Cartographier les odonymes de Paris cités dans les romans du XIXème siècle

Ludovic Moncla[1], **Mauro Gaio**[2], **Thierry Joliveau**[3]

1. *INSA Lyon, CNRS, LIRIS UMR 5205, France*
   ludovic.moncla@liris.cnrs.fr
2. *Laboratoire LIUPPA, Université de Pau et des Pays de l'Adour, France*
   mauro.gaio@univ-pau.fr
3. *Université de Saint-Etienne, UMR EVS, France*
   thierry.joliveau@univ-st-etienne.fr

RÉSUMÉ. *Cet article propose une méthodologie pour cartographier les empreintes spatiales des romans et des auteurs sur la base de tous les odonymes extraits des romans. Nous présentons une manière originale d'explorer l'espace parisien et les paysages fictifs en parcourant de manière interactive et simultanée l'espace géographique et le texte littéraire. Notre projet consiste à construire une plate-forme capable d'extraire, cartographier et analyser les occurrences des odonymes dans des romans dans lesquels l'action se déroule en totalité ou en partie à Paris. Cette plate-forme sera utilisée dans plusieurs domaines, tels que le tourisme culturel, la recherche urbaine et l'analyse littéraire.*

ABSTRACT. *This paper propose a methodology to map the spatial fingerprints of novels and authors based on all the odonyms extracted of the novels. We present an original way to explore Parisian space and fictional landscapes by interactively and simultaneously browsing geographical space and literary text. Our project involves building a platform capable of retrieving, mapping and analyzing the occurrences of odonyms in novels in which the action occurs wholly or partly in Paris. This platform will be used in several areas, such as cultural tourism, urban research, and literary analysis.*

MOTS-CLÉS : *recherche d'information géographique; humanités numériques; cartographie; reconnaissance d'entités nommées*

KEYWORDS: *geographical information retrieval; digital humanities; mapping; named entity recognition*

1. Introduction

Grâce au travail pionnier de Moretti (Moretti, 1999) la cartographie est désormais utilisée pour donner une représentation romanesque du lieu dans le but de permettre de nouvelles interprétations des romans. De nouvelles questions ont émergé sur la relation entre littérature et cartographie (Engberg-Pedersen, 2017). Repérer manuellement les endroits mentionnés dans un ouvrage est une tâche fastidieuse et longue et les technologies numériques permettent de simplifier considérablement la manière d'extraire l'information spatiale des textes littéraires et de visualiser les lieux et l'espace dans les récits (Gregory *et al.*, 2015 ; Cooper *et al.*, 2016). Nous présentons dans cet article nos premières expérimentations dans le cadre du développement d'une plate-forme capable d'extraire, localiser, cartographier et analyser les lieux de Paris mentionnés dans les romans. Cette plate-forme a pour vocation d'intéresser un large public: urbanistes, historiens, experts littéraires, touristes culturels ou habitants curieux des lieux perdus et existants décrits dans les romans.

2. Combiner deux approches pour l'annotation des odonymes

Les méthodes automatiques de reconnaissance d'entités nommées (en particulier pour l'extraction des noms de lieux) dans les documents textuels ont été abordées dans de nombreux travaux de recherche. (Melo, Martins, 2017) propose un inventaire de méthodes et de systèmes existants dans ce domaine. Par ailleurs, comme indiqué par (Gritta *et al.*, 2017) la nouvelle génération de géoparseurs doit utiliser davantage d'informations pour comprendre la signification du contexte.

Notre proposition met en oeuvre l'annotation automatique des noms de lieux [1]. Elle enrichit l'outil de reconnaissance des entités nommées de la plateforme Perdido (Moncla *et al.*, 2014) implémentée par une cascade de transducteurs selon les principes des grammaires de construction (Yannick-Mathieu, 2003). Notre solution annote sémantiquement les entités nommées étendues (ENE) et les entités spatiales nommées étendues (ESNE), telles que définies par (Gaio, Moncla, 2017) ainsi que leurs relations spatiales associées, couvrant ainsi la plupart des formes utilisées pour exprimer les odonymes. Notre objectif n'est pas de construire un processus entièrement automatique (du texte à la carte), mais de proposer de nouveaux outils aux experts (géographes, historiens, etc.) pour explorer un corpus de romans. Dans ce contexte, nous avons proposé la combinaison de l'approche d'annotation automatique implémentée au sein de la plateforme Perdido (Moncla *et al.*, 2014) avec une approche textométrique (requêtes CQL au sein de la plate-forme TXM (Heiden, 2010)) permettant l'in-

1. Nous nous concentrons en particulier sur 14 catégories d'odonymes parmi les plus cités dans les romans : allée, avenue, boulevard, cour, galerie, impasse, parvis, passage, place, pont, port, quai, rue, square.

teraction avec des utilisateurs pour l'évaluation quantitative, la correction et l'amélioration de l'annotation automatique réalisée (Moncla *et al.*, 2017).

Notre corpus expérimental comprend 31 romans français centrés sur Paris et couvrant différentes périodes entre 1830 et 1913. Les résultats (tab. 1) montrent que certaines étapes du processus (telle que la reconnaissance des odonymes) peuvent être semi-automatiques en utilisant les méthodes TAL implémentées dans Perdido et complétées par des interactions humaines, permises par TXM.

TABLE *1. Evaluation de la reconnaissance automatique des odonymes*

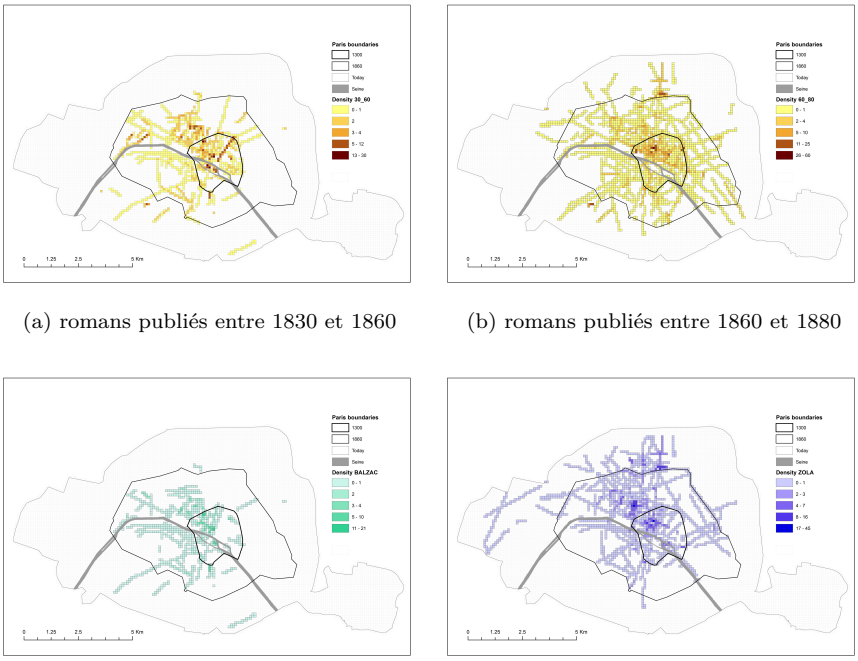
	CQL-TXM	Perdido
occurrences des odonymes trouvées à tort (faux positif)	286	88
occurrences des odonymes non trouvées (faux négatif)	11	117
occurrences des odonymes trouvées (vrai positif)	3573	3467
Total d'occurrences trouvées	3859	3555
Précision	0.926	0.975
Rappel	0.997	0.967
F-score	0.960	0.971

3. Proposer un rendu géographique adapté

Tous les odonymes valides ont été localisés en consultant des ressources géo-historiques (atlas de rues et ressources Web [2]). Nous avons pu localiser 3433 références d'odonymes dans les 31 romans, associées à 712 routes (existantes (634) ou disparues (78)). Pour la création des cartes nous avons utilisé un SIG construit à partir du Plan Vasserot (1810-1836) [3] pour les rues datant d'avant 1850 et du réseau de rues existants sur le site ParisOpendata [4] pour les rues d'après 1850. Cela nous a permis de construire une première représentation de l'empreinte spatiale des romans et des auteurs en fonction de la distribution des odonymes extraits du texte.

Après différents tests de représentations en implantation ponctuelle et linéaire, une cartographie de la densité des occurrences sur une grille régulière apparaît comme un bon compromis entre le respect de la nature linéaire de l'information et l'objectif de visualiser la structure spatiale du phénomène. Ici par exemple, l'indice de densité de la route est calculé pour chaque cellule par 1 ha carré au prorata de la partie de la longueur de chaque route dans la cellule. Ensuite, les index pour chaque route sont additionnés dans la cellule (fig. 1). De cette manière les valeurs quantitatives absolues sont perdues, mais nous éliminons le biais lié à la longueur de la route et à l'accumulation de symboles ponctuels dans les zones denses. Nous avons ajouté sur les cartes trois limites de

2. http://geohistoricaldata.org/
3. Digitalisé dans le cadre du projet Alpage : http://alpage.huma-num.fr/
4. https://opendata.paris.fr/



(a) romans publiés entre 1830 et 1860

(b) romans publiés entre 1860 et 1880

(c) dans les romans de Balzac du corpus

(d) dans les romans de Zola du corpus

Figure 1. Densité des citations de noms de rues

Paris : le dernier mur médiéval environnant, les douze arrondissements de Paris avant l'annexion de localités périphériques en 1859, et les vingt arrondissements encore en place aujourd'hui.

Le nombre élevé de références dans le vieux Paris est bien sûr dû à sa permanence tout au long de la période. Si le vieux centre a été radicalement transformé après 1852 par Haussmann, il reste un lieu où les écrivains situent leurs histoires, même si l'Île de la Cité disparaît des romans. En conséquence, une densité plus faible d'occurrences dans les zones périphériques n'est pas inattendue. Certaines routes mentionnées par Zola n'existaient pas lorsque Balzac était en vie. La comparaison de deux cartes basées sur l'année de publication des romans met ainsi en évidence la dynamique temporelle des espaces nommés dans les romans (fig. 1a et 1b). Les romans publiés avant 1860 suivent l'extension de la ville en dehors de la cité médiévale et restent principalement dans les limites de 1860. Entre 1860 et 1880, les romans de notre échantillon se sont répandus dans les nouveaux domaines de l'urbanisation. Comme le montre les figures (1c) et (1d), les cartes peuvent également aider à comparer les étendues spatiales de différents romans ou auteurs. Nous avons proposé aussi l'élaboration d'une sorte de signature spatiale d'un roman qui combine diffé-

rentes mesures de la répartition géographique des odonymes cités (enveloppe convexe, ellipse de déviation standard, barycentre) (fig. 2).

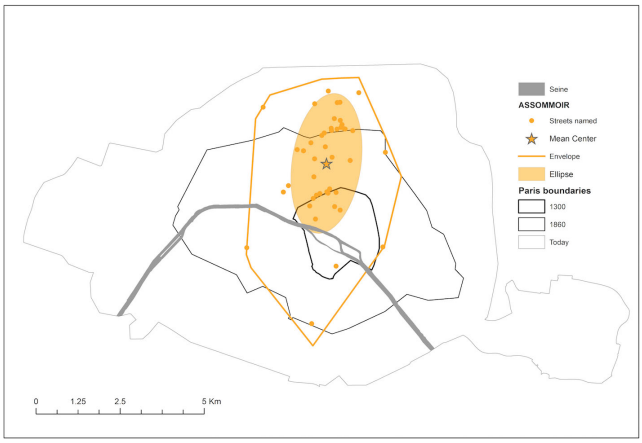


Figure 2. Empreinte spatiale du roman l'Assommoir (Zola)

4. Conclusion

Dans cet article, nous avons proposé le développement d'une plate-forme pour récupérer et afficher géographiquement les odonymes des romans. Ces odonymes ont été modélisés sous la forme de motifs lexico-syntaxiques. Les motifs sont basés sur des mots simples, sur une combinaison de mots ou sur un groupe de mots avec des propriétés structurées. Ces motifs sont annotés automatiquement pour être ensuite requêtés via un outil de textométrie. Nous avons également développé une approche cartographique originale pour visualiser et analyser les résultats.

Bibliographie

Cooper D., Donaldson C., Murrieta-Flores P. (2016). *Literary mapping in the digital age*. Routledge.

Engberg-Pedersen A. (2017). *Literature and cartography: Theories, histories, genres*. MIT Press.

Gaio M., Moncla L. (2017). Extended named entity recognition using finite-state transducers: An application to place names. In *9th international conference on advanced geographic information systems, applications, and services*. Nice, France.

Gregory I., Donaldson C., Murrieta-Flores P., Rayson P. (2015). Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, vol. 9, n° 1, p. 1–14.

Gritta M., Pilehvar M. T., Limsopatham N., Collier N. (2017). What's missing in geographical parsing? *Language Resources and Evaluation*.

Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In *24th Pacific Asia Conference on Language, Information and Computation*, p. 389-398.

Melo F., Martins B. (2017). Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, vol. 21, n° 1, p. 3–38.

Moncla L., Gaio M., Joliveau T., Lay Y.-F. L. (2017). Automated geoparsing of paris street names in 19th century novels. In *Proceedings of the 1st acm sigspatial workshop on geospatial humanities*. ACM.

Moncla L., Renteria-Agualimpia W., Nogueras-Iso J., Gaio M. (2014). Geocoding for texts with fine-grain toponyms. In *22nd ACM SIGSPATIAL international conference on advances in geographic information systems*, p. 183–192. Dallas,TX, USA, ACM.

Moretti F. (1999). *Atlas of the european novel, 1800-1900*. London, UK, Verso.

Yannick-Mathieu Y. (2003). La Grammaire de Construction. *Approches syntaxiques contemporaines*, n° 48, p. 43–56.