

Constitution d'une base de connaissances sur la réorganisation du maillage administratif de la France lors de la Révolution



Contexte

L'Institut National de l'Information Géographique et Forestière (IGN) produit pour sa mission de service public des référentiels géographiques avec une certaine profondeur temporelle destinés à l'analyse des évolutions du territoire national. Le laboratoire LASTIG travaille ainsi depuis de nombreuses années avec des historiens pour constituer des référentiels de données spatio-temporelles décrivant les transformations du territoire dans le temps long. L'une de ces collaborations reconstitue la création et l'évolution du maillage communal français depuis le 18^e siècle. L'historique des transformations communales depuis 1795 ainsi que la cartographie des paroisses religieuses de l'Ancien Régime sont très avancés. Le début de la période Révolutionnaire (1789-1794) est peu couvert, pourtant il a vu les paroisses être réorganisées pour former les futures communes. Cette réorganisation est mal connue et aucune base de données spatio-temporelle la décrivant n'existe encore. Les informations permettant la constitution d'une telle base existent pourtant sous la forme de textes légaux promulgués par l'Assemblée Nationale Constituante et les assemblées de district en 1790-1791 (voir figure 1). **Ce stage porte sur l'extraction automatique et la géolocalisation des informations contenues dans ces textes historiques pour construire une base de connaissances géohistoriques des remembrements paroissiaux sous la Révolution française**, premier chaînon manquant entre paroisses d'Ancien Régime et premières communes.

Un premier stage a abouti à la mise en place d'une chaîne de traitement montrant la faisabilité de l'extraction, de la structuration et de la géolocalisation automatique des informations à partir d'un échantillon de textes (Keller, Abadie, Dumenieu, Baciocchi & Kergosien, 2018). Cette première proposition présente toutefois plusieurs limites :

- La relative variabilité des textes est encore mal prise en compte, réduisant la générnicité de la chaîne d'extraction.
- Les textes décrivent l'évolution des paroisses à deux niveaux de granularité : le territoire et les lieux de culte. Seul le niveau du territoire est actuellement pris en compte.
- La géolocalisation est gênée par la variabilité des formes écrites des toponymes.

<p style="text-align: center;">Art. 25.</p> <p style="text-align: center;"><i>Département du Puy-de-Dôme. District de Clermont.</i></p> <p>« L'église du ci-devant monastère des minimes de la ville de Clermont sera conservée comme oratoire de la paroisse cathédrale. Les paroisses du district de Clermont, hors la ville, chef-lieu de ce district, seront au nombre de 54 dont suit l'état :</p> <p>« Paroisses de :</p> <p>« Allaingnat.</p> <p>« Aubière, à laquelle sera réuni Pérignat-les-Sarliève. Il y aura à Pérignat un oratoire.</p> <p>« Aulnat.</p> <p>« Authezat, dont est distrait le territoire de</p>	<p>Corent-la-Sauvetat, qui continuera d'en faire partie, aura une succursale.</p> <p>« Beaumont.</p> <p>« Blanzat, qui comprendra Serre et les Mauvaises.</p> <p>« Bourg-Lasticq, à laquelle sera réunie la paroisse de Saint-Sulpice, distraction-faite des hameaux de Lasticq, Méauzat et Granges.</p> <p>« Briffon, à laquelle sera réunie la paroisse de Tortebesse, qui formera une succursale.</p> <p>« Cebasal.</p> <p>« Ceyrot.</p> <p>« Chamalières, qui joindra à son territoire celui du hameau de Villars.</p> <p>« Chanonat, à laquelle sera réunie la paroisse de Jussat.</p> <p>« Cournon, dont les deux paroisses sont réunies sous le nom et dans l'église de Saint-Martin.</p> <p>« Crest (le).</p>
--	---

Figure 1 – Un extrait de décret de l'Assemblée Nationale portant sur la réorganisation des paroisses du département du Puy de Dôme. Extrait de l'article 25 - 1^{er} juin 1791

Objectif du stage

L'objectif du stage est double :

1. améliorer la chaîne d'extraction automatique des informations spatio-temporelles à partir des textes,
2. améliorer l'approche de géolocalisation des lieux cités dans les textes (on parle d'entités spatiales nommées ou ESN).

Dans les deux cas, on s'appuie sur une base de données de toponymes produite par vectornisation de la carte de Cassini (voir figure 2), quasi contemporaine des textes et présentant un niveau de détail équivalent. Les toponymes de la carte de Cassini sont ainsi très susceptibles d'être mentionnés dans les textes sur les remembrements paroissiaux.

Cette base de toponymes est exploitée par la chaîne d'extraction des informations spatio-temporelles pour faciliter l'identification des portions de textes désignant des entités spatiales nommées, c'est-à-dire des noms de lieux éventuellement accompagnés de descripteurs (p.ex. "l'église paroissiale de Bourg-Lasticq", "le hameau de Laveix"). Ne disposant pas d'un corpus pré-annoté, cette chaîne de traitement a été implémentée à l'aide de lexiques et de patrons lexico-syntaxiques. Ceux-ci ont cependant été définis pour un corpus restreint et manquent parfois de généralité :

- Si la majorité des entités spatiales mentionnées sont des toponymes seuls ou éventuellement accompagnés d'un descripteur, certains textes comportent néanmoins des entités spatiales étendues nécessitant un traitement adéquat.
- L'extraction des relations spatio-temporelles entre paroisses repose sur une approche essentiellement lexicale et reste largement perfectible. Elle gagnerait notamment à traiter l'expression "ci-devant", très utilisée pendant la Révolution, qui fait référence au caractère révolu de l'information mentionnée à sa suite (p.ex. "la partie du village de Jouffreits ci-devant dépendant de la paroisse de Charbonnières-lès-vieilles").



Figure 2 – Extrait de la feuille 12 de la carte de Cassini sur la Creuse. Les lieux nommés en gras et symbolisés par un clocher sont des chefs-lieux de paroisses; le clocher représente l'église paroissiale. Source gallica.bnf.fr / BnF <https://gallica.bnf.fr/ark:/12148/btv1b53095185n/>

- L'extraction des informations sur les lieux de culte permettrait de compléter les informations relatives au territoire des paroisses, notamment dans le cas des paroisses urbaines jusqu'ici non traitées car ne figurant pas sur la carte de Cassini.

La base des toponymes de Cassini est aussi utilisée pour géolocaliser les ESN extraites des textes. À chaque mention d'ESN extraite du texte sont associés les lieux de la carte de Cassini qui peuvent lui correspondre à l'aide d'une mesure de similarité de chaînes de caractères. Chacune de ces listes de lieux candidats est ensuite ordonnée d'après la distance de chaque lieu à la médiane marginale de l'ensemble des lieux candidats des autres ESN présentes dans le même article que celle en cours de désambiguïsation. L'analyse des résultats obtenus via cette approche révèle deux principaux points d'amélioration :

1. La mesure de similarité de chaînes de caractères utilisée s'appuie sur une distance d'édition, ce qui présente des limites importantes dès lors que les ESN et les toponymes sont orthographiés différemment. Or l'hétérogénéité des graphies de toponymes est encore très présente au XVIII^e siècle. Le passage à une similarité fondée sur la phonétique des toponymes est une piste d'amélioration majeure.
2. La médiane marginale est une approximation du point central de la distribution spatiale des lieux candidats qui a l'avantage de la simplicité. Le classement des candidat pourrait être amélioré par une meilleure approximation comme la médiane géométrique.

Compétences et formation requises

Compétences et connaissances

- Extraction d'informations à partir de textes
- Résolution d'Entités Spatiales Nommées
- Apprentissage automatique
- Données géographiques vectorielles
- Web de données
- Un intérêt pour les données anciennes et la linguistique est un plus

Formation

Master 2 ou troisième année d'école d'ingénieur en informatique, traitement automatique des langues ou en géomatique avec une forte composante informatique.

Selon le profil du candidat, l'un ou l'autre des deux principaux objectifs du stage pourra être plus particulièrement développé.

Informations pratiques

Durée et période de stage

5 mois, printemps-été 2020

Lieu du stage

Equipe LaSTIG/STRUDEL de l'Institut National de l'Information Géographique et Forestière (IGN), à Saint-Mandé (métro 1, station Saint Mandé). Le stage se déroulera dans l'équipe STRUDEL menant des recherches en géomatique sur les structures spatio-temporelles pour l'analyse des territoires.

Indemnités de stage

Stage gratifié selon la législation française.

Modalités de candidature

Envoyer CV et lettre de motivation ciblée sur le sujet par email au format PDF et en un seul fichier aux encadrants listés ci dessous.

Encadrement du stage

Nathalie Abadie [STRUDEL/IGN] : nathalie-f.abadie[at]ign.fr

Éric Kergosien [GERiiCO/SID/Université Lille 3] : eric.kergosien[at]univ-lille3.fr

Bertrand Duménieu [CRH/EHESS] : bertrand.dumenieu[at]ehess.fr

Stéphane Baciocchi [CRH/EHESS] : stephane.baciocchi[at]ehess.fr

Références

Keller, A., Abadie, N., Dumenieu, B., Baciocchi, S. & Kergosien, E. (2018). Vers la construction d'une base de connaissances sur la réorganisation territoriale française à la Révolution. In *Conférence Sagéo 2018 Atelier Exces*, Récupérée à partir de <https://hal.archives-ouvertes.fr/hal-02399176/>