Congrès AFSP Aix 2015

ST n° 4 Investiguer le bureau de vote. Réflexions épistémologiques et mutualisation des expériences de terrain

Soumaya Yahiaoui (LIA, Université d'Avignon) yahiaoui.soumaya@gmail.com
Didier Josselin (UMR ESPACE 7300, CNRS, SFR Agor@ntic) didier.josselin@univ-avignon.fr
Eric San Juan (LIA, Université d'Avignon, SFR Agor@ntic) eric.sanjuan@univ-avignon.fr
Christèle Marchand-Lagier (MCF de science politique, LBNC EA 3788-UAPV, SFR Agor@ntic, chercheure associée au CHERPA EA 4261-Science-po Aix) christele.marchand@univ-avignon.fr

Délimitation géométrique semi-automatique des bureaux de vote à partir de textes juridiques et d'information géographique numérique : enjeux et difficultés

Semi-automatic spatial partitionning of polling boundaries using legal texts and numerical geographical information: an issue

Résumé/Abstract

Nous proposons une méthode automatique de délimitation des bureaux de vote à partir des textes juridiques et de données géographiques. En effet, pour les communes françaises de taille moyenne possédant plusieurs bureaux, il n'existe pas de méthode permettant l'obtention d'une cartographie numérique exhaustive, pérenne et aisément modifiable dans le temps. Or, l'exploitation de données électorales à l'échelle des bureaux de vote est désormais une source essentielle d'information en sciences politiques. Il s'agit d'améliorer autant que faire se peut la qualité et la pertinence des découpages réalisés, à partir desquels sont opérés les processus électoraux et générés les statistiques électorales. Utilisant une suite géomatique libre (QGIS/PostGIS), traitant des expressions régulières et se basant sur des données de géolocalisation courantes (réseaux et adresses), ce travail produit des résultats prometteurs de vérification et de (re)construction automatique de géométries dont les apports et les limites sont discutés dans cet article.
Mots-clés : extraction d'information géographique, expressions régulières, limites des bureaux de vote, géométrie, topologie

Using free GIS (QGIS/POSTGIS), regular expression processing and geographical data about networks and adresses, we propose an automatic computation to verify and to (re)draw polling boundaries from official legal texts. Indeed, middle size French towns miss a reusable method to build this important geographical information in the long term. Also, it is of essential interest to handle pertinent and accurate spatial partitioning of polling areas, because those are the basic entities for electoral process and information. Contributions and limitations of the results obtained in this work are presented and discussed.
Keywords : geographical information retrieval, regular expressions, polling station boundary, geometry, topology

Cette communication s'appuie sur la collaboration interdisciplinaire entre informaticiens, géographes et politistes menée dans le cadre de la SFR Agor@nTIC[1] de l'Université d'Avignon depuis 2012. Elle s'inscrit dans la suite des travaux présentés lors du workshop *Le bureau de vote. Méthodes, outils et conditions d'enquête. Pour une approche renouvelée des comportements électoraux* organisé à l'université d'Avignon en novembre 2013. Ces travaux initiaux qui s'intéressaient à la spatialisation des comportements abstentionnistes sur un bureau de vote avignonnais (BV 220 La Croix rouge), s'ils ont montré leurs limites, ont permis d'attirer l'attention sur l'artificialité des délimitations du bureau de vote étudié. Le BV 220 est positionné à 81 % sur l'IRIS 122 Rotondes Barbières et à 18 % sur l'IRIS 127 Croix des oiseaux situé en ZUS. L'IRIS 122, pourtant classé parmi les IRIS les plus pauvres de l'agglomération avignonnaise[2], rassemble, sur le territoire du BV 220, une partie d'habitations pavillonnaires, habitat certes modeste mais caractérisant des petits propriétaires, dont on peut supposer que la situation économique est plus favorisée que celles des habitants de logements sociaux qui composent assez largement le reste du territoire du BV. Habiter d'un côté ou de l'autre d'une rue, dans la partie sud proche de la rocade et des quartiers défavorisés ou au nord en limite de l'intra-muros produit des comportements électoraux très contrastés sur le même bureau de vote (bureau socialiste caractérisé par un fort vote en faveur du FN). Le relevé de listes d'émargement sur ce bureau effectué lors des Présidentielle et législatives 2012 et à nouveau lors des municipales 2014 (complété là par un relevé de boîtes aux lettres), les QSU réalisés en 2012 ainsi que la conduite d'enquêtes étudiantes sur la vie associative (Centre social), sociale (écoles) et économique (enquête auprès des commerçants ou entreprise en ZFU) de ce territoire en 2013, nous ont permis de creuser cette artificialité en distinguant des formes de sociabilités concurrentes.

Il ne s'agit donc pas de faire de la spatialisation un but en soi, malgré tout son intérêt méthodologique, mais bien de contribuer à une réflexion plus générale sur les manières d'appréhender les comportements électoraux dans leurs contextes de production pour « en proposer des systèmes explicatifs « à géographie variable » » (Gombin, Rivière, 2012). De ce point de vue, les délimitations administratives des bureaux de vote constituent bien un des « contextes », parmi d'autres, de production des comportements électoraux. Cette notion de contexte reste relativement large en sciences sociales allant de la prise en considération de l'étendue des expériences socialisatrices (Lahire, 1999) jusqu'à l'environnement proprement écologique (Braconnier 2010) qui entoure ces comportements. La notion de contexte peut ainsi couvrir invariablement celle de contexte familial, amical, social, territorial, environnemental et il n'est pas certain que géographes et politistes y mettent toujours exactement la même chose : « Sauf à les réifier (inclination récurrente de Siegfried à Lévy), les environnements résidentiels importent donc moins que l'intensité de la vie sociale et les réseaux de sociabilité qui structurent le contexte local » (Lehingue, 2011, p. 134).

L'organisation du territoire, à travers ses découpages administratifs, informe sur les interactions entre les populations, les lieux et les événements et notamment sur les pratiques électorales associées au suffrage universel depuis 1848 (Ihl, 2002; Garrigou, 2002). Ces informations permettent de réaliser des analyses descriptives ou statistiques à l'échelle des entités administratives. Les délimitations des bureaux de vote constituent donc un cadre de référence législatif avec lequel il faut compter, sachant que la diffusion des résultats électoraux par le ministère se fait désormais à cette échelle. La question du découpage de ces bureaux de vote est centrale pour plusieurs raisons :

- Les délimitations des BV sont des limites légales (référence : textes Légifrance) et

1 Nous remercions le Laboratoire d'Informatique d'Avignon et le Laboratoire Biens, Norme, Contrats d'avoir cofinancé le stage de fin d'étude d'ingénierie en informatique industrielle et automatique de Soumaya Yahiaoui dans le cadre de ce projet interdisciplinaire de la SFR Agor@antic. (Structure Fédérative de Recherche : http://blogs.univ-avignon.fr/sfr-agorantic/)

2 Diagnostic territorial réalisé en mars 2013 par le cabinet d'étude COMPAS pour le conseil régional PACA. C'est essentiellement l'habitat collectif au abord de la rocade sud qui fait descendre l'indice de développement territorial, utilisé par le cabinet Compas et sur lequel nous reviendrons, à<-130. De la même manière, c'est la juxtaposition habitat collectif/habitat pavillonnaire qui maintient l'IDT entre [-130 ;-50[sur l'Iris 127 pourtant positionné sur la ZUS Croix des oiseaux Saint-Chamand.

produisent en ce sens des effets (comptabilisation des résultats électoraux dans ces limites) ;

- Souvent méconnues des individus, elles découpent, recoupent ou distinguent des espaces inégalement appropriés par les citoyens, inscrits, mal-inscrits ou non inscrits (Braconnier, Dormagen, 2007) ;
- Ces délimitations se calent imparfaitement sur les limites administratives que sont les communes puisque la France compte 65000 bureaux de vote pour 36000 communes. Les 6000 communes les plus grandes ont plusieurs bureaux de vote (Jadot et al, 2010) ;
- Ces limites sont amenées à être modifiées d'une élection à l'autre à la faveur de manœuvres électoralistes dont les récentes élections départementales sont un exemple. De ce point de vue, le relevé des listes d'émargement sur le BV n°220 n'a pas pu être renouvelé en 2015 ;
- Ces limites découpent des espaces urbains et péri-urbains en profonde transformation qui impacte les pratiques électorales (Ravenel, Buléon, Fourquet, 2003 ; Jardin, 2014).

Les délimitations des bureaux de vote interrogent donc la démocratie dans sa forme représentative (Gaxie, 1996), puisque la notion même de représentation nécessite le découpage de sous-espaces. Ce découpage constitue un réel enjeu démocratique, d'autant plus dans un contexte français marqué, depuis plusieurs décennies, par une croissance régulière de l'abstention à tous les scrutins (Braconnier, Dormagen 2007). Cet enjeu démocratique interroge au moins autant la science politique que la géographie (Bussi, 2007) qui peuvent y trouver les moyens de rapprochements. Pouvoir disposer de partitions territoriales fiables dont les entités spatiales sont représentatives de l'électorat est en effet un objectif évident pour toute démocratie. En sciences politiques, la compréhension des processus électoraux à partir de l'analyse des données électorales est largement dépendante des découpages territoriaux de recensement des votes. La forme et la composition de la partition spatiale considérée impacte directement les processus électoraux, dans leur déroulement comme dans leurs résultats, notamment ceux qui concernent les scrutins où les représentations politiques restent locales mais où le leadership est déterminé par la dominance des couleurs politiques (c'est-à-dire, *in fine*, l'ensemble des élections sauf l'élection présidentielle).

Connaître précisément les limites des bureaux de vote nous semble donc être une première étape nécessaire pour repérer ensuite de quelle manière ces limites se calent ou pas sur les espaces de vie vécus par les électeurs (rue, quartier, IRIS…). Les travaux pionniers de Cartelec prennent déjà au sérieux le découpage territorial des bureaux de vote en proposant de faire de ces derniers « une nouvelle couche d'information géographique » (Jadot et al., 2010). Les outils développés dans le cadre de ce projet ANR posent deux séries de questions auxquelles nous pensons pouvoir apporter des éléments de réponse :

- La première concerne les informations à partir desquelles les fonds de cartes produits par Cartelec ont été réalisés. Ces derniers ont pu être questionnés à partir de notre connaissance du terrain avignonnais. La carte des BV présente en effet certaines incohérences. Nous avons donc choisi ce terrain d'expérimentation pour proposer une ou plusieurs alternatives aux problèmes soulevés selon nous par la cartographie Cartelec.
- La seconde interroge la correspondance entre délimitation des BV et délimitation des IRIS : en réaffectant proportionnellement à la surface du bureau les données sociales INSEE à l'IRIS, on supprime les particularités infra IRIS pouvant rassembler des territoires très contrastés comme nous l'avons suggéré pour le BV 220. À ce titre et au-delà de cet article, notre projet vise à terme à obtenir un découpage électoral composé de bureaux de vote aux délimitations exactes et les plus représentatives possibles de l'électorat. Il vise également, à terme, à mutualiser les informations de l'INSEE produites sous forme de carroyage ou par IRIS avec celles dont nous pouvons disposer par ailleurs (échantillons, recensements locaux)

afin d'étudier la relation entre les électeurs et leurs territoires, des points de vue spatial et sociologique.

Globalement, notre approche conduit à d'autres questionnements. Faut-il reconsidérer complètement la géométrie des bureaux de vote en les calquant sur les découpages de l'INSEE par carroyage[3] ? Faut-il au contraire les construire en fonction des densités locales de population votante ? Comment combiner les différentes couches d'informations (IRIS, bureaux de vote) pour saisir les tenants et les aboutissants de la sociologie et de la géographie du vote ? Tels sont les enjeux de cette recherche à plus long terme.

Les résultats exposés dans cet article concerne le premier volet de la recherche, à savoir la construction d'une méthode (semi-)automatique de délimitation des bureaux de vote. Il s'agit d'extraire l'information géographique des limites des bureaux de vote à partir des textes juridiques, afin de réaliser une cartographie respectant au maximum la géométrie et la topologie des limites des objets géographiques en question, à savoir, les bureaux de vote.

1. Contexte

Pour diverses raisons (manque d'information, coût de mise en œuvre, négligence sur la qualité des transcriptions et des supports, etc.), la correspondance entre diverses informations sur la population (e.g. recensement de terres agricoles ou de propriétés foncières) et les entités administratives correspondantes est parfois difficile à valider (Josselin et al., 1999; 2000). Notamment, une partie non négligeable de l'information est inscrite dans des textes, par exemple historiques ou juridiques. C'est le cas des données électorales, qui sont censées regrouper notamment le nombre de votants par bureau de vote et pour lesquels il n'existe pas encore, pour l'ensemble des communes de France, de méthode générique de cartographie numérique exhaustive, pérenne et adaptable aux changements sur le long terme.

Dans le cas particulier de la délimitation des bureaux de vote, on dispose en effet de textes juridiques de Légifrance décrivant assez précisément la façon dont un législateur ou un technicien territorial peut dessiner à main levée ou digitaliser, sur un support cartographique (parfois papier), les limites des bureaux de vote, à partir d'une référence de réseaux nommés. Pour environ 30 000 communes, le problème de délimitation ne se pose pas car la commune n'a qu'un unique bureau de vote couvrant son emprise. En revanche, pour le reste, les limites des bureaux peuvent être plus complexes. Des grandes villes comme Nantes ou Lyon ont, par exemple, numérisé avec précision et rendu public ce découpage.

Comme nous l'avons déjà souligné, le projet ANR Cartelec a comblé en très grande partie l'absence de cartographie pour bon nombre des communes du territoire français (Jadot et al., 2010) (cf. Figure 1). Ce projet s'est basé sur un ensemble de documents hétérogènes (fichiers de tableur, cartes papier ou numériques, découpages préexistants) pour reconstruire la géométrie des bureaux de vote à l'échelle française (Beauguitte, Colange, 2013). Le travail de recomposition de la géométrie a été colossal et partiellement automatisé. La méthode paraît cependant délicate à reproduire dans le temps, ce qui risque de grever à terme la qualité et l'exhaustivité de l'information. Il peut ainsi manquer un certain nombre de communes de taille moyenne possédant seulement quelques bureaux de vote. De plus, même en cas d'obtention d'une carte complète à une date *t*, l'évolution de l'emprise des bureaux de vote (fusion, séparation, modification de contour) peut rendre rapidement la cartographie obsolète. Qui plus est, beaucoup de communes de taille moyenne ne disposeront jamais des moyens nécessaires pour héberger un SIG et digitaliser les contours de leurs bureaux. Ces arguments sont en faveur d'une méthode plus légère et automatisée de mise à jour des

3La maille (200m sur 200 m) offrant un traitement partiel des données INSEE à un niveau d'agrégation bien inférieur à celui de l'IRIS et offrant la perspective d'envisager des ré-agrégations plus pertinentes à l'échelle du BV.

découpages selon les besoins, se basant sur les textes juridiques de référence.

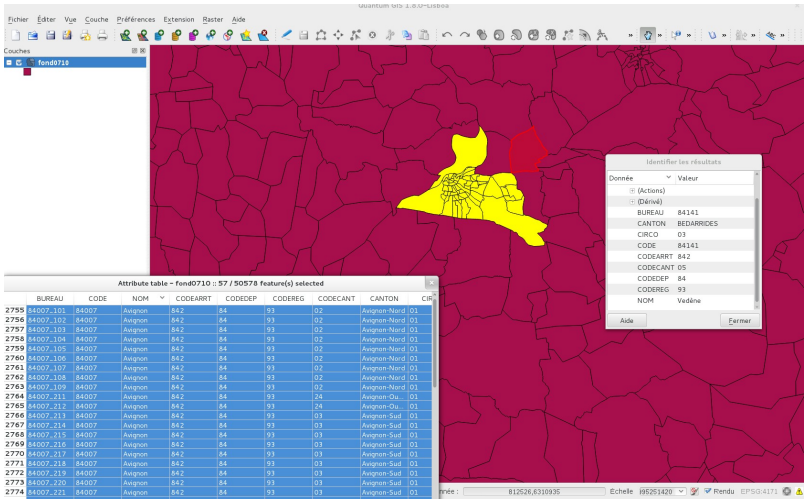

Figure 1. Extrait de la base de données géographiques CARTELEC en 2014 : les bureaux de vote d'Avignon (en clair au centre) sont bien délimités ; la commune de Vedène, au nord-est d'Avignon, qui possède plusieurs bureaux de vote, est considéré d'un seul bloc.

Une contrainte délicate à respecter pour la carte des bureaux de vote est d'assurer une partition spatiale, afin que chaque citoyen puisse voter en un seul et unique lieu. Pour des raisons de risque potentiel d'invalidation des résultats électoraux, il est préconisé que la qualité du découpage ne souffre aucune erreur (ACE, 2013). Les informations que nous manipulons sont publiques, mais sensibles, d'ailleurs assez difficiles à trouver en dépit de leur caractère public. Elles contiennent des expressions textuelles régulières, généralement partagées dans les textes se référant aux différents bureaux de vote d'un secteur géographique donné. La loi impose en effet que chaque bureau possède un « périmètre géographique ». De fait, il doit exister un texte issu des échanges entre les mairies et la préfecture, écrit de façon assez uniforme par les rédacteurs et ensuite validé par la préfecture dans un arrêté. À ces informations s'ajoutent des données géographiques détaillées : les adresses et les réseaux. Notre méthode vise à terme la généricité et l'exactitude afin d'être appliquée à toute commune, quelle que soit sa taille, pour peu qu'elle comporte plus d'un seul bureau de vote. La méthode développée peut ainsi potentiellement intéresser les services de l'état (français, voire étranger), par sa capacité de généralisation et de production automatique de partitions spatiales des bureaux de vote et son adaptabilité au contexte local.

2. Objectifs et approche

L'extraction d'informations géographiques par l'analyse des textes doit permettre *in fine* d'automatiser et d'accélérer la construction de cartes électorales en série. Dans cette perspective, nous proposons tout d'abord de travailler sur des textes juridiques simples, bien structurés. Le but est de construire une méthode de partitionnement la plus efficace et désambiguisée possible, respectant scrupuleusement la géométrie des bureaux de vote. La partition devra être de granularité fine, correspondre fidèlement à la population votante et au découpage administratif réel. Une telle expérimentation peut par ailleurs offrir les moyens de tracer les limites de l'échelle pertinente d'analyse des résultats électoraux, en calant ces découpages sur les espaces de vie vécus par les citoyens (partition d'un bureau de vote, regroupement de plusieurs bureaux).

La qualité de la partition spatiale est cruciale : l'ensemble doit couvrir le territoire (pas de trou) et aucun polygone ne doit se superposer à un autre. On pressent l'importance des contraintes topologiques. Celles-ci interviennent également au niveau des polygones des bureaux de vote eux-

même, puisqu'à nouveau, leurs géométries doivent être complètes (pas d'interruption de continuité), précises (conservation de toutes les coordonnées des points), justes (pas d'écart de géocodage) et concises (pas de géométrie superflue). Ces contraintes peuvent constituer, en plus des éléments issus des textes juridiques, des points d'ancrage forts pour guider la construction de la partition. Une des pistes consiste à comparer nos résultats à des partitions existant dans d'autres documents. En complément, à partir des textes propres aux communes non référencées dont les bureaux de vote ont été agrégés, il serait possible d'améliorer la précision de la partition, en jouant notamment sur les limites partagées des bureaux de vote limitrophes (point non traité dans cet article).

Dans cet article, nous posons la première brique de ce projet, qui consiste en l'analyse de corpus de textes juridiques en vue de la cartographie des limites des bureaux de vote. Notre approche permet aussi potentiellement de détecter des erreurs, des inexactitudes ou des ambiguïtés dans les textes eux-mêmes, mais également dans les bases de données géographiques manipulées. Nous verrons qu'en dépit de la rigueur d'écriture supposée des textes juridiques, il n'est pas si aisé de délimiter automatiquement les bureaux de vote, et que, finalement, notre méthode se révèle très complémentaire de l'approche développée dans le projet Cartelec.

L'objet de cet article est donc :

- de proposer une méthode pour extraire les limites des bureaux de vote, associant des informations textuelles et géographiques ;
- d'évaluer la capacité de la méthode à détecter des erreurs dans les textes et dans les informations géographiques ;
- de présenter les difficultés méthodologiques rencontrées pour l'obtention d'une délimitation parfaite d'un bureau de vote.

3. Recherche et extraction d'information

3.1. Recherche d'expressions régulières dans les textes

Les expressions régulières (Watt, 2005, Goyvaerts, Levithan, 2012) sont de puissants outils pour faire des recherches dans les chaînes de caractères (Friedl, 2006; Fitzgerald, 2012). Elles sont utilisées par exemple à des fins d'analyse géographique (Leidner, Lieberman, 2011). Le terme *regex* renvoie à un mini-language permettant de trouver des motifs génériques dans un texte ou une chaîne de caractères et d'effectuer des substitutions de chaînes par motif. Les expressions régulières ont un certain nombre de déclinaisons, mais la plupart sont plus ou moins compatibles avec celles du langage Perl, comme celles de Python, Java, Racket et PHP. La plupart des langages de programmation offrent un module d'expressions régulières soit directement inclus dans le langage, soit sous forme de paquet additionnel. Leur utilisation principale est la reconnaissance de motif, c'est-à-dire la recherche de (sous-)chaînes de caractères dans une autre chaîne à partir d'un motif donné. Un motif est une représentation générique d'un ensemble de chaînes de caractères, un peu de la même manière qu'une fonction représente un ensemble d'entrées/sorties. L'écriture de motifs est la principale difficulté de l'utilisation des expressions régulières.

3.2. Le domaine de la recherche et de l'extraction d'informations géographiques

L'extraction d'information dans les textes est un vaste domaine de recherche (Manning et al., 2008), notamment en géomatique (Sallaberry, 2013). Jones et Purves (2008) identifient une série de problèmes clés liés à l'extraction et à la recherche d'information géographique : la lecture délicate des toponymes et l'ambiguïté des noms dans les adresses et les lieux (Martins et al., 2010), les

problèmes de terminologie vague ou de sémantique floue, l'indexation textuelle et spatiale (Gaio et al., 2012). Ces difficultés interviennent sous différentes formes dans notre problématique.

Dans le domaine de l'analyse spatiale, les travaux portent notamment sur la recherche d'informations localisées dans l'espace (Bilhaut et al., 2007), jusqu'à l'amélioration des ontologies spatiales (Tien Nguyen et al., 2009). Beaucoup de ces travaux mettent l'accent sur l'extraction d'informations géographiques. Peu finalement la reconstruisent *a posteriori* ; par exemple, citons ici les travaux originaux de représentation d'itinéraires virtuels à partir de récits de voyage (Gaio et al., 2008).

La mise en correspondance de l'information extraite avec des bases de données existantes relève du géocodage dans les SIG. Les problèmes d'appariement à ce niveau sont bien connus et, dans la pratique, on sait qu'il est difficile d'obtenir une association parfaite entre une source de données alpha-numérique et géographique (Widlocher et al., 2004). L'enjeu de ce travail est de réduire ces erreurs autant que faire se peut, dans le contexte sensible du suffrage universel et de la représentation cartographique des données électorales, qui ne souffre en théorie aucune inexactitude. Notre approche s'appuie sur les deux premières des préconisations énoncées par Leidner et Lieberman (2011), notamment dans la phase qualifiée de « geoparsing » :

- appariement direct de noms dans une base avec les mots dans le texte (dans notre cas, à cause de l'insuffisance des adresses inscrites dans les listes électorales ne permettant pas un recouvrement correct) ;
- système à base de règles et en particulier à base d'expressions régulières, adapté car le texte a suffisamment de régularités ;
- extension à terme vers l'apprentissage (donc approximatif) si les deux précédentes approches sont insuffisantes (aspect non développé ici).

4. Types de données manipulées

4.1. Exemples d'expressions régulières sous Python

L'utilisation des expressions régulières en langage Python requiert le module « RE ». Par exemple, on peut importer un paragraphe du corpus :

>>>sample_text="Comprenant les électeurs demeurant sur la partie du territoire délimitée par la rue Ferruce côté pair, la rue Puits de la Reille côté pair, la rue Balance côté pair, la place Puits des Boeufs côté pair, la place de l'Horloge côté pair, la rue des Marchands côté impair, la place Carnot côté impair."

Dans ce paragraphe, on cherche à identifier les rues et les places :

>>>re.findall("(rue.?|place.*?)côté",sample_text)*

La liste obtenue est ensuite insérée dans une base de données :

['rue Ferruce ', 'rue Puits de la Reille ', 'rue Balance ', 'place Puits des Boeufs ', "place de l'Horloge ", 'rue des Marchands ', 'place Carnot ']

4.2. Caractéristiques des textes juridiques

Nous travaillons sur les bureaux de vote de la ville d'Avignon, dont voici un exemple. La régularité du texte juridique est perçue à travers les éléments récurrents suivants :

– Structure globale du corpus :

BUREAU N° 101 ; Lieu de vote : HOTEL DE VILLE - PLACE DE L'HORLOGE

– Structure des paragraphes pour la délimitation de ce bureau de vote :

« Comprenant les électeurs demeurant sur la partie du territoire délimitée par la rue Ferruce côté pair, la rue Puits de la Reille côté pair, la rue Balance côté pair, la place Puits des Boeufs côté pair, la place de l'Horloge côté pair, la rue des Marchands côté impair, la place Carnot côté impair, la rue Armand de Pontmartin côté impair, la rue Sainte Catherine côté impair, la rue Lafare côté impair, la rue du Grand Paradis côté impair, la place Saint Joseph côté impair, la rue Palapharnerie côté impair du 15 à la fin, du quai de la Ligne à la porte du Rhône. »

Pour l'ensemble du corpus de textes traité, un certain nombre de termes permettent d'instancier les adresses et les géométries : les types de lieux (rue, place, quai, voie ferrée), les lieux eux-mêmes (gare S.N.C.F) ou les intersections de géométries linéaires (carrefours). D'autres termes n'ont pas d'emprise spatiale explicite : par exemple, la ligne imaginaire reliant deux intersections. A noter que nous observons dans les textes des différences subtiles d'écriture, potentiellement source d'erreur : *côté pair*, *(côté pair)* ou *(côté pair)*.

4.3. Données géographiques

Dans ce travail, nous manipulons essentiellement deux types de données (réseaux routiers et fichiers d'adresses) :

- La *carte Navteq des routes* couvre l'ensemble du réseau routier sur Avignon. Sa table attributaire fournit des informations sur le nom de la route (rue, place, quai, etc.), ce qui constitue une base de données de référence pour pouvoir identifier, sur la carte, les limites géographiques extraites du texte (cf. Figure 2). Des tests réalisés avec des données Open Street Map n'ont par ailleurs pas été concluants.
- La *base de données IGN des adresses* fournit, entre autres, des informations sur le numéro de chaque bâtiment, le nom de la rue et de quel côté le point est situé (droite/gauche). Elle nous sert de point de comparaison pour (in)valider les géométries extraites des textes.

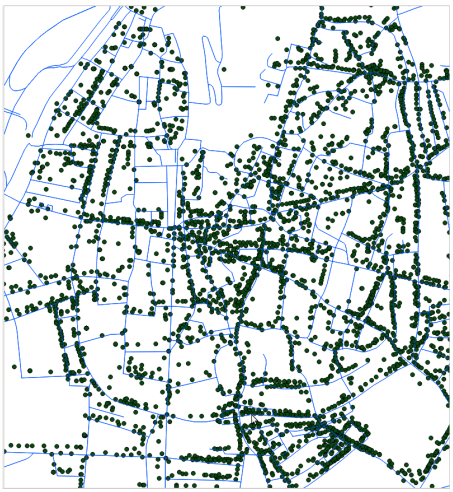

Figure 2. Extrait du réseau Navteq (lignes en bleu)
et de la BD adresses de l'IGN (semis de points) sur un secteur d'Avignon

5. Association d'un Système d'Information Géographique et d'un module de traitement des expressions régulières

L’environnement de travail est principalement composé d’un système d’information géographique. Dans notre cas, il s’agit de Quantum Gis et d’un système de gestion de base de données, en l’occurrence, PostgreSQL/PostGIS. La méthode développée s’appuie, d’un côté, sur QGis pour l’extraction des données géométriques nécessaires au traçage des limites des bureaux et, d’un autre côté, sur l’outil PostgreSQL pour le stockage des informations à la fois textuelles et géographiques.

Dans les grandes lignes, la démarche se décompose ainsi (cf figure 4 pour davantage de détails) :

1. Préparer une base de données PostgreSQL/PostGIS où seront stockés le réseau routier et les adresses des bâtiments sur Avignon ;

2. Extraire l’ensemble des routes qui délimitent le bureau de vote à partir du texte et les stocker dans cette même base ;

3. Ayant identifié les routes sur le réseau, tracer le contour du bureau et générer la géométrie directement dans un fichier shape sous QGIS ;

4. Corriger les éventuelles erreurs de géométrie (post-traitement).

La solution proposée pour l’extraction automatique des géométries des bureaux est développée sous Python et réalisée grâce au module RE (figure 3). Ce dernier définit plusieurs fonctions utiles ainsi que des objets propres pour modéliser des expressions. Parmi ces fonctions, nous avons utilisé *compile()* dont le rôle est de compiler une expression régulière, exécutée à plusieurs reprises par un appel dans d’autres fonctions. La fonction *findall()* permet de chercher toutes les occurrences de l’expression dans le texte et renvoie une liste des éléments correspondants.

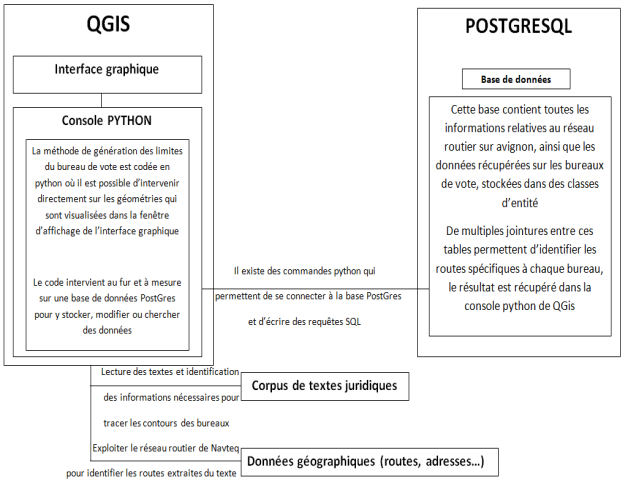

Figure 3. Structure de l’environnement logiciel

Le schéma de séquence de traitement en Figure 4 déroule les différentes étapes techniques de la méthode utilisée. La structure de la base de données PostgreSQL avec ses principales tables, dont la gestion est codée sous Python, est présentée dans la Figure 5. D’autres classes intermédiaires sont créées et modifiées au fur et à mesure de l’exécution du code et selon les besoins. Elles permettent, par exemple, de stocker, pour un bureau donné, l’ensemble des routes qui le délimitent ainsi que leurs coordonnées.

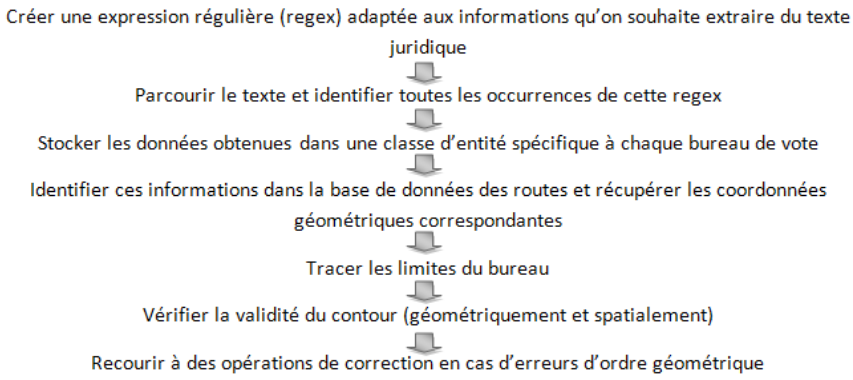

Figure 4. Les étapes de la méthode

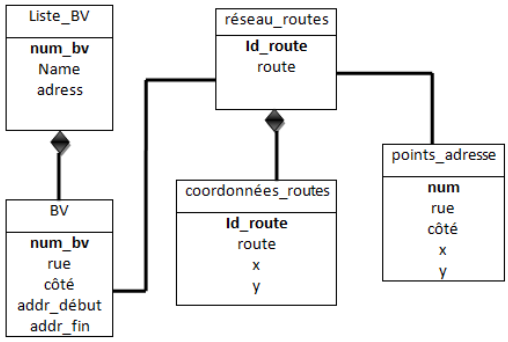

Figure 5. Principales classes du modèle conceptuel

6. Résultats : des polygones dont la cohérence géométrique est variable

Les premiers résultats de cette recherche résident dans la proposition d'une méthode d'extraction des contours des bureaux de vote, également dans l'identification de problèmes spécifiques et des pistes de résolution. Beaucoup de tests ont été réalisés sur les bureaux de vote de la ville d'Avignon. Nous ne présentons ici que quelques cas particulièrement parlants pour illustrer les problèmes rencontrés.

Parmi les 57 bureaux de vote d'Avignon, les exemples que nous présentons ont les numéros suivants : 101, 214 et 219. Dans la base de données géographiques Cartelec, ils apparaissent comme sur la figure 6.

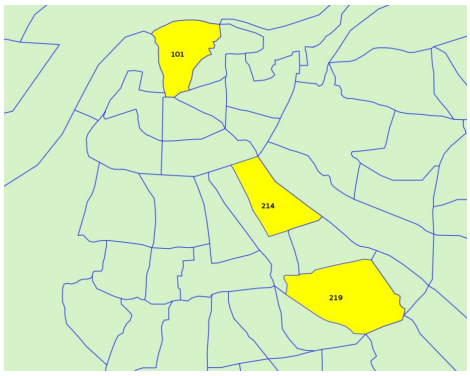

Figure 6. Les bureaux de vote 101, 214 et 219 d'Avignon dans la base Cartelec

– Exemple 1 : le bureau n° 101

Dans la géométrie du bureau de vote 101, les erreurs présentes sont soit des lignes qui dépassent, soit des trous où deux rues successives ne sont pas connectées. Ce dernier cas s'explique par deux types d'erreurs présents selon les cas :

- provenant du texte où la connexion manquante est une rue qui n'est pas citée dans la description du texte juridique ;

- due à une faute de construction ou d'affectation d'adresse dans le réseau Navteq.

On voit ici l'importance de la qualité des données manipulées, textuelles et géographiques, pour permettre un bon appariement géométrique.

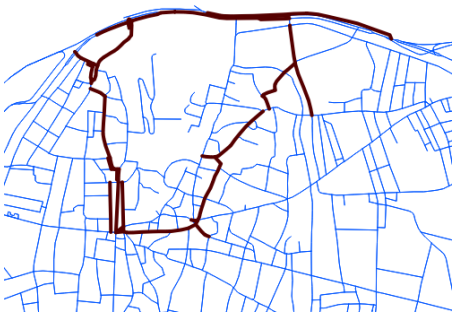

Figure 7. Géométrie extraite du bureau de vote 101

– Exemple 2 : le bureau n° 214

« Comprenant les électeurs demeurant sur la partie du territoire délimitée par le boulevard Saint Michel côté pair du numéro 24 à la fin, l'avenue Pierre Sémard côté pair du numéro 2 au numéro 74, le boulevard de la Première Division Blindée côté pair, l'avenue des Sources côté impair du numéro 1 au numéro 49. »

La géométrie du bureau de vote 214 (cf. figure 8) présente également des débordements de lignes. Mais on remarque aussi un autre type d'erreur géométrique où la bifurcation d'une rue résulte en un petit contour fermé adjacent à celui du bureau. En effet, dans les nombreux bureaux que nous avons étudiés, nous nous sommes aperçus que souvent des diverticules de voiries ou des impasses (chemins peu carrossables ou routes en cul-de-sac) conservent de façon erronée le nom de la rue sur laquelle ils débouchent, générant des excroissances géométriques sur le polygone du bureau de vote. On voit dans cet exemple toute la variété des géométries superflues, qui peuvent être linéaires ou polygonales, attachées ou non à la juste géométrie du bureau de vote (celle qu'il faut conserver *in fine*).

Figure 8. Géométries extraites des bureaux de vote 214 et 219

– Exemple 3 : le bureau n°219

« Comprenant les électeurs demeurant sur la partie du territoire délimitée par l'avenue Capitaine de Vaisseau l'Herminier côté pair, le boulevard Montesquieu côté pair, l'avenue de la Croix des Oiseaux côté pair du numéro 12 à la fin, l'avenue des Sources côté impair du numéro 59 à la fin. »

La géométrie du bureau de vote 219 (cf. figure 8) présente quant à elle un paradoxe. D'un côté, l'algorithme a généré un contour bien fermé, permettant d'obtenir l'emprise surfacique exacte du bureau de vote, en dépit de 5 arêtes pendantes qui ont été conservées. En effet, le croisement du texte avec la base de données géographiques permettait difficilement de trouver l'intersection pour couper la géométrie au bon endroit, à cause notamment de la méconnaissance de la localisation exacte des numéros des voies. Cet aspect devra être complété à terme par l'usage de la BD adresse de l'IGN. D'un autre côté, en comparant cette géométrie à celle du même bureau de vote dans la base Cartelec (Figure 6), on constate que les deux surfaces ne coïncident absolument pas. Ce problème est dû à un écart dans les versions mises à jour, la géométrie extraite du texte juridique par notre algorithme représentant aujourd'hui une délimitation plus juste du périmètre réel du bureau de vote 219.

7. Discussion sur la méthodologie

L'analyse détaillée des trois exemples et des autres bureaux de vote d'Avignon montre l'existence de plusieurs types d'erreur provenant :

– du caractère équivoque des indications dans les textes juridiques, probablement causés par des ambiguïtés humaines ou une application toute relative des directives d'écriture ;

– des erreurs d'écriture ou de dénomination des voies, présentes dans les bases de données géographiques ;

– d'incompatibilité des versions à cause de mises à jour incohérentes de part et d'autre des données (texte *vs* géographie) ;

– des difficultés d'ordre géométrique : précision insuffisante, géométries discontinues à cause de polylignes non adjacentes, séquence d'arêtes non ordonnées, incertitude de localisation des numéros des voies ou intersections introuvables ; certaines de ces erreurs ne sont pas encore résolues par nos algorithmes.

À noter également les différences de qualité (Goodchild, Jeansoulin, 1998) entre les géométries produites par notre algorithme et celles présentes par Cartelec. En ce qui concerne Cartelec, les limites des bureaux de vote ont des géométries épurées, probablement pour alléger la base au niveau national et pour permettre également une cartographie lisible. Nous qualifierons cette géométrie d'exacte (chaque point est correctement localisé), cohérente (tous les polygones des bureaux de vote sont fermés et partagent des arêtes adjacentes avec les bureaux limitrophes), mais relativement peu précise (la géométrie des limites des bureaux est simplifiée et peut potentiellement amener à des erreurs de localisation des votants) et pas totalement complète (quelques communes n'ont pas leurs bureaux de vote correctement identifiés). Dans notre cas, la géométrie créée dépend entièrement de la précision des données géographiques et de leur compatibilité, en termes de grain, avec les éléments présents dans les textes juridiques. De fait, nous obtenons potentiellement des résultats plus précis (visibles en comparant les contours respectifs des bureaux de vote) et à terme plus complets (si nous exploitons tous les textes juridiques en faisant l'hypothèse qu'ils soient tous accessibles et écrits en bonne et due forme). Mais cette qualité a un coût : probablement celui d'une moindre exactitude ou consistance (à cause d'un excès de géométrie pas forcément utile) et certainement celui de risque d'incohérence entre le textuel et le spatial (car la précision contraint le processus d'appariement à plus de rigueur et tolère moins l'erreur). Devant ce dilemme, des choix

(voire des compromis) devront être faits pour construire une géométrie précise, fiable et utilisable en cartographie, qui plus est aisément modifiable dans le temps.

8. Conclusion et perspectives

Ce travail constitue une première étape vers l'automatisation de l'extraction de contours de bureaux de vote à partir de textes juridiques. Il vient compléter les travaux du projet Cartelec. Si les textes juridiques semblent bien formatés en apparence, leur ressemblance ne permet pas encore de construire des géométries sans ambiguïté, à cause des problèmes de qualité des données évoqués ci-dessus et du caractère inachevé des algorithmes actuellement développés. De plus, les formats d'écriture varient sensiblement d'une ville à une autre, rendant délicate l'application d'une méthode calquée sur l'analyse d'un échantillon de textes pas forcément représentatif. Toutefois, on peut aisément imaginer l'intérêt d'un tel outil lorsqu'il construira automatiquement des géométries justes. Enfin, dans une certaine mesure, notre outil peut d'ores et déjà, dans son état actuel, aider les juristes et cartographes à vérifier la qualité des textes juridiques et/ou des données géographiques, puisqu'il permet de pointer des problèmes d'incohérence entre les deux. À noter que les algorithmes développés pourraient être avantageusement ré-utilisés à terme pour toute autre construction similaire de partition territoriale basée sur des textes juridiques (cartes scolaires, par exemple). De même, on pourrait inverser le processus en générant automatiquement des textes juridiques homogènes à partir de découpages géométriques, basés sur des critères objectifs de forme et de composition des bureaux de vote est ancrés dans les réseaux et les adresses des voies.

Dans un premier temps, nous allons nous atteler au nettoyage des géométries superflues en post-traitement. Jusqu'à maintenant, le fichier d'adresses ponctuelles ne nous a permis que de vérifier les adresses et l'adéquation des géométries. Nous allons maintenant directement l'exploiter pour tenter de réduire les géométries superflues (arêtes pendantes) au moment de la construction du polygone. S'il reste encore ce type de géométries, dès lors qu'il existe déjà un polygone fermé central, leur élimination est largement envisageable. C'est plus difficile lorsque le polygone n'est pas fermé. En ce qui concerne le « rebouchage » des trous, il est nécessaire que les géométries correspondantes soient essentiellement associées à une unique voie car dans ce cas, la géométrie peut être facilement complétée par un calcul de contiguïté. Dans le cas contraire, on pourra s'appuyer sur les limites des bureaux de vote adjacents ou éventuellement sur des calculs de plus courts chemins sur le réseau (hypothèses). Le problème de non uniformité des styles de textes (mots-clés, syntaxe) nous amène également à envisager des méthodes d'apprentissage adaptant la structure d'extraction d'information géographique dans les textes au contexte local et aux spécificités des rédacteurs des textes juridiques. Mais ne nous leurrons pas : il est très probable que la qualité de la partition obtenue ne puisse atteindre 100 % dans tous les cas. Dès lors, une approche manuelle pourra permettre de vérifier et compléter cette première base de données des bureaux de vote.

Si l'obtention d'une partition territoriale respectant parfaitement les textes juridiques décrivant les bureaux de vote est un objectif méthodologique d'importance, la façon dont cette partition « sert » le processus politique est un problème connexe qu'il ne faut pas éluder. En effet, la « qualité » d'un découpage électoral réside à la fois dans sa consistance géométrique, mais, et surtout, dans sa capacité à représenter les populations locales dans un processus électoral démocratique. Tout processus électoral, se déroulant par étapes successives et agrégeant des dominances politiques à différentes échelles emboîtées, reste intimement lié à la structure du découpage administratif des bureaux de vote. La forme et la taille de ces derniers, le nombre d'électeurs qu'ils regroupent et leur localisation respective dans l'espace des quartiers influencent directement les résultats mêmes des élections locales. Dès lors, la maîtrise des méthodes de construction de partition territoriale, incluant découpages électoraux, devient une arme à double tranchant. D'un côté, on peut rechercher à obtenir une partition la plus objective possible, s'appuyant sur des critères d'optimalité et de représentativité

politique dans le processus d'agrégation électorale. D'un autre côté, les découpages électoraux peuvent être construits à façon et biaiser fortement la représentativité électorale, en servant tel ou tel homme politique local. Rechercher la partition territoriale pertinente des bureaux de vote comme outil de mesure objectif des relations entre la société et le vote, construire des méthode d'association de ces sources d'information avec celles issues des découpages administratifs, connaître l'ampleur de l'influence de la partition territoriale sur les résultats électoraux et la marge de manœuvre dont disposent les acteurs (politiques, scientifiques, producteurs de données sociales) pour orienter ou évaluer objectivement le résultat électoral : tels sont quelques-uns des défis majeurs que nous identifions à l'interface entre en science politique et géographie quantitative.

Références

ACE. (2013). *The ACE encyclopaedia: Boundary delimitation. Technical report.* The Electoral knowledge network, www.aceproject.org (224 pages)

Beauguitte L., Colange C. (2013). Cartelec - Analyser les comportements électoraux à l'échelle du bureau de vote. Technical report. ANR.

Bilhaut F., Dumoncel F., Enjalbert P., Hernandez N. (2007). Indexation sémantique et recherche d'information interactive. le moteur géosem. In Proceedings of CORIA, Saint-Etienne, 28-30 mars 2007 , p. 65-76.

Braconnier, C. (2010), *Une autre sociologie du vote : les électeurs dans leurs contextes. Bilan critique et perspectives*, Paris : Lextenso/Laboratoire d'études juridiques et politiques (LEJEP).

Braconnier C., Dormagen J.Y. (2007), *La Démocratie de L'abstention : Aux origines de la démobilisation électorale en milieu populaire*. Folio Actuel. Folio,

Bussi M. (2007), « Pour une géographie de la démocratie », *L'Espace Politique* [En ligne], 1 | 2007-1, mis en ligne le 01janvier 2007 (http://espacepolitique.revues.org/243)

Fitzgerald M. (2012). *Introducing regular expressions*. Oreilly.

Friedl J. E. F. (2006). *Mastering regular expressions*. O'Reilly.

Gaio M., Nguyen V., Sallaberry C. (2012). Typage de noms toponymiques à des fins d'indexation géographiques. *Revue Traitement Automatique des Langues*, Vol. 53 (2), pp. 143-176.

Gaio M., Sallaberry C., Etcheverry P., Marquesuzaa C., Lesbegueries J. (2008). A global process to access documents' contents from a geographical point of view. *Journal of Visual Languages & Computing*, Vol. 19, pp. 3-23.

Garrigou A. (2002). *Histoire sociale du suffrage universel en france*, 1848-2000. Paris, Seuil.

Gaxie, D. (1996), *La démocratie représentative*, seconde édition, Paris, Montchrestien, Clefs politiques

Gombin J., Rivière J. (2012), La carte et le sondage, *Métropolitiques*, http://www.metropolitiques.eu/La-carte-et-le-sondage.html

Goodchild M., Jeansoulin R. (Eds.), (1998), *Data quality in geographic information*. Hermès.

Goyvaerts J., Levithan S. (2012). *Regular expressions cookbook.* O'Reilly.

Ihl O. (2002). Une ingéniérie politique. Augustin Cauchy et les élections du 23 avril 1848. *Genèses*, Vol. 49(4), pp. 4-28.

Jadot A., Bussi M., Colange C., Freire-Diaz S. (2010). Un outil d'analyse électorale en cours de création. CARTELEC, un SIG au niveau des bureaux de vote français. Le monde des cartes. *Revue*

du comité français de cartographie, Vol. 205, pp. 81-98.

Jardin A., (2014), « Le vote intermittent. Comment les ségrégations urbaines influencent-elles les comportements électoraux en Ile-de-France ? », *L'Espace Politique* [En ligne], 23 | 2014-2, mis en ligne le 04 juillet 2014, consulté le 02 juin 2015. URL : http://espacepolitique.revues.org/3082 ; DOI : 10.4000/espacepolitique.3082

Jones C. B., Purves R. S. (2008). Geographical information retrieval. *International Journal of Geographical Information Systems*, Vol. 22(3), pp. 219-228.

Josselin D., Bolot J., Chatonnay P. (2000). Optimisation de découpage territoriaux. Proposition de méthodes d'aggrégation spatiale dirigée. *Revue Internationale de Géomatique*, SIG, Aménagement du Territoire et Environnement, Cassini 2000 , Vol. 10, No. 3-4, pp. 383–409.

Josselin D., Janin C., Bolot J. (1999). Proposition d'une lecture terrotoriale des flux agricoles. *Revue de Géographie de l'Est*, Vol. TOME XXXIX - 4, pp. 207-216.

Lahire B., (1999), « De la théorie de l'habitus à une sociologie psychologique », in B. LAHIRE (dir.), *Le travail sociologique de Pierre Bourdieu. Dettes et critiques*, Parsi, Editions de la découverte.

Lehingue, P., *Le vote. Approches sociologiques de l'institution et des comportements électoraux*, Paris, La découverte, 2011, 287 p.

Leidner J. L., Lieberman M. D. (2011). Detecting geographical references in the form of place names and associated spatial natural language. *SIGSPATIAL Special*, Vol. 3, No. 2, pp. 5-11.

Manning C. D., Raghava P., Schutze H. (2008). *Introduction to information retrieval.* Cambridge University Press.

Martins B., Anastacio I., Calado P. (2010). A machine learning approach for resolving place references in text. In *Lecture notes in geoinformation and cartography*. Springer Verlag.

Ravenel (Loïc), Buléon (Pascal), Fourquet (Jérôme), « Vote et distances aux villes lors des présidentielles 2002 », *Espaces, Population, Société*, 2003, n° 3, p. 469-482.

Sallaberry C. (2013). *Geographical information retrieval in textual corpora* (FOCUS Geographical information systems series, Ed.). ISTE WILEY.

Tien Nguyen V., Gaio M., Sallaberry C. (2009). Recherche de relations spatio-temporelles : une méthode basée sur l'analyse de corpus textuels. In TIA'09 Toulouse 18-20 novembre 2009.

Watt A. (2005). *Beginning regular expressions*. WROX.

Widlocher A., Faurot E., Bilhaut F. (2004). Multimodal indexation of contrastive structures in geographical documents. In Actes RIAO 2004, Avignon, p. 550-570.