
Reconnaissance et transcription des informations textuelles figurant sur des cartes anciennes

Mots-clés : Reconnaissance de texte dans des documents, reconnaissance optique de caractères, reconnaissances de texte manuscrit, traitement d'images, scans de cartes anciennes, apprentissage profond, programmation Python.

A. L'extraction de texte dans les cartes anciennes

Il existe des solutions de plus en plus performantes pour détecter et transcrire du texte dans des scènes ou dans des documents. Peu à peu, celles-ci sont également appliquées à des cartes scannées pour en extraire et transcrire les écritures. En effet, extraire les noms de lieux contenus dans les cartes anciennes permettrait d'indexer spatialement les grands corpus de cartes numérisés par différentes institutions, mais pas nécessairement géoréférencées. Par ailleurs, c'est un moyen de constituer ou d'enrichir automatiquement des gazetiers utiles à la géolocalisation d'autres ressources (cartographiques ou textuelles).

Nous souhaitons appliquer des méthodes d'extraction et de reconnaissance des toponymes à différents types de cartes anciennes :

- des cartes topographiques du territoire (carte de Cassini ou carte d'état major),
- des plans à grande échelle, comme ceux du cadastre napoléonien, qui contiennent des toponymes et éventuellement des numéros de parcelles. Ces derniers permettent de localiser les entités géographiques mentionnées dans les matrices cadastrales.

Les principales difficultés d'extraction sont liées à la grande variabilité des styles de cartes, aux multiples polices ou graphies d'écriture, au placement et à l'orientation des écritures dans les cartes, ainsi qu'aux recouvrements entre le texte et les objets géographiques représentés. Par ailleurs, les modèles à base de réseaux de neurones profonds utilisés pour détecter du texte dans les images nécessitent de disposer de grands corpus annotés pour être entraînés. Or, il n'existe pas de corpus de ce type, utilisable pour entraîner des modèles de reconnaissance du texte dans des cartes et sa production semble difficilement réalisable car chaque type de carte est différent, les styles des fonds et des textes changent.



Extrait de la feuille N°1 de la Carte de Cassini (1:86400),
1756. [Assemblage IGN](#). Source : BNF.



Extrait de la minute de la carte d'état-major (1:40 000),
feuille de Melun. 1832. Source : IGN.

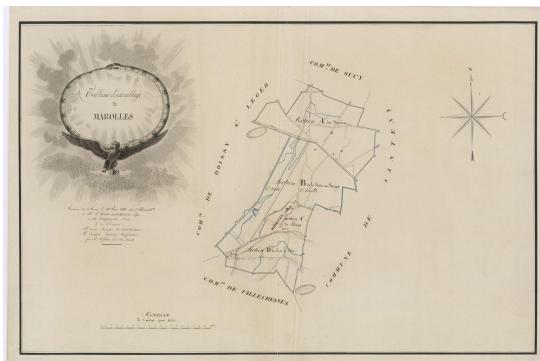
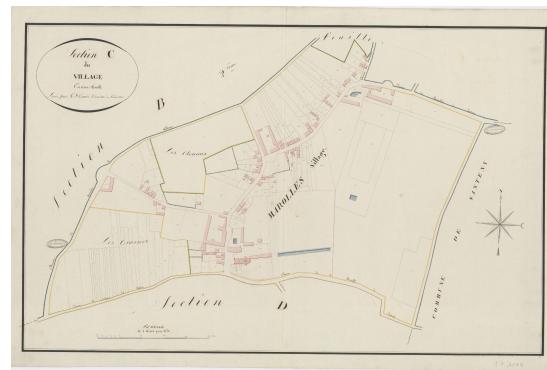


Tableau d'assemblage (1:10 000) des plans cadastraux de la commune de Marolles-en-Brie. 1810.
Source : Archives Départementales du Val de Marne.

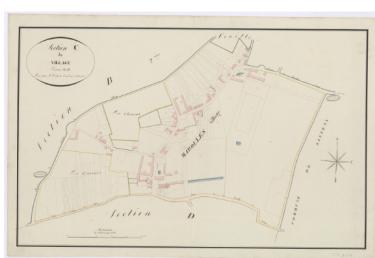


Plan cadastral (extrait d'atlas, échelle 1:2500) de la section de C de Marolles-en-Brie. 1810.
Source : Archives Départementales du Val de Marne.

B. Objectifs du stage

L'objectif de ce stage est de produire des modèles entraînés avec peu d'exemples et capables de s'adapter à des types de cartes totalement inédits. Pour cela, les pistes suivantes devront être explorées :

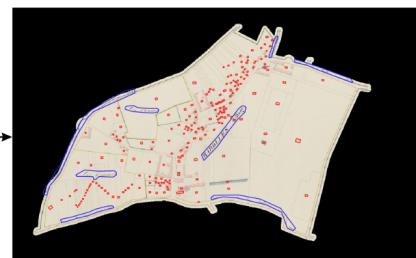
- Utilisation de modèles de détection de texte dans des scènes naturelles pour détecter les écritures dans des cartes ;
- Génération d'un jeu de données synthétique à partir du cadastre moderne (transfert de style des cartes du cadastre ancien) ;
- Proposition d'une approche de découpage de la carte selon les limites de la zone cartographiée.



Plan numérisé



Masque de la zone cartographiée



Détection et classification des numéros et des toponymes.

Illustration des étapes possibles de la chaîne de traitement.

C. Productions attendues

- Corpus d'entraînement et de test pour la détection, la classification et la reconnaissance des écritures dans les cartes anciennes;
- Modèles entraînés + le code déposé sur un Github;
- Mémoire de stage.

D. Compétences et formation requises

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

Compétences et connaissances :

- Extraction d'informations dans des images (scans de cartes, images aériennes, images au sol type StreetView, etc.) : *segmentation sémantique d'images, OCR, HTR, etc.*
- Données géographiques structurées,
- Développement Python,
- Un intérêt pour la cartographie historique 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 quatre chercheuses et chercheurs :

- Nathalie Abadie [STRUDEL/IGN] : nathalie-f.abadie@ign.fr
- Solenn Tual [STRUDEL/IGN] : solenn.tual@ign.fr
- Joseph Chazalon [LRDE/EPITA] : joseph.chazalon@epita.fr

Durée et période de stage : 5 mois, printemps-été 2024.

Rémunération : stage gratifié 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) ou Ecole Nationale des Sciences Géographiques (ENSG), Champs-sur-Marne (RER A, station Noisy-Champs).