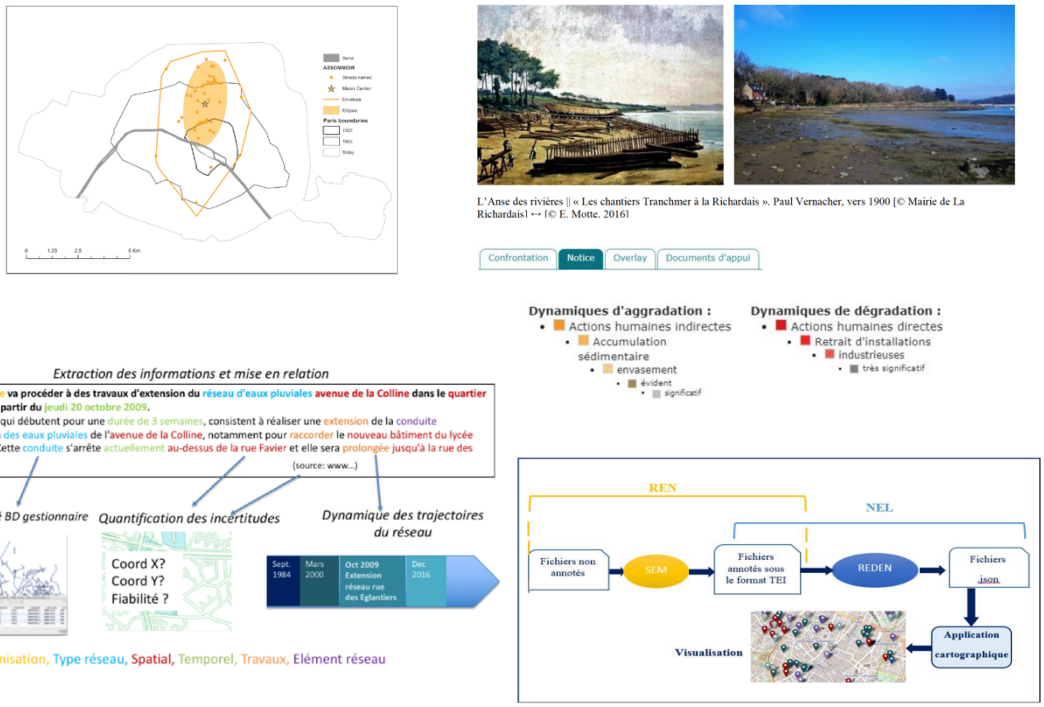
Atelier Humanités Numériques Spatialisées (HumaNS'2018)

Carmen Brando, Francesca Frontini, Mathieu Roche

Atelier *Humanités Numériques Spatialisées* (HumaNS'2018)

Conférence SAGEO'2018

6 novembre 2018, Montpellier



Carmen Brando, Francesca Frontini, Mathieu Roche (Eds.)

SAGEO 2018
Spatial Analysis and GEOmatics

SAGEO 2018
Spatial Analysis and GEOmatics

Spatial Analysis and GEOmatics (SAGEO'2018)

Atelier "Humanités Numériques Spatialisées" (HumaNS'2018)

6 novembre 2018, Montpellier, France

Éditorial

L'atelier **Humanités Numériques Spatialisées** (HumaNS'2018), organisé dans le cadre de la conférence SAGEO'2018, a pour objectif de rassembler les communautés francophones de géomaticiens et de spécialistes en sciences humaines et sociales (SHS) qui expoitent des *méthodes computationnelles* pour explorer une perspective spatiale et spatio-temporelle. Aujourd'hui, les méthodes et les outils de la géomatique occupent une place importante dans le paysage des *Humanités Numériques* (HN).

Ces dernières années, des groupes d'intérêt spécial (GIS) en *Geo-humanities* se sont constitués et deviennent très actifs au sein des infrastructures de recherche pour les SHS tels que DARIAH[1] et des associations internationales en humanités numériques comme *Alliance of Digital Humanities Organizations* (ADHO[2]). De même, les conférences en HN et géomatique organisent davantage d'ateliers en HN spatialisées comme celui que nous proposons. À titre d'exemple, nous pouvons citer l'atelier *APlace4Places* soutenu par le GIS Geo-Humanities d'ADHO en 2016, l'ACM SIGSPATIAL Workshop on *Geospatial Humanities* en 2017 et enfin, la conférence *Spatial Humanities* en 2018.

Les thèmes de l'atelier sont :

- la détection et la représentation de l'information spatiale dans les textes à partir de méthodes de traitement automatique des langues (TAL), etc. ;
- l'annotation de l'information spatiale à partir de données de SHS (enquêtes, récits, presse, littérature grise, etc.) ;
- la création de ressources géo-historiques dédiées et Web des Données ;

[1] https ://www.dariah.eu/
[2] https ://adho.org/

- la visualisation de l'information spatio-temporelle pour l'analyse textuelle en SHS ;
- l'analyse spatiale diachronique en Lettres et Histoire ;
- l'exploitation d'ontologies spatio-temporelles pour modéliser l'information historique.

Chaque article soumis sous forme de résumés étendus a été évalué par deux membres du Comité Scientifique de HumaNS'2018 que nous remercions. Au final, 10 communications ont été présentées à travers trois sessions :

- Observer le territoire et valoriser le patrimoine sur le Web ;
- Nommer, partager et cartographier les lieux anciens ;
- Extraire et intégrer d'informations géographiques.

Responsables du Comité Scientifique de HumaNS'2018

- *Carmen Brando, Ecole des hautes études en sciences sociales (EHESS Paris)*
- *Francesca Frontini, Praxiling UMR 5267 CNRS - Université Paul-Valéry Montpellier 3*
- *Mathieu Roche, Cirad, TETIS*

Comité Scientifique

- Nathalie Abadie (Lastig, IGN)
- Nicolas Béchet (IRISA)
- Francesco Beretta (UMR 5190 - LARHRA, ENS Lyon, CNRS)
- Delphine Bernhard (EA 1339, Linguistique, langues, parole, Univ. de Strasbourg)
- Sandra Bringay (LIRMM, Univ. Paul-Valéry Montpellier III)
- Sascha Diwersy (Praxiling, Univ. Paul-Valéry Montpellier III)
- Bertrand Dumenieu (CRH UMR 8558, EHESS)

- Cédric Lopez (EMVISTA)
- Frédérique Mélanie-Becquet (CNRS, Lattice CNRS-ENS)
- Eric Mermet (CNRS, CAMS-EHESS/ISC)
- Jérôme Pasquet (TETIS, Univ. Paul-Valéry Montpellier III)
- Maria Susana Seguin (Univ. Paul-Valéry Montpellier III - ENS de Lyon - IUF)
- Maguelonne Teisseire (TETIS, Irstea)

Organisateurs et sponsors de SAGEO'2018

Expérimentation de méthodes d'extraction d'informations géographiques pour les documents historiques

Katherine McDonough[1], **Ludovic Moncla**[2], **Matje van de Camp**[3]

1. *Department of History and Center for Interdisciplinary Digital Research, Stanford University, USA*
   kmcdono2@stanford.edu
2. *INSA Lyon, CNRS, LIRIS UMR 5205, France*
   ludovic.moncla@liris.cnrs.fr
3. *De Taalmonsters, Tilburg, Netherlands*
   matje@taalmonsters.nl

RÉSUMÉ. *Dans cet article, nous nous intéressons à deux aspects peu étudiés dans les travaux de recherche en TAL : traiter des documents historiques en français et traiter des structures textuelles complexes au-delà du texte courant ou des listes de noms de lieux. Notre méthodologie s'appuie sur l'évaluation des résultats de deux outils de reconnaissance d'entités nommées spatiales dans le cadre de l'analyse de documents du début de l'époque moderne et structurés à la manière de dictionnaires.*

ABSTRACT. *In this article, we address two gaps in NLP research: working with historical French and working with complex textual structures moving beyond running text or lists of place names. Our methodology is based on the evaluation of the results of two spatial named entity recognition tools in the context of early modern document analysis structured as dictionaries.*

MOTS-CLÉS : *recherche d'informations géographiques, traitement automatique des langues (TAL), reconnaissance d'entités nommées, humanités numériques*

KEYWORDS: *geographic information retrieval, natural language processing (NLP), named entity recognition, digital humanities*

HumaNS'2018 Atelier SAGEO - Montpellier, 6 novembre 2018

1. Introduction

Les recherches sur l'analyse de dictionnaires géographiques se sont jusqu'ici concentrées sur des projets utilisant des corpus en anglais moderne, publiés principalement à partir de la fin du XVIIIe siècle. Ces projets dépendent de lexiques et de ressources géographiques spécifiques à l'époque considérée qui permettent d'améliorer l'identification des noms de lieux et leur localisation. Contrairement aux érudits du monde classique et du monde moderne, les chercheurs contemporains travaillant sur le début de la période moderne (1400-1800) manquent de telles ressources. Ce projet est à l'interface entre les sciences humaines (histoire, géographie), le traitement automatique des langues (TAL) et les sciences de l'information géographique. Nous nous appuyons sur des travaux récents sur les espaces et les lieux du siècle des Lumières (Safier, 2014 ; Withers, Mayhew, 2011) ainsi que sur divers projets en humanités numériques portant sur la période des Lumières (Edelstein, 2016 ; Comsa *et al.*, 2016). Contrairement aux études sur cette période qui sont souvent axées sur l'analyse de petits groupes d'élites (dirigeants politiques, scientifiques, philosophes, voyageurs), cette recherche est une étape vers la compréhension de la mobilité de communautés locales et cosmopolites à travers l'information géographique.

Dans cet article, nous nous intéressons à deux aspects peu étudiés dans les travaux de recherche en TAL : 1) traiter des documents historiques en français et 2) traiter des structures textuelles complexes au-delà du texte courant ou des listes de noms de lieux. Le texte numérisé de l'Encyclopédie [1] de Diderot et d'Alembert (Morrissey, Roe, 2017) édité entre 1751 et 1772 est un exemple de corpus historique du genre dictionnaire que l'on se propose d'étudier.

2. Comparaison d'outils d'annotation pour l'adaptation à des corpus historiques français

L'adaptation d'un outil de reconnaissance d'entités nommées spatiales pour le français nous permettra d'explorer la structure et le contenu de l'information géographique au sein de documents du début de l'époque moderne de manière automatique. Cela nous permettra également de développer des techniques répondant aux challenges de la recherche d'informations spatiales communs à tous les états anciens des langues : identification des variantes de noms de lieux, association de variantes de noms à un même lieu et désambiguïsation de différents lieux, et détermination des types de relations entre ressources géographiques (officielles ou participatives) et descriptions textuelles de lieux historiques.

Notre méthodologie s'appuie sur l'évaluation des résultats de deux outils de reconnaissance d'entités nommées spatiales dans le cadre de l'analyse de

1. *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres. http://encyclopedie.uchicago.edu*

documents du début de l'époque moderne et structurés à la manière de dictionnaires. La reconnaissance d'entités nommées spatiales combine les tâches de reconnaissance d'entités nommées et d'association de ces entités à une localisation (coordonnées géographiques). Nous avons testé l'outil Edinburgh Geoparser (EG) (Alex *et al.*, 2015) et l'outil Perdido (Gaio, Moncla, 2017). Le premier a été développé pour analyser des documents en langue anglaise et le second pour des documents rédigés en langue française, tout les deux pour les langues modernes. Nous avons sélectionné EG car il s'agit d'un outil reconnu et très utilisé dans les projets d'annotation d'informations géographiques. Nous étudions ici ces deux outils en parallèle non pas pour critiquer leur performances, mais pour montrer aux spécialistes des sciences humaines 1) à quel point les outils d'annotation automatique peuvent être différents selon leur conception et 2) pourquoi adapter un tel outil au contexte issu du texte est une préoccupation méthodologique fondamentale pour l'analyse géographique de textes.

EG comme Perdido sont conçus comme une chaîne de traitement comprenant une étape de pré-traitement (tokenization, lemmatisation et analyse morpho-syntaxique) et un système de règles (patrons linguistiques, lexiques. ...) pour l'annotation des entités nommées. Comme (Won *et al.*, 2018) nous avons obtenu la version d'EG préparé pour le projet *Reassembling the Republic of Letters* mais avec un analyseur grammatical du français. EG prend donc en compte les catégories grammaticales du français, mais implémente la reconnaissance des entités nommées à partir de règles développées pour l'anglais. C'est un problème reconnu par les chercheurs travaillant avec des langues autres que l'anglais (voire même pour l'anglais d'avant le vingtième siècle). (Il est, bien sûr, possible de modifier les règles et les lexiques de EG.) Perdido, au contraire, est conçu spécifiquement pour la langue française. Par ailleurs, les noms de lieux construits autour d'un ou plusieurs mots peuvent être intégrés dans d'autres types d'entités (personnes, fonctions, ...). Dans le cas d'EG, le balisage interne du nom de lieu n'est pas conservé une fois que l'entité est imbriquée et l'information sur la nature spatiale du nom est perdue. À la différence, Perdido a été spécialement développé pour annoter les entités nommées étendues (ENE) telles que définies par (Gaio, Moncla, 2017) et conserver l'information de chaque niveau d'imbrication.

3. Expérimentations

Le corpus étudié comprend les 14445 articles de l'Encyclopédie appartenant à la catégorie Géographie (Morrissey, Roe, 2017). Ceux-ci sont fournis au format TEI, un premier pré-traitement supprime l'en-tête et la structuration afin de faciliter le travail avec les outils d'annotation. Nous n'avons ni modifié ni modernisé le langage étant donné les recherches antérieures sur les textes anglais de l'époque moderne pour lesquels la modernisation est jugé superflus (Won *et al.*, 2018). Nous avons constitué un corpus d'évaluation composé de 100 articles sélectionnés aléatoirement parmi l'ensemble du corpus. Ces articles

ont été annoté manuellement par K. McDonough en utilisant l'outil GeoViz [2] (McDonough, Camp, 2017). Le corpus d'évaluation comprend environ 30000 mots pour 2151 occurrences de noms de lieux.

Nos expérimentations ont été conçues principalement pour identifier et évaluer les faiblesses des deux méthodes testées. Nous avons intentionnellement utilisé des méthodes qui n'ont pas été conçues spécifiquement pour ce contexte (une pratique de plus en plus courante à mesure que les chercheurs en sciences humaines cherchent à utiliser des méthodes d'extraction d'informations géographiques). Nous avons ainsi acquis des informations précieuses pour guider les futures adaptations de Perdido en réfléchissant à ce qui a fonctionné, à ce qui n'a pas fonctionné et pourquoi. Les résultats de notre évaluation mesurent le rappel, la précision et le F1-score pour EG et Perdido par rapport à l'annotation manuelle (tab. 1). Le rappel mesure le nombre d'entités identifiées par l'outil comme étant un lieu par rapport à tous les noms de lieux existants. La précision est le nombre d'entités correctement identifiées sur le nombre total d'entités de lieu trouvées par l'outil. Le score F1 est la moyenne pondérée de ces deux mesures.

TABLE *1. Evaluation de la reconnaissance des entités nommées spatiales*

	Recall	Precision	F1-score
EG	9.20%	94.64%	16.78%
PERDIDO	55.58%	75.71%	64.10%

Comme attendu le rappel de EG est faible compte tenu de l'absence d'adaptations spécifiques à la langue et au format (9,2 %). Mais Perdido n'obtient pas un rappel aussi haut qu'espéré (seulement 55,58%) par rapport à d'autres tests réalisés avec des romans français du XIX^{e} siècle (rappel à 99,7% pour les noms de rues parisiens) (Moncla *et al.*, 2017). La précision des résultats obtenue par EG est élevée, 94,64% contre seulement 75,71% avec Perdido et s'explique par le grand nombre d'entités non identifiées. En guise de comparaison, l'utilisation de EG sur un corpus de correspondances anglaises (lettres Hartlib) a permis d'obtenir une précision de 53,8%, un taux de rappel de 52,4% et un F1-score de 53,1% (Won *et al.*, 2018). Comme pour les lettres Hartlib, les noms de lieux peuvent apparaître dans l'Encyclopédie en plusieurs langues (grec, espagnol, anglais, allemand et autres), ce qui perturbe le processus de reconnaissance. Perdido produit un nombre important d'erreurs de catégorisation des entités nommées, et a tendance à privilégier la catégorie spatiale. Ce problème se pose en particulier lorsqu'il y a très peu de contexte pouvant servir à la désambiguïsation et que le nom identifié existe dans les ressources géographiques (ici, il est questions de l'Alexandria Digital Library). En revanche comme l'illustre l'exemple (1), lorsque des éléments du contexte, et en particulier ceux compo-

2. http://geoviz.taalmonsters.nl

sants l'ENE, sont présents alors la catégorisation des noms de personnes par exemple pose moins de problème.

(1) a. *le roi de* *Chypre*

b. *le chevalier de* *Jaucourt*

Les lexiques, les priorités et les règles d'interrogation des ressources géographiques nécessitent d'être améliorés afin d'adapter Perdido à l'analyse de documents historiques ne comprenant pas exclusivement des informations géographiques. Un problème anticipé et confirmé par nos expérimentations est celui de la couverture spatiale mais également temporelle des ressources géographiques interrogées. En revanche, les expérimentations ont également permis de mettre en évidence l'importance dans le classement des résultats retournés par les ressources géographiques qui peut être basé sur des heuristiques qui ne sont pas adaptées au corpus étudié.

4. Identification des entités complexes et imbriquées

Comme ont pu le montrer nos expérimentations, le contexte associé aux entités nommées et en particulier les informations impliquées dans la construction des ENE est primordiale pour la catégorisation et la désambiguïsation des entités nommées. Par ailleurs, les ENE peuvent nous permettre de désambiguïser et d'associer un nom de lieu à un lieu spécifique ou de tenir compte de l'ambiguïté grâce à certaines expressions locatives. Elles capturent la portée de l'information géographique telle qu'elle a été façonnée dans une période où les localisations précises étaient difficiles à mesurer et pas nécessairement utiles.

Nos efforts pour identifier les ENE au sein des articles de l'Encyclopédie reflètent l'importance de la manière dont les lieux peuvent être intégrés dans les relations sociales et spatiales. Cela nous permettra de compléter, voire de remplacer, la recherche de coordonnées géographiques avec d'autres types d'informations qui exploitent les relations contextuelles. Parmi les 1721 entités annotées par Perdido, 396 (23%) sont imbriquées au sein d'une ENE. De plus, 51 ENE font référence à un nom de personne parmi lesquelles 33 étendent un nom de lieu (65%). Les noms de lieux imbriqués dans les noms de personne sont généralement perdus, même lorsque ces lieux peuvent constituer des références significatives au contexte spatial du texte.

5. Conclusion et perspectives

La méthodologie proposée est généralisable : les articles de l'Encyclopédie ne sont pas les seules à contenir des associations culturelles, sociales ou géogra-

phiques entre des lieux et d'autres entités. Notre principale contribution à la conception des outils de reconnaissance des entités nommées spatiales consiste à améliorer l'identification de ces entités complexes. Alors que la plupart des outils n'identifieront pas un nom de lieu en tant qu'entité de lieu s'il est imbriqué dans un autre type d'entité, nous souhaitons de notre côté capturer et conserver ces informations. Cela comprend: a) les noms de lieux associés à des relations spatiales, b) des noms de lieux imbriqués au sein d'une autre entité (par exemple une personne ou une institution), c) les entités ambigues, et d) des entités qui ne peuvent pas être géolocalisées (ex : mythique ou extraterrestre). La conception initiale de Perdido adaptée pour l'annotation des ENE nous semble un point important et les différentes faiblesses identifiées de la méthode nous permettrons de faire évoluer l'outil afin d'obtenir de meilleurs résultats pour l'annotation de corpus historiques. Nos expérimentations ont montré les limites de l'utilisation d'une méthode unique basée sur des règles pour l'extraction d'informations. Les principaux problèmes sont dus aux spécificités de la langue utilisée dans les textes historiques en français classique. En effet, Perdido et EG utilisent tous deux une analyse morphosyntaxique basée sur des modèles non adaptés à ce langage. Ainsi, une amélioration importante consiste à construire de nouveaux modèles à l'aide de l'Encyclopédie et d'autres textes de référence de style classique, que ce soit pour l'analyse grammaticale ou pour la reconnaissance des entités nommées. La combinaison d'approches symboliques et statistiques (apprentissage automatique) semble une perspective intéressante pour une meilleure adaptabilité de l'outil au corpus ainsi que pour l'amélioration des résultats.

Bibliographie

Alex B., Byrne K., Grover C., Tobin R. (2015). Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal of Humanities and Arts Computing*, vol. 9, n° 1, p. 15–35.

Comsa M. T., Conroy M., Edelstein D., Edmondson C. S., Willan C. (2016). The french enlightenment network. , vol. 88, n° 3, p. 495–534.

Edelstein D. (2016). Intellectual history and digital humanities. *Modern Intellectual History*, vol. 13, n° 1, p. 237–246.

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.

McDonough K., Camp M. van de. (2017). Mapping the Encyclopedie: working towards an early modern digital gazetteer. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, p. 16–22. ACM.

Moncla L., Gaio M., Joliveau T., Le Lay Y.-F. (2017). Automated Geoparsing of Paris Street Names in 19th Century Novels. In *1st ACM SIGSPATIAL Workshop on Geospatial Humanities.* Redondo Beach, CA, United States.

Morrissey R., Roe G. (2017). Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc., eds. Denis Diderot and Jean le Rond d'Alembert. University of Chicago: ARTFL Encyclopédie Project (Autumn 2017 Edition).

Safier N. (2014, février). The Tenacious Travels of the Torrid Zone and the Global Dimensions of Geographical Knowledge in the Eighteenth Century. *Journal of Early Modern History*, vol. 18, n° 1-2, p. 141–172.

Withers C. W. J., Mayhew R. J. (2011, décembre). Geography: Space, Place and Intellectual History in the Eighteenth Century. *Journal for Eighteenth-Century Studies*, vol. 34, n° 4, p. 445–452.

Won M., Murrieta-Flores P., Martins B. (2018). Ensemble named entity recognition (ner): Evaluating ner tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*.

Un outil numérique pour observer les évolutions du littoral à travers les œuvres d'art : une application aux rivages de la Rance (Bretagne)

A digital tool to observe coastal evolutions throughout artworks: implementation for the Rance estuary (Brittany, French Channel)

Motte Edwige 1[1], McInnes Robin 2[2]

1. LETG, UMR CNRS 6554, Département de Géographie, Université Rennes 2, Place du recteur Henry Le Moal, 35000 Rennes
edwige.motte@gmail.com

2. Geography and Environment department, University of Southampton University Road, Southampton SO17 1BJ
rgmcinnes@btinternet.com

RESUME. Cette communication vise à exposer une initiative numérique innovante de production et de diffusion des connaissances environnementales basées sur la mobilisation d'œuvres d'art comme outil d'observation spatio-temporel. Dans le cadre d'un projet pilote portant sur l'estuaire de la Rance maritime en Bretagne, nous présentons la mise en œuvre d'un « observatoire iconographique » des dynamiques géomorphologiques et paysagères intervenues sur les rivages au cours des deux cents dernières années. Cet observatoire – visant à être développé à plus grande échelle et enrichi par la participation conjointe de musées, galeries d'art, centres d'archives et collectionneurs d'une part, et de citoyens souhaitant emprunter le chemin des artistes d'hier pour re-photographier les vues initiales d'autre part - est matérialisé sous la forme d'un site web interactif permettant la géolocalisation, la visualisation et l'interprétation des couples diachroniques d'images. L'utilisateur s'y voit également offrir la possibilité de synthétiser les observations réalisées à l'échelle du territoire en générant dynamiquement des cartes thématiques et des résumés statistiques. Afin de garantir la fiabilité des observations visuelles issues de l'interprétation des images, des cartes anciennes et sources écrites géo référencées complètent la base de

données. Cette démarche interdisciplinaire, à la croisée entre science, art et société, s'inscrit dans une perspective de connaissance et de valorisation à double résonance des patrimoines : - patrimoine artistique par la mobilisation de l'héritage iconographique ; patrimoine paysager par l'analyse géographique

MOTS-CLES : représentations artistiques – paysage – géomorphologie – patrimoine – changements côtiers – dynamiques sédimentaires – humanités numériques – base de données spatiale – sciences participatives

ABSTRACT. This paper will introduce an innovative digital project aiming at producing and disseminating environmental knowledge by using artworks as tools of spatial observation. In the general frame of a pilot project realised on the Rance estuary in Brittany (France Channel coast), we present the implementation of an « iconographic observatory » of coastal landscape evolutions over the last two hundred years. This observatory – which aims to be developed at larger scale and supplied by the joined collaboration of museums, art galleries, archive centres and art collectors on the one hand, and citizens wishing to re-walk the path of yesterday's artists by re-photographing the original views on the other hand -- is to be found on an interactive website allowing the location and visualization of the diachronic pairs of images. It also provides users with the possibility of putting together the observations made on the scale of the territory by generating dynamic topical maps and statistical summaries. To guarantee the reliability of the visual observations, geo-referenced old maps and written sources will complete the database. This interdisciplinary approach, in between science, art and society, aims at increasing the knowledge and value of two kinds of heritage: first the artistic heritage through iconography, and secondly the landscape heritage through geographical analysis.

KEYWORDS: artworks – landscape – geomorphology – heritage – coastal changes - sediment dynamics – digital humanities – spatial database – citizen sciences

1. Introduction

Les dispositifs de suivi mis en place pour appréhender l'évolution des paysages n'offrent souvent qu'un recul limité. En effet, l'imagerie généralement mobilisée à ces fins - prise de vue in situ dans le cadre de la mise en place d'observatoire photographique, photographies aériennes ou satellitaires - ne permet pas d'observation antérieure à quelques décennies. Il est donc nécessaire de se tourner vers d'autres sources d'information pour bénéficier d'un regard rétrospectif élargi.

Plusieurs études réalisées dans des champs disciplinaires variés ont récemment démontré que les représentations artistiques pouvaient être efficacement mobilisées en tant qu'outils de connaissance des évolutions environnementales (Zerefo 2007 ; Camuffo 2010 ; Borchia et Nesci 2011 ; Metzger et Desarthe 2017).

Sur la base d'une méthodologie préalablement explorée dans le cadre du Projet Arch-Manche, Archéologie, Art et Patrimoine côtier (Mc Inness ; Motte, 2014), cette communication vise à présenter un « observatoire iconographique » des dynamiques géomorphologiques du littoral réalisé à l'échelle des rivages de la Rance en Bretagne.

2. Territoire d'étude

L'estuaire de la Rance correspond à la partie avale – estuarienne – d'un petit fleuve côtier d'une longueur de 106 km qui prend sa source dans les monts du Méné, en Côtes d'Armor, et se jette dans la Manche entre Dinard et Saint-Malo. L'industrialisation ancienne de l'estuaire (Chaigneau-Normand, 2002), poussée à son paroxysme depuis l'implantation de l'usine marémotrice en 1966, en fait un géosystème très singulier au sein duquel l'homme contribue en grande partie aux transformations produites (formes bâties) et induites (dynamiques sédimentaires) du paysage. Cet espace littoral remarquable est aujourd'hui inscrit dans une démarche de labellisation à travers un projet de Parc naturel régional portant sur le territoire Rance-Côte d'Emeraude.

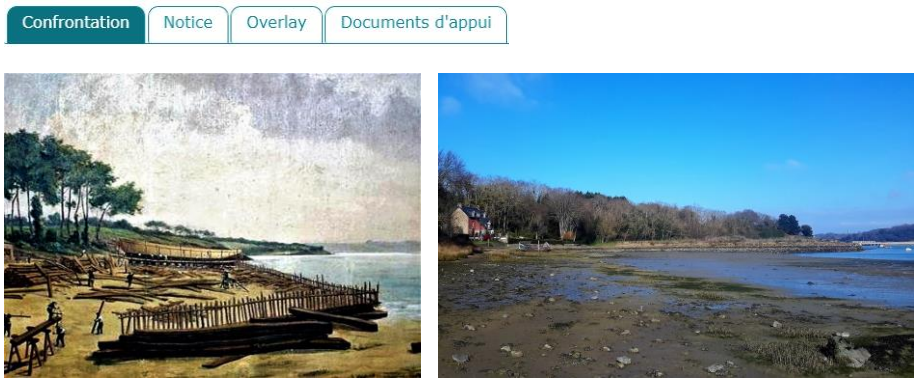
3. L' « observatoire iconographique » : enjeux, matériel et méthodes

L' « observatoire iconographique » vise à mettre en évidence, par la mobilisation et l'interprétation d'un corpus iconographique dense, les principales dynamiques paysagères – architecturales et sédimentaires notamment - intervenues sur les rivages de l'estuaire au cours des derniers siècles.

Plus de 200 images historiques – peintures, gravures, cartes postales anciennes - ont été collectées. Ces images, qui couvrent le territoire de façon homogène, permettent une rétrospective a minima séculaire de l'état passé des rivages de l'estuaire.

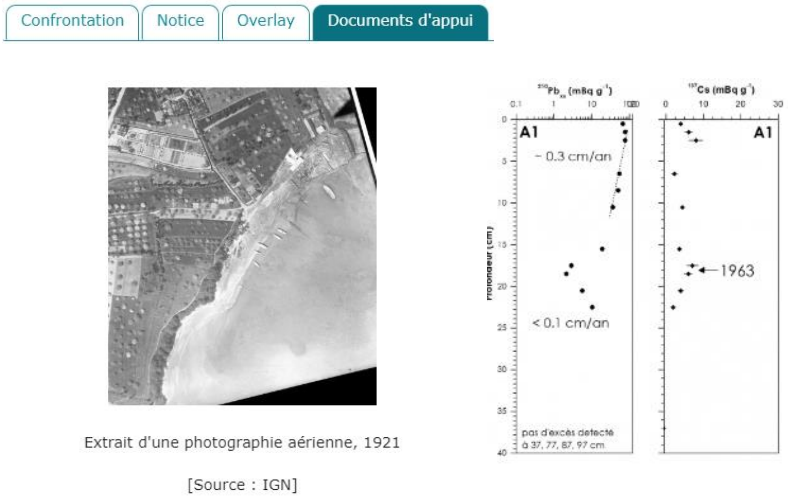
Afin de diffuser les résultats de ce travail au près d'un large public, et d'inscrire la démarche dans une logique pérenne et participative, une interface de diffusion sous forme de site web interactif a été élaborée. D'ores et déjà accessible à l'adresse suivante : http://www.geocompart.com, elle permet de visualiser de façon dynamique - sous forme de cartes ou de graphiques interactifs - l'ensemble des observations réalisées et offre la possibilité aux utilisateurs de soumettre de nouveaux documents en vue de participer à son enrichissement (Le mode d'emploi de la plateforme est présenté sur la page d'accueil sous l'onglet « Naviguer sur le site »).

4. Un cas d'étude en guise d'illustration



L'Anse des rivières || « Les chantiers Tranchmer à la Richardais ». Paul Vernacher, vers 1900 [© Mairie de La Richardais] ↔ [© E. Motte. 2016]

L'observation diachronique du paysage à partir de ce couple d'images figurant une portion de l'Anse des Grandes Rivières à la Richardais permet de rendre compte de la disparition des installations liées à l'atelier de construction navale autrefois implanté sur la grève - observation ici attestée par une photographie aérienne de 1921. Par ailleurs, la comparaison des clichés suggère une modification significative de la nature de l'estran : la plage, vraisemblablement sableuse si l'on en croit la représentation de Vernacher, est aujourd'hui vaseuse et peu praticable.



Extrait d'une photographie aérienne, 1921

[Source : IGN]

Différents documents d'appui confortent ces observations : plusieurs plans et prises de vue aériennes anciennes attestent de l'existence des chantiers navals ; l'analyse radio chronologique d'un échantillon prélevé sur la grève apporte des éléments de précision quant à la chronologie locale de l'envasement.

Conclusion

En plus de permettre un élargissement de l'échelle temporelle des connaissances, cette démarche, qui participe à consolider l'émergence d'un syncrétisme entre images artistiques et études environnementales, présente de nombreux avantages :

- L'approche iconographique peut susciter la curiosité des non-spécialistes (Piveteau, 1965) et constituer en ce sens un vecteur de sensibilisation aux enjeux des changements environnementaux.

- Les représentations artistiques sont susceptibles de jouer un rôle-clef dans la reconnaissance patrimoniale de sites archéologiques et de certains reliefs littoraux (« Géomorphosites ») en participant de leur valeur esthétique et culturelle (Panizza, 2001; Giusti, 2012; Portal, 2013 ; 2014 ; Reynard et al., 2016).

- La relation sensible à l'espace et aux territoires suggérée par l'entrée artistique peut influencer le rapport des sociétés au paysage, et par conséquent orienter les décisions à prendre en matière d'aménagement et/ou de valorisation territoriale (Stocker et Kennedy, 2013).

Perspectives

Dans la perspective d'élargir cette démarche d'observatoire iconographique à un territoire plus étendu (rivages de Bretagne, de la Manche voir métropolitains) et de permettre l'interopérabilité des données contenues avec celle des portails de structures partenaires (organismes d'Etat, collectivités territoriales, infrastructures numériques de recherches, institutions culturelles), une réflexion sur les modalités d'accès et d'usage de l'information publiée et un travail de formalisation des contenus pour leur intégration au Web de données est à mener.

Remerciements

Les auteurs remercient la Fondation de France qui a participé au financement de ce travail au travers du programme « Quels littoraux pour demain ».

Bibliographie

Borchia R., Nesci O. (2011). *The Invisible Landscape. Discovering the real landscapes of Piero della Francesca*, *Il Lavoro Editoriale, Italie*.

Camuffo D. (2010). Le niveau de la mer à Venise d'après l'œuvre picturale de Véronèse, Canaletto et Bellotto. *Revue d'Histoire moderne et contemporaine*. nº. 57-3, p. 92–110.

Chaigneau-Normand M. (2002). *La Rance industrieuse, espace et archéologie d'un fleuve côtier*, PUR, Rennes.

Giusti C. (2012). Les sites d'intérêt géomorphologique : un patrimoine invisible ?, *Géocarrefour*, nº 3-4, p. 151–156.

McInnes R. (2008). *Art as a tool in support the understanding of coastal change*, The crown estate, London

Metzger A., Desarthe J. (2017). Regarde s'il pleut. Effets d'inondation dans la peinture française (1856-1910), *Communications* (Numéro spécial : Le temps qu'il fait). vol. 2, nº 101, p. 119-141.Motte E. (2014). L'usage de représentations artistiques des rivages comme outils de connaissance de l'évolution du littoral: exemples bretons, *Revue d'Histoire maritime*, nº 18, p. 339–358.

Panizza M. (2001). Geomorphosites: Concepts, methods and examples of geomorphological survey, *Chinese Science Bulletin,* vol. 46, nº.1, p. 4 - 5.

Piveteau J L. (1965). Peinture et géographie, *Le Globe. Revue genevoise de Géographie*, Vol. 105, n° 1, p. 9–10.

Poole HA. Garwood D.A. (2018). "Natural allies": Librarians, archivists, and big data in international digital humanities project work, *Journal of documentation*, vol. 74 nº. 4 ? p. 804-826.

Portal C. (2013). Du socle au paysage : essai pour un nouveau regard sur les reliefs, *Projet de paysage*, nº. 8 [Revue en ligne] URL > http://geomorphologie.revues.org/337 (Consulté le 13_11_2016)

Portal C. (2014). Appréhender le patrimoine géomorphologique. Approche géohistorique de la patrimonialité des reliefs par les documents d'archives. L'exemple du Parc National de Killarney (Kerry, Irlande), *Géomorphologie, relief, processus, environnement*, vol. 20, nº 1, p.15-26.

Reynard E., Coratza P., Hobléa F. (2016). Current Research on Geomorphosites, *Geoheritage,* vol. 8, nº.1, p. 1–3.

Stocker L., Kennedy D. (2013). Artistic representations of the sea and coast : implications for sustainability, *Landscapes: the Journal of the International Centre for Landscape and Language*. vol. 4, nº. 2, p. 96-123.

Zerefos C.S., Gerogiannis V., Balis D., Zerefos S. (2007). Atmospheric effects of volcanic eruptions as seen by famous artists and depicted in their paintings. *Atmospheric Chemistry an Physics.* vol. 7, p. 4027–4042.

Mégadonnées, données liées et fouille de données pour les réseaux d'assainissement

Thierry Bonnabaud La Bruyère[1], **Nanée Chahinian**[1], **Carole Delenne**[1], **Laurent Deruelle**[2], **Mustapha Derras**[2], **Francesca Frontini**[3], **Rachel Panckhurst**[3], **Mathieu Roche**[4], **Lucile Sautot**[4], **Maguelonne Teisseire**[4]

1. *HSM, Univ. Montpellier, CNRS, IRD, Montpellier, France*
   nanee.chahinian@ird.fr
2. *Berger Levrault, Montpellier, France*
3. *Praxiling UMR 5267, CNRS, Univ. Paul-Valéry Montpellier 3, France*
   francesca.frontini@univ-montp3.fr,rachel.panckhurst@univ-montp3.fr
4. *UMR 9000 TETIS, Cirad, Irstea, CNRS, AgroparisTech, Univ. Montpellier Maison de la Télédétection, Montpellier, France*

RÉSUMÉ. *Le projet "Mégadonnées, données liées et fouille de données pour les réseaux d'assainissement" (MeDo) a pour objectif de tirer profit des mégadonnées disponibles sur le web pour renseigner la géométrie et l'historique d'un réseau d'assainissement, en combinant différentes techniques de fouille de données et en multipliant les sources analysées. Par l'amélioration des connaissances sur le réseau d'assainissement, ce projet contribue à une meilleure gestion du patrimoine hydraulique existant et de la ressource en eau et permettra d'analyser les interactions entre les politiques de développement urbain et les enjeux liés à la gestion de l'eau.*

ABSTRACT. *The "Megadata, Linked Data and Data Mining for WasteWater Networks" (MeDo) project aims to use Web big data for learning about geometry and history of wastewater networks, by combining different data mining techniques and multiplying analysed sources. The improved knowledge will lead to enhanced management of hydraulic heritage and water resources and allow analysis of interactions between urban development policies and water management related challenges.*

MOTS-CLÉS : *Fouille de données, mégadonnées, TALN, réseaux d'assainissement*

KEYWORDS: *Data mining, Big Data, NLP, wastewater networks*

1. Introduction

Les réseaux d'assainissement font partie intégrante de l'architecture urbaine. Durant le siècle passé, il était de coutume pour chaque opérateur de poser, d'entretenir et d'archiver les données relatives à son réseau (Rogers *et al.*, 2012). Avec les différentes politiques de privatisation/affermage/régie publique, les données ont souvent changé de lieu de stockage et de propriétaire. Pour certaines, notamment les plus anciennes qui n'étaient pas numérisées, l'information est généralement perdue.
L'objectif du projet « MeDo » est de tirer profit des mégadonnées disponibles sur le web pour renseigner à la fois la géométrie du réseau et les données « annexes » pouvant servir aux gestionnaires. Un premier défi consiste à traduire des informations textuelles non structurées en données quantitatives et en connaissance structurée du réseau. Le processus proposé s'appuiera sur des méthodes automatiques d'extraction d'information (EI) afin d'identifier des informations spatio-temporelles (Zenasni *et al.*, 2016) et thématiques (Frontini *et al.*, 2012) à partir des données textuelles en prenant en considération les typologies textuelles parfois non standard rencontrées (Roche *et al.*, 2016). L'analyse des incertitudes liées à ces informations est le deuxième aspect innovant du projet. En effet, il est nécessaire de transformer une information approximative en connaissance incertaine.

Bien que des applications de fouille de données existent déjà dans plusieurs domaines (médical, financier, reconnaissance de la parole et synthèse vocale), leur utilisation dans le domaine de l'eau est moins répandue (Chahinian *et al.*, 2016). Une exploitation originale des mégadonnées disponibles sur le web suppose une approche multidisciplinaire regroupant la linguistique et le traitement automatique du langage naturel (TALN) appliqué au français, l'informatique et l'hydrologie. Une mise en correspondance suivant les trois dimensions spatiales, thématiques et temporelles des données disponibles permettra de compléter les données de la base de référence et de restituer la dynamique locale du réseau au regard des différents acteurs impliqués et des ressentis des usagers, dans un objectif de prise de décision adaptée.

2. Matériels et méthodes

La chaîne de traitement proposée comporte une première phase de collecte de documents via le web pour la constitution d'un corpus exploitable. Les documents sont ensuite convertis en format texte, afin d'être exploités pour l'extraction d'informations sémantiques et spatiales.

Les documents sont récupérés à partir d'une succession de requêtes Google avec filtrage des résultats, ceux-ci étant limités aux dix premiers. Le programme

marque une pause de 90 secondes entre chaque requête afin de ne pas être bloqué par le serveur. La requête est de la forme :

```
"VILLE" AND ("EXPRESSION1" AND "EXPRESSION2") -MOTCLE1 -MOTCLE2
 -site:SITE1 -site:SITE2
```

Les textes contenant les expressions EXPRESSION1 et EXPRESSION2 en lien avec VILLE sont extraits. Certains mots-clés (bricolage, devis, plomberie, etc.) et certains sites (youtube.com, pagesjaunes.fr, etc.) sont exclus afin de ne conserver que les résultats les plus pertinents. La conversion des documents en texte brut se fait avec PDFMiner ou HTML2Text (deux modules Python), selon le type du fichier récupéré.

2.1. Extraction d'informations sémantiques

Une ontologie spécifique au domaine de l'hydraulique des réseaux d'assainissement est établie à partir d'une sélection par les experts de termes présents dans quatre lexiques disponibles sur le web [1].

Afin d'établir un corpus de comparaison et d'entraînement (gold standard), les documents trouvés sur le web sont classés automatiquement par genre textuel (officiel, presse, scientifique et social).

2.2. Extraction d'informations spatiales

L'extraction des entités spatiales et temporelles s'appuie sur la chaîne développée dans le projet Cart'Eaux (Delenne *et al.*, 2017) enrichie selon les besoins des experts et la base de référence développée. La mise en correspondance permet d'identifier des descripteurs et de construire des trajectoires retraçant la dynamique du réseau par une analyse a posteriori du cycle d'évolution de celui-ci (Fig.1).

Dans la chaîne de traitement mise en place, l'outil brat [2] est utilisé pour annoter les documents pour la phase d'apprentissage. Cette phase permet de valider l'extraction des entités nommées effectuée à l'aide de spaCy [3] (bibliothèque Python) et celle des entités temporelles faite avec Heideltime [4] (programme Java).

1. http://www.sivalodet.fr/accueil/lexique/var/lang/FR/rub/2921.html, http://www.hevia.fr/assainissement-lexique, https://www.siaap.fr/glossaire/lexique/B/, https://www.eau-anjou.fr/raccourcis/glossaire
2. http://brat.nlplab.org/
3. https://spacy.io/
4. https://heideltime.ifi.uni-heidelberg.de/heideltime/

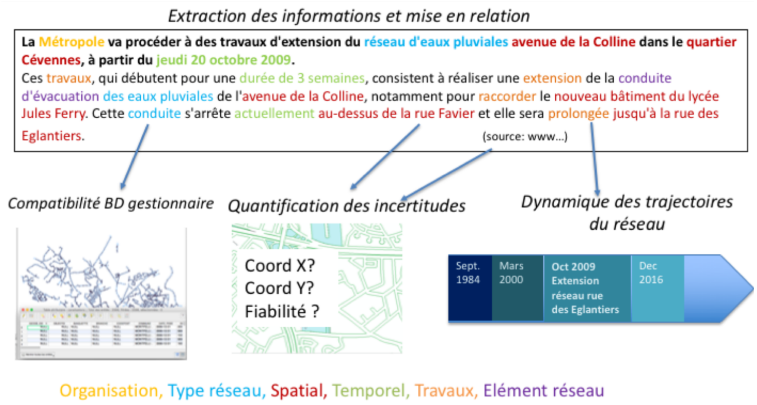


FIGURE *1. Exemple d'extraction des entités sémantiques et spatiales.*

3. Conclusions et Remerciements

Au-delà du web, le prototype développé dans le cadre du projet MeDo permet l'investigation de tout type d'archives numériques disponibles. La chaîne méthodologique mise en place est générique puisque son utilisation repose sur des mots-clés experts, communs à l'ensemble de la communauté des hydrologues, hydrauliciens, aménageurs et gestionnaires de l'eau. À l'avenir, nous pourrions approfondir cette recherche afin d'adapter le prototype à d'autres réseaux (naturels ou artificiels) ailleurs dans le monde, et ce au-delà de l'espace francophone.

Le projet MeDo, d'une durée de 24 mois, bénéficie du soutien de la **Région Occitanie-Pyrénées-Méditerranée** à travers le dispositif **« Recherche et Société(s) 2017 »**.

Bibliographie

Chahinian N., Piat-Marchand A., Bringay S., Teisseire M., Boulogne E., Deruelle L. *et al.* (2016). How can big data be used to reduce uncertainty in stormwater modelling? In *Spatial Accuracy 2016, 5-8 Juillet 2016.*

Delenne C., Chahinian N., Bailly J.-S., Bringay S., Commandre B., Chaumont M. *et al.* (2017). Cart'Eaux: an automatic mapping procedure for wastewater networks using machine learning and data mining. In *2017 AGU Fall Meeting.* New-Orleans, United States.

Frontini F., Aliprandi C., Bacciu C., Bartolini R., Marchetti A., Parenti E. *et al.* (2012). GLOSS, an Infrastructure for the Semantic Annotation and Mining of Documents in the Public Security Domain. In *Proceedings of the Workshop on Exploring and Exploiting Official Publications-EEOP2012 Istanbul May 2012.*

Roche M., Verine B., Lopez C., Panckhurst R. (2016). La néographie dans un grand corpus de SMS français : 88milSMS. In J. G. Palacios, G. D. Sterck, D. Linder, N. Maroto, M. S. Ibáñez, J. T. del Rey (Eds.), *La neología en las lenguas románicas Recursos, estrategias y nuevas orientaciones.*, p. 279-302. Peter Lang, Frankfurt.

Rogers C., Hao T., Costello S., Burrow M., Metje N., Chapman D. *et al.* (2012). Condition assessment of the buried utility service infrastructure – a proposal for integration. *Journal of Tunnelling and Underground Space Technology*, vol. 28, p. 331–344.

Zenasni S., Kergosien E., Roche M., Teisseire M. (2016). Extracting new spatial entities and relations from short messages. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, p. 189-196.

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*

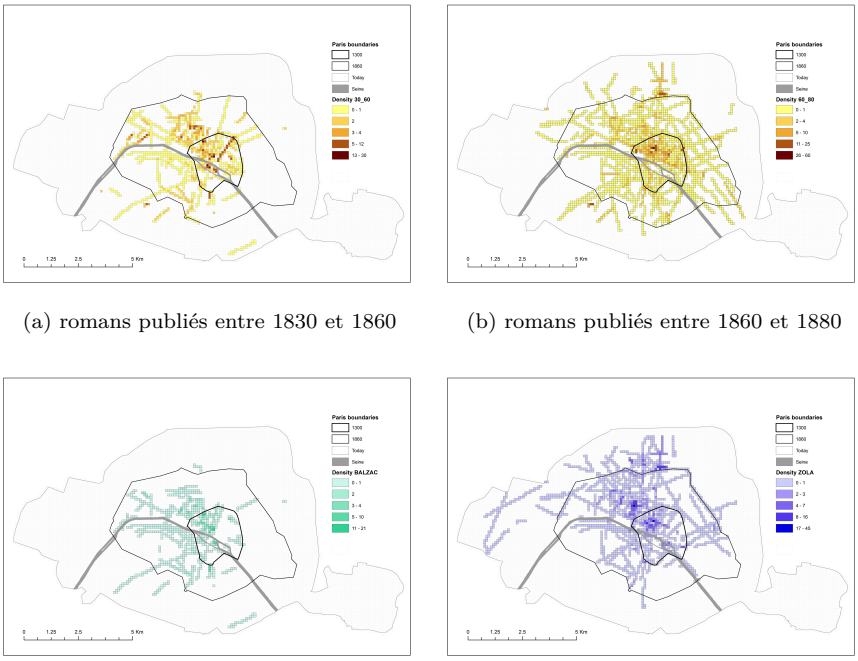
	CQL-TXM	Perdido
occurrences des odonymes trouvées à tort (faux positif)	286	88
occurrences des odonymes non trouvées (faux négatif)	11	117
occurrences des odonymes trouvées (vrai positif)	3573	3467
Total d'occurrences trouvées	3859	3555
Précision	0.926	0.975
Rappel	0.997	0.967
F-score	0.960	0.971

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).

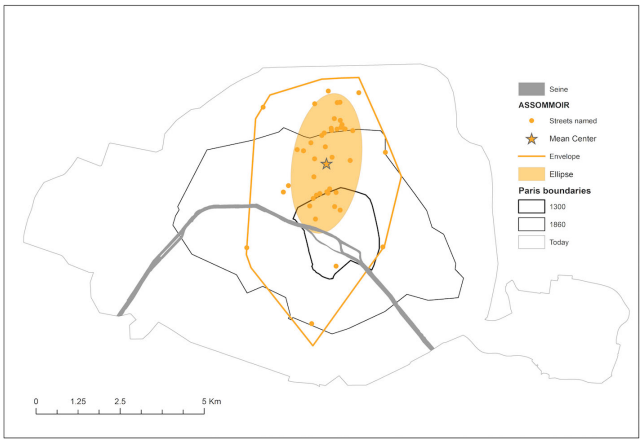


Figure 2. Empreinte spatiale du roman l'Assommoir (Zola)

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.

Visualiser l'évolution des toponymes anciens à partir de sources de données patrimoniales

Perrine Pittet[1], Damien Vurpillot[1], Béatrice Markhoff[2], Marion Lamé[3], Benoist Pierre[1].

1. Intelligence des Patrimoines, UMR 7323 CESR, Université de Tours
{damien.vurpillot|perrine.thuringer|benoist.pierre}@univ-tours.fr

2. LIFAT, Université de Tours
markoff@univ-tours.fr

3. UMR 7324 CITERES-LAT, Université de Tours
marion.lame@univ-tours.fr

Résumé. L'étude de l'évolution des noms de lieux offre un outil de recherche essentiel afin de mieux appréhender l'organisation des territoires et leurs transformations au cours du temps. La normalisation des toponymes, par des cartes et des nomenclatures officielles, relève d'un effort relativement récent à l'échelle du territoire français (fin 18ème siècle). Pour les périodes plus anciennes, le croisement de sources issues de plusieurs domaines de recherche devient nécessaire afin de retracer l'évolution des toponymes avec une marge d'incertitude plus ou moins grande, amenée à fluctuer avec l'apport de nouvelles informations. Cet article introduit le développement d'un outil s'appuyant sur un modèle sémantique pour visualiser l'évolution des toponymes anciens à partir de sources de données patrimoniales.

Abstract. The evolution of place names offers an essential research field to better understand the organization of territories and their transformations over time. Standardization of place names, through maps and official nomenclatures, is a relatively recent effort across french territory (end of the 18th century). For older periods, crossing sources coming from several fields of research becomes necessary in order to trace the evolution of toponyms with a margin of uncertainty more or less important and that is intended to fluctuate with the contribution of new information. This article introduces a semantic model to visualize the evolution of ancient place names from heritage data sources.

Mots-clés : données patrimoniales, évolution des toponymes anciens, CIDOC CRM, CRMgeo, GeoSPARQL, projet HeritageS

Keywords: heritage data, ancient place names evolution, CIDOC CRM, CRMgeo, GeoSPARQL, HeritageS project

1. Introduction

Le projet HeritageS, financé par le programme ARD Intelligence des Patrimoines[1] a mis, parmi ses objectifs, la valorisation des données de la recherche interdisciplinaire sur les patrimoines naturel et culturel du Val de Loire. Cela se traduit par le développement d'une plateforme numérique de données hétérogènes et d'un écosystème d'applications et de services numériques à destination tant des chercheurs que des acteurs socio-économiques du territoire. Ce challenge central du projet vise à rendre interopérables des données patrimoniales produites par près de 33 laboratoires au carrefour de nombreux domaines de recherche qui rencontrent parfois des difficultés légitimes à communiquer entre eux. L'évidente hétérogénéité sémantique de ces données nécessite une approche d'intégration sémantique (Noy, 2004). Le choix s'est porté sur une intégration hybride (Wache et al., 2001) fondée sur l'ontologie de référence pour les ressources patrimoniales, CIDOC CRM (Doerr, 2003), et sur ses extensions. Le modèle de connaissance dérivé de cette ontologie doit permettre à la plateforme d'alimenter les applications et services de l'écosystème et doit ainsi s'adapter pour répondre à leurs besoins spécifiques.

Dans ce contexte, un outil de visualisation en ligne à dimension spatio-temporelle est en développement au sein de l'écosystème HeritageS. Il permettra d'appréhender l'évolution spatio-temporelle de ressources toponymiques sur un territoire en croisant les sources de données patrimoniales intégrées à la plateforme. Il s'agira de faire converger des mentions textuelles de lieux avec la représentation spatiale qui coïncide le mieux en fonction de la période concernée et des éléments de référence disponibles pour inférer le rapprochement toponymique. Fonctionnellement, l'outil devra permettre d'afficher pour chaque lieu visualisé à une date t sur une carte : (1) le centroïde du lieu symbolisé par un point accompagné du toponyme de référence à la date t; (2) l'emprise spatiale du lieu symbolisée par un polygone autour du centroïde; (3) un encart comprenant : les liens vers une copie numérique des documents administratifs publics ou privés à portée historique qui ont permis de restituer le toponyme de référence et l'emprise spatiale à cette date ainsi que les variations orthographiques du toponyme. Cette démarche s'inscrit dans le processus plus complet de création d'une base de connaissances où sont regroupés et recoupés les données scientifiques produites issues de l'interprétation des toponymes par les experts du domaine et les référentiels utiles à ces interprétations. Cette démarche précède l'étape qui vise à lever des ambiguités d'interprétation qui n'avaient pu être résolues de manière satisfaisante dans la première phase d'interprétation.

2. Problématique liée aux données patrimoniales

Les données patrimoniales considérées ici sont des données numériques multimédia issues de projets de recherche interdisciplinaires. Ces données sont de

[1] *Intelligence des Patrimoines,* https://intelligencedespatrimoines.fr

trois types : des sources primaires numérisées et étudiées dans le cadre des projets de recherche (ex : numérisations 2D et 3D de documents et d'objets anciens), des données dérivées produites par ces projets à partir de ces sources primaires (ex : transcriptions, analyses, modèles), et leurs métadonnées associées (ex: notices). Toutes ces données se trouvent intégrées dans la plateforme HeritageS, en outre, les métadonnées sont alignées au CIDOC CRM. Néanmoins, n'alimenteront l'outil de visualisation que les sources primaires dont le contenu présente des informations toponymiques pertinentes, c'est-à-dire éventuellement rapprochées dans les métadonnées associées à des toponymes normalisés. Ces données constituent le corpus que nous cherchons à croiser au référentiel toponymique dont la plateforme HeritageS dispose par ailleurs, afin de proposer de renforcer ou de confirmer un rapprochement toponymique.

Un référentiel toponymique est défini ici comme un document scientifique créé au sein du projet HeritageS, faisant autorité et qui référence l'intitulé ainsi que l'emprise spatiale des différents toponymes issus de documents administratifs publics ou privés, faisant quant à eux autorité lors de leur création. Ce référentiel toponymique est construit à partir de deux types de sources primaires: celles en provenance du corpus lui-même et celles extérieures à celui-ci, mais présentes et exploitables au sein du projet HéritageS. Il s'agit dans tous les cas de (1) documents administratifs cartographiques : cartes, cadastres et autres représentations graphiques conventionnelles de données administratives concrètes ou abstraites référencées géographiquement, reconnues dans le cadre du projet comme faisant autorité ; (2) des documents administratifs textuels : chartes, cartulaires, bullaires, et autres documents administratifs référençant des toponymes, reconnus comme faisant autorité toujours dans le cadre du projet.

Cet ensemble de documents ainsi organisés nous permet de définir deux types de toponymes : (1) des toponymes de référence : toponymes présents dans le référentiel toponymique (ex: "Beaumont" était le toponyme de référence de Beaumont-la-Ronce selon la carte de Cassini[2] de 1793); (2) des toponymes alternatifs : toponymes apparaissant dans les documents primaires du corpus présentant des variations orthographiques ou des formes grammaticales (ex : pluriel, déclinaisons latines) différentes des toponymes de référence (ex: "Beaumont la Ronse" dans le poème Le poème de Ronsard[3] en 1555). Nous considérons donc qu'à une date t un lieu est identifié par un toponyme de référence, défini par son emprise spatiale sur une période donnée et qu'un toponyme de référence peut être associé à un ou plusieurs toponymes alternatifs.

Cependant, l'étude des toponymes à partir de ces ressources comporte plusieurs inconvénients. D'une part, les toponymes qui apparaissent dans les sources primaires du corpus ne sont pas toujours mis en lien avec un toponyme de référence dans les métadonnées associées. Or on trouve de nombreuses variations orthographiques et formes grammaticales pour un même lieu, parfois dans un même

[2] Carte de Cassini : https://www.geoportail.gouv.fr/donnees/carte-de-cassini
[3] Le Voyage de Tours ou les amoureux, Pierre de Ronsard, 1555

document, notamment pour les documents en français ancien issus de corpus du Moyen-Âge et de la Renaissance. Par exemple, nous trouvons “Beaumont” et “Beaumont de la Ronce” dans le même acte de vente en 1476[4]. Lorsqu'ils sont rapprochés à un toponyme de référence dans les métadonnées, celui-ci n'est généralement pas issu d'un référentiel toponymique contemporain de la source primaire mais d'un référentiel plus récent comme Geonames[5].

D'autre part, les toponymes de référence ne sont pas toujours géoréférencés (Timár *et al., 2014*) et c'est souvent le cas pour des référentiels anciens. Le cas échéant, il ne s'agit souvent pas de référentiels cartographiques contemporains de la source primaire du corpus mais des référentiels plus récents, et/ou approximatifs (ex: rattaché au centroïde de la commune), ce qui ne permet pas de connaître la véritable emprise spatiale du lieu à cette époque.

3. Approche

Pour définir le modèle qui servira à l'outil de visualisation, nous avons choisi d'étendre le modèle d'intégration sémantique de la plateforme HeritageS par le biais de l'extension CRMgeo (Hiebel *et al.,* 2017) du CIDOC CRM et de l'ontologie Geosparql (Battle & Kolas, 2011) (cf. Fig.1). L'apport du CRMgeo est le concept de volume spatio-temporel (cf. Fig.1 en bleu E92_SpaceTimeVolume) qui permet de définir un lieu par une projection spatiale et une projection temporelle. Il devient donc possible de décrire l'évolution d'un lieu comme la succession de plusieurs volumes spatio-temporels. L'ontologie Geosparql (cf. Fig.1. en vert) apporte la possibilité de spécifier la géométrie de l'emprise spatiale de lieu dans des formats compatibles avec la plupart des systèmes d'information géographiques. Son alignement avec le CIDOC CRM se fait par le biais d'une relation d'héritage entre le concept geo:SpatialObject et le concept E53_Place. Enfin, afin de modéliser l'alignement de toponymes à des documents administratifs produits à la même époque, nous avons choisi de définir des périodes toponymiques et spatiales de référence. Une période toponymique de référence correspond à la période hypothétique d'usage d'un toponyme de référence. Celle-là démarre à la date de création du premier document administratif connu dans lequel ce toponyme apparaît et se termine à la date de création du document suivant référençant un nouveau toponyme. La période spatiale de référence correspond à la période hypothétique d'existence de l'emprise spatiale d'un lieu. Celle-là démarre à la date de création du premier document cartographique administratif connu qui la définit géographiquement et se termine à la date de création du document cartographique suivant définissant une emprise spatiale différente. Nous avons donc créé deux concepts ToponymReferencePeriod et SpatialReferencePeriod qui héritent du concept E4_Period (cf. Fig.1 en rouge).

[4] AD37, Minute notariale du 5 décembre 1476, Jaloignes, Jehan
[5] GeoNames : http://www.geonames.org/

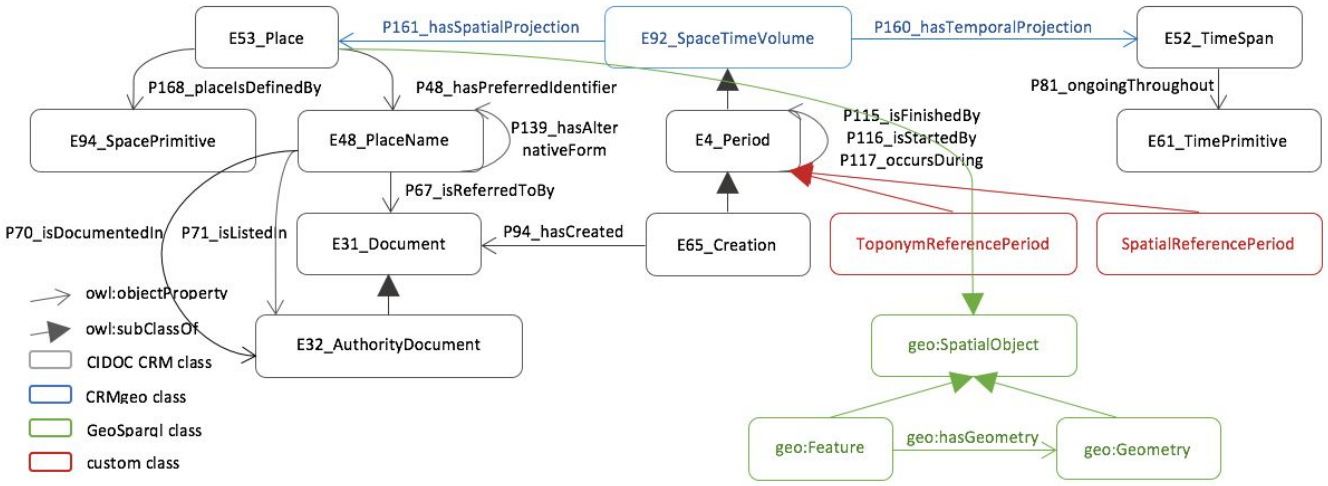


FIGURE *1. Modèle sémantique CIDOC-CRM dédié à l'évolution des toponymes*

4. Modèle

Un lieu se trouve représenté par un volume spatio-temporel (E92_SpaceTimeVolume) défini spatialement (P161_hasSpatialProjection) par une projection spatiale (E53_Place) et temporellement (P160_hasTemporalProjection) par une projection temporelle (E52_TimeSpan). Un toponyme de référence (E48_PlaceName) identifie (P48_isPreferredIdentifierOf) une projection spatiale et apparaît dans une ou plusieurs sources primaires du corpus (E31 Document) elles-mêmes listées (P71_isListedIn) dans le référentiel toponymique (E32_AuthorityDocument). Une projection spatiale est décrite géographiquement (P70_isDocumentedIn) par un document administratif cartographique (E31_Document) et est définie dans le référentiel toponymique (E32_AuthorityDocument). Un document , retenu pour le corpus ou externe au corpus, fut créé (P94_hasCreated) par une activité de (E65_Creation) à une date donnée (E52_TimeSpan). Un toponyme de référence peut disposer d'un ou plusieurs toponymes alternatifs (P139_hasAlternativeForm). Un toponyme (de référence ou alternatif) peut être cité (P67_isReferredToBy) par un ou plusieurs documents présents dans le corpus (E31_Document). La projection spatiale du lieu est définie (P168_placeIsDefinedBy) par une primitive spatiale (E94:SpacePrimitive). La projection temporelle du lieu est définie (P81_onGoingThroughout) par une primitive temporelle (E61_TimePrimitive). Un toponyme de référence identifie (P48_hasPreferredIdentifier) la projection spatiale pendant une période de référence (ToponymReferencePeriod). Cette période de référence démarre (P116_isStartedBy) à la création (E65_Creation) du document administratif (E31 Document) dans

lequel le toponyme de référence (E48_PlaceName) apparaît et se termine (P115_isFinishedBy) à la création (E65_Creation) du premier document administratif (E31 Document) lui succédant dans le temps, et dans lequel apparaît un toponyme différent, les deux toponymes étant indexés dans (E32_Authority_Document). La projection spatiale d'un lieu (E53_Place) est limitée dans le temps sur une période spatiale de référence (SpatialReferencePeriod). Cette période spatiale démarre à la création (E65_Creation) du document administratif cartographique (E31 Document) dans lequel l'emprise spatiale du toponyme de référence est décrite, et se termine à la création (E65_Creation) du premier document administratif cartographique (E31 Document) lui succédant et dans lequel apparaît le changement de l'emprise spatiale précédemment définie, les deux emprises spatiales étant toutes deux indexées dans le référentiel toponymique (E32 Authority Document).

Afin d'aligner un toponyme de référence et une projection spatiale à un toponyme cité dans un document source, deux règles sont définies: (R1) Un toponyme alternatif est associé à un toponyme de référence si la date de création de la source primaire où il est cité est incluse (P117_occursDuring) dans la période de référence de ce toponyme. (R2) Un toponyme alternatif est associé à une projection spatiale si la date de création de la source primaire où il est cité est incluse dans la période spatiale de référence.

Remerciements

Les auteurs remercient Johann Forte et Xavier Rodier pour leur aide et leurs conseils dans l'élaboration de cette proposition.

Références

Battle, R., & Kolas, D. (2011). Geosparql: enabling a geospatial semantic web. Semantic Web Journal, 3(4), 355-370.

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine, 24(3), p. 75.

Hiebel, G., Doerr, M., & Eide, Ø. (2017). CRMgeo: A spatiotemporal extension of CIDOC-CRM. International Journal on Digital Libraries, 18(4), 271-279.

Noy N. F. (2004). Semantic integration: a survey of ontology-based approaches. ACM Sigmod Record, 33(4), p. 65-70.

Timár, G., Mészáros, J., & Molnár, G. (2014). A simple solution for georeferencing the Cassini map series of France. In 9th International Workshop on Digital Approaches to Cartographic Heritage.

Wache, H. et al.(2001). Ontology-based integration of information-a survey of existing approaches. IJCAI-01 workshop: ontologies and information sharing. vol. 2001 p. 108-117.

Annexe

Nous illustrons l'utilisation du modèle avec l'exemple du site de Beaumont-la-Ronce. L'objectif est ici de trouver quel est le toponyme de référence associé au toponyme "Beaumont la Ronse" cité dans le poème "Le Voyage de Tours" de Pierre de Ronsard datant de 1555. Dans les ressources d'HeritageS, le référentiel toponymique ayant précédé la création de ce poème est le registre de chancellerie de Louis XII, datant de 1479 dans lequel le toponyme de référence est "Beaumont de la Ronce". Le référentiel toponymique lui succédant est la carte de Capitaine datant de 1790 dans laquelle le toponyme de référence est "Beaumont la Ronce". Nous décrivons l'ensemble de ces informations à l'aide du modèle présenté plus haut (cf. Fig. 2). La période toponymique de référence délimitée par les deux référentiels s'étend de 1479 à 1790 (cf. Fig. 2 en rouge). L'application de la règle (R1) permet d'associer le toponyme "Beaumont la Ronse" au toponyme de référence "Beaumont de la Ronce" (cf. Fig. 2 en violet).

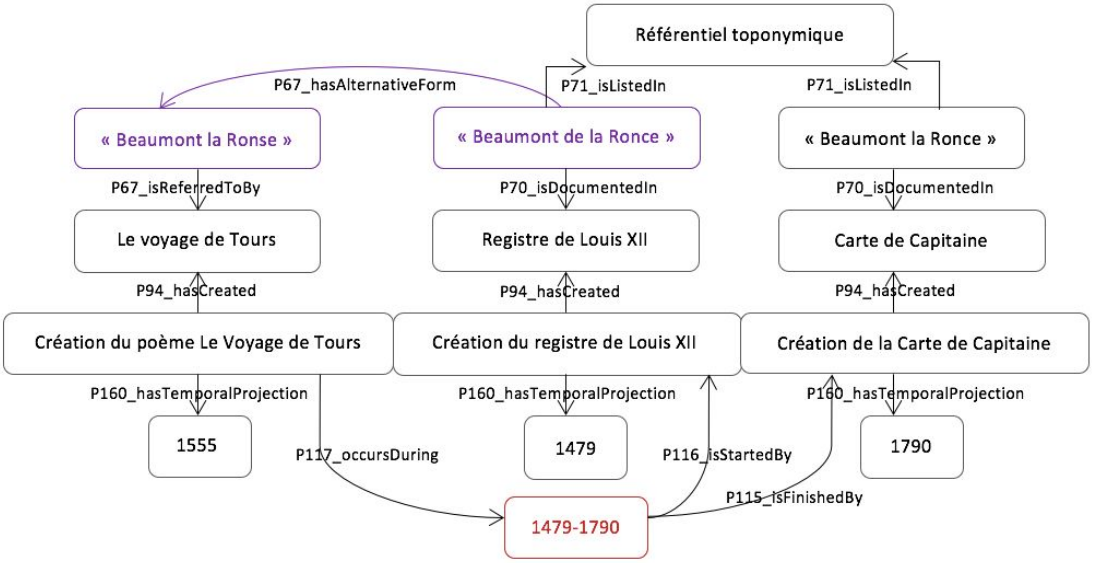


Figure 2. Modèle appliqué à l'exemple de Beaumont-la-Ronce

La période spatiale de référence pourra être de la même manière définie à partir des documents administratifs cartographiques produits à la même époque ou sur une période la plus proche possible dans le temps. La règle (R2) du modèle sera alors appliquée pour associer la projection spatiale restituée au toponyme. Le volume

spatio-temporel de Beaumont-la-Ronce en correspondant au lieu désigné par le poème de Ronsard en 1555 pourra alors être construit. Selon le même processus, d'autres volumes spatio-temporels de Beaumont-la-Ronce seront définis à partir d'autres données patrimoniales. Leur évolution pourra être visualisée dans l'outil cartographique en cours d'élaboration, sans que cela préjuge d'emplois pour le moment inconnus et absents des sources disponibles.

L'Atlas Historique du Limousin. Questions et méthodes vers un système de données spatio-temporelles contributif

Juliette Morel

Criham, Université de Limoges
Faculté des Lettres et Sciences Humaines
Campus Vanteau, 39E rue Camille Gérin, 87036 Limoges, France
Juliette.morel@unilim.fr

Nous proposons de présenter à l'atelier SAGEO HumaNS'2018 l'Atlas Historique du Limousin (AHL), et plus précisément les enjeux méthodologiques impliqués par le développement d'un outil de contribution directe pour l'alimentation de la base de données spatio-temporelle de cet Atlas historique.

L'Atlas Historique du Limousin est un projet mené au Laboratoire Criham de l'Université de Limoges et a bénéficié depuis 2014 des financements de la Région Limousin et de l'Institut de Recherche des Sciences de l'Homme et de la Société de l'Université de Limoges. Il a pour objectif de rassembler des données spatio-temporelles et des sources historiques sur l'histoire du Limousin, et de proposer un outil d'interrogation spatiale, temporelle et/ou thématique sur le web pour les représenter et les comparer. L'Atlas a été mis en ligne en 2015 et agrémenté d'une interface de *webmapping* sur l'histoire de Limoges en 2017. Ses perspectives d'évolution concernent actuellement le mode d'alimentation de la base de données géo-historique. L'objectif est de promouvoir la contribution directe de différents acteurs grâce aux outils du web 2.0 pour permettre une croissance importante de la quantité de données et donc un potentiel de croisements élargi et inédit. Nous entendons solliciter des contributions de deux types : d'une part des contributions spécialisées liées aux champs d'expertise des collaborateurs plus ou moins directs de l'AHL, et d'autre part des contributions du grand public sur des aspects précis et encadrés. Cette ouverture à la contribution engendre des questionnements méthodologiques, techniques, voire déontologiques. Ces questionnements et les statistiques présentés dans la suite de l'article se fondent notamment sur une enquête menée en 2018 auprès des collaborateurs de l'AHL, contributeurs spécialisés

potentiels (contributions de premier type), pour évaluer les besoins et les problèmes liés au développement d'outils contributifs[1].

TABLE 1 : Récapitulatif du contexte de développement des outils contributifs de l'AHL

Contributeurs potentiels	Historiens du Criham (23% du total des contributeurs spécialisés) ou d'autres laboratoires de recherche (11%) ; Archéologues opérant en Limousin (8%) ; Sociétés savantes (15%) ; Dépôts d'archives et collectivités territoriales (archives municipales et départementales, ville de Limoges, *etc.*) (27%) ; le grand public.
Données	Des données et sources de données hétérogènes localisée (90%) et datées (100%) : littérature grise, plans anciens, photographies, descriptions littéraires et documentaires, données archéologiques, données issues de recherches historiques, données statistiques, démographiques, *etc.*
Implémen-tations réalisées ou prévues	1e étape : Une base de données relationnelle spatio-temporelle et documentaire et une plateforme d'interrogation web (réalisée et en ligne) ; 2e étape : des outils de contribution/versement automatique de données spatio-temporelles dans la base de données (en cours). 3e étape : intégration des données à un entrepôt de données intégré au web des données (à développer).
Média de visualisation et consultation	Un moteur de recherche spatial, temporel et thématique. Webmapping et datavisualisation dynamique pour la visualisation des sélections (réalisé et en ligne. À rendre automatique / à adapter à l'intégration automatique de données).

1. Hétérogénéité géographique et qualité disparate des données historiques collectées

D'après l'enquête menée auprès des contributeurs potentiels de l'AHL, les jeux de données historiques dont l'emprise correspond au Limousin ne concernent que 40% des contributions potentielles et la granularité de la localisation est variable : 46% des données sont localisées à l'échelle infra-communale, 36% à l'échelle communale, et 9% respectivement à l'échelle départementale et à l'échelle régionale. La base de données spatio-temporelle contributive de l'AHL ne pourra donc offrir une couverture géographique continue ni homogène. Les méthodes de géolocalisation (notamment en termes d'implantation cartographique et de précision de la géométrie) sont différentes d'un chercheur ou d'une étude à l'autre ; et ce d'autant plus dans un contexte interdisciplinaire. Les géométries issues de travaux archéologiques, souvent très précises mais correspondant à un état partiel des objets historiques (puisque

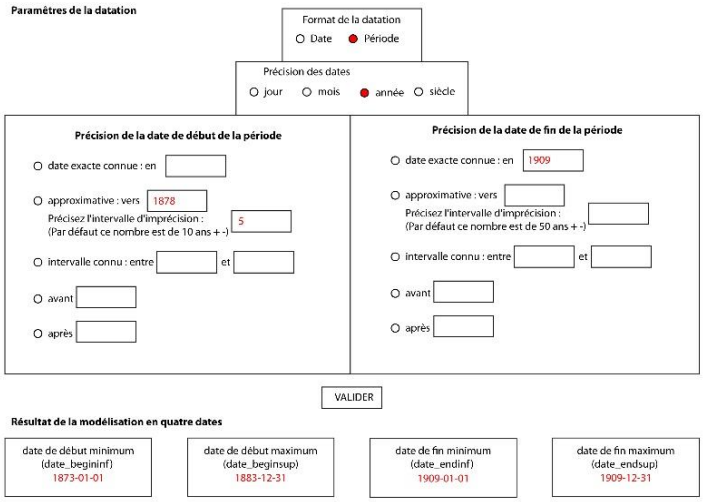
[1] Cette enquête est un questionnaire web (créé avec le site *Drag'n'Survey*) envoyé à 130 personnes ayant des liens avec l'Atlas Historique du Limousin (voir la liste des contributeurs potentiels dans la Table 1) ou l'Atlas Historique de la Nouvelle-Aquitaine (qui a vocation à fusionner dans les années à venir avec l'AHL). 33 personnes interrogées ont déclaré avoir des données géo-historiques sur le Limousin et répondu à la totalité du questionnaire. Les pourcentages indiqués dans le reste de l'article concernent ces 33 personnes.

correspondant à l'état des découvertes matérielles), sera par exemple difficile à comparer à la localisation des camps de résistants en Corrèze pendant la seconde guerre mondiale, agrégés à la commune faute d'archives plus précises.

À cette hétérogénéité s'ajoutent la difficulté intrinsèque de la géolocalisation des données historiques – celles-ci se limitant aux informations dont témoignent les traces conservées (écrites ou matérielles), forcément partielles –, ainsi que les cas de contributions incomplètes. L'enjeu sera donc de renseigner la cause de l'incomplétude des données : si les données n'existent pas du fait d'une absence de sources, si les données n'ont pas encore été intégrées à la base de données mais qu'elles pourront l'être plus tard, ou si le contributeur ne possède pas les données sur ce territoire et/ou à cette époque parce que ceux-ci sortent de son cadrage spatio-temporel. Nous entendons répondre à ces diverses sources d'imperfection par un sourçage précis, systématique et obligatoire des données, de leur(s) source(s) et de leur intégration dans la BDD : qui a produit les données ? dans quel cadre institutionnel ou scientifique ? à quel moment ? qui les a entrées dans la base ? y-a-t'il eut des mises à jour ?

De la consultation de nos différents collaborateurs est également ressortie une difficulté concernant la formalisation numérique de la datation, problème récurrent de la modélisation en histoire. Pour répondre à cette difficulté et pour que chaque contributeur puisse formaliser l'aspect temporel de ses données de manière autonome mais standardisée, nous travaillons à la construction d'un outil de traduction d'une datation complexe en une modélisation en quatre dates (voir figure 1).

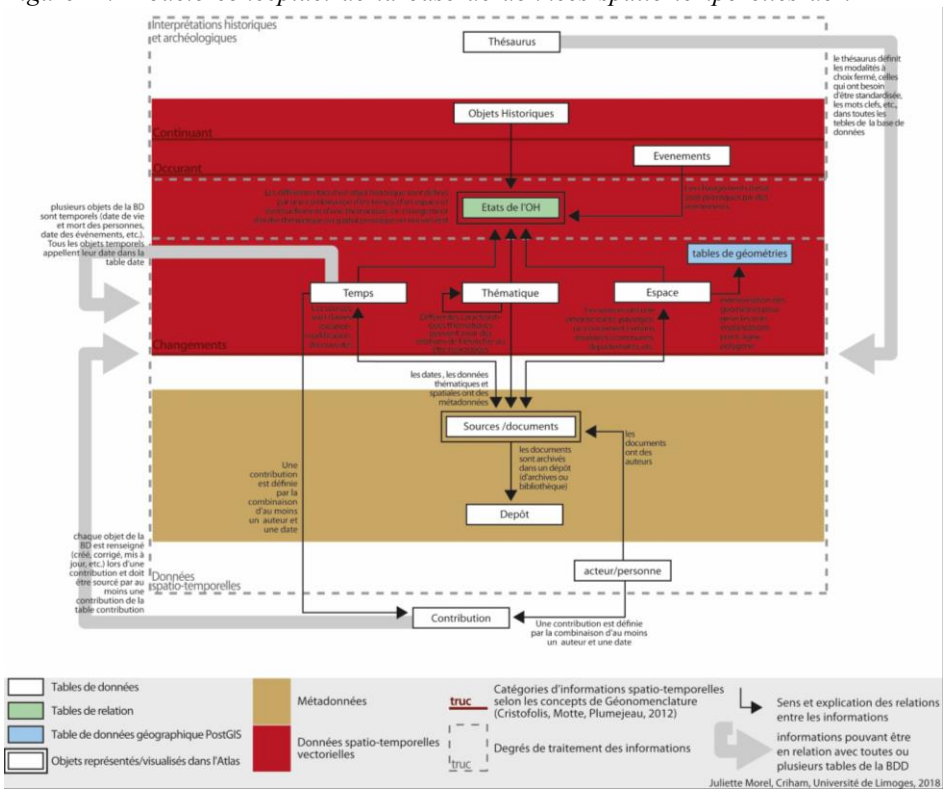
Figure 1 : Prototype d'interface pour l'outil de modélisation automatique des datations



2. Un modèle pour prendre en charge l'hétérogénéité des données et des sources

En conséquence de ces diverses contraintes, le modèle de données sur lequel repose l'Atlas doit être équilibré entre flexibilité, pour accueillir l'hétérogénéité des données, et généricité, pour pouvoir les comparer. Nous avons choisi de construire un modèle de données spatio-temporel classique autour de la triade de Peuquet (Peuquet, 1994), c'est-à-dire que nous définissons les entités spatio-temporelles comme une combinaison d'attributs de trois types : spatial, temporel et thématique. Les dimensions temporelle et spatiale sont obligatoires – ce sont les deux points communs, donc de comparaison, minimaux entre toutes les données – et la dimension thématique est libre et capable d'absorber la diversité des travaux de recherche sur l'histoire du Limousin. Le changement historique est modélisé par la relation entre des éléments « continuant » (identité des objets historiques malgré leur modification lorsque l'analyse historique estime qu'il y a permanence), des états changeant (définis par la combinaison : temporel + spatial + thématique) et des événements « occurrant », expliquant ponctuellement le passage d'un état à l'autre des objets historiques (Plumejeaud, Cristofoli, Motte, 2012).

Figure 2 : Modèle conceptuel de la base de données spatio-temporelles de l'AHL



La fixation d'un tel modèle contraint les contributeurs à adapter leurs informations à une structure de données prédéfinie et générique, qui n'est pas spécifique à leurs données. L'enquête que nous avons menée révèle qu'entre 20 et 60% de nos contributeurs potentiels – selon qu'ils sont historiens universitaires ou professionnels de l'information spatiale et/ou historique – possèdent des informations géo-historiques sous forme numérique (tableur ou fichier SIG) et que 30 à 60% d'entre eux (toujours selon la même distinction de corps professionnels) sont prêts à adapter le format de leurs informations au modèle conceptuel imposé.

3. L'enjeu de la participation

Comme dans tout projet de sciences participative, l'enjeu tient précisément dans l'ampleur de la participation qu'il rencontrera. L'enquête et les chiffres précédemment cités indiquent que le nombre de contributeurs spécialisés est limité, mais que leur motivation et la pertinence de leur contribution est grande. C'est exactement ce que nous attendons des contributions spécialisées dans le cadre de l'Atlas Historique du Limousin.

Comme nous le disions en introduction, un deuxième vivier de contributeurs est visé : le grand public. Il s'agit cette fois de solliciter une connaissance locale, liée à la mémoire collective, à la pratique du territoire et à l'investissement citoyen dans l'histoire locale. Cette sollicitation se fera concernant des aspects très précis de l'histoire et de la représentation du territoire Limousin, pour nous aider à préciser l'estimation spatiale et temporelle et à numériser un grand nombre d'informations : photographies anciennes et paysages à localiser et dater, cartes anciennes à vectoriser, *etc.* Comme dans le cas des grands programmes de production collective de connaissance (*Wikipédia*, *OpenStreetMap*), seule une participation significative (variant selon l'échelle et l'ambition du projet) garantit la qualité des informations collectées. Ici comme ailleurs, nous serons sans aucun doute soumis à la loi du « 90-9-1 » (Meriskay, Roche, 2011). Il faut donc tout mettre en œuvre pour rendre simple et attractif les outils permettant de telles contributions. Voilà pourquoi nous attachons un soin particulier à leur graphisme, leur ergonomie et leur pédagogie, en nous inspirant de l'outil *Footprint checker* créé par la *New-York Public Library* pour inviter le grand public à vectoriser ou à valider la vectorisation des bâtiments présents sur les anciennes cartes de New-York[2].

4. Conditions d'intégration et de diffusion des données

L'ouverture à la contribution – qu'elle soit spécialisée ou grand public – implique d'expliciter des conditions d'intégration et de diffusion des données collectées et de

[2] Outil *Footprint checker* de la *New-York Public Library*, en ligne, consulté le 25/10/2018, https://buildinginspector.nypl.org/footprint

s'assurer de leur acceptation par les contributeurs. Or tous les contributeurs ne sont pas d'accord sur le sujet, et il semble nécessaire d'adapter au cas par cas le degré de diffusion des données pour ne pas risquer d'en « perdre ». Celui-ci peut aller du contributeur qui permet le téléchargement de ces données sans conditions de réutilisation (60% des contributeurs spécialisés qui se sont prononcés sur cette question), jusqu'à l'usager qui souhaite seulement visualiser ses données de manière temporaire dans le *webmapping* sans les intégrer à la base de données, en passant par le contributeur qui refuse le téléchargement (33.3%) ou qui n'accepte qu'une consultation restreinte par ses seuls collègues (6.7%).

Pour résumer, notre intention est de multiplier les possibilités de contributions en multipliant les conditions d'intégration et de diffusion et les interfaces de saisie, pour les adapter aux divers données, usages et publics. Les contributeurs pourront :

- Intégrer des données à la BDD en dessinant les géométries une par une sur une carte et en renseignant les modalités temporelles et thématiques concernant chaque objet. Le contributeur sera considéré comme l'auteur des données et définira leurs conditions de réutilisation selon les combinaisons de la licence *Creative Commons* (cette solution intéresse 46,7% des contributeurs spécialisés qui se sont prononcés);
- Intégrer des données à la BDD en uploadant des tableurs ou fichiers SIG complexes correspondant à une structure uniformisée préalablement imposée et documentée. Le contributeur sera considéré comme l'auteur des données et définira leurs conditions de réutilisation selon les combinaisons de la licence *Creative Commons* (26,7%) ;
- Ajouter temporairement des données géographiques au *webmapping* de l'AHL pour les visualiser avec/superposées aux données de l'AHL[3];
- Proposer des campagnes de numérisation participative précises au grand public en utilisant des outils ad-hoc simples et ergonomiques (la première concernera la localisation des prises de vue des photographies anciennes à Limoges).

Conclusion : perspectives et limites de la contribution dans le projet AHL

Les choix méthodologiques effectués pour répondre aux problématiques posées par l'ouverture de l'Atlas Historique du Limousin à la contribution prennent le parti de la flexibilité face à la diversité des contributeurs, des données et des usages ; celle-ci nous paraissant nécessaire à l'attractivité et donc à la réussite du projet contributif. Néanmoins, ces choix pourront devoir être reconsidérés selon les directions que

[3] Comme le propose par exemple *Atlas Historique des territoires politiques*, qui permet de « Glissez-déposez un fichier texte [csv] dans [une zone] pour projeter vos points sur la carte », GEO-LARHRA, LARHRA – UMR5190, CNRS, ISH Lyon, mis en ligne entre 2012 et 2015, consulté le 25/10/2018, http://geo-larhra.ish-lyon.cnrs.fr/?q=atlas-historique/regroupement-de-territoires/evolution-des-territoires-en-europe.

prendra l'Atlas dans le futur, comme l'intégration de ses données au web sémantique. Une telle intégration pose la question suivante : une fois que toutes les données de l'AHL seront moissonnables, comment maîtriser les hybridations et les éventuelles déformations résultant de la circulation des données ? La question des conditions juridiques de la diffusion des données collectées devra alors être reposée et uniformisée.

Plus globalement, c'est la pérennité institutionnelle de l'Atlas Historique du Limousin qui pourrait être remise en doute – et qui redéfinirait les choix méthodologiques. Le projet contributif de l'AHL vise à accumuler des données Géo-historiques sur le long terme. La pérennité nécessaire à cet objectif est garantie par l'écosystème du web qui est promis à durer (à condition d'actualisations techniques régulières) ; mais elle est confrontée aux modes de financement à court terme de la recherche universitaire. Le projet AHL est en effet financé jusque début 2019. Il doit fusionner en 2019 avec l'Atlas Historique de la Nouvelle-Aquitaine, dont le financement (Région Nouvelle-Aquitaine) court lui jusqu'en 2021. Mais après cela, comment garantir l'actualisation technique et la modération scientifique minimales nécessaires au fonctionnement de l'aspect contributif de l'Atlas ?

Bibliographie

Butez C. (2013). Conception d'un atlas historique numérique et d'une plateforme de travail collaborative à partir de la méthode SyMoGIH. Partie 1 - Naissance et conception d'un système d'information géo-historique collaboratif. *Géomatique Expert*, n°91, pp 30-35.

Joliveau T., Noucher M., Roche S. (2013). La cartographie 2.0, vers une approche critique d'un nouveau régime cartographique. *L'Information géographique*, vol. 77, n° 4, pp. 29-46.

Mericskay B., Roche S. (2011). Cartographie 2.0 : le grand public, producteur de contenus et de savoirs géographiques avec le web 2.0. *Cybergeo : European Journal of Geography* [En ligne], Science et Toile, document 552, mis en ligne le 20 octobre 2011, consulté le 08 avril 2018. URL : http://journals.openedition.org/cybergeo/24710 ; DOI : 10.4000/cybergeo.24710

Peuquet D. (1994). « It's about Time: A Conceptual Framework for the Representation of Temporal Dynamics. Geographic Information Systems », *Annals of the Association of the American Geographers*, n°84 (3), pp. 441-461.

Plumejeaud C., Cristofoli P., Motte C. (2012). « De l'étude des nomenclatures territoriales à la modélisation des dynamiques des territoires administratifs en France », *Revue Internationale de Géomatique* – n° 1/2012, pp. 1-5

Atlas Historique du Limousin (2017), Criham, http://www.unilim.fr/atlas-historique-limousin/

Le *Dictionnaire topographique. Une API pour les toponymes anciens français*

Olivier Canteaut[1], Vincent Jolivet[2], Julien Pilla[3]

1. École nationale des chartes
65 rue de Richelieu, 75002 Paris, France
olivier.canteaut@chartes.psl.eu

2. École nationale des chartes
65 rue de Richelieu, 75002 Paris, France
vincent.jolivet@chartes.psl.eu

3. École nationale des chartes
65 rue de Richelieu, 75002 Paris, France
julien.pilla@chartes.psl.eu

RÉSUMÉ. Le Dictionnaire topographique *est une ressource de premier plan pour les historiens et les toponymistes : il compte près de 400 000 entrées et hiérarchise plus de 980 000 toponymes anciens attestés, datés et référencés. Depuis 2009, le CTHS numérise les différents volumes, pour en proposer une édition numérique. Une nouvelle application est en cours de développement. Adossée à une API documentée, elle offre un accès normalisé aux données, et tire parti du liage des données au référentiel INSEE pour localiser les toponymes. L'objectif de cette API est de favoriser les remplois de cette ressource importante, mais aussi d'en poursuivre l'enrichissement en offrant aux chercheurs une interface pour corriger et compléter le contenu au gré de leurs découvertes. Cette communication vise à faire connaître cette ressource essentielle pour l'étude toponymique : nous présenterons l'histoire de cette entreprise éditoriale hors norme, détaillerons les étapes de la numérisation, de la restructuration et de l'enrichissement des données. Nous présenterons enfin l'API et l'application associée qui rend possible l'exploitation de nouvelles relations au sein du* Dictionnaire*, et qui surtout permettra de revitaliser une entreprise éditoriale inachevée.*

ABSTRACT. The Dictionnaire topographique *is a leading resource for historians and toponymists: it has nearly 400,000 entries and ranks more than 980,000 ancient toponyms that have been documented, dated and referenced. Since 2009, the CTHS has been digitising the various volumes, in order to offer a digital edition. A new application is being developed. Its documented API provides standardized access to data, and uses data binding to the INSEE repository to locate place names. The objective of this API is to promote the re-use of this*

important resource, but also to continue to enrich it by providing researchers with an interface to correct and complete the content as they discover it. This paper aims to promote this essential resource for toponymic research: we will present the history of this extraordinary publishing initiative, detailing the steps involved in digitization, restructuring and data enrichment. Finally, we will present the API and the associated application that makes it possible to exploit new relationships within the Dictionnaire*, and above all, to revitalize an unfinished editorial initiative.*

Mots-clés : données géohistoriques, toponymes, France, API, application Web, Web de données, gazetier

Keywords: geohistorical data, toponyms, history, France, API, Web application, Linked Open Data, gazetteer

---

1. Introduction

Entreprise éditoriale au long cours lancée par le Comité des travaux historiques – aujourd'hui Comité des travaux historiques et scientifiques (CTHS) –, le *Dictionnaire topographique* a eu pour mission de compiler tous les toponymes anciens et modernes de la France. Au total, 35 tomes (pour 35 départements) ont été publiés de 1861 à 2008. Même si la couverture (plus du tiers du territoire métropolitain) n'est pas à la hauteur de l'ambition nationale initiale, le *Dictionnaire topographique* est une ressource de premier plan pour les historiens et les toponymistes : il compte près de 400 000 entrées et hiérarchise plus de 980 000 toponymes anciens attestés, datés et référencés. Depuis 2009, le CTHS numérise les différents volumes, pour en proposer une édition numérique, enrichie progressivement au fil des numérisations. Le corpus est complet depuis 2018, et à cette occasion, une nouvelle application est en cours de développement. Celle-ci n'est plus une simple édition numérique ; adossée à une API documentée, elle offre un accès normalisé aux données, et tire parti du liage des données au référentiel de l'INSEE pour localiser les toponymes. L'objectif de cette API est de favoriser les remplois de cette ressource importante, mais aussi d'en poursuivre l'enrichissement en offrant aux chercheurs une interface pour corriger et compléter le contenu au gré de leurs découvertes. Cette communication vise à faire connaître cette ressource essentielle pour l'étude toponymique.

2. Le *Dictionnaire topographique de la France*

2.1. Une entreprise éditoriale

Lancé en 1859 sur proposition de l'historien Léopold Delisle, le *Dictionnaire topographique de la France* avait pour ambition de doter les savants d'un dictionnaire géographique « de la France ancienne et moderne » utile à l'étude de l'histoire et de la géographie des provinces françaises. Il s'agissait de recenser les noms de lieux fournis par la géographie physique, les noms des lieux habités et ceux qui se rapportent à la « géographie historique » (anciennes circonscriptions, fiefs, abbayes, vieux chemins, etc.), en indiquant pour chacun de ces lieux sa nature (ferme, hameau, moulin, etc.), sa localisation (commune d'appartenance), diverses données historiques (ressort judiciaire, ecclésiastique) et surtout les différentes graphies de son nom au cours des siècles, dûment datées et référencées. En raison de l'ampleur de la tâche, le Comité opta rapidement pour le principe d'un volume par département, le tout devant à terme – et en théorie – être unifié par un index général.

Les débuts furent prometteurs : dix-neuf dictionnaires parurent entre 1861 et 1884, dus principalement aux archivistes départementaux, parfois à d'autres érudits, correspondants locaux du Comité. Le mouvement se poursuivit à un rythme plus modéré, à raison de deux à quatre dictionnaires par décennie jusqu'aux années 1920, les derniers volumes publiés étant ceux de la Sarthe et de la Seine-et-Marne (années

1950), de la Seine-Maritime (années 1980) et, dernier en date, celui de la Saône-et-Loire (2008). Ce sont aujourd'hui trente-cinq départements qui sont couverts, représentant plus du tiers du territoire national métropolitain.

2.2. Une base de toponymes anciens

L'appellation *Dictionnaire topographique* reflète imparfaitement le contenu des dictionnaires édités. Ils se conforment tous au modèle éditorial prescrit par L. Delisle, composé de trois parties : 1. une introduction consacrée à la géographie historique, et comprenant une description physique du département (justifiant le qualificatif « topographique ») ; 2. le corps même du dictionnaire, composé des notices historiques des toponymes (nature et localisation administrative du lieu, liste des formes anciennes référencées) ; 3. l'index général des formes anciennes.

Le cœur du dictionnaire, la donnée numérisée et (re)structurée pour construire l'application, consiste donc en une liste considérable et ordonnée de toponymes historiques attestés : cette particularité se comprend aisément dans la mesure où les principaux contributeurs ont été des archivistes et des historiens qui en travaillant à ce long recensement des formes toponymiques anciennes se conformaient à la prescription initiale de L. Delisle les enjoignant de « n'écarter aucun nom qui ait un caractère d'ancienneté et qui présente un intérêt historique ou philologique ». L'ensemble compile plus de 980 000 toponymes, typés, datés et référencés.

3. Une donnée structurée pour la recherche

C'est cette collection de toponymes qui a fait l'objet d'une campagne de numérisation depuis 2009, conduite par le CTHS, en partenariat notamment avec avec l'École des chartes, le centre d'onomastique des Archives nationales et l'UMR ARTEHIS (Dijon). L'objectif initial était de proposer une réédition numérique des dictionnaires imprimés « offrant des possibilités d'interrogation et d'exploitation nouvelles : interrogation conjointe de plusieurs dictionnaires, interrogations croisées, recherche par type de lieu... ». Ce travail de numérisation (OCR, restructuration XML des notices selon un schéma simple) a été rendu possible par la (relative) homogénéité éditoriale des différents volumes, qui ont tous respecté assez fidèlement les prescriptions éditoriales initiales. Dix unités de sens furent identifiées : article, vedette, définition, localisation, typologie, forme ancienne, commentaire, date, référence et renvoi. Les enrichissements typographiques (italique, majuscules, exposants) et la structure éditoriale (paragraphes et les numéros de pages) ont également été balisés. C'est sous cette forme que les fichiers sont distribués[1]. Une première application de consultation a été développée et est accessible sur le site du CTHS[2].

En 2018, le projet a été redéfini de manière à tirer parti du liage des données avec d'autres référentiels pour enrichir la base et rendre possible la géolocalisation

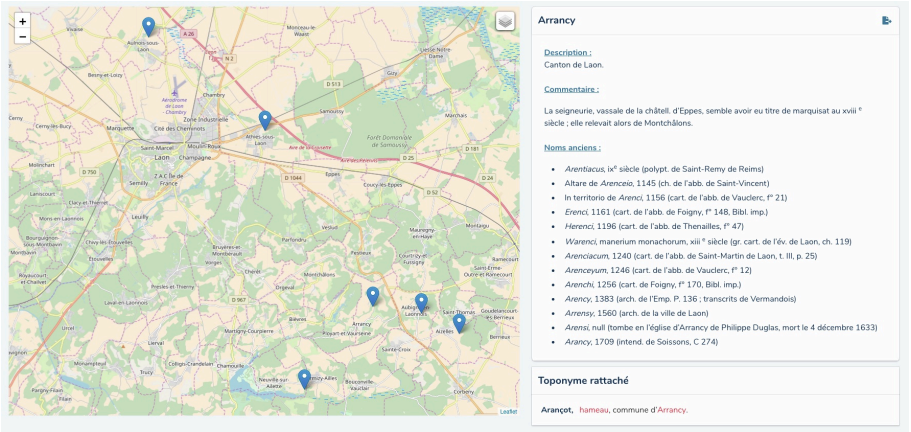
[1] https://github.com/chartes/dico-topo

[2] http://cths.fr/dico-topo/

des toponymes historiques. Une première campagne d'enrichissement a été menée avec l'aide de C. Burette, étudiante du master TNAH de l'École des chartes : 15 852 entrées du *Dictionnaire* correspondant à des communes (toponymes de type *commune*) et 283 877 entrées rattachées (des toponymes localisés dans une commune) ont été liées au référentiel de l'INSEE et ont pu être géolocalisées. Autrement dit, 75 % des noms de lieux contenus dans la base sont localisés au niveau de leur commune de rattachement. Pour les 25 % restants, un important effort d'annotation est à prévoir pour affiner leur localisation dont la granularité, du fait du découpage éditorial des dictionnaires (un tome par département), est a minima départementale.

4. Une API et une application Web

L'initiative lancée en 2009 de numérisation et de partage des données s'inscrivait dans le mouvement de l'*Open Data* : la volonté était de partager les ressources avec la communauté scientifique. La prise de conscience récente que l'enrichissement des données est conditionné à leur liage à d'autres jeux de données (INSEE, IGN, etc.) a redéfini considérablement le projet d'édition numérique initial. Il ne s'agit pas seulement de donner à lire, mais de donner accès aux données selon une méthode normalisée et documentée. L'API définie est conforme à la spécification JSON API et permet d'accéder aux données relatives à chaque toponyme (localisation, formes anciennes, coordonnées géographiques, code INSEE, etc.). L'idée n'est pas seulement de partager les données, mais de permettre leur exploitation par des applications tierces – construire par exemple un service d'identification des toponymes anciens.



Une application de consultation[3] adossée à cette API tire parti du liage des données en offrant du rebond vers d'autres référentiels et en offrant, grâce au service géoportail de l'IGN, une carte des toponymes localisés. Sur le modèle de Pleiades, l'application permet à un utilisateur authentifié de corriger une entrée ou de saisir de nouvelles formes anciennes attestées des toponymes. Notre ambition à travers ces fonctionnalités est de relancer une initiative éditoriale centenaire et malheureusement interrompue depuis 2008 et d'envisager – à terme – une couverture plus exhaustive du territoire métropolitain.

5. Conclusion

La communication présentera donc un projet éditorial centenaire, une initiative numérique d'une décennie et un projet de *Linked Open Data* récent. La première version de l'application présentée est une véritable preuve de concept. Le travail à entreprendre reste considérable : améliorer le liage des données avec d'autres référentiels (notamment GeoNames, Pleiades, WHG) et surtout fédérer la communauté des chercheurs, historiens et archivistes qui pourra contribuer à l'enrichissement de cette base précieuse pour étudier la toponymie ancienne.

[3] https://github.com/chartes/dico-topo-app

Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle

Aicha SOUDANI[1,2] Aicha.soudani@hotmail.fr
Yosra MEHERZI[1,2] Yosra.meherzi@gmail.com
Asma BOUHAFS[1] Asma_bouhafs@yahoo.com
Francesca FRONTINI[3] Francesca.frontini@univ-montp3.fr
Carmen BRANDO[4] Carmen.brando@ehess.fr
Yoann Dupont[2] Yoa.dupont@gmail.com
Frédérique Mélanie-Becquet[2] Frederique.melanie@ens.fr

(1) *ECSTRA, IHEC, Université de Carthage*
(2) *Lattice UMR 8094, Université Paris 3-Sorbonne Nouvelle*
(3) *Praxiling UMR 5267, Université Paul-Valéry de Montpellier*
(4) *CRH UMR 8558 / Plateforme géomatique, EHESS*

Résumé

Dans cet article, nous proposons une chaîne de traitement reposant sur deux outils existants, l'un pour la reconnaissance des entités nommées, et l'autre pour la résolution des entités nommées. Par la suite, l'évaluation et l'adaptation de ces systèmes à l'analyse des textes issus de la littérature française du 19ème siècle sont présentés. Le résultat fourni par notre chaîne de traitement propose une visualisation projetant les entités nommées de type Lieu sur une carte et nous montrons enfin l'intérêt de ce travail pour les humanités numériques.

Abstract

In this article, we propose a processing pipeline relying on two existing tools, one for named-entity recognition, and the other for named-entity linking. We first present the evaluation and the adaptation of these systems to the analysis of texts from 19th Century French literature. We then show the result provided by the pipeline, namely a projection of the place entities onto a map. Finally, we discuss the interest of this work for the digital humanities.

Mots-clés : reconnaissance des entités nommées, résolution des entités nommées, cartographie.
Keywords : entity recognition, named entity linking, cartography.

Introduction

Les outils du traitement automatique du langage naturel (TAL) occupent une place importante dans le spectre des humanités numériques contribuant notamment à l'analyse d'œuvres littéraires numérisés sous l'angle des entités nommées (EN) [Lecluze et al, 2014]. Dans ce papier, nous nous intéressons particulièrement à la reconnaissance des entités nommés (REN) et la résolution des entités nommées (NEL), tâches répandues en TAL, afin d'enrichir des textes issus du canon littéraire français du 19ème siècle. D'une part, la REN consiste à identifier et catégoriser des expressions linguistiques comme les noms de personne, de lieu, et d'institution, d'autre part, la NEL vise à déterminer l'identité des entités, mentionnées dans le texte, à partir d'une base de connaissances (BC) telle que DBpedia, Wikidata, Yago, Geonames.

Dans ce travail, nous proposons une chaîne de traitement reposant sur deux outils REN et NEL qui seront adaptés à l'analyse de textes de la littérature française. Nous avons constitué un *gold standard* composé de deux chapitres du roman « Le ventre de Paris » d'Emile Zola et le premier chapitre du roman « César Birotteau » d'Honoré de Balzac. Nous avons également développé une application en ligne afin de proposer un rendu dynamique projetant les lieux repérés dans les textes sur une carte. Le restant de ce papier se décompose en trois parties. Les deux premières parties se consacrent à la présentation des outils REN et NEL sélectionnés pour notre

étude, et la dernière partie présente une conclusion évoquant l'intérêt de ce travail pour des projets en humanités numériques en cours.

1. Chaîne de traitement proposée

La chaîne de traitement proposée est illustrée en Figure 1, elle intègre le système « SEM » [Dupont 2017] pour la tâche REN et « REDEN » [Frontini et al 2015 ; Brando et al 2016] pour la tâche NEL. Le choix du format XML/TEI, incontournable pour l'édition numérique de textes, est le choix d'encodage fait pour l'entrée de notre chaîne. Il est donc indispensable que les deux outils supportent ce format. Pour REDEN, la question ne se pose pas. Par contre, il a fallu adapter SEM afin de donner support au format XML/TEI.

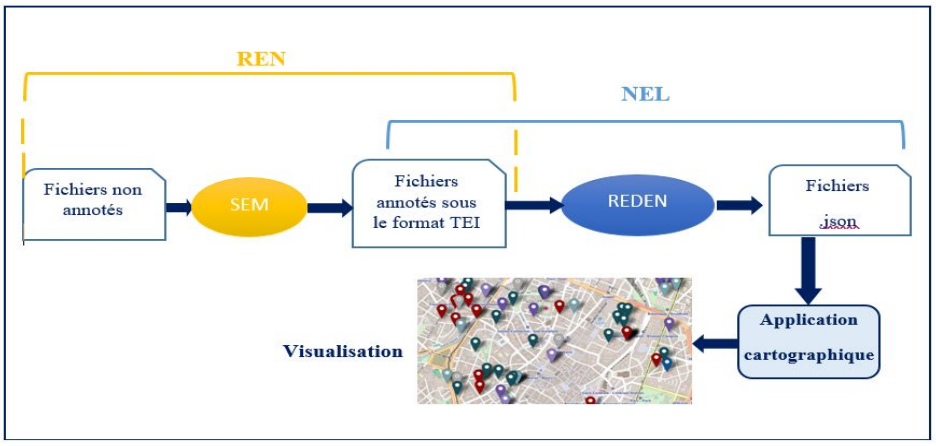


Figure 1. Chaîne de traitement proposée

1.1. Adaptation et évaluation du système SEM

SEM est un système de REN qui s'appuie sur une approche d'apprentissage supervisée [Raymond et al., 2010]. L'apprentissage du système repose sur l'entraînement d'un modèle à partir d'exemples afin de reproduire une tâche de prédiction. Un modèle pour SEM est disponible, entraîné à partir de textes journalistiques du French TreeBank [Sagot et al., 2012]. Pour ce travail, nous avons entraîné un modèle REN pour SEM à partir des textes littéraires français que nous avons manuellement annotés (le *gold standard*) par deux annotateurs distincts.

L'accord inter-annotateur donne les valeurs suivantes: 0,88 pour le rappel, 0,96 pour la précision et 0,91 pour la F-mesure. La proximité des valeurs de l'accord inter-annotateur, tous proches de 1, déduit la similarité de l'annotation produite par les deux experts.

Nous avons évalué les performances de ce modèle en termes de rappel, de précision et de F-mesure, ainsi qu'analysé les erreurs d'annotation récurrentes. Les expérimentations sur SEM concernent deux types d'entités nommées, à savoir personnes et lieux, et sont effectués sur le *gold standard.* Elles se divisent en trois grandes parties, décrites ci-dessous.

(1) Évaluation de SEM avec le modèle French Tree Bank :

Cette expérience montre que le modèle entraîné sur des textes journalistiques contemporains n'est pas suffisamment portable sur des textes littéraires (cf. tab 1). En particulier, les EN de type « Person » pose plus de problèmes avec des résultats de F-mesure variables qui peuvent atteindre un minimum de 10%. Cela vient du fait que les noms de personnes correspondent à des noms fictifs. Cependant, pour la reconnaissance des lieux, les résultats sont meilleurs avec des valeurs de F-mesure qui varient entre 24% et 33%. Ceci est dû au fait que les noms des lieux représentent des lieux réels existants et donc connus et appris par le modèle et présents dans le dictionnaire. Le tableau 1 suivant montre les résultats des expériences.

Tableau 1. Résultats expérimentations SEM

	SEM modèle FrenchTreeBank : Zola		SEM modèle FrenchTreeBank : Balzac		Adaptation 1	Adaptation 2
	Location	Person	Location	Person		
Précision globale	0,46	0,2	0,25	0,14	1	0,7
Rappel global	0,25	0,08	0,28	0,08	0,69	0,26
F-mesure globale	0,33	0,12	0,26	0,10	0,82	0,38

Trois types d'erreurs ont été remarquées : (1) erreur de type d'EN, (2) annotation partielle, (3) absence d'annotation. En particulier, la non reconnaissance de déclencheurs de lieux était à l'origine d'une partie des erreurs. Pour cette raison, l'adaptation au domaine a nécessité de mettre au point un dictionnaire de déclencheurs de lieux (boulevard, rue ...) afin d'améliorer la phase de réentraînement.

(1) Adaptation au domaine :

Adaptation 1 : La première adaptation consiste à effectuer un entraînement sur un extrait du roman « Le ventre de Paris » et une évaluation sur une autre partie du même roman. Les résultats de cette expérience nous montrent qu'un modèle entraîné et testé sur le même auteur, pour notre cas Zola, donne des résultats assez bons avec une précision globale de 1, un rappel global qui atteint 0,69 et une F-mesure entre de 0,82 (voir tab 1).

Ces résultats s'expliquent par le fait qu'un chapitre est un domaine de référence très restreint, dans lequel les mêmes noms de personne et de lieu ont tendance à se répéter. Donc entraîner sur une partie du chapitre produit de bons résultats d'annotation sur l'autre.

Adaptation 2 : C'est un entraînement sur un extrait du roman « Le ventre de Paris » (Zola) et une évaluation sur un roman d'un auteur différent, ici « César Birotteau » (Balzac).

Le même modèle d'apprentissage issu du roman de Zola (Adaptation 1) et testé sur celui de Balzac, donne des mesures assez faibles. Une précision de 0,7, un rappel de 0,26 et une F-mesure de 0,38. Ces résultats se justifient par le fait que le modèle est moins portable d'un auteur à l'autre.

(2) Progression :

Dans cette expérience le corpus d'entraînement composé d'extraits de Zola est progressivement augmenté afin de trouver la taille optimale de l'échantillon d'apprentissage. L'échantillon d'apprentissage est composé respectivement d'1/3 , 2/3 et 3/3 d'un chapitre et les modèles sont testés sur deux extraits différents de Zola et de Balzac.

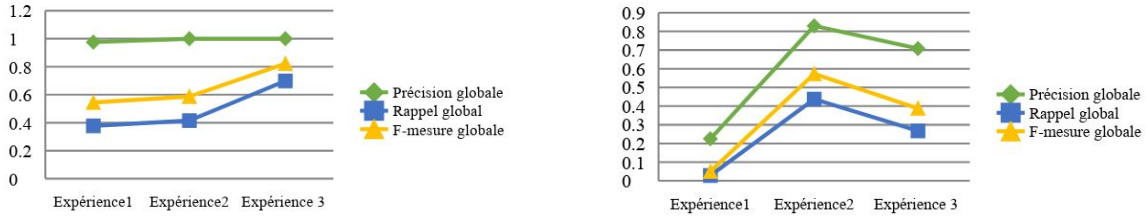


Figure 2. Progressions respectives pour Zola et Balzac

Pour le test sur le texte de Zola, en augmentant la taille du corpus d'entraînement la performance augmente progressivement. Cependant en effectuant un entraînement avec un texte de Zola et le testant sur le texte de Balzac, les résultats s'améliorent d'abord mais baissent quand le corpus d'entraînement est composé de 3 extraits. Ceci peut s'expliquer par le fait que le corpus d'entraînement est devenu trop adapté à Zola ce qui pourrait entraîner un problème de sur-apprentissage et manque de généralisation [Douglas M., 2004].

1.2. Adaptation et évaluation du système REDEN

REDEN [Frontini et al, 2015, Brando et al, 2016], est un outil de NEL fondé sur la théorie de graphes permettant la résolution d'EN s'appuyant sur des sources du Web de données. Cet outil prend en entrée un fichier XML/TEI tagué en entités nommées et produit en sortie le même fichier enrichi avec un identifiant, un IRI (acronyme pour

Internationalized Resource Identifier), pour chaque entité et un NIL pour les entités sans référent. Nous nous sommes intéressés uniquement aux lieux et avons adapté trois BC (DBpedia, Bnf, WikiData) à partir de requêtes SPARQL personnalisées. Chaque requête est relative à une seule base de connaissance et produit un dictionnaire qui est utilisé par REDEN pour rechercher les candidats. Nous avons enfin évalué REDEN à partir de textes annotés par SEM et pour chaque BC. Les métriques d'évaluation dépendent de la phase NEL concernée, à savoir (1) Recherche des candidats, (2) Sélection du bon candidat. Ces métriques étendent celles du REN (voir [Brando et al, 2016] pour les détails). Les résultats obtenus lors des expérimentations sont présentés dans le tableau 2.

Tableau 2. Résultats des expérimentations REDEN

Les mesures / BC	Candidate Precision	Candidate recall	NIL precision	NIL recall	Disambiguation accuracy	Taux D'ambigüité	Overall linking accuracy
DBpedia	1	0,816	0,367	1	none	0	0,834
BNF	0,760	0,630	0,580	0,972	1	0,005	0,7
Wikidata	0,912	0,830	0,440	1	1	0.29	0,85

Les métriques de la phase (1) nous permettent de déterminer son efficacité.
La « Candidate Precison », est de 1 pour la BC DBpedia, 0,912 pour Wikidata et 0,76 pour la Bnf. Donc en général, REDEN est capable de trouver l'IRI approprié dans la BC parmi un ensemble de candidats non vides.
En outre, pour le « Candidate recall » nous remarquons que les résultats avec DBpedia (0,816) et Wikidata (0,83) sont supérieurs aux résultats obtenus avec Bnf (0,63). Ceci se traduit par le fait que REDEN trouve plus de référents corrects avec DBpedia et Wikidata comparé à BnF. De plus, puisque nous nous intéressons aux toponymes dans la littérature issue du 19éme siècle, certains lieux ont changé de nom, par exemple « le fort de Bicêtre » existe dans la BNF sous le nom de « château de Bicêtre ». Aussi, les mentions sont des villes fictives comme « Plassans » qui existe dans DBpedia et Wikidata mais pas dans BNF.

Pour, évaluer la capacité de l'algorithme à produire des annotations NIL correctes pour les mentions n'ayant pas de réfèrent dans le *gold*. Nous observons que le résultat du « NIL Precision » est légèrement meilleur avec la base BNF 0,580 comparé à Wikidata 0,44 et DBpedia 0,367. Cependant, les résultats restent faibles, ceci s'explique par le fait que REDEN traite la phase de recherche des candidats avec l'algorithme des mesures de correspondance parfaites entre chaînes de caractères (*exact string match*).
D'autre part, le « NIL Recall » est élevé pour les trois BC, 1 pour DBpedia et Wikidata, 0,972 pour BnF. Ce qui se traduit par le fait, que comparé au NIL dans le *gold*, REDEN retourne des NILs corrects.
Quant aux mesures de la phase (2), le résultat obtenu pour la « Désambiguïsation accuracy » pour DBpedia est vide, comparé à la Bnf et Wikidata 1. Il est important de noter que cette mesure est intéressante quand on a des ensembles de candidats de taille supérieure à 1. Le taux d'ambiguïté est nul pour DBpedia, de 0,05 pour la Bnf et de 0,29 pour Wikidata. Ce qui signifie que la moyenne des ensembles de candidats ayant plus de 2 candidats est faible pour DBpedia ainsi que Bnf et légèrement plus importante avec Wikidata.
Enfin, la mesure de « Overall Linking » obtenue est de 0,834 pour DBpedia, 0,85 pour Wikidata et 0,7 pour BnF. REDEN a donc été efficace. Cette mesure essaie d'évaluer l'efficacité globale du système, et non par phase, et exprime la fiabilité de REDEN pour la tâche de résolution des entités nommées. Nous pouvons ainsi conclure que les trois bases sont efficaces pour la tâche de NEL, et donnent de résultats corrects. Cependant, Wikidata est légèrement meilleur que DBpedia et BnF.

Conclusion

Inspirés par le projet *Venice Time Machine,* développé à l'Ecole Polytechnique Fédérale de Lausanne par [Di Leonardo., 2015], notre travail s'avérera utile pour la création d'un volet littéraire d'un *Paris Time Machine*. En effet, ces projets ambitieux visent à la reconstruction du passé des grandes villes d'Europe à travers l'extraction d'informations à partir des documents historiques, y compris littéraires. Notre travail est très proche des travaux de [Boeglin et al., 2016], qui proposent une méthode pour localiser, cartographier et analyser les occurrences de lieux citées dans un corpus de 31 romans du 19ème siècle dont l'action se situe pour tout ou partie à Paris. Notre contribution est particulièrement d'avoir produit un modèle pour la reconnaissance d'EN dans les textes littéraires français ainsi que des BC du Web de données pertinentes pour la tâche NEL dans ce même contexte.

Afin d'illustrer nos propos, les figures 3 et 4, proposent deux rendus cartographiques possibles à partir de la chaîne de traitement proposée. Etant donné que les BC utilisés répertorient les coordonnées de localisation pour chaque entité de type lieu, il est donc possible, après la phase NEL, de rapatrier cette information (et d'autres) automatiquement par le biais des IRI. La première vue montre les lieux qui ont été mentionnés dans les romans et les marque avec un indicateur sur la carte. La deuxième vue nous permet de visualiser sur une carte pour chaque mention de lieu repérée dans le texte le nombre d'occurrences de cette EN dans les romans.

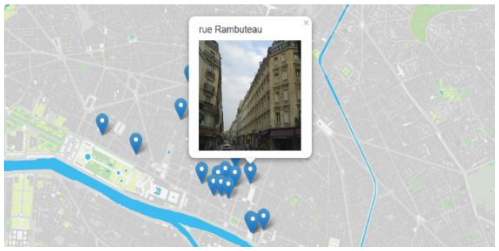


Figure 3. Vue numéro 1 de la cartographie

Figure 4. Vue numéro 2 de la cartographie

Références

Boeglin et al., 2016 : Boeglin N., Depeyre M., Joliveau T., Le Lay.F . Pour une cartographie romanesque de Paris au XIXe siècle. Proposition méthodologique. Conférence Spatial Analysis and GEOmatics, Actes de la conférence SAGEO'2016 - Spatial Analysis and GEOmatics Dec 2016, Nice, France, 2016.

Brando et al., 2015 : Brando.C, Frontini F., Ganascia J.: "Linked Data for toponym linking in French Literary texts", in Proceedings of the 9th Workshop on Geographic Information Retrieval; 2015.

Brando et al., 2016 : Brando C., Abadie N., Frontini F. : « Evaluation de la qualité des sources du Web de Données pour la résolution d'entités nommées ». Revue Ingénierie des Systèmes d'Informations, 2016.

Di Leonardo et al., 2015 : Di Lenardo, Isabella, Kaplan, Frédéric. Venice Time Machine : Recreating the density of the past, Digital Humanities 2015, Sydney, June 29 - July 3, 2015.

Douglas et al., 2014 : Douglas M. Hawkins,The Problem of Overfitting,School of Statistics, University of Minnesota, Minneapolis, Minnesota, 2014.

Dupont 2017 : Dupont Yoann. Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique. RECITAL, 2017, p. 42.

Frontini et al., 2015 : Francesca F., Brando C., and Ganascia J. : "Domain adapted named-entity linker using Linked Data » ; in Workshop on NLP Applications : Completing the Puzzle co-located with the 20th International Conference on Applications of Natural Language to Information Systems ;2015.

Lecluze et al., 2014 : Lecluze C., et Lejeune G., « DEFT2014, analyse automatique de textes littéraires et scientifiques en langue française » 21ème Traitement Automatique des Langues Naturelles, Marseille, 2014.

Raymond et al., 2010 : Raymond C., et Fayolle J., . Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. Dans Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10), Montréal, Canada, 2010.

Sagot et al., 2012 : Sagot B., Richard M., Stern R. . Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. Antoniadis G., Blanchon H., Sérasset G., . Traitement Automatique des Langues Naturelles (TALN), Jun 2012, Grenoble, France. 2 - TALN, 2012, Actes de la conférence conjointe JEP-TALN-RECITAL 2012. <hal-00703108>

Oronce Fine : une plateforme pour l'annotation sémantique de ressources cartographiques sur le Web de données*

Pandolfi Benoît[1], Brando Carmen[1], Mermet Eric[1], Verdier Nicolas[1], Léa Hermenault[1]

1. Plateforme Géomatique, EHESS
54 boulevard Raspail 75006 Paris, France
{prenom.nom}@ehess.fr

Résumé. Ce texte présente la plateforme collaborative en ligne Oronce Fine. Cette plateforme conforme aux technologies et standards du Web sémantique a pour objectif d'être le lieu d'accueil des données des projets scientifiques issus des disciplines, terrains et temporalités divers en sciences humaines et sociales. A travers cet outil le projet Oronce Fine se propose principalement de faciliter l'intégration et le partage de documents anciens intégrant de multiples dimensions spatio-temporelles (cartes et plans anciens, données géographiques format vecteur, images satellite, photographies, textes) et enrichis par des vocabulaires du Web de données. Elle a enfin pour objectif d'être un outil de travail collaboratif pour entre autres des historiens et des archéologues, ainsi que d'être consultable par un large public.

Abstract. This proposal presents the Oronce Fine online collaborative platform. This platform complies with Semantic Web technologies and standards and aims to accommodate data for scientific projects on various disciplines, fields and time horizons in the humanities and social sciences. Through this tool, we intend to facilitate the integration and sharing of historical documents integrating multiple spatial and temporal dimensions: old maps and plans, geographical vector data, satellite images, photographs, texts, semantically enriched with vocabularies of the Web of data. Finally, it aims to be a collaborative working tool for historians and archaeologists, among others, as well as to be accessible to a wide audience.

Mots-clés : annotation sémantique, web de données, données anciennes.
Keywords : semantic annotation, web of data, geo-historical data.

.

[1]* Avec le soutien financier de l'EHESS et du projet PSL Oronce-Fine, Semantic-enabled platform for the publication, integration and exploration of geo-historical resources-Oronce Fine

1. Introduction

La production et la valorisation de données de la recherche est un enjeu pour la reproductibilité et la visibilité des travaux en sciences humaines et sociales (SHS). Dans ce contexte, notre projet vise à constituer et déployer une plateforme collaborative en ligne, baptisée Oronce Fine, construite conformément aux technologies et standards du Web sémantique afin de faciliter l'intégration et le partage de documents anciens hétérogènes intégrant des dimensions spatio-temporelles multiples.

La plateforme Oronce Fine sera le lieu d'hébergement des données des projets scientifiques sur disciplines, terrains et temporalités divers, il est d'ores et déjà un outil de travail collaboratif pour des historiens du Centre de recherche historique (CRH UMR 8558 CNRS/EHESS), ainsi que pour des archéologues rattachés à la Maison de l'Archéologie et de l'Ethnologie (Nanterre). A ce stade, nous nous concentrons particulièrement sur trois cas d'étude : les itinéraires et lieux habités du Chemin de Saint-Jacques dans le nord de l'Espagne, les paysages de l'agriculture d'hier et d'aujourd'hui en Île-de-France, et les études africaines et l'histoire de la "Sémiologie graphique" vues au travers des fonds du Laboratoire de graphique dirigé par Jacques Bertin.

La valeur ajoutée de la plateforme Oronce Fine est de fournir un service de dépôt de ressources cartographiques géoréférencées, de description à partir des métadonnées normalisées (Dublin Core, INSPIRE) et d'annotation sémantique compatible avec le modèle Web Annotation Data Model (W3C). La description et l'annotation s'appuient sur les vocabulaires du Web des données et relient automatiquement à des sources et référentiels externes de données spatialisées sémantisées (Geonames, Getty TGN). La présentation décrira l'entreprise de conception et de développement de notre prototype qui s'appuie sur l'outil ouvert de gestion et de publication de contenus, Omeka-S[2].

2. Outil de gestion de contenus patrimoniaux

Le choix a été fait de ne pas développer un système entier, mais d'utiliser des logiciels existants et ouverts. Ces logiciels répondent chacun à une partie des besoins des différents cas d'étude. Les différentes fonctionnalités sont reliées dans un système centralisé, ici Omeka S. Celui-ci est une application inspirée de *Content Management Systems* et il est développé en code ouvert par le Centre Roy Rosenzweig de l'Université George Mason (Etats-Unis). Omeka-S centralise les données géoréférencées y compris les métadonnées, les données sont également

[2] https://omeka.org/s/

diffusées en flux WMS (*Web map service*) grâce à un serveur cartographique comme GeoServer[3] et via le protocole OAI-PMH[4], très répandu dans le monde des bibliothèques numériques. Une seule interface administrateur permet la publication de sources, enrichies sémantiquement par l'utilisateur, et rend possible la gestion de plusieurs sites. Cette fonctionnalité convenait parfaitement au type de plateforme envisagé dans le projet.

3. Géoréférencement et intégration

Les sources cartographiques suivent un processus prédéfini de géoréférencement et d'intégration à Omeka S. Ce processus est présenté en figure 1. Ces sources regroupent des documents se présentant sous diverses formes : plans, cartes, images, diapositives, articles, cartes postales, témoignages, etc. Le processus de géoréférencement et d'intégration se déroule en trois phases. Durant la première phase les documents sont numérisés et mis en forme en mode hors ligne par les équipes de recherche, puis envoyés sur un serveur externe d'hébergement des données. Lors de la seconde phase, les données sont normalisées et intégrées à Omeka S ; un identifiant est attribué par document au sein du système, de type ARK [5] dans les meilleurs des cas. La troisième et dernière phase consiste à employer le module d'annotation sémantique d'Omeka S.

Dans le détail, les documents sont, dans un premier temps, numérisés localement. Dans le cadre des trois cas d'études précédemment cités, il s'agit de plus de 250 cartes des fonds de l'ancien Laboratoire de graphique de l'EHESS qui ont été numérisés et plus de 3000 plans, cartes et diapositives du fonds de J. Passini (Passini, 1984), dans le cadre du projet Chemin de Saint Jacques. L'ensemble des documents sont alors décrits par leurs métadonnées dans un tableur préformaté. Celui-ci contient des champs correspondant à la description normée des documents et de leur contenu. Les données pouvant être géoréférencées sont traitées sous le logiciel QGis[6], sur des fonds disponibles sur le Web comme OpenStreetMap en WGS84. Une fois géoréférencée, l'emprise des contenus est extraite par vectorisation. Le vecteur ainsi obtenu permet de découper les images géoréférencées pour ne garder que le contenu spatialisé. Cette emprise est aussi stockée en WKT (*well-known text*)[7] pour servir plus tard au processus d'annotation spatiale sous Omeka S.

Toutes les données sont ensuite transférées sur le serveur et font l'objet d'un retraitement. Les données-images sont stockées à la fois sous forme brute, afin d'être disponible au téléchargement, et sous forme normalisée afin d'être intégrées

[3] http://geoserver.org
[4] Acronyme anglais pour *Open Archives Initiative Protocol for Metadata Harvesting*
[5] http://www.bnf.fr/fr/professionals/issn_isbn_other_identifiers/a.ark_en.html
[6] https://www.qgis.org
[7] Il s'agit d'une représentation textuelle de la géométrie d'une entité spatiale

aux sites diffusés par Omeka S. Les images géoréférencées sont, de leur côté, tuilées pour ensuite être intégrées à un serveur Geoserver et diffusées en WMS (*Web Map Service*). Le tableur contenant les métadonnées subit, lui aussi, des modifications car les champs préformatés sont mappés sur des ontologies choisies : Dublin Core, Bibliographic Ontology et INSPIRE pour les cas d'études précédemment cités. Le tableur est ensuite transformé en CSV et stocké sur le serveur.

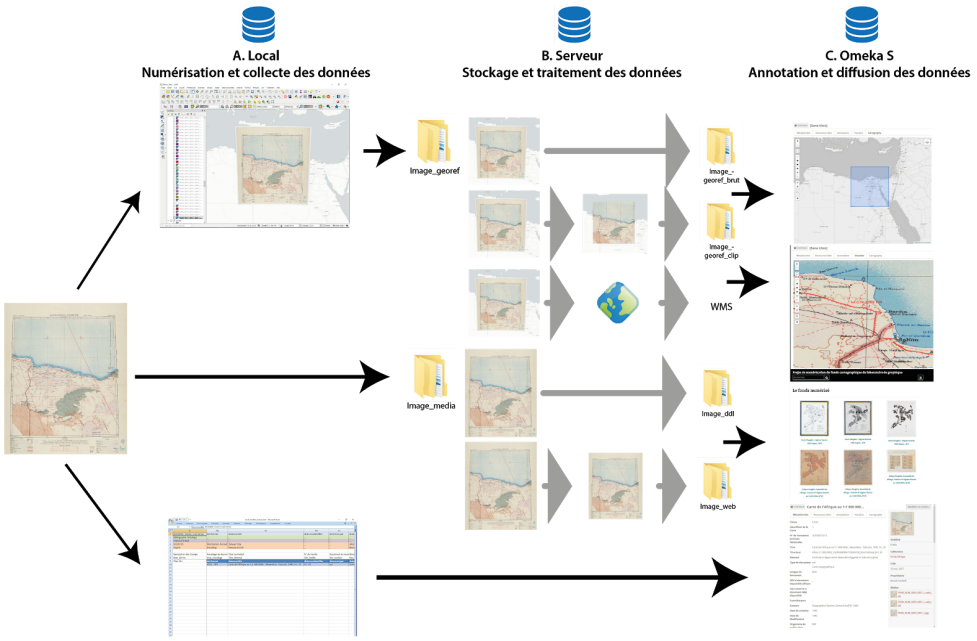


FIGURE *1. Processus de géoréférencement et d'intégration des sources à la plateforme Oronce Fine*

À l'issue de cette phase de traitement ces éléments sont intégrés à Omeka S. Les métadonnées du CSV deviennent des items décrits par des métadonnées. Les images retraitées sont associées à ces items sous forme de media, et les images brutes rendues disponibles par des liens de téléchargement. Sous Omeka S grâce à deux modules développés dans le cadre du projet Oronce Fine il est désormais possible d'annoter à la fois le contenu des images (media) lui-même et d'annoter spatialement les données. Pour finir Omeka S, diffuse sous forme de Sites web les items, les medias, les annotations et les WMS dans les sites propres à chaque cas d'étude. Les métadonnées sont aussi diffusées selon le protocole OAI-PMH.

4. Module d'annotation sémantique

Un module d'annotation sémantique a été développé sous forme de plugin Omeka S[8] par D. Berthereau. Cette implémentation s'appuie sur les recommandations de la W3C qui conseille l'utilisation du *Web annotation Data model* (WADM) ainsi que l'ontologie associée, le *Web Annotation Vocabulary* (WAC). L'annotation est plus précisément définie comme une ressource composée d'un *body* qui correspond au commentaire ou à la ressource descriptive ainsi que d'une *target* qui constitue la ressource concernée par ce qui est décrit dans le *body* de l'annotation. L'annotation peut également être caractérisée par des propriétés additionnelles comme un type parmi celles répertoriées dans la WAC et un identifiant IRI (*Internationalized resource identifier*). En ce qui concerne la dimension spatiale, la *target* de l'annotation fait souvent référence à une géométrie exprimée en WKT afin de démarquer la zone concernée par l'annotation. Le *body* d'une annotation peut aussi pointer à une IRI d'un gazetier sémantisé externe comme Géonames ou bien interne au sien d'Omeka S, afin d'identifier le lieu dont il est question dans l'annotation. La figure 2 illustre deux annotations sur le plan du cadastre d'Estella, qui fait partie du fonds Chemin de St Jacques sur l'interface d'annotation que nous proposons.

FIGURE *2.* Deux annotations sur le Plan du cadastre de Estella (Chemin de St. Jacques), l'une porte sur un extrait du Chemin de St Jacques (en bleu) et l'autre désigne l'entité spatiale Église du Saint Sépulcre (en violet).

Un extrait de ces mêmes annotations sérialisées en JSON-LD (oa constitue le préfixe de la WAC) sont présentées ci-dessous.

***motivation**: "linking",*
***body**: [{ type: "o:Item",*
value: "http://psig.huma-num.fr/omeka-s/api/items/634145"} …] , ← Église du Saint Sépulcre
***target**: { … type: "oa:Selector", format: "application/wkt", value: "POLYGON*

[8] https://github.com/Daniel-KM/Omeka-S-module-Annotate et https://github.com/Daniel-KM/Omeka-S-module-Cartography

((1746.719783 3882, 1744.71958 3788, 1952.740654 3776, 1958.741262 3870, 1746.719783 3882))" }}}

Cet ensemble de spécifications décrit un modèle structuré permettant aux annotations d'être partagées et utilisées entre applications via un API REST (*representational state transfer*). L'avantage de l'outil Omeka S est la possibilité d'ajouter simplement une ontologie quelconque si l'on souhaite étendre le modèle d'annotation avec des nouveaux champs de description. Dans d'autres systèmes d'annotation de cartes comme Recogito[9], la géométrie des entités n'est pas une information qui appartient à l'annotation. Au contraire, ce choix est fait dans un outil d'annotation (non sémantique) et de géoréférencement de cartes comme Maphub[10]. Pour Recogito, c'est le gazetier (ou référentiel) qui fournit l'accès aux descriptions des entités géographiques et permet de tracer leur évolution dans le temps, à condition néanmoins que le gazetier pertinent existe. Ceci représente un choix intéressant qui mérite d'être discuté largement avec la communauté en humanités spatiales.

Conclusion

La plateforme Oronce Fine est une plateforme capable de fournir un service de dépôt de données. Son originalité tient en ce que ces dernières sont décrites à partir de métadonnées normalisées et qu'elles peuvent être annotées grâce à un module développé dans le cadre de ce projet. La démonstration du fonctionnement de ce nouvel outil permet de réaliser une première évaluation critique des fonctionnalités de la plateforme et de quelle manière elle répond concrètement aux besoins de production et de valorisation de données de la recherche en SHS.

Bibliographie

Grassi M., Morbidoni C., Nucci M., Fonda S., Piazza F., « Pundit: augmenting web contents with semantics ». Literary Linguist Computing, 28 (4): 640-659. doi: 10.1093/llc/fqt060, 2013.

Passini J., 1984, Villes médiévales du chemin de Saint-Jacques-de-Compostelle (de Pampelune à Burgos) Villes de fondation et villes d'origine romaine. Éditions Recherche sur les Civilisations, "Mémoire" n° 47, 183 p.

Simon R., Barker E., Isaksen L., De Soto Canamares P., « Linked Data Annotation Without the Pointy Brackets: Introducing Recogito 2 », Journal of Map & Geography Libraries Vol. 13 , Issue 1, 2017

[9] https://recogito.pelagios.org/
[10] http://maphub.github.io/

Vers une semi-automatisation du processus d'intégration de plan cadastral ancien dans une base de données multi-dates

Jean-Michel Follin[1], Elisabeth Simonetto[1]

1. Laboratoire GeF, Ecole Supérieure des Géomètres et Topographes, le Cnam
1 boulevard Pythagore, 72000 Le Mans, France
prenom.nom@lecnam.net

RESUME. Les recherches en géomatique s'intéressant au patrimoine géohistorique connaissent un essor depuis plusieurs années comme en attestent les nombreux projets et les conférences et groupes de travail dédiés. Elles se nourrissent de la disponibilité d'un nombre croissant de documents anciens géographiques. Le cadastre napoléonien, diffusé par les archives départementales, fournit en particulier la description la plus détaillée du territoire français dans sa globalité durant le 19ème siècle. Grâce à cette donnée l'évolution du parcellaire cadastral sur une période de deux siècles peut être étudiée en lien avec les transformations affectant le territoire. Dans cette perspective les travaux présentés ici portent sur une chaîne de traitement semi-automatique permettant de vectoriser les planches scannées de cadastre ancien, de les géoréférencer, de les assembler puis de les intégrer dans une base de données multi-époques. Dans cet atelier, nous abordons les solutions adoptées plus particulièrement pour les étapes de vectorisation et de mosaïquage avec une application sur les données de trois époques (1813, 1850 et 1972-74) d'une commune rurale du sud de la Sarthe pour laquelle les premiers résultats obtenus seront présentés.

ABSTRACT. Studies of geo-historical heritage have recently picked up momentum, as evidenced by the many projects, conferences and working groups dedicated to it. They thrive on an increased availability of ancient geographical documents. For instance, in France, the Napeoleonic cadastre, archived in the "departemental" libraries, provides the most accurate representation of France's whole territory during the 19th century. It is thus now possible to study two centuries of cadastral parcels evolution driven by territorial transformations. This article details a semi-automatic processing chain tailored to vectorize the old scanned cadaster sheets, geo-reference them, assemble them, and finally feed them into a historical database. We focus especially here on the two steps of vectorization and tiling, using a three date dataset (1813, 1850, 1972-74) covering one rural municipality located in the south of the Sarthe "departement". First results will be presented and discussed.

MOTS-CLÉS : Cadastre ancien, vectorisation, mosaïquage de données vectorielles

KEYWORDS: Old cadastre, vectorization, tiling of vector data

1. Introduction

Actuellement, de plus en plus de documents anciens, en particulier géographiques (cartes, plans, clichés aériens), sont dématérialisés et mis à disposition sous format numérique. Ils constituent une source d'information précieuse sur le territoire, notamment pour les chercheurs en sciences humaines. Dans de nombreux départements français, les services des archives diffusent ainsi leurs fonds documentaires numérisés en ligne dont notamment les plans cadastraux anciens scannés. Or, selon la période à laquelle ils ont été dressés, ces derniers contiennent des informations très variées, et qui sont plus ou moins détaillées ou précises (Clergeot P. et Berteau G., 2008).

Afin d'étudier l'évolution du territoire à travers les changements opérant au cours du temps dans la forme de son parcellaire cadastral, nous proposons une chaîne méthodologique reproductible, semi-automatique et basée sur des outils libres comprenant une première étape de préparation des données (vectorisation, géoréférencement et mosaïquage de feuilles cadastrales anciennes) et une deuxième étape d'intégration des informations dans une base de données spatio-temporelle modélisant l'évolution des parcelles au cours du temps qui sera utilisée pour mener des analyses (quantification et qualification des changements, mise en relation avec d'autres facteurs, identification de motifs récurrents, ...)

Nous allons brosser un état de l'art succinct sur les mises en œuvre de bases de données cartographiques multi-dates avant de faire une présentation globale de notre démarche et des difficultés rencontrées.

2. État de l'art

Les études menées sur des données cartographiques anciennes se sont considérablement développées ces dernières années en France et dans le monde. Elles démontrent l'intérêt récent pour les bases de données géographiques historiques et leur utilisation à des fins analytiques ou prospectives. En France, nous pouvons notamment citer les projets de recherche GéoPeuple (Costes *et al.*, 2012), GeoHistoricalData (Cura *et al.*, 2018), ALPAGE (Noizet et Grosso, 2012), MODE RESPYR (Herrault *et al.*, 2013) et ModelSpace / Architerre (Le Couédic *et al.*, 2012). Le développement de chaînes de traitements semi-automatiques des plans anciens est un sujet très actif comme en témoignent par exemple les travaux de Iosifescu *et al.* (2016) et Arteaga (2013) qui emploient plusieurs outils (tels que GDAL et ImageMagick), nécessitent des paramétrages manuels lors du prétraitement et peuvent faire appel à une démarche collaborative dans l'étape de validation ("Building Inspector" de NYPL).

Ces différents projets ne s'intéressent pas nécessairement au même type de document et ne partagent pas notre objectif de semi-automatisation du processus complet depuis l'image jusqu'à la base de données historiques.

3. De la vectorisation au mosaïquage : propositions méthodologiques

3.1. La chaîne de traitement semi-automatique

La première étape est la vectorisation semi-automatique des images correspondant aux planches cadastrales anciennes. Deux méthodes ont ici été évaluées : l'une basée sur la transformée de Hough dite probabiliste, proposée dans la bibliothèque scikit-image en langage Python, et l'autre sur une méthode dite du « suivi de chemin » du logiciel GRASS. La deuxième étape est le géoréférencement dont l'objectif est de pouvoir rendre superposables, avec une qualité optimale, le parcellaire vectorisé ancien au parcellaire numérique récent. Elle est précédée par une procédure manuelle de sélection des points de liaison qui servent au calcul d'un modèle géométrique de transformation des coordonnées. En reprenant la méthodologie de (Herrault *et al.*, 2013) la fonction « ridge par noyau gaussien » s'est avérée être la meilleure parmi celles évaluées à l'aide de l'erreur moyenne quadratique pour les plans les plus anciens (Follin *et al.*, 2016). La troisième étape est le mosaïquage par traitement topologique des feuilles cadastrales vectorisées et géoréférencées qui a pour finalité de contrôler la cohérence topologique de l'assemblage. En effet en sortie du traitement de géoréférencement des défauts topologiques (chevauchement, trou) entre les parcelles voisines des différentes feuilles cadastrales sont constatés. La quatrième étape est la création de la base de données multi-dates qui permet de modéliser les changements d'état (division, fusion) au cours du temps et constitue le support des analyses menées sur l'évolution du parcellaire cadastral.

3.2. Application

Nous avons appliqué notre méthode sur des feuilles du cadastre de la commune d'Aubigné-Racan (Sarthe) couvrant trois époques : 1813, 1850 et 1972-74. Les données de 1813, et dans une moindre mesure celles de 1850, sont difficiles à traiter en raison du jaunissement du papier lié au temps et de nombreuses traces laissées par l'homme qui compliquent la vectorisation automatique. Les planches numérisées de 1972-74 se présentent sous forme d'images binaires en noir et blanc qu'il est plus aisé de traiter (figure 1). Les caractéristiques sont résumées dans la table 1.

FIGURE 1. *Extraits de feuilles cadastrales d'Aubigné-Racan (1813, 1951 et 1972) couvrant une zone d'environ 300 mètres sur 400 mètres*

TABLE *1. Propriétés des feuilles scannées d'Aubigné-Racan*

Année	1813	1850	1972-1974
Échelle	1/2500	1/2000	1/2000
Effectifs de planches	22	45	44
Pas du scanner	200 dpi	200 dpi	400 dpi
Taille d'un pixel dans l'image (µm)	127	127	63,5
Taille d'un pixel sur le terrain (cm)	31,8	25,4	12,7

Dans cet atelier, nous présenterons en détail les performances et les limites des méthodes proposées pour les étapes de vectorisation et de mosaïquage.

Concernant la vectorisation, les deux méthodes évaluées visent à extraire des segments de contour des parcelles qui seront converties en polygones dans un deuxième temps. Elles se basent sur des étapes identiques pour le prétraitement (conversion en niveaux de gris de valeur 0 ou 255, application d'un masque dessiné manuellement et détection de points de contour avec un seuillage par hystérésis à seuils adaptatifs localement suivi d'une suppression des objets isolés puis d'une squelettisation) et le post-traitement (simplification et nettoyage topologique des lignes puis transformation de celles-ci en objets surfaciques après un passage en mode raster pour procéder à un étiquetage des composantes connexes). Le seuillage par hystérésis à seuils adaptatifs localement qui hybride le seuillage simple par moyennes glissantes et celui par hystérésis permet d'obtenir un prétraitement robuste en limitant sa sensibilité aux variations de paramètres.

L'évaluation des deux méthodes - basée sur la transformée de Hough probabiliste (THP) et dite de « suivi de chemin » (SdC) - pour chaque époque a porté sur les résultats obtenus avant les étapes de transformation des polylignes en objets surfaciques (table 2). Elle a consisté à comparer les linéaires obtenus à ce stade avec des données de référence, saisies manuellement, par superposition.

TABLE *2. Comparaison pour trois planches et pour les deux méthodes des probabilités de détection et de fausses alarmes obtenues pour les données linéaires*

Méthode	**Probabilité**	**Planche de 1813**	**Planche de 1850**	**Planche de 1972**
THP	**Détection**	60%	93%	88%
	Fausses alarmes	50%	6%	13%
SdC	**Détection**	55%	82%	90%
	Fausses alarmes	59%	16%	10%

Les erreurs en sortie de la vectorisation, qui restent importantes pour les plans les plus anciens, nécessitant des interventions manuelles. Néanmoins le temps global de traitement (processus automatique et correction manuelle) est inférieur au temps nécessaire à une vectorisation entièrement manuelle du plan (table 3).

TABLE *3. Durée nécessaire pour une vectorisation semi-automatique (THP, SdC) et une vectorisation manuelle (M) calculée pour chacune des planches*

	Planche de 1813			**Planche de 1850**			**Planche de 1972**		
Méthode	**THP**	**SdC**	**M**	**THP**	**SdC**	**M**	**THP**	**SdC**	**M**
Temps total	34'	24'	1h15'	17'	12'	50'	13'	10'	30'

Les deux méthodes évaluées offrent donc des résultats de qualités sensiblement équivalentes mais la deuxième présente l'avantage d'être un peu plus rapide, de ne nécessiter aucun paramètre et est donc jugée plus autonome et simple d'utilisation.

L'étape du mosaïquage vise à corriger les erreurs observées entre des parcelles voisines appartenant à des planches différentes après les étapes de vectorisation et de géoréférencement. La feuille cadastrale décrit des espaces cadastrés (les parcelles sur lesquelles sont situés des bâtiments) et non cadastrés (la voirie, les cours d'eau), ces derniers permettant de délimiter les îlots (définis comme des regroupements de parcelles contiguës). Notre méthode de mosaïquage permet de réaliser deux corrections : suppression des chevauchements et comblement des trous entre les parcelles des différentes planches qui devraient être contiguës. Elle permet également de construire des polygones modélisant les espaces non cadastrés (ENC) qui doivent être distingués des trous qui correspondent à des défauts topologiques. Cette étape s'appuie sur des géotraitements tels que l'union, la fusion de polygones et le calcul d'enveloppe convexe par triangulation de Delaunay. Elle requiert un traitement des triangles superflus, coupant les concavités du contour.

Si elle permet de traiter en grande partie les problèmes rencontrés à l'issue de l'étape précédente, nous avons constaté quelques erreurs en sortie de ce processus (duplications de géométries ou création d'artefacts) qui sont liées à l'étape de vectorisation. Par ailleurs elle ne traite pas tous les problèmes topologiques que l'on pourrait rencontrer (par exemple des parcelles en contact alors qu'elles sont normalement séparées par un ENC).

4. Conclusion et perspectives

Nous avons conçu une chaîne méthodologique semi-automatique basée sur des outils open-source pour la vectorisation, le géoréférencement et le mosaïquage du cadastre ancien. Elle permet d'obtenir les entités pour construire une base de données modélisant les parcelles ainsi que les espaces non cadastrés. Notre méthode a été appliquée sur des planches scannées d'une commune rurale du sud de la Sarthe sur lesquelles de bons résultats ont été obtenus avec un temps d'intervention manuelle limité. Cependant il reste des points à améliorer qui peuvent concerner l'automatisation du processus ou la qualité des résultats. Nous pouvons citer notamment, pour la vectorisation semi-automatique, l'application d'une transformée de Hough non probabiliste. Plus globalement une modélisation par graphe d'adjacence granulaire qui permet de considérer différents niveaux détails pourrait

améliorer l'ensemble du processus : contrôler et corriger les résultats de la vectorisation ou guider les corrections à apporter lors du mosaïquage.

Remerciements

Nous tenons à remercier les anciens étudiants de l'esgt stagiaires du laboratoire GeF, Maïté Fahrasmane, Charlotte Odie et Jean-Marc Beveraggi qui ont fait avancer ce projet, les collègues Marie Fournier et Mathieu Bonnefond ainsi que Gilles Berteau de la DGFiP.

Bibliographie

Arteaga, M. G. (2013). Historical map polygon and feature extractor. In Proceedings of the 1st ACM SIGSPATIAL Int. Workshop on MapInteraction, pp. 66-71.

Clergeot P. et Berteau G. (2008). Du cadastre napoléonien au cadastre en ligne sur Internet : 1ère partie. Revue XYZ n° 119, pp 49-59.

Costes B., Grosso E. et Plumejeaud C. (2012). Géoréférencement et appariement de données issues des cartes de Cassini – Intégration dans un référentiel topographique actuel. Actes de SAGEO 2012, Liège, Belgique.

Cura, R., Dumenieu, B., Abadie, N., Costes, B., Perret, J., and Gribaudi, M. (2018). Historical collaborative geocoding. ISPRS Int. J. of Geo-Information, 7(7), pp. 262-290.

Follin J.-M., Fahrasmane M., Simonetto E., (2016). An open-source based toolchain for the georeferencing of old cadastral maps. Actes OGRS2016, Pérouse, Italie.

Herrault P-A., Sheeren D., Fauvel M., Monteil C., Paegelow M., (2013). A comparative study of geometric transformation models for the historical 'Map of France' registration. Geographia Technica n° 1, pp. 34 - 46.

Iosifescu I, Tsorlini A, Hurni L, (2016). Towards a comprehensive methodology for automatic vectorization of raster historical maps. e-Perimetron vol. 11, n°2, pp. 57-76.

Le Couédic M., Leturcq S., Rodier X., Hautefeuille F., Fieux E. et Jouve B. (2012). Du cadastre ancien au graphe. Les dynamiques spatiales fiscales et modernes. Revue ArchéoSciences 36. Presses universitaires de Rennes. pp 71-84.

Noizet H. et Grosso E. (2012). Mesurer la ville : Paris de l'actuel au Moyen âge. Les apports du système d'information géographique d'ALPAGE. Revue du comité français de cartographie, n°211, pp. 85-100