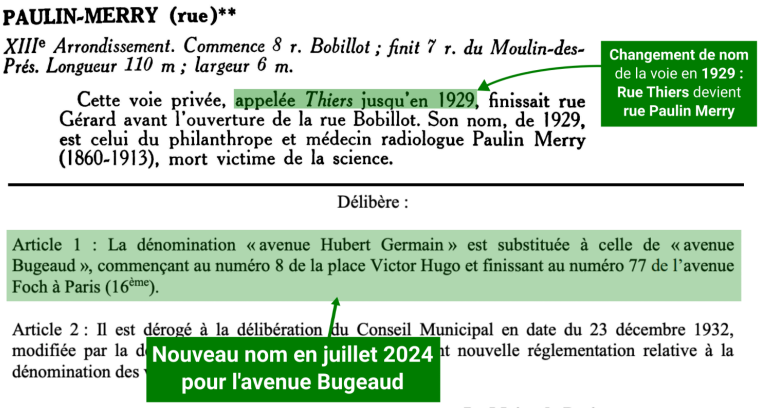
Extraction et structuration des informations d'évolution des rues dans des corpus textuels à l'aide de grands modèles de langue

Contexte et objectifs

Ce stage s'inscrit dans un ensemble de travaux visant à proposer une méthodologie générique et reproductible pour la construction d'un graphe de connaissances géohistorique des voies et des adresses à partir des documents historiques et de données publiées sur le Web. Or de nombreuses informations sur les rues anciennes se présentent plutôt sous la forme de textes. De nombreux corpus textuels, qui décrivent l'évolution des rues et des adresses (création, disparition, renommage, extension d'une rue...), sont aujourd'hui accessibles en ligne. Les informations qu'ils renferment renseignent à la fois sur l'état des rues ou des adresses à une période donnée et sur les événements qui conduisent à leur évolution. Mais à ce stade, ces informations ne sont pas structurées et sont donc difficilement exploitables pour retrouver l'état d'une rue ou d'une adresse à une date donnée ou bien encore reconstituer leur généalogie de façon automatique.

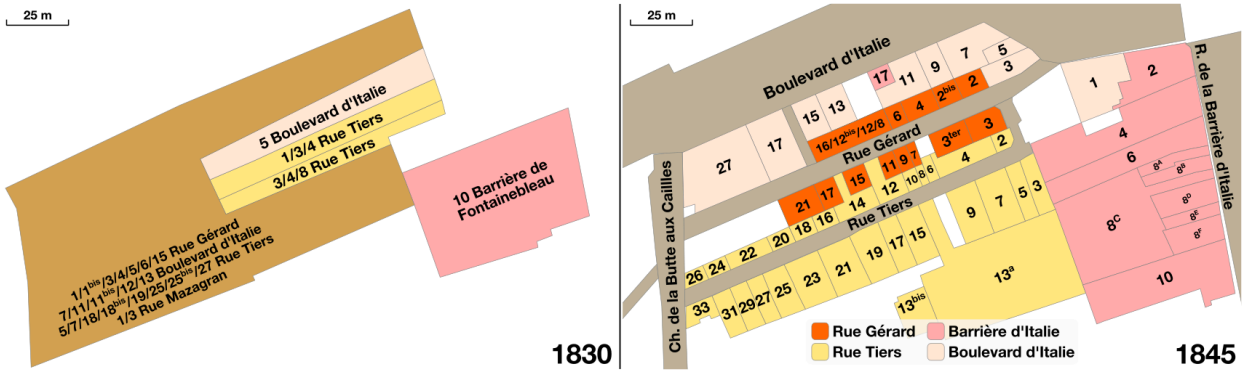
Depuis quelques années, les grands modèles de langue (Large Language Models) sont de plus en plus utilisés pour reconnaître et structurer des données à partir de textes et alimenter des graphes de connaissances. Ce stage vise à explorer et adapter ces approches fondées sur des LLM pour produire des données structurées sur l'évolution des rues de Paris depuis le 17^{e} siècle à partir de corpus textuels décrivant la ville et ses voies (dictionnaires des rues de Paris, bulletins officiels, pages Wikipedia...).



Extraits d'un dictionnaire et d'un bulletin officiel de la Ville de Paris décrivant des évolutions de voies parisiennes

Verrous scientifiques et productions attendues

De nombreuses approches ont été proposées au cours des dernières années, pour peupler des ontologies à partir de textes. En particulier, les grands modèles de langues ont permis des avancées notables, avec des performances de reconnaissance d'entités nommées très élevées. Une première étape consistera donc à dresser un état de l'art des approches existantes pour identifier les plus pertinentes pour le cas d'application du stage ; on s'intéressera donc plus particulièrement à la reconnaissance et l'extraction d'événements d'évolution et des entités concernées.



Snapshots décrivant le quartier de la Butte aux Cailles en 1830 et en 1845, obtenus après population de l'ontologie

Une deuxième étape visera à rassembler un corpus de textes pertinents et à préparer un jeu de données de référence, en annotant un échantillon de ce corpus (sur un quartier de Paris par exemple), et en produisant les triplets correspondants. Celui-ci sera mis à profit pour entraîner et évaluer les approches identifiées lors de l'étape précédente.

Enfin, il s'agira de tester et évaluer les approches identifiées sur le corpus constitué pour peupler l'ontologie PeGazUs[1].

Renseignements pratiques

Profil recherché : Master 2 ou diplôme d'ingénieur en informatique, sciences de l'information géographique ou humanités numériques.

Compétences et connaissances:

- Données géographiques structurées.
- Web de données, graphes de connaissances.
- Traitement automatique du langage naturel, recherche d'informations dans des textes, LLM.
- Programmation Python.
- Un goût pour l'histoire urbaine est un plus.

Durée et période de stage : 5 mois, au cours du printemps et de l'été 2025.

Lieu du stage : Équipe LaSTIG/Strudel, École Nationale des Sciences Géographiques, 6-8 avenue Blaise Pascal, 77420 Champs-sur-Marne (RER A, station Noisy-Champs).

Indemnités de stage : Stage gratifié selon la législation française.

Modalités de candidature : Envoyer un CV, une lettre de motivation adaptée au stage et les relevés de notes des 2 dernières années d'études par email, au format PDF et en un seul fichier aux encadrants.

Encadrement du stage :

- Charly Bernard (LaSTIG - Strudel, univ. Eiffel, IGN-ENSG, charly.bernard@ign.fr)
- Nathalie Abadie (LaSTIG - Strudel, univ. Eiffel, IGN-ENSG, nathalie-f.abadie@ign.fr)

[1] Charly Bernard, Solenn Tual, Nathalie Abadie, Bertrand Duménieu, Julien Perret, Joseph Chazalon. PeGazUs: A knowledge graph based approach to build urban perpetual gazetteers. International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), Nov 2024, Amsterdam, Netherlands. Pp.364-381. (https://hal.science/hal-04721538/)