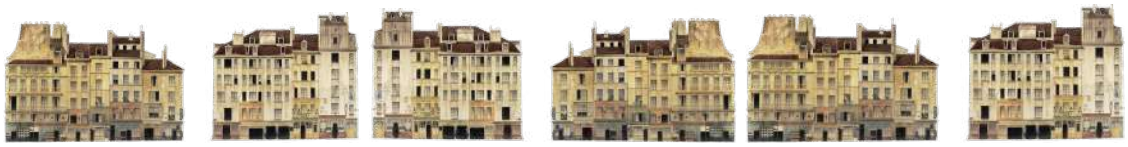
Structuration et analyse spatio-temporelles des entrées des annuaires du commerce parisien du 19ème siècle :

applications aux photographes et aux marchands d'art.



Mots-clés : résolution d'entités nommées, graphe géohistorique, analyse spatio-temporelle multi-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 début du XXe 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).

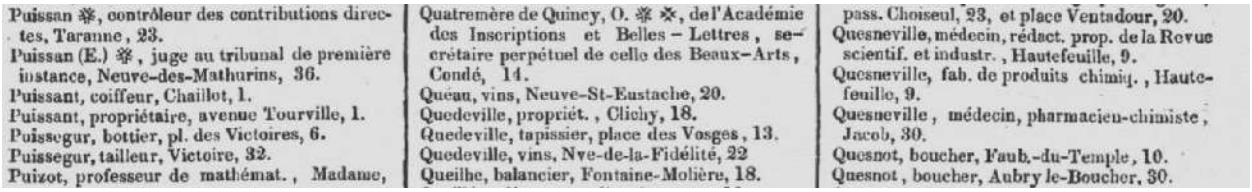

Puissan ✻, contrôleur des contributions directes, Taranne, 23.
Puissan (E.) ✻, juge au tribunal de première instance, Neuve-des-Mathurins, 36.
Puissant, coiffeur, Chaillot, 1.
Puissant, propriétaire, avenue Tourville, 1.
Puissegur, bottier, pl. des Victoires, 6.
Puissegur, tailleur, Victoire, 32.
Puizot, 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 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.

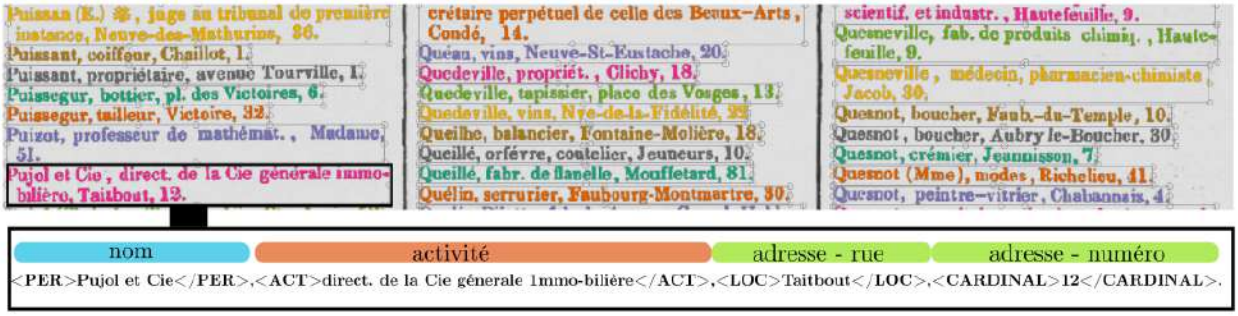


Fig 2 : en haut, les entrées segmentées avec leur texte extrait par OCR; en bas les entités nommées détectées de l'entrée "Pujol et Cie, [...]".

Un premier stage [2] a permis d'affiner l'approche d'extraction d'entités nommées dans les entrées d'annuaires et d'en proposer une autre pour l'appariement des entrées représentant un même commerce d'une année à l'autre. Celle-ci a été appliquée aux métiers de la photographie. Une interface de visualisation a été développée pour explorer ces données spatio-temporelles géocodées et comprendre les dynamiques d'évolution des commerces liées à la photographie (mobilité des commerces, transmissions/reprises de commerces, évolution des techniques photographiques, etc.).

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 deux objectifs qui pourront être plus ou moins approfondis selon le profil du candidat ou de la candidate.

1) Assurer la reproductibilité de l'approche d'appariement des entrées proposée lors du stage précédent et **proposer une approche pour passer d'un graphe spatial et temporel à un véritable graphe géohistorique** (c.-à-d. passer d'un modèle de *snapshots* temporels appariés à un modèle spatio-temporel). Lorsque c'est possible, lier les commerces avec des informations et ressources externes pour enrichir les connaissances disponibles pour l'analyse spatio-temporelle.

2) Proposer **une approche pour l'analyse spatio-temporelle de l'évolution des commerces étudiés.** Celle-ci devra permettre de **dégager d'éventuelles logiques individuelles ou collectives** de localisation (proximité avec la clientèle, avec des commerces du même type, des fournisseurs, des détaillants, etc.), de déménagement, de transmission, de reprise, de publicité, de changement/modernisation des produits, etc., tout en considérant les forts changements de Paris durant la période (croissance de la population, densification urbaine, etc.).

C. Verrous scientifiques

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

- ➔ Fusionner automatiquement les états successifs d'un même commerce ne présentant pas de changement de propriétés, caractéristiques d'une évolution à traiter comme un événement, et assigner à la ressource résultante les valeurs de propriétés, les attestations et les temps valides appropriés ;

- ➔ Caractériser, détecter automatiquement et représenter les événements qui surviennent au cours de l'existence des commerces au cours du temps.

Ceci nécessitera, entre autres, de proposer une ontologie adaptée pour représenter les commerces et leurs évolutions.

D. Compétences et formation requises

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

Compétences et connaissances :

- Données géographiques structurées, données spatio-temporelles,
- Analyse spatio-temporelle,
- Graphes de connaissances, modèles pour le Web sémantique (OWL, RDF, etc.)
- Développement Python ou R,
- 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 participant au projet SoDUCo :

- Nathalie Abadie [STRUDEL/IGN] : nathalie-f.abadie[at]ign.fr
- Solenn Tual [STRUDEL/IGN] : solenn.tual[at]ign.fr
- Julie Gravier [CRH-CAMS/EHESS] : juliecatherine.gravier[at]gmail.com

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).