF et en un seul fichier** aux encadrants listés ci-dessous.

Encadrement & contacts : Le stage se déroulera dans l'équipe STRUDEL du laboratoire LASTIG de l'IGN, menant des recherches en géomatique sur les structures spatio-temporelles pour l'analyse des territoires.

Vous serez encadré.e par trois chercheuses et chercheurs participant au projet SoDUCo :

- Nathalie Abadie [STRUDEL/IGN] : nathalie-f.abadie[at]ign.fr
- Bertrand Duménieu [CRH/EHESS] : bertrand.dumenieu[at]ehess.fr
- Joseph Chazalon [LRDE/EPITA] : joseph.chazalon[at]lrde.epita.fr

Durée et période de stage : 5 mois, printemps-été 2023.

Gratification de stage : selon la législation française (environ 550€ net / mois).

Localisation : Institut National de l'Information Géographique et Forestière (IGN), Saint-Mandé (métro 1, station Saint Mandé).