HAL
open science

Mise en correspondance de données textuelles hétérogènes fondée sur la dimension spatiale

Jacques Fize

To cite this version:

Jacques Fize. Mise en correspondance de données textuelles hétérogènes fondée sur la dimension spatiale. Autre [cs.OH]. Université Montpellier, 2019. Français. NNT : 2019MONTS099 . tel-02494342

HAL Id: tel-02494342

https://theses.hal.science/tel-02494342

Submitted on 28 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Informatique

École doctorale Information Structures Systèmes (I2S)

Unité de recherche UMR TETIS

Mise en correspondance de données textuelles hétérogènes fondée sur la dimension spatiale

Présentée par Jacques Fize

Le 12 novembre 2019

Sous la direction de Mathieu Roche et Maguelonne Teisseire

Devant le jury composé de

Julien Velcin, Professeur, Université Lumière Lyon 2	Rapporteur
Christian Sallaberry, Université de Pau et des Pays de l'Adour	Rapporteur
Célia de Costa Pereira, Maître de Conférences, Université Nice Sophia Antipolis	Examinatrice
Madalina Croitoru, Professeur, Université de Montpellier	Examinatrice
Mathieu Roche, Chercheur, TETIS, CIRAD	Co-Directeur
Maguelonne Teisseire, Chercheur, TETIS, IRSTEA	Co-Directrice

UNIVERSITÉ DE MONTPELLIER

Jacques Fize : *Mise en correspondance de données textuelles hétérogènes fondée sur la dimension spatiale*, , © 11 février 2020

I'm still standing after all this time

— Elton John

À ma famille.

RÉSUMÉ

Avec l'essor du Big Data, le traitement du Volume, de la Vélocité (croissance et évolution) et de la Variété de la donnée concentre les efforts des différentes communautés pour exploiter ces nouvelles ressources. Ces nouvelles ressources sont devenues si importantes, que celles-ci sont considérées comme le nouvel « or noir ». Au cours des dernières années, le volume et la vélocité sont des aspects de la donnée qui sont maitrisés contrairement à la variété qui elle reste un défi majeur. Cette thèse présente deux contributions dans le domaine de mise en correspondance de données hétérogènes, avec un focus sur la dimension spatiale.

La première contribution repose sur un processus de mise en correspondance de données textuelles hétérogènes divisé en deux étapes : la *georepresentation* et le *geomatching*. Dans la première phase, nous proposons de représenter la dimension spatiale de chaque document d'un corpus à travers une structure dédiée, la Spatial Textual Representation (STR). Cette représentation de type graphe est composée des entités spatiales identifiées dans le document, ainsi que les relations spatiales qu'elles entretiennent. Pour identifier les entités spatiales d'un document et leurs relations spatiales, nous proposons une ressource dédiée, nommée Geodict. La seconde phase, le *geomatching*, consiste à mesurer la similarité entre les représentations générées (STR). S'appuyant sur la nature de la structure de la STR (i.e. graphe), différents algorithmes de *graph matching* ont été étudiés. Pour évaluer la pertinence d'une correspondance, nous proposons un ensemble de 6 critères s'appuyant sur une définition de la similarité spatiale entre deux documents.

La seconde contribution repose sur la dimension thématique des données textuelles et sa participation dans le processus de mise en correspondance spatiale. Nous proposons d'identifier les thèmes apparaissant dans la même fenêtre contextuelle que certaines entités spatiales. L'objectif est d'induire certaines des similarités spatiales implicites entre les documents. Pour cela, nous proposons d'étendre la structure de la STR à l'aide de deux concepts : l'*entité thématique* et de la *relation thématique*. L'entité thématique représente un concept propre à un domaine particulier (agronome, médical) et représenté selon différentes orthographes présentes dans une ressource terminologique, ici un vocabulaire. Une relation thématique lie une entité spatiale à une entité thématique si celles-ci apparaissent dans une même fenêtre contextuelle. Les vocabulaires choisis ainsi que la nouvelle forme de la STR intégrant la dimension thématique sont évalués selon leur couverture sur les corpus étudiés, ainsi que leurs contributions dans le processus de mise en correspondance spatiale.

ABSTRACT

With the rise of Big Data, the processing of Volume, Velocity (growth and evolution) and data Variety concentrates the efforts of communities to exploit these new resources. These new resources have become so important that they are considered the new "black gold". In recent years, volume and velocity have been aspects of the data that are controlled, unlike variety, which remains a major challenge. This thesis presents two contributions in the field of heterogeneous data matching, with a focus on the spatial dimension.

The first contribution is based on a two-step process for matching heterogeneous textual data : *georepresentation* and *geomatching*. In the first phase, we propose to represent the spatial dimension of each document in a corpus through a dedicated structure, the Spatial Textual Representation (STR). This graph representation is composed of the spatial entities identified in the document, as well as the spatial relationships they maintain. To identify the spatial entities of a document and their spatial relationships, we propose a dedicated resource, called Geodict. The second phase, *geomatching*, computes the similarity between the generated representations (STR). Based on the nature of the STR structure (i.e. graph), different algorithms of *graph matching* were studied. To assess the relevance of a match, we propose a set of 6 criteria based on a definition of the spatial similarity between two documents.

The second contribution is based on the thematic dimension of textual data and its participation in the spatial matching process. We propose to identify the themes that appear in the same contextual window as certain spatial entities. The objective is to induce some of the implicit spatial similarities between the documents. To do this, we propose to extend the structure of STR using two concepts : the *thematic entity* and the *thematic relationship*. The thematic entity represents a concept specific to a particular field (agronomic, medical) and represented according to different spellings present in a terminology resource, in this case a vocabulary. A thematic relationship links a spatial entity to a thematic entity if they appear in the same window. The selected vocabularies and the new form of STR integrating the thematic dimension are evaluated according to their coverage on the studied corpora, as well as their contributions to the heterogeneous textual matching process on the spatial dimension.

PUBLICATIONS

Vous pourrez trouver ci-dessous une liste des publications des travaux effectués durant la thèse :

JOURNAL

Fize, Jacques, Mathieu Roche et Maguelonne Teisseire (2019a). « Exploitation de l'hétérogénéité dans les données textuelles. Utilisation de données produites à Madagascar ». fre. In : *Recherche d'information, Document et Web Sémantique* 2.1, p. 5.

Fize, Jacques, Mathieu Roche et Maguelonne Teisseire (2019c). « Matching Heterogeneous Textual Data Using Spatial Features ». en. In : *Intelligent Data Analysis (Accepté le 9 juillet 2019)*, p. 24.

ACTES DE CONFÉRENCES ET WORKSHOPS INTERNATIONAUX

Fize, Jacques, Mathieu Roche et Maguelonne Teisseire (2018). « Gemedoc : A Text Similarity Annotation Platform ». In : *Natural Language Processing and Information Systems*. Sous la dir. de Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais et Farid Meziane. Cham : Springer International Publishing, p. 333–336.

Fize, Jacques, Mathieu Roche et Maguelonne Teisseire (2019b). « Mapping Heterogeneous Textual Data : a Multidimensional Approach based on Spatiality and Theme ». In : *Proceedings of the 6th International Conference on INTERNET SCIENCE*. Perpignan, France (Accepté le 21 août 2019), p. 8.

Fize, Jacques et Gaurav Shrivastava (2017). « GeoDict : an integrated gazetteer ». In : *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017) at the Conference on Computational Semantics (IWCS)*. Montpellier, France, p. 11.

Fize, Jacques, Maguelonne Teisseire et Mathieu Roche (2018). « Matching Heterogeneous Textual Data Using Spatial Features ». en. In : *SSTDM 2018 : International Workshop on Spatial and Spatiotemporal Data Mining at the 18th IEEE International Conference on Data Mining (ICDM 18)*, p. 1389–1396.

ACTES DE CONFÉRENCES ET WORKSHOPS NATIONAUX

Fize, Jacques, Mathieu Roche et Maguelonne Teisseire (2017). « Spatial Textual Representation (STR) ou comment représenter la spa-

tialité des donnnées textuelles ». In : *Spatial Analysis and GEOmatics 2017*. INSA de rouen. Rouen, France, p. 14.

Fize, Jacques, Maguelonne Teisseire et Mathieu Roche (2017). « Gemedoc : Un outil pour annoter les correspondances entre les documents ». fr. In : *EXCES - EXtraction de Connaissances à partir de donnéEs Spatialisées, Spatial Analysis and GEOmatics 2017*, p. 6.

Lucile, Sautot, Eric Chraibi, **Fize, Jacques**, Sébastien Peillet, Ludovic Journaux et Flavie Cernesson (2019). « Qui a peur du changement climatique ? » In : *Actes de la conférence SAGEO'2019 (Spatial Analysis and GEOmatics), Clermont Ferrand, Novembre 2019 (Accepté le 26 juillet 2019)*, 12 p.

REMERCIEMENTS

Trois ans ! Au début du doctorat, cela semblait si loin ! Puis, trois années se sont écoulées, si vite. Et pourtant, le grand jour est arrivé, avec lui l'heure du bilan des souvenirs, des personnes rencontrées et des réalisations effectuées. Même si cela peut sembler cliché, ces trois années furent pour moi une expérience enrichissante tant au niveau professionnel que personnel.

Tout d'abord, je tiens à remercier mes deux directeurs de thèse Mathieu Roche et Maguelonne Teisseire. Je vous remercie de m'avoir soutenu, mais aussi supporté pendant ces trois ans. Merci d'avoir été présents activement durant ce voyage initiatique dans le monde de la recherche. Merci pour les bons moments passés au cours de voyages, soirées organisées ou simplement autour d'une boisson chaude et de sucreries venant des quatre coins du monde. Merci à tous les deux pour votre gentillesse et votre bienveillance.

Aux membres de mon jury, je souhaite remercier Julien Velcin (Rapporteur) et Christian Sallaberry (Rapporteur) d'avoir validé ces travaux et de m'avoir permis de soutenir. Je remercie Célia de Costa Pereira (Examinatrice) ainsi que Madalina Croitoru (Examinatrice et présidente du jury) qui m'ont fait l'honneur de faire partie de mon jury.

Comme le dit Bilbon, "Trois ans, ce fut un temps un peu court à passer en compagnie de si excellents et si admirables chercheurs !". Si le doctorat est une période difficile, j'ai passé d'excellents moments en compagnie des chercheurs, doctorants, stagiaires, et personnels de la MTD.

Pour commencer, je tiens à remercier mes collègues de bureau. Un grand merci à toi Dav pour ses deux ans de colocation. J'espère que mon aide en Python à fait de toi un véritable *Pythoneer* ;) Nous nous sommes rencontrés sur la dernière et difficile année de la thèse, merci Christian pour les bons moments passés à Prades-le-Lez au bord de la garrigue. Merci de m'avoir partagé l'amour que tu portes à ton pays, le Costa Rica (j'arrive bientôt !). Enfin, même si on ne se connait que depuis un an, merci Eric. J'ai adoré partager avec toi recettes de cuisines, musiques, jeux et parties de Smash.

Autrefois padawans, maintenant jedi, vous êtes les prochains docteurs. Merci Arthur pour les bons moments passés, pour ton aide sur l'analyse de données spatiales. Merci Sarah V. pour ta bonne humeur, tes fous rires inopinés et les petits gâteaux que tu ramenais dans le bureau. Enfin, Guilhem, merci l'artiste ;)

N'oublions pas les maîtres jedi ! Un grand merci à Marky Mark (a.k.a Marc). Grâce à toi, j'ai pu entrevoir en avant-première ce qu'était la dernière année de thèse. Merci pour ton aide mais aussi pour les bons moments autour d'une cervoise ou bien d'une partie de Smash. Merci à Lucile ! Merci pour tous ces moments de franches rigolades tant sur des sujets politiques que tes cochons d'inde. Merci pour l'aide que tu m'as apporté. Merci de m'avoir initié aux joies du potager. Et même si le "potager des informaticiens" n'a pas donné beaucoup de récoltes, je pense que l'on peut être fier d'avoir réussi à faire pousser Ze melon. Merci Sarah Z. et Linda pour les bons moments mais aussi de m'avoir fait découvrir les spécialités culinaires du Maghreb.

Enfin, merci aux padawans Milo, Martin, Hugo. Merci aux contrebandiers : Sebastien, Gaurav, Vincent, Maxime. (Promis, j'en ai fini avec les analogies Star Wars ! ;))

Du côté du CIRAD, je n'oublie pas Sophie avec qui j'ai beaucoup discuté et traité de la publication des données de recherches. Merci Annie, Nathalie et Laurence de m'avoir guidé dans les méandres administratifs, toujours avec gentillesse et bienveillance (et sucreries :)). Merci Jean-Philippe Tonneau pour sa franchise et son aide dans les travaux entrepris sur les données de Madagascar. Merci Christian, alias le Baron, pour nous avoir partagé ta bonne humeur.

Je n'oublie pas les copains de toujours : Jean-Alexis, Tanguy, Pia, Bastien, Luiz !

Les meilleurs pour la fin ? :) Je tiens à remercier ma famille que j'aime et je les remercie d'avoir été là pour m'aider dans les bons et les mauvais moments. Merci papa d'avoir toujours pris le temps de m'aider et m'accompagner dans mes doutes alors que tu croulais sous le travail. Merci maman de m'avoir aidé dans la correction de ce manuscrit, mais aussi pour ta gentillesse et ta bonne humeur contagieuse. Enfin, merci ma soeur pour ta gentillesse, ton empathie mais aussi pour les récits de tes expériences quotidiennes abracadabrandesques. Cette thèse, je vous la dédie.

Merci. Jacques

TABLE DES MATIÈRES

Introduction

1 INTRODUCTION 3
1.1 Contexte 3
1.2 Problématiques 4

I DIMENSION SPATIALE

1 MISE EN CORRESPONDANCE SPATIALE DE DONNÉES TEXTUELLES HÉTÉROGÈNES 13
1.1 Mise en correspondance de données 13
1.2 Applications du processus de mise en correspondance proposée dans le domaine du Traitement Automatique des Langues Naturelles 15
1.3 Processus de mise en correspondance proposé 16
1.4 Une correspondance sur la dimension spatiale : définition de la similarité 17
2 GEOREPRESENTATION 21
2.1 Représentation de la spatialité 21
2.2 Geoparsing : Identification des entités spatiales dans les documents 31
2.3 Geocompletion : Extraction des relations spatiales . . . 39
2.4 Transformation de la STR 41
3 GEODICT : UN NOUVEL INDEX GÉOGRAPHIQUE 45
3.1 Sources utilisées pour construire GEODICT 46
3.2 Processus de création de Geodict 47
3.3 Statistiques 49
3.4 Conclusion et perspectives 50
4 GEOMATCHING : APPARIEMENT DES STR 51
4.1 La théorie des graphes 51
4.2 Le Graph Matching 53
4.3 Algorithmes Structure-based 54
4.4 Algorithmes Pattern-based 56
5 PROTOCOLES D'ÉVALUATION ET EXPÉRIMENTATIONS 63
5.1 Données utilisées 63
5.2 Évaluation de la georepresentation 65
5.3 Évaluation de la mise en correspondance : GeoMatching 68
5.4 Résultats et Discussions 78
5.5 Discussions 85

II DIMENSION THÉMATIQUE

1 INTÉGRATION DE L'INFORMATION THÉMATIQUE DANS LA STR 95

1.1 Exploiter l'information contextuelle entre thème et espace 96
1.2 Entité thématique 99
1.3 Relation thématique 112
2 RÉSULTATS ET DISCUSSIONS 123
2.1 Couverture des terminologies 123
2.2 Mise en correspondance spatiale 125
2.3 Discussions 126

III CONCLUSION ET PERSPECTIVES
1 CONCLUSION 133
2 PERSPECTIVES 135
2.1 Évolution des représentations pour la mise en correspondance 135
2.2 Visualisation des correspondances à l'analyse de corpus 136
2.3 Évaluation des contributions par un comité d'experts . 136

Appendix
A APPENDIX 139
A.1 Classification de Geonames 139
A.2 Outils développées 140
A.3 Algorithmes 143
A.4 Enquête auprès des experts 147
A.5 Enquête auprès des experts 148
A.6 Enquête auprès des experts 149
A.7 Enquête auprès des experts 150
A.8 Enquête auprès des experts 151
A.9 Enquête auprès des experts 152
A.10 Enquête auprès des experts 153
A.11 Enquête auprès des experts 154
A.12 Enquête auprès des experts 155
A.13 Enquête auprès des experts 156
A.14 Enquête auprès des experts 157

BIBLIOGRAPHIE 159

TABLE DES FIGURES

Figure 1 Hétérogénéité dans les données textuelles. . . 5
Figure 2 Illustration d'un réseau de correspondances entre les documents du corpus *AgroMada*. . . . 7
Figure 3 Applications générale des méthodes de mise en correspondance de données 14
Figure 4 Processus usuel de mise en correspondance de données. 15
Figure 5 Processus de mise en correspondance spatiale proposé. 17
Figure 6 Projection de points affichée sur l'interface pour l'experience de Montello et al., 2003. 19
Figure 7 Échelle utilisée dans l'expérimentation de Montello et al., 2003. 19
Figure 8 Illustration du *feature-effect* décrit par Montello et al., 2003. 20
Figure 9 Dimension multi-échelle spatiale possible à l'aide des relations d'inclusion (arc rouge) et d'adjacence (arc vert). 25
Figure 10 Informations associées à l'entité spatiale Paris. 27
Figure 11 Illustration de la grammaire de logique spatiale : RCC-8 (Randell, Cui et Cohn, 1992). . 27
Figure 12 Illustration de la relation d'inclusion. 28
Figure 13 Relation d'adjacence. 28
Figure 14 Exemple d'une Spatial Textual Representation générée à partir d'un document du corpus *AgroMada* (*c.f.* Section 5.1.2). Les arcs verts correspondent aux relations d'adjacence et les arcs rouges correspondent aux relations d'inclusion. 30
Figure 15 Processus de Georepresentation. 30
Figure 16 Fonctionnement global d'un algorithme de Geoparsing. Source : (Gritta et al., 2018). 31
Figure 17 Illustration de l'algorithme MostCommon. L'entité sélectionnée par l'algorithme est encadrée en rouge. 35
Figure 18 Illustration du fonctionnement de l'algorithme SharedProp sur le toponyme *Paris*. En rouge, le toponyme *France* qui n'est associé qu'à une entité spatiale. L'entité sélectionnée par l'algorithme est encadrée en rouge. 36
Figure 19 Calcul du score d'inclusion pour chaque entité candidate d'un toponyme ambiguë (Paris). . . 37

Figure 20 Graphe de cooccurrence utilisée dans l'algorithme de désambiguïsation WikiCooc. 38
Figure 21 Illustration du fonctionnement de l'algorithme WikiCooc. L'entité sélectionnée par l'algorithme est encadrée en rouge. 38
Figure 22 Exemple de positionnement de différentes classes d'entités spatiales selon leur granularité (de la plus fine à la plus grossière). 42
Figure 23 Une STR (à gauche) et sa forme généralisée (à droite). La flèche grise relie une entité à l'entité qui la remplace au cours du processus de généralisation. Ici, la limite de la généralisation est fixée au niveau départemental. 43
Figure 24 Une STR (à gauche) et sa forme étendue (à droite) avec les paramètres (n=1 et r=50km). . 44
Figure 25 Exemple de clique. 52
Figure 26 Appariement exact de graphe. 53
Figure 27 Appariement inexact de graphe. 54
Figure 28 Calcul de la GED $BP(G_1, G_2)$ entre deux graphes G_1 et G_2 à l'aide de la matrice de coût C_{G_1,G_2}. 57
Figure 29 Exemple de représentation sac-de-mots. 58
Figure 30 Représentation *Bag of Cliques* (à droite) d'un graphe (à gauche). 58
Figure 31 Carte des lieux mentionnés dans le corpus *PadiWeb*. 64
Figure 32 Entités spatiales et leur fréquence d'apparition dans le corpus *AgroMada*. 65
Figure 33 Protocole d'évaluation de correspondances. . . 69
Figure 34 Illustration du critère Entités Partagées (ESP). 72
Figure 35 Illustration de l'application du critère d'Entités Proches (EP). 72
Figure 36 Illustration du critère d'Emprise Spatial Significative (ESS). 73
Figure 37 Illustration du critère d'Emprise Spatial Stricte (ESSC). 73
Figure 38 Illustration du critère de Proximité Moyenne (PM). 73
Figure 39 Illustration du critère PEP. Ici la valeur du critère $PEP = 2 * \frac{8}{4+5} = 0.88$. 74
Figure 40 Valeurs cumulées de MAP@n de la combinaison dominante pour chaque mesure de similarité évaluée (*PadiWeb*). 81
Figure 41 Valeurs cumulées de MAP@n de la combinaison dominante pour chaque mesure de similarité évaluée (*AgroMada*). 82
Figure 42 Illustration de la sparsité de certaines STR. . . 86

FIGURE 43 Distributions des classes des entités présentes dans les différents corpus. 87
FIGURE 44 Exemple de correspondances entre des doublons retournées par ClassicBOW. 88
FIGURE 45 Illustration de la différence entre contexte local et global. 96
FIGURE 46 Illustration de relation récurrente entre thème et espace dans un corpus. 98
FIGURE 47 Informations partagées entre deux documents en prenant conjointement les dimensions thématique et spatiale. 99
FIGURE 48 Interaction entre les trois méthodes d'extraction de vocabulaire de corpus. 106
FIGURE 49 Valeurs de cohérence obtenues sur différents nombres de *topics* extraits par la LDA sur le corpus *AgroMada*. 108
FIGURE 50 Triptyque commun dans le domaine d'extraction de relation. 113
FIGURE 51 Exemple d'arbre de dépendance généré à partir de la phrase "Le chat a griffé le chien". 114
FIGURE 52 Extrait du document avec une table dans sa forme originale et brute (texte). 115
FIGURE 53 Exemple de relation thématique entre l'entité spatiale *Lac Alaotra* et l'entité thématique *SCV*. 116
FIGURE 54 Exemple de relation thématique entre deux entités spatiales 116
FIGURE 55 Exemple d'une STR intégrant les relations thématiques. 120
FIGURE 56 Propogation de l'information thématique sur la forme généralisée de la STR 121
FIGURE 57 Propagation de l'information thématique sur la forme étendue de la STR. 122
FIGURE 58 Gains de temps obtenus sur l'implémentation Cython des algorithmes implémentés dans GMatch4py. 141

LISTE DES TABLEAUX

Table 1 Données disponibles pour chaque entrée de Geodict. Chaque variable est accompagnée de sa source et d'un exemple de valeur. 46
Table 2 Comparaison avec d'autres index géographiques. 49
Table 3 Comparaison du nombre d'entités dans Geodict et *Geonames*. La comparaison est effectuée à l'aide des classes générales définies par *Geonames* et utilisées dans Geodict. 49
Table 4 Performances des différentes méthodes de NER (Reconnaissance d'entités nommées) sur le *geotagging* sur les deux échantillons de corpus. . . 78
Table 5 Performances des différents algorithmes de désambiguïsation (*geocoding*) sur les deux échantillons des corpus. 79
Table 6 Les combinaisons (type de STR, mesure de similarité) les plus performantes selon différentes pondérations de critères (*AgroMada*). 83
Table 7 Les combinaisons (type de STR, mesure de similarité) les plus performantes selon différentes pondérations de critères (*PadiWeb*). 84
Table 8 Valeurs moyennes obtenues pour chaque critère selon la forme de la STR (*AgroMada*). 89
Table 9 Valeurs moyennes obtenues pour chaque critère selon la forme de la STR (*PadiWeb*). 89
Table 10 Choix de mesures en fonction des besoins de mise en correspondances. 90
Table 11 Exemple de requêtes proposés par les systèmes d'auto-complétion des deux moteurs de recherche : Google et Qwant (date : 24 Juin 2019). 97
Table 12 Informations sur l'entité thématique *parcelle* sur *Wikidata* 100
Table 13 Ensemble des vocabulaires utilisés pour la détection d'entités thématiques dans les corpus étudiés. 101
Table 14 Extrait de la terminologie provenant du dictionnaire de l'agroécologie. 103
Table 15 Extrait de la terminologie sur les maladies infectieuse. 103
Table 16 Extrait du Dictionnaire du Développement Durable. 104
Table 17 Exemple de distributions de probabilités sur les mots dans différents *topics* retournés par la LDA. 107

Table 18 Le mot-clé *Viral infections* apparait dans le motif NOM + VERBE + NOM_2 + ADP. 108
Table 19 Mots-clés extraits avec l'outil Biotex sur le document illustré dans la Figure 46. 109
Table 20 Comparaison entre les résultats obtenues à l'aide du modèle de la LDA avec et sans l'utilisation des mots-clés extraits par Biotex. 110
Table 21 Extrait du vocabulaire obtenu par Biotex sur les corpus de documents de *PadiWeb*. 111
Table 22 Extrait du vocabulaire du corpus *AgroMada* obtenu par Biotex. 112
Table 23 Extrait du vocabulaire obtenu par l'approche BiotexLDA sur *AgroMada*. 112
Table 24 Parcours d'un document à l'aide d'une fenêtre de taille fixe. 117
Table 25 Statistiques de couverture des différents vocabulaires choisis pour le corpus *PadiWeb*. 124
Table 26 Statistiques de couverture des différents vocabulaires choisis pour le corpus *AgroMada*. . . . 124
Table 27 Statistiques de couverture obtenues sur les STRs extraites des documents de *PadiWeb*. 125
Table 28 Statistiques de couvertures obtenues sur les STRs extraites des documents d'*AgroMada*. . . 125
Table 29 Les combinaisons (type de STR, mesure de similarité) les plus performantes selon différentes pondérations de critères (*AgroMada*). 127
Table 30 Les combinaisons (type de STR, mesure de similarité) les plus performantes selon différentes pondérations de critères (*PadiWeb*). 128
Table 31 Extrait de la classification utilisé par Geonames 139
Table 32 Algorithmes implémentés dans GMatch4Py . 142

LISTINGS

Listing 1 Exemple de fichier de configuration de GEODICT 143

ABRÉVIATIONS

EP Entités Proches
ESP Entités Spatiales Partagés
ESS Emprise Spatiale Significative
ESSC Emprise Spatiale Stricte
GIR Geographic Information Retrieval
GED Graph Edit Distance
QSR Qualitative Spatial Representation
QCN Qualitative Constraint Network
SIG Système d'Information Géographique
STR Spatial Textual Representation
LDA Latent Dirchlet Analysis
LSA Latent Semantic Analysis
MCS Most Common Subgraph
PM Proximité Moyenne
PEP Pourcentage d'Entités spatiales Partagées

INTRODUCTION

1 INTRODUCTION

1.1 CONTEXTE

Durant les dernières décennies, nous avons assisté à la diffusion des systèmes d'informations dans les différentes communautés scientifiques, industrielles et plus récemment, avec plus d'intensité, dans la sphère publique. À la base de ces systèmes d'informations, la Donnée, qui par son volume et sa variété présente de nouvelles opportunités mais aussi de nouveaux défis. Ces nouveaux volumes de données sont caractérisés généralement par les 3V du *Big Data* (Gandomi et Haider, 2015 ; Sagiroglu et Sinanc, 2013) : le Volume, la Vélocité et la Variété. La Volumétrie et la Vélocité reposent sur l'actualisation, le stockage, et le transfert de données. Différents domaines tels que le *datawarehousing* se spécialisent dans l'optimisation du stockage et de l'accès à ces volumes. À travers la diversité de domaines, différentes structures de données sont produites. Cette diversité est représentée par le troisième V du BigData, la Variété ou Hétérogénéité de la Donnée. Dans les travaux menés au cours de cette thèse, nous nous sommes concentrés sur la dimensions hétérogène de la donnée.

La définition de l'hétérogénéité de la donnée est très large et dépend de son contexte d'utilisation. Plusieurs contributions scientifiques s'appuient sur l'utilisation de données hétérogènes au sein de modèles d'analyse. Par exemple, les modèles de prédiction météorologique reposent sur différentes sources de données (i.e. capteurs) avec une fréquence de capture et des unités de mesures différentes. Dans cet exemple, l'hétérogénéité est fondé sur la diversité des variables utilisées. Dans nos travaux, nous nous intéressons sur la mise en correspondance de données textuelles hétérogènes. Plus particulièrement, nous nous focalisons sur la mise en correspondance de données hétérogènes selon différents aspects des données : thématique, spatial et temporel. Ici, l'hétérogénéité est associée à la variété du conteneur, du structurant de la données. Nous caractérisons comme données textuelles hétérogènes, toutes données représentées à l'aide de chaînes de caractères, ce qui inclue : articles de presse, articles scientifiques, tableaux, ...

1.2 PROBLÉMATIQUES

Nos travaux se concentrent sur la mise en correspondance de données textuelles hétérogènes. Durant cette thèse, les différentes contributions doivent répondre à deux problématiques suivantes : **la mise en correspondance de données** et **la prise en compte de l'hétérogénéité de la donnée**.

1.2.1 *Mise en correspondances de données*

La mise en correspondance de données consiste à mettre en relation deux unités de données faisant référence à un ou plusieurs mêmes objets (Christen, 2012). Dans nos travaux, nous nous focalisons sur la mise en correspondance de données hétérogènes selon trois de leur propriétés ou dimensions : la *thématique*, la *spatialité* et la *temporalité*. La dimension thématique regroupe les sujets abordés dans la donnée. La dimension spatiale se concentre sur la configuration spatiale de la donnée. Enfin, la dimension temporelle situe la donnée dans le temps.

L'intégration de l'hétérogénéité dans le processus de mise en correspondance permet d'augmenter la couverture de données exploitables, notamment pour des systèmes de Recherche d'Information (Salton et Buckley, 1991 ; Shaw, Mian et Yadav, 2002). Dans le cas des données textuelles, différents domaines d'applications comme la génération de résumé automatique (Allahyari et al., 2017 ; Portet et al., 2009), les systèmes de Question-Réponse (Voorhees, 1999, 2001, 2003), ou encore l'analyse de corpus (Kuang, Choo et Park, 2015 ; Mehler, 2008) bénéficient de telles contributions.

La prise en compte du caractère hétérogène des données engendre des verrous scientifiques qui sont étudiées dans cette thèse.

1.2.2 *Données textuelles hétérogènes*

L'hétérogénéité d'une entité est définie [1] par la nature des différents éléments la composant. Dans nos travaux, les données hétérogènes regroupent plusieurs unités de données (fichier, entrée dans une base de données, page web) chacune différenciée selon plusieurs caractéristiques : *la structure, le format, la précision et la source*.

La **structure** correspond à la disposition de l'information dans une unité de données. Par exemple, une structure {abstract, introduction, état de l'art, méthodes, résultats, discussion, conclusion} est récurrente dans la littérature scientifique. Le **format** correspond à la méthode de représentation de la donnée dans la mémoire de l'ordinateur. Le format regroupe la compression de données, l'utilisation de format propriétaire ou ouvert, etc. La **précision** correspond à l'information fournie par la donnée. Par exemple, il existe en cartographie plusieurs

1. Source : https://www.cnrtl.fr/definition/hétérogène

versions des données nécessaires pour dessiner les contours d'un pays. La **source** correspond à la provenance de la donnée.

Dans nos travaux, nous travaillons sur une catégorie spécifique des données, les données textuelles. Les données textuelles sont regroupées dans plusieurs ensembles appelés **corpus**. Pour définir les unités de données traitées dans nos travaux, nous proposons d'utiliser la définition proposée dans(FREED et BORENSTEIN, 1992) : "Un texte est utilisé pour représenter une information textuelle à l'aide de plusieurs combinaisons de caractères (lettre, symbole) définies par un language" [1]. Pour plus de facilité, nous utiliserons la notion de *document* pour désigner les unités de données textuelles mises en correspondance. De manière simplifiée, nous considérons un document comme une séquence de *token*, i.e. mots.

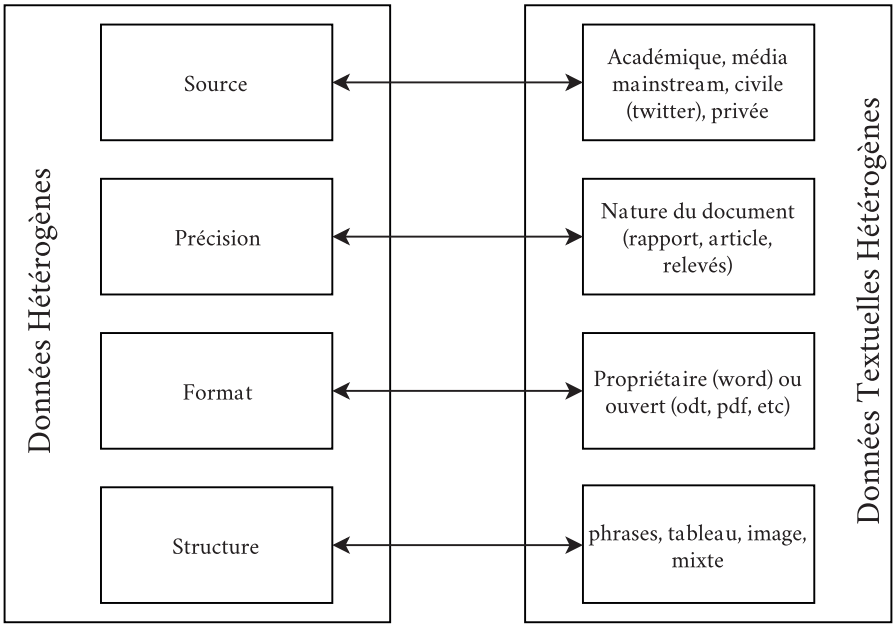


FIGURE 1 – Hétérogénéité dans les données textuelles.

Reprenant les notions précédentes, l'hétérogénéité dans les données textuelles s'exprime de différentes manières (voir Figure 1). La structure correspond à l'organisation des données dans un document. Par exemple, une brève (*news*) est généralement représentée par un enchaînement de paragraphes (texte brut), contrairement à un rapport technique intégrant en plus des tableaux, graphiques, etc. Lié à ce dernier, le format d'un document est associé à sa compression, son stockage en mémoire et sa mise en forme (*language de balisage*). Une autre dimension du format de la donnée réside dans l'accessibilité de la donnée brute avec les formats propriétaires (*.doc* de MICROSOFT) ou ouverts (*OpenDocument* initié par OPENOFFICE). Dans un document, la précision s'appuie sur la nature du document et la variété d'infor-

*(FREED et BORENSTEIN, 1992) définit un **texte brut**, comme compréhensible sans l'utilisation préalable d'un programme informatique.*

1. En anglais : A text is used to represent textual information in a number of character sets and formatted text description languages in a standardized manner.

mations présentes. Un *tweet* permet de diffuser moins d'informations (248 caractères) qu'un article scientifique. La source d'un document indique sa provenance. La source d'un document peut aussi contraindre le format, la précision et la structure d'un texte.

CONTRIBUTION ET PLAN DE LA THÈSE

Au cours de cette thèse, deux contributions ont été développées autour du processus de mise en correspondance de données textuelles hétérogènes. La première contribution repose sur un processus de mise en correspondance sur la dimension spatiale des données textuelles hétérogènes. La seconde contribution se concentre sur l'aspect thématique de la données et son intégration dans le processus de mise en correspondance spatiale.

La première contribution repose sur la conception d'un **processus de mise en correspondance spatiale** en deux étapes. À la base de ce processus, nous proposons une représentation de la spatialité d'un document, la Spatial Textual Representation, qui s'appuie sur les entités spatiales identifiées (lieux) et les relations spatiales (adjacence et inclusion) qu'elles entretiennent. Pour palier l'absence d'information spatiale qui peut nuire à l'estimation de la similarité spatiale entre deux documents, différentes transformations des entités spatiales sont proposées. Pour mesurer la similarité entre les représentations générées à partir des documents d'un corpus, nous proposons d'utiliser des algorithmes provenant du domaine d'appariement de graphe, ou *graph matching*. Chacune des étapes du processus de mise en correspondance, de la génération de la représentation d'un document à sa mise en correspondance est évaluée sur différents jeux de données ou corpus. Pour s'assurer de la généricité de notre approche, l'intensité de l'hétérogénéité des jeux de données choisis est variée. Dans notre cas, l'intensité de l'hétérogénéité correspond au nombre de groupes de documents représentés par une structure, un format, une source et une précision particulière. À l'aide d'un outil de visualisation, la Figure 2 illustre un extrait du réseau des correspondances identifiées entre des documents à l'aide de l'approche proposée.

Dans la seconde contribution, nous proposons **d'étendre l'information contenue dans la STR en s'appuyant sur la dimension thématique** des données. Plus spécifiquement, nous nous intéressons aux transferts de propriétés spatiales entre une entité spatiale et une thématique selon leur proximité contextuelle. Par exemple, dans le contexte de données agronomiques, certaines thématiques associées aux techniques d'agricoles peuvent être associées à un lieu particulier. Pour cela, nous proposons une nouvelle forme de la STR intégrant deux nouveaux concepts qui sont l'entité thématique et la relation thématique. L'entité thématique correspond à un concept appartenant à un

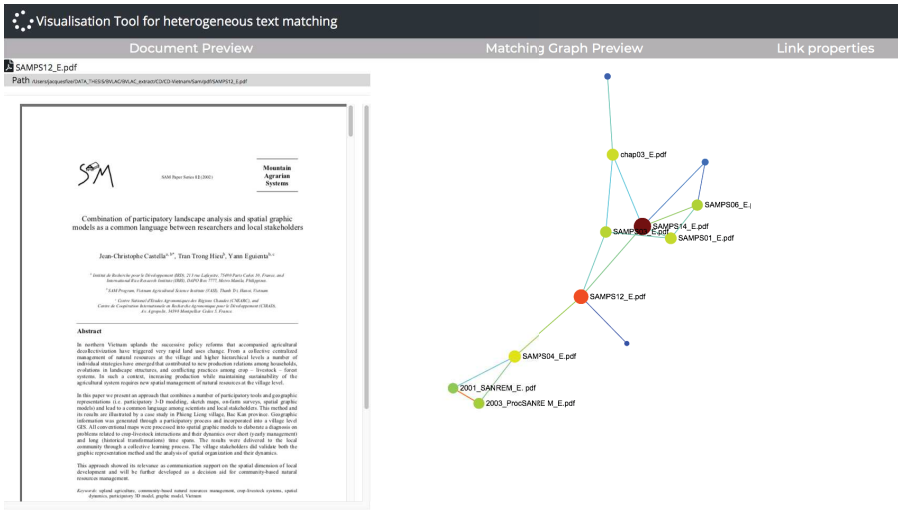


FIGURE 2 – Illustration d'un réseau de correspondances entre les documents du corpus *AgroMada*.

domaine particulier (médical, agronomique) qui est représenté dans une ressource terminologique. Différentes ressources terminologiques sont sélectionnées ou générées en adéquation avec les thématiques des jeux de données étudiées. Pour représenter l'appartenance à un même contexte, la relation thématique associe une entité spatiale à une entité thématique si elles apparaissent dans une même fenêtre contextuelle.

Le mémoire est organisé de la manière suivante. La Partie i détaille le processus de mise en correspondance de données textuelles hétérogènes sur la dimension spatiale. Dans la Partie ii, nous présentons le processus d'intégration de la thématique dans le processus de mise en correspondance. Enfin, nous concluons par une synthèse des contributions et les perspectives envisagées dans la Partie iii.

Première partie

DIMENSION SPATIALE

Dans nos travaux sur la mise en correspondance de données textuelles hétérogènes, notre première proposition se concentre sur la dimension spatiale des données. Pour cela, nous proposons un processus de mise en correspondance en deux phases. La première phase consiste à représenter chaque document d'un corpus selon un même modèle. Nous proposons la Spatial Textual Representation (STR), une structure graphe intégrant les relations topologiques entre les entités spatiales détectées dans un document. La deuxième phase consiste à mesurer et associer les STRs générées à l'aide de mesures de similarités adéquates. Différentes mesures du domaine de *graph matching* sont étudiées. Enfin, une évaluation des différentes phases du processus de mise en correspondance est proposée sur deux corpus de documents : *PadiWeb* et *AgroMada*.

INTRODUCTION

Dans un processus de mise en correspondance classique, chaque objet comparé est associé à un ensemble de variables. Généralement, ce processus s'effectue entre deux bases de données où l'hétérogénéité s'exprime dans le format des données utilisées (MM/DD/YY, MM/DD/YYYY, etc.) et les noms de variables (surname, family name). Chaque variable est associée à une donnée qui peut avoir un caractère spatial (code postal, nom de ville), temporel (date) ou thématique.

Dans nos travaux, nous nous focalisons sur la mise en correspondance d'informations spatiales entre des données textuelles hétérogènes, i.e. documents (*c.f.* Section 1.2.2), regroupés dans un même corpus. Nous proposons un processus de mise en correspondance suivant deux phases : la **georepresentation** (Section 2) et le **geomatching** (Section 4). Dans la première phase, les informations spatiales sont extraites des documents puis intégrées dans une représentation dédiée : la **Spatial Textual Representation**. Cette structure de type graphe est conçue pour inventorier les entités spatiales et les relations de topologie (adjacence, inclusions) qu'elles entretiennent. Différentes transformations de cette représentation sont proposées pour répondre aux problèmes de détection de l'information spatiale dans les documents. Dans la seconde phase, différentes mesures de similarités sont étudiées pour mesurer la qualité des correspondances. Enfin, un protocole d'évaluation s'appuyant sur différents critères de similarité spatiale est proposé.

Cette partie est organisée de la manière suivante. Avant de présenter les différentes phases du processus général, nous introduisons les concepts de mise en correspondance et de similarité spatiale dans le chapitre 1. Le chapitre 2 définit la représentation de la spatialité choisie (STR) suivie par une explication des processus sur lesquels s'appuie sa génération. Une fois défini, nous proposons de présenter les différentes mesures de similarité étudiées dans le cadre du processus de *geomatching* dans le chapitre 4. Le protocole d'évaluation, les résultats obtenues puis les discussions sont présentés dans le chapitre 5.

1

MISE EN CORRESPONDANCE SPATIALE DE DONNÉES TEXTUELLES HÉTÉROGÈNES

Un processus de mise en correspondance est utilisé pour relier différentes représentations d'un même objet ou entité dans deux bases de données différentes (e.g. données de recensement). Ce processus s'effectue en plusieurs étapes dont le pré-traitement, la mesure de similarité et la mise en relation des données. Dans la conception des différentes étapes du processus, la définition de la similarité entre deux représentations est essentielle.

Dans nos travaux, nous proposons de mettre en correspondance des données textuelles hétérogènes selon leur dimension spatiale. Par conséquent, nous proposons une définition des critères essentiels permettant d'établir une similarité spatiale. En s'appuyant sur cette définition, nous proposons un processus de mise en correspondance s'appuyant sur la construction d'une représentation de la spatialité d'un document, la Spatial Textual Representation (STR), des mesures de similarité adéquates ainsi qu'un protocole d'évaluation dédié.

Ce chapitre est organisé de la manière suivante. Dans la section 1.1, nous présentons le processus général de mise en correspondance de données présent dans la littérature. Dans la section 1.2, nous présentons les bénéfices de l'exploitation de l'hétérogénéité et de la spatialité dans la mise en correspondance de données textuelles. Dans la section 1.3, le processus de mise en correspondance sur la dimension spatiale proposé est introduit. Enfin, dans la section 1.4, nous proposons une définition de la similarité spatiale sur laquelle s'appuie le processus de mise en correspondance.

1.1 MISE EN CORRESPONDANCE DE DONNÉES

La mise en correspondance de données (*data matching* en anglais) consiste à identifier et mettre en relation des enregistrements provenant de différentes sources et qui se réfèrent aux mêmes entités (Christen, 2012). Ces méthodes sont principalement utilisées dans la *détection de doublons*, i.e. deux entrées avec des informations identiques et se référent à une même entité ; ainsi que la *fusion de données*, i.e. deux entrées qui désignent une même entité avec des informations différentes. La Figure 3 illustre l'association des informations concernant *Anne Onyme* de deux bases de données, DB_1 et DB_2.

Un processus de mise en correspondance entre deux sources de données peut s'effectuer selon 5 étapes (voir Figure 4). Dans une

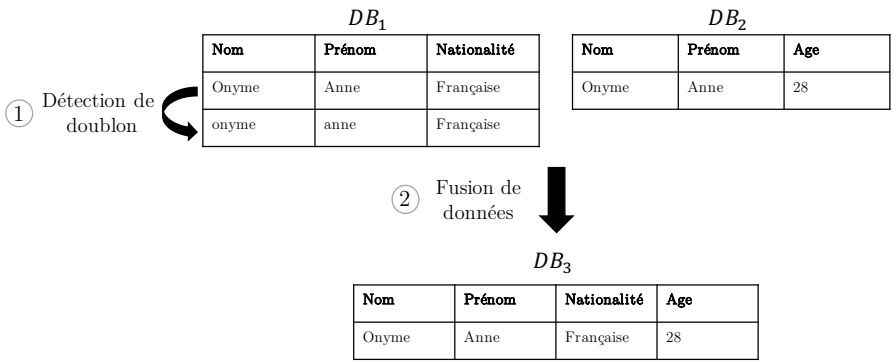


FIGURE 3 – Applications générale des méthodes de mise en correspondance de données

première étape, les données des différentes sources sont **uniformisées** (format des dates, codes postaux, numéros de téléphone) pour maximiser l'identification de caractéristiques similaires. Dans un processus de mise en relation, chaque entrée e_i de la base de données DB_1 est comparée avec chaque entité e_j dans DB_2. En conséquence, plus le nombre d'enregistrements dans DB_1 et DB_2 augmente, plus le temps de calcul augmente drastiquement (complexité O(mn)). Pour cela, l'étape d'**indexation** consiste à réduire le nombre de couples d'enregistrements à comparer. L'étape suivante consiste à mesurer la similarité entre les enregistrements des deux bases de données. Dans la dernière étape, les valeurs de similarité entre les différents enregistrements sont utilisées pour valider une correspondance.

Les processus de mise en correspondance entre des bases de données sont largement utilisés dans différents domaines d'application. Dans le domaine du recensement (ROGOT, SORLIE et JOHNSON, 1986 ; WINKLER, 2006), ces méthodes sont utilisées pour fusionner différentes informations concernant des individus ou réaliser un recensement d'une population quand certains individus n'apparaissent pas dans l'ensemble des bases de données nationales (santé, judiciaire, imposition). Dans le domaine biomédical (KHO et al., 2015), les données produites pour une personne sont diverses et dispersées dans les bases de données de différents établissements. Par exemple, dans le cadre de l'étude d'une maladie, la mise en correspondance de données sur une population permet d'obtenir des informations sur les symptômes majoritaires, le nombre de cas déclarés selon une région ou une période donnée. Dans le cadre de la sécurité nationale (MOHAMMED, CLARKE et F. LI, 2018 ; PHUA et al., 2012), des organisations telles que le gouvernement américain, ont accès à une variété de sources d'informations concernant différents individus : empreinte digitale, extrait bancaire, location de voiture, casier judiciaire, etc. Différents systèmes de mise en correspondance sont utilisés pour identifier les irrégularités dans ces données et faciliter la détection des utilisations de fausse identités.

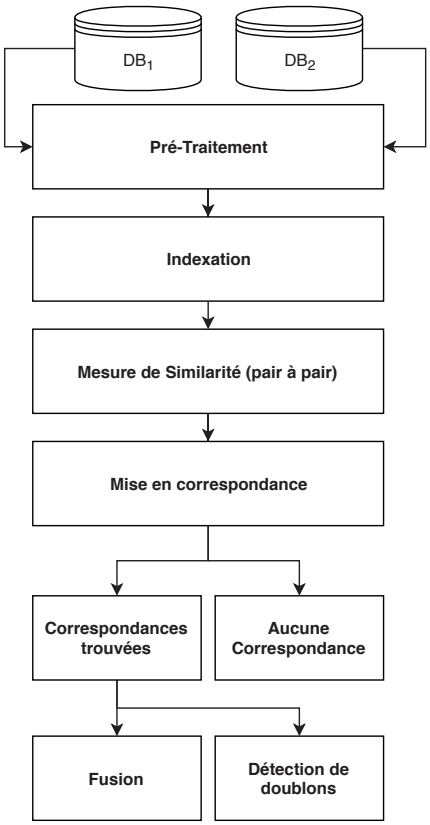


FIGURE 4 – Processus usuel de mise en correspondance de données.

Dans les exemples précédents, l'hétérogénéité relève du formatage des données ou encore des différences d'orthographes des variables sur des données structurées (e.g. tables de base de données). Dans nos travaux, nous nous focalisons sur la mise en correspondance de données *textuelles hétérogènes* sur la *dimension spatiale*. En conséquence, différentes applications sont rendues possibles quelque soit le domaine. Dans la section suivante, nous présentons les applications en Traitement du Langage Naturel pouvant bénéficier de ces contributions.

1.2 APPLICATIONS DU PROCESSUS DE MISE EN CORRESPONDANCE PROPOSÉE DANS LE DOMAINE DU TRAITEMENT AUTOMATIQUE DES LANGUES NATURELLES

La mise en correspondance de données textuelles hétérogènes permet d'augmenter la couverture des types de documents utilisés dans des tâches classiques en Traitement Automatique du Langage Naturel (TALN). Dans la tâche de résumé automatique (ALLAHYARI et al., 2017 ; PORTET et al., 2009), ce type d'approche permet d'augmenter la variété et la précision des informations retranscrites. L'accès à une donnée plus variée et connectée favorise l'extraction automatique de connaissances. De telles connaissances favorisent l'amélioration de

systèmes comme Question Réponse (Voorhees, 2003). Les mesures de similarité développées peuvent être intégrées dans des systèmes de Recherche d'Information (RI) (Shaw, Mian et Yadav, 2002). Enfin, l'identification de correspondances entre les documents d'un corpus peut être intégrée dans un processus de classification de documents (e.g. techniques de clustering (Kuang, Choo et Park, 2015)).

Dans l'analyse de données textuelles traitant de phénomènes comme des catastrophes naturelles, des épidémies ou encore les résultats d'élections : l'extraction d'informations de différentes catégories, thématique, spatiale et temporelle est essentielle. Par exemple, la mise en correspondance de données sur la dimension spatiale permet de regrouper (*document clustering*) des informations apparaissant dans des régions d'intérêts communes. En rajoutant la dimension hétérogène, la variété des données associées permet d'accéder à une information plus riche et de produire des analyses plus pertinentes.

Dans nos travaux, nous avons choisi de nous focaliser sur la mise en correspondance de ces données à partir de la dimension spatiale. Pour cela, nous proposons le processus de mise en correspondance décrit dans la section suivante.

1.3 PROCESSUS DE MISE EN CORRESPONDANCE PROPOSÉ

Dans nos travaux, nous proposons un processus de mise en correspondance divisé en deux phases : la **Georepresentation** et le **Geomatching**.

1. GEOREPRESENTATION. L'étape de georepresentation génère une représentation de la spatialité de chaque document d'un corpus selon un même modèle. Dans nos travaux, nous proposons la Spatial Textual Representation (STR) (*c.f.* Section 2.1.3), une représentation s'appuyant sur deux informations : les entités spatiales et les relations spatiales qu'elles entretiennent. Afin de générer cette représentation, les entités spatiales sont extraites du document à l'aide d'un processus de *geoparsing*. Puis, les entités extraites sont connectées à l'aide de deux relations spatiales. Différentes transformations (*c.f.* Section 2.4) sont proposées pour obtenir une représentation qui répond à différents critères essentiels de la similarité spatiale.
2. GEOMATCHING. L'étape de geomatching consiste à mesurer la similarité entre les représentations de chaque document. Pour mesurer la similarité entre les représentations, différentes approches ont été étudiées. En adéquation avec la nature de la STR (i.e. graphe), des mesures de similarité appartenant aux domaines du *graph matching* ou appariement de graphe (Riesen, Jiang et Horst Bunke, 2010) ont été sélectionnées.

La Figure 5 illustre le passage de la donnée d'entrée (corpus) par la phase de Georepresentation (génération STR), puis par la phase de Geomatching permettant de créer un réseau des correspondances entre les textes.

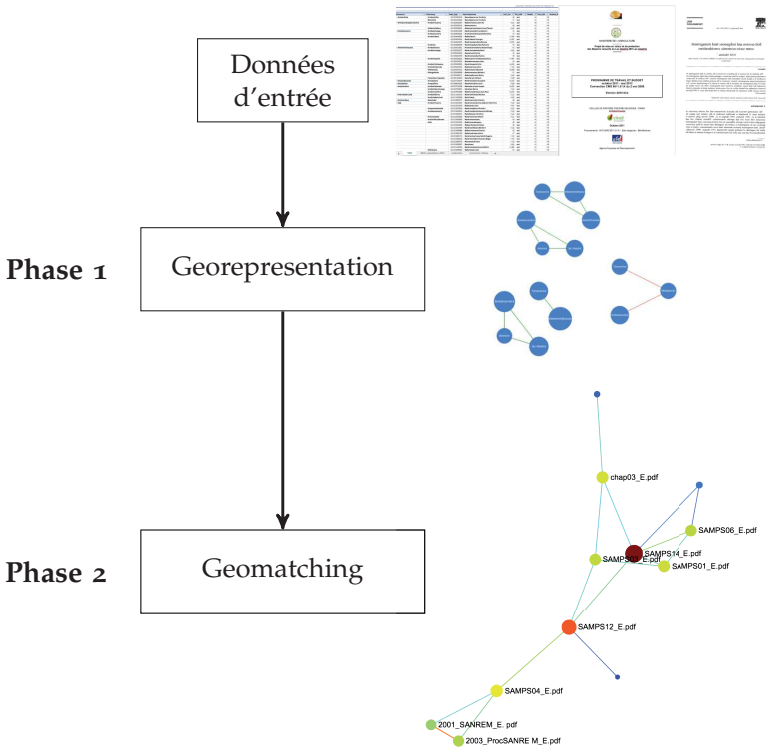


FIGURE 5 – Processus de mise en correspondance spatiale proposé.

Avant de présenter le processus de mise en correspondance, nous présentons la notion de similarité spatiale sur laquelle il s'appuie.

1.4 UNE CORRESPONDANCE SUR LA DIMENSION SPATIALE : DÉFINITION DE LA SIMILARITÉ

Dans la mise en correspondance de données, deux aspects sont essentiels : la représentation de la donnée et la mesure de similarité. Dans nos travaux, la conception et la définition de chaque concept s'appuie sur une notion essentielle : la **similarité spatiale**. Dans nos travaux, la similarité spatiale est associée à la question suivante : *Quels sont les critères permettant d'affirmer que deux documents s'appuient sur une configuration spatiale proche ?*

*La **configuration spatiale** est l'ensemble des références spatiales propre à un texte.*

La notion de similarité générale entre deux objets a fait l'objet de nombreuses contributions. Parmi ces contributions, nous proposons de présenter celle des travaux de (D. LIN, 1998), dans lesquels l'auteur énonce trois hypothèses pour définir la similarité entre deux objets :

— **La première hypothèse** stipule que la similarité entre deux objets est liée à leur *commonality*. La définition de *commonality* donnée par le dictionnaire Oxford est la suivante : *un état dans lequel deux objets partagent des fonctions ou des attributs* [1].

— **La deuxième hypothèse** précise que la similarité entre deux objets est liée aux différences qu'ils entretiennent. Plus il existe de différences entre deux objets, moins ils sont similaires.

— Enfin, **la troisième hypothèse** repose sur le fait que la valeur maximale de similarité entre deux objets A et B est atteinte, si A et B sont identiques.

Cette première définition donne les éléments clés concernant l'évaluation de la similarité entre deux objets. Cependant, certains aspects propres à la dimension spatiale doivent être intégrés. Dans ce sens, nous présentons deux *lois* de la géographie (Montello et al., 2003 ; Tobler, 1970) liées à la notion de similarité spatiale.

THE FIRST LAW OF GEOGRAPHY (Tobler, 1970). En 1970, l'auteur de (Tobler, 1970) propose un modèle de simulation de la croissance urbaine de la région de Détroit (États-Unis). Dans cette approche, il insiste sur l'importance d'établir un modèle intégrant les informations d'un lieu et des lieux voisins. Fort de ce constat, il "*invoque*" la *première loi de la géographie* (en anglais : *the First Law of Geography*) :

> *Every thing is related to everything else, but near things are more related than distant things.*

Dans le cadre de la mesure de la similarité spatiale entre deux documents : deux documents avec des configurations spatiales proches ont une plus forte probabilité d'être associés.

THE FIRST LAW OF COGNITIVE GEOGRAPHY (Montello et al., 2003). Poursuivant la théorie développée dans (Tobler, 1970), les auteurs de (Montello et al., 2003) proposent une nouvelle *loi* intitulée *la première loi de la géographie cognitive* qui stipule que :

> *[Les] entités proches sont plus similaires que si elles étaient éloignées[. . .]*

Dans leurs travaux, les auteurs présentent les fondements de la loi proposée et l'expérience menée pour prouver son application. Au cours des différentes phases de l'expérience, un participant est confronté à une série de projection de points comme illustré dans la Figure 6. Sur les différentes projections, trois points sont mis en évidence : les points 1, 2 et le point de référence A. La différence entre les deux phases de l'expérience repose sur les conditions d'affichage (taille de l'interface, dispersion des points, etc.).

1. https://en.oxforddictionaries.com/definition/commonality

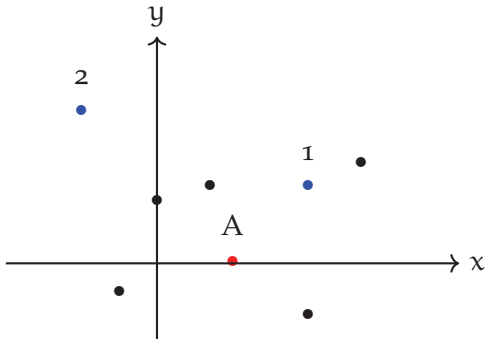


FIGURE 6 – Projection de points affichée sur l'interface pour l'experience de MONTELLO et al., 2003.

Pour chaque projection, le participant doit indiquer sur une échelle (voir Figure 7) le rapport entre la similarité du couple de points (A, 1) et celle du couple (A, 2). Les deux bornes de cette échelle sont les suivantes :

— **À gauche de l'échelle.** Les individus du couple $(A, 1)$ sont plus similaires que ceux du couple $(A, 2)$.
— **Au centre de l'échelle.** Les individus des couples $(A, 1)$ et $(A, 2)$ sont similaires de manière égale.
— **À droite de l'échelle.** Les individus du couple $(A, 2)$ sont plus similaires que ceux du couple $(A, 1)$.

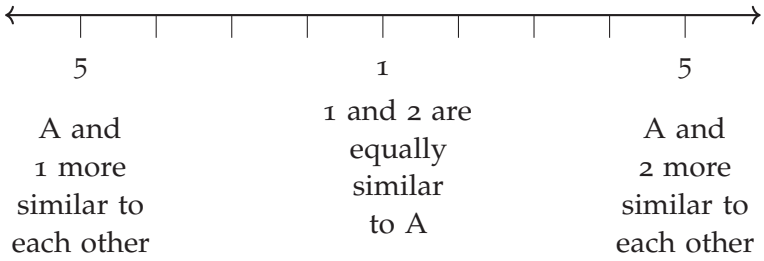


FIGURE 7 – Échelle utilisée dans l'expérimentation de MONTELLO et al., 2003.

Si les résultats des expériences valident la loi de *Montello et al.*, ils mettent à jour un autre phénomène, le *feature-effect*. Dans la projection de la Figure 6, aucune structure significative n'est observable. Par conséquent, la différence de similarité entre $(A, 1)$ et $(A, 2)$ est évaluée sur la distance entre les points. Inversement, dans la projection de la Figure 8, le point 1 est le plus proche de A, mais l'individu le plus similaire est représenté par le point 2. **Ici, la similarité entre deux entités n'est pas fondée sur la distance entre les points mais sur l'appartenance à une structure commune (i.e. cluster de points).** Ce *feature effect* est défini comme un phénomène où l'appartenance à une même structure dépasse le critère de distance dans l'évaluation de la similarité entre deux individus.

Dans le contexte de la mesure de spatialité entre deux documents composés de plusieurs entités spatiales (*c.f.* Section 2.1.2), l'existence

de *clusters* permet la détection de focus géographiques (Amitay et al., 2004), *i.e. ensemble d'entités spatiales pertinent dans l'analyse ou la comparaison.*

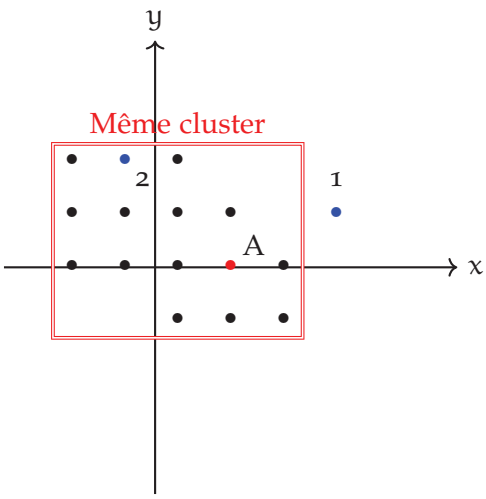


Figure 8 – Illustration du *feature-effect* décrit par Montello et al., 2003.

Définition de la similarité spatiale. En s'appuyant sur les contributions de (D. Lin, 1998 ; Montello et al., 2003 ; Tobler, 1970), nos travaux sur la mise en correspondance de données sur la dimension spatiale s'appuient sur trois critères de similarité :

— **L'identification de références spatiales identiques**, *i.e. les données partagent un ensemble d'entités spatiales (c.f. Définition 2.1.2)* ;

— **L'identification de références spatiales proches**, *i.e. les données possèdent des entités différentes mais proches spatialement* ;

— **L'identification de groupes de références spatiales similaires,** *i.e. les données partagent des groupes d'entités proches spatialement (cluster).*

Reprenant le processus de mise en correspondance proposé (*c.f.* Section 1.3), le prochain chapitre présente le processus de représentation de l'information spatiale des données textuelles hétérogènes, ou **Géroreprésentation.**

2 GEOREPRESENTATION

En Traitement Automatique du Language Naturel (TALN), il est courant de mesurer la similarité entre deux documents à l'aide d'une représentation commune (e.g. sac-de-mots). Dans la mise en correspondance de données, cette représentation commune s'appuie sur l'uniformisation des variables (date, nom, nombre). Dans nos travaux, nous proposons des méthodes de mise en correspondance sur la spatialité des données textuelles. Par conséquent, une représentation commune de la spatialité est nécessaire. Dans ce chapitre, nous présentons le modèle de représentation de la spatialité d'un document, la Spatial Textual Representation (STR), intégrant les entités spatiales (*c.f.* section 2.1.2) reliées par deux relations spatiales (inclusion, adjacence).

Ce chapitre est organisé de la manière suivante. La section 2.1 présente la représentation de la spatialité des documents, la Spatial Textual Representation (STR). Dans les sections 2.2 et 2.3, nous présentons le processus d'extraction d'entités spatiales et d'extraction de relations spatiales permettant de générer la STR. Enfin, différentes transformations de la STR sont proposées en section 2.4.

2.1 REPRÉSENTATION DE LA SPATIALITÉ

2.1.1 *État de l'art*

Pour introduire le modèle de représentation, nous proposons un état de l'art des représentations spatiales dans deux domaines : la recherche d'information géographique (GIR) et la représentation spatiale qualitative (QSR).

2.1.1.1 *Geographical Information Retrieval (GIR)*

Geographical Information Retrieval, ou GIR, est une extension de la Recherche d'Information qui se concentre sur la dimension géographique des données. Si la majorité des travaux en SIG (Science de l'Information Géographique) s'appuient sur des données structurées stockées dans des bases de données géographiques, le GIR s'intéresse tout particulièrement à l'exploitation de documents non-structurés présents sur le Web (Jones et Purves, 2008). L'exploitation des informations géographiques de ces données se retrouve face à plusieurs verrous (Jones et Purves, 2008) comme la détection de références géographiques dans les documents, la liaison entre ces références, une représentation géographique des documents, l'indexation des docu-

ments, la définition de nouvelles mesures de similarité et l'évaluation des systèmes de recherche (*geographic relevance*). De plus, plusieurs des systèmes GIR proposés intègrent l'information thématique (Adams, 2018 ; Michael D Lieberman, Samet et Sankaranarayanan, 2010 ; Purves, Clough, Jones, Arampatzis et al., 2007) en s'appuyant sur l'utilisation récurrente de termes géographiques et non-géographiques (Sanderson et Kohler, 2004) dans les requêtes utilisateurs.

Dans les approches proposées, une majorité s'appuie sur l'extraction et la résolution de toponymes (nom de lieu) dans les documents (Amitay et al., 2004 ; Michael D Lieberman, Samet et Sankaranarayanan, 2010 ; Purves, Clough, Jones, Arampatzis et al., 2007 ; Woodruff et Plaunt, 1994). Dans (Woodruff et Plaunt, 1994), les auteurs proposent de générer une représentation à partir des polygones des différents lieux présents dans un document. Déjà, les auteurs identifient les problèmes récurrents en *geoparsing*, notamment la couverture du référentiel utilisé. Dans (Amitay et al., 2004), les auteurs proposent une représentation d'un document selon son *focus géographique*, i.e. lieu central et connecté aux autres lieux mentionnés dans un document, *e.g.* $\mathtt{focus}(\mathtt{Paris}, \mathtt{Cherbourg}, \mathtt{Montpellier}) = \mathtt{France}$. Comme (Amitay et al., 2004), les auteurs de (Michael D. Lieberman et al., 2007) proposent un système de recherche d'information géographique, STEWARD, qui repose sur la création d'un focus géographique. Contrairement à (Amitay et al., 2004), la définition du focus s'appuie sur la structure du document. Ici, un lieu est défini comme focus géographique s'il co-existe avec un maximum de lieux dans ce même document. Enfin, dans (Purves, Clough, Jones, Arampatzis et al., 2007), les auteurs présentent un système de Recherche d'Information, appelé SPIRIT, qui s'appuie sur une indexation sur une grille spatiale. Chaque cellule de la grille correspond à une *bounding-box* délimitée par des coordonnées sur le globe. Une cellule est associée à un document si celui-ci possède une empreinte spatiale[1] similaire.

Ici, le ***référentiel*** *correspond à une base de données contenant des informations géographiques pour un nombre de lieux c.f. Section 3*

D'autres méthodes, appartenant au domaine du *language modeling* (Melo et Martins, 2017 ; Roller et al., 2012 ; Wing et Baldridge, 2014), utilisent la totalité du document pour déterminer son empreinte spatiale. Un *language model* est représenté par une distribution de probabilité des séquences de mots dans un document. Dans le cadre d'un système de Recherche d'Information (RI), ces probabilités sont utilisées pour calculer la probabilité $P(Q|D)$ que les termes d'une requête Q soit associés à un document D. Dans un contexte spatial, ces approches sont utilisées pour calculer la probabilité $P(Q|(D, E)$ que les termes d'une requête Q soient utilisés dans un document D associé à une empreinte géographique E (point, région, polygone).

1. Une empreinte spatiale peut correspondre à des coordonnées (latitude-longitude) ou une structure plus complexe (polygone, multipolygone) (*c.f.* Section 2.1.2)

2.1.1.2 *Qualitative Spatial Representation*

Un des verrous des approches proposées en GIR repose sur l'identification et la résolution des toponymes (*geoparsing*). À cette fin, d'autres approches d'analyse de la spatialité dans les documents s'appuient sur une analyse qualitative. Le *Qualitative Spatial Reasoning* est un sous-domaine de la représentation de connaissances s'appuyant sur la définition de relations spatiales qualitatives. Ces relations qualitatives, ou *Qualitative Spatial Calculus* (WALLGRÜN, FROMMBERGER et al., 2009), relient différents objets appartenant à un domaine particulier. Parmi les représentations proposées, le Qualitative Constraint Network (QCN) propose de relier différentes entités selon des relations spatiales qualitatives. Parmi les relations utilisées, nous retrouvons la *topologie* (e.g. inclu dans, égal), la *distance* (e.g. proche, loin), *l'orientation* (e.g. au nord de, en face de) (RENZ, 2002 ; AL-SALMAN, DYLLA et FOGLIARONI, 2012) ou encore le *mouvement* (WU, 2015).

Les raisonnements appliqués sur ces représentations sont multiples. Les auteurs de (MONCLA, GAIO et al., 2016 ; AL-SALMAN, DYLLA et FOGLIARONI, 2012 ; WALLGRÜN, WOLTER et RICHTER, 2010) proposent de générer des représentations dans l'analyse de description d'itinéraire ou d'un lieu en particulier. Une fois générées, ces représentations peuvent être analysées et mises en correspondance avec des données géographiques dans une base de données SIG[1] (SALLABERRY et al., 2008 ; AL-SALMAN, DYLLA et FOGLIARONI, 2012 ; WALLGRÜN, WOLTER et RICHTER, 2010). D'autres approches (BELOUAER, BROSSET et CLARAMUNT, 2016 ; WALLGRÜN, WOLTER et RICHTER, 2010) proposent de générer de nouvelles représentations (i.e. carte). À l'aide d'une ontologie spatiale, les auteurs de (BELOUAER, BROSSET et CLARAMUNT, 2016) proposent de construire un réseau spatial permettant de générer une carte.

Si ces structures permettent de s'affranchir de référentiels (index géographiques), il est nécessaire de vérifier la cohérence spatiale des relations extraites. Par exemple, si un individu A est au nord de B, B est au nord de C, alors C ne peut être au nord de A. Selon la taille[2] (ou la densité) du réseau formé par ces relations, ce processus peut s'avérer coûteux (WALLGRÜN, WOLTER et RICHTER, 2010).

2.1.1.3 *Approche choisie*

Le premier critère de la similarité spatiale (*c.f.* Section 1.4) repose sur le partage d'individus identiques entre deux configurations spatiales des documents d'un corpus. Dans le domaine de Recherche d'Information Géographique (GIR), l'indexation des documents s'effectue au travers des toponymes présents dans les documents. Ces toponymes sont associés à un identifiant ou des coordonnées (*latitude-longitude*)

1. Systèmes d'Information Géographique
2. Voir définition dans le chapitre 4

qui sont ensuite utilisés dans l'étape d'indexation. Contrairement aux approches extrayant un focus géographique (Amitay et al., 2004 ; Michael D. Lieberman et al., 2007), nous choisissons d'utiliser l'ensemble des lieux mentionnés formant la configuration spatiale d'un document. La représentation proposée dans nos travaux s'articule autour des entités spatiales (*c.f.* Définition 2.1.2) associées aux toponymes d'un document. Une entité spatiale est associée à un identifiant unique, différentes représentations géométriques (centroïde, polygone), une classe et d'autres informations additionnelles. L'unicité de l'entité spatiale permet de respecter le premier critère de la similarité spatiale (*c.f.* Section 1.4) qui consiste à estimer quelles sont les entités spatiales partagées entre deux configurations spatiales, ici représentées par les documents.

Les deux critères restants – i.e. la proximité spatiale et les groupes d'entités – sont regroupés dans la notion de topologie spatiale (Egenhofer et Franzosa, 1991), i.e. l'organisation spatiale des entités dans l'espace. Pour intégrer cette information, nous profitons de la variété d'informations associées avec les entités spatiales d'un document (coordonnées, polygone ou variable spécifique [1]) pour les relier. Deux catégories de relations peuvent être identifiées : les relations quantitatives et les relations qualitatives. Une relation de distance est une relation quantitative. Une relation de direction est une relation qualitative.

Pour faire ressortir la topologie, *i.e. connexion entre les entités formant un motif*, l'utilisation de relations quantitatives telle que la distance entre les entités a été envisagée. Cependant, ce type de relation n'est pas pertinent compte tenu du calcul de la distance entre certaines entités spatiales avec une emprise spatiale différente (quelle est la distance entre la ville de Paris et la Belgique ?). Par conséquent, nous proposons de privilégier les relations qualitatives qui permettent de représenter les différences d'échelle et la proximité entre les entités spatiales. Pour permettre de représenter la topologie spatiale – i.e. proximité et cluster– existante des entités présentes dans un document, nous avons sélectionné deux relations : l'une indiquant l'appartenance d'une entité à la surface d'une autre entité (*inclusion*) et la deuxième indiquant *l'adjacence* entre deux entités. Dans la Figure 9, nous illustrons les différentes informations spatiales représentées dans la représentation proposée à l'aide des entités et des relations spatiales.

Nous avons choisi d'intégrer les informations spatiales associées à ces deux concepts dans une représentation de type graphe. Ce choix de représentation repose principalement sur la nature des informations spatiales utilisées, et plus particulièrement celles des relations spatiales comme dans les travaux de (Wallgrün, Wolter et Richter, 2010). De

1. Il existe dans certaines bases de données comme *Wikidata*, des propriétés indiquant l'existence de relations spatiales entre différentes entités.

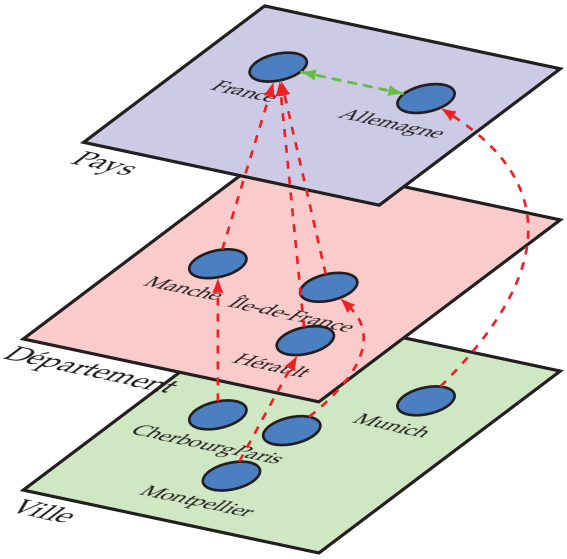


FIGURE 9 – Dimension multi-échelle spatiale possible à l'aide des relations d'inclusion (arc rouge) et d'adjacence (arc vert).

plus, les structures graphes sont évolutives et permettent de stocker des informations : thématiques (*c.f.* Partie ii) et temporelles.

2.1.2 *Entité Spatiale et Relation Spatiale*

Pour représenter l'information spatiale d'un document, nous avons choisi de proposer un modèle s'appuyant sur deux concepts : **l'entité spatiale** et **la relation spatiale**.

2.1.2.1 *Entité spatiale*

Dans *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems* (NYERGES et al., 1995), les auteurs proposent une définition des primitives à la base de la connaissance spatiale. À la base de cette théorie, *l'occurrence* est définie comme "un phénomène appartenant à notre réalité". Ce phénomène est associé à 4 primitives : *son identité*, *sa localisation*, *son emprise* et *sa temporalité*.

1. IDENTITÉ. L'identité d'une occurrence est associée à une *étiquette* (ou *label*) qui lui permet de se différencier des autres, *e.g. identifiant d'une base de données*.
2. LA LOCALISATION. La localisation — une primitive fondamentale de la connaissance spatiale — indique la position de l'occurrence dans un environnement, *e.g. des coordonnées en latitude-longitude*.
3. L'EMPRISE. L'emprise d'une occurrence est définie à partir d'une classification construite selon différentes variables *e.g. démographie, nombre d'habitants, surface d'occupation, politique, etc.*, *e.g. ville, mégalopole, région, etc.*

4. LA TEMPORALITÉ. Une occurrence n'est ni éternelle, ni immuable. La temporalité d'une occurrence prend en compte les changements subis comme : changement de nom, de localisation ou de son empreinte géographique *e.g. émigration/immigration, agrandissement*. Par exemple, la ville de Paris a subi au cours des années une évolution de sa surface d'occupation mais aussi de toponyme (Lutèce → Paris).

Dans (GOODCHILD et HILL, 2008), les auteurs se réfèrent à un lieu géographique ou *geographic place*. Un lieu géographique est défini selon trois propriétés : un *nom*, une *classe* et une *empreinte géographique*. Reprenant la définition de (NYERGES et al., 1995), nous remarquons que chacune des propriétés peut être associée à une des primitives spatiales. Le nom (identité) est défini selon son utilisation commune à un temps donné (temporalité). La classe (emprise) définit la nature du lieu *e.g. ville, lac, route* mais aussi sa place dans la hiérarchie administrative d'un pays *e.g. région, département, etc.*. Enfin, l'empreinte géographique (localisation) correspond aux coordonnées du lieu géographique.

En s'appuyant sur les travaux de (NYERGES et al., 1995) et (GOODCHILD et HILL, 2008), nous proposons de définir une **entité spatiale** de la manière suivante. Une **entité spatiale** est une entité définie par 5 attributs : un identifiant, un toponyme, une empreinte spatiale, une classe et d'autres informations additionnelles. La Figure 10 montre les différents attributs associés à l'entité spatiale de Paris (France). Les attributs d'une entité spatiale sont définis de la manière suivante :

1. Un **identifiant unique** correspond à une suite de chiffre (ou caractères) unique *e.g. 2988507 ↔ Paris*.
2. Le **toponyme** est le nom d'une entité spatiale, *e.g Paris*.
3. Une **empreinte spatiale** est représentée par des coordonnées exprimées en *latitude-longitude* (*e.g. 48,7508 ; 27905*) ou par une géométrie plus complexe (e.g. Polygone, MultiPolygone, etc.).
4. Une **classe** qui définit la nature de cette entité, *e.g. ville, région, pays, etc.*
5. Des **informations additionnelles** telles que le pays d'appartenance, les entités voisines, la taille de la population, les toponymes alternatifs (alias).

2.1.2.2 *Les relation spatiales*

Une **relation spatiale** lie deux entités dans un environnement spatial commun. Ces relations peuvent être **quantitatives**, *e.g. relation pondérée par la distance entre deux entités*, ou bien **qualitatives**, *e.g. inclusion, adjacence, au nord de, proche de*. Une relation spatiale peut être dirigée, *e.g. une relation d'inclusion est dirigée, une relation de distance est non-dirigée*.

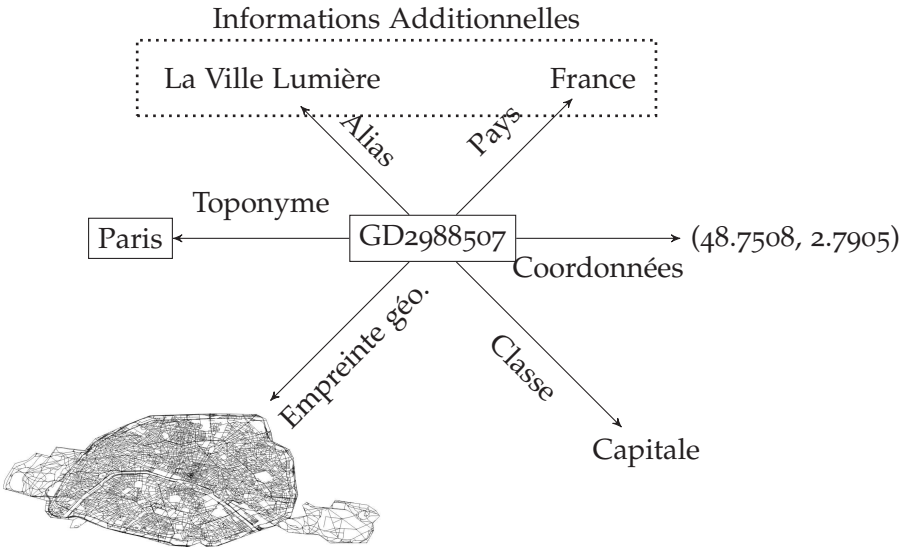


FIGURE 10 – Informations associées à l'entité spatiale Paris.

Dans notre approche, nous proposons de connecter les entités spatiales selon deux types de relations : **adjacence** et **inclusion**. Elles permettent d'intégrer des informations sur la topologie dont les différences d'échelles spatiales et la proximité entre les entités. Ces deux relations sont inspirées de la grammaire spatiale RCC-8 (RANDELL, CUI et COHN, 1992) illustrée dans la Figure 11. Dans cette grammaire, différentes relations sont établies selon la position des différents éléments X et Y. Les deux relations choisies sont issues du rassemblement des relations dans RCC-8 sur trois catégories : **l'égalité** avec {X EQ Y}, **l'adjacence** avec { X *DC* Y, X *PO* Y, X *EC* Y} et **l'inclusion** avec {X *TPP* Y, X *NTPP* Y, *XTPPi* Y, *NTPPi* Y}. Les différentes relations choisies sont définies dans les paragraphes suivants.

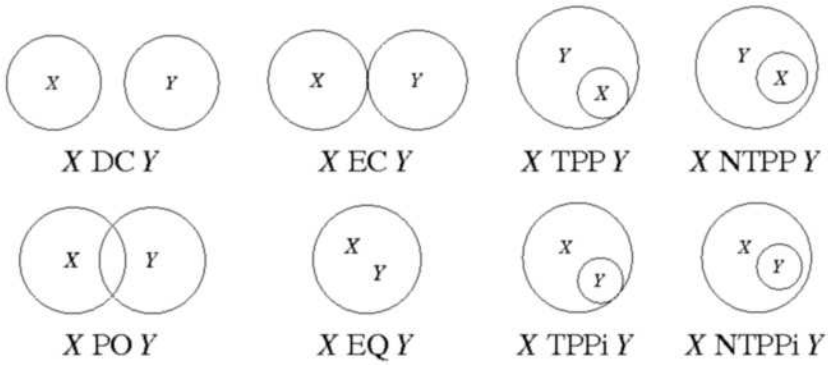


FIGURE 11 – Illustration de la grammaire de logique spatiale : RCC-8 (RANDELL, CUI et COHN, 1992).

INCLUSION. Une entité spatiale est incluse dans une autre entité si et seulement si son empreinte géographique (géométrie) est

contenue dans celle de la seconde entité. Dans la Figure 12, la ville de *Paris* est incluse dans la région *Ile-de-France* et la région *Ile-de-France* est incluse dans le pays *France*. Si une entité A est inclue dans une entité B et que B est inclue dans C, alors A est inclue dans C. De manière générale, la relation d'inclusion entre deux entités est définie selon le découpage administratif d'un territoire. Toutefois, une relation d'inclusion peut être aussi définie sur son appartenance à une formation terrestre [1], e.g. île, montagne, lac. Par exemple, la Statue de la Liberté se situe dans l'état de New York mais aussi sur l'île *Liberty Island*.

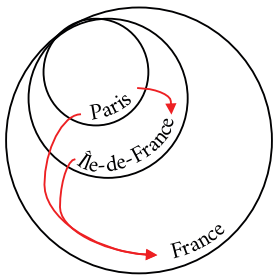


FIGURE 12 – Illustration de la relation d'inclusion.

ADJACENCE. Deux entités sont adjacentes si leurs empreintes géographiques sont adjacentes ou très proches. Par exemple, la Belgique et la France sont deux entités spatiales adjacentes, tout comme la ville de Paris et la ville de Saint-Denis (voir Figure 13).

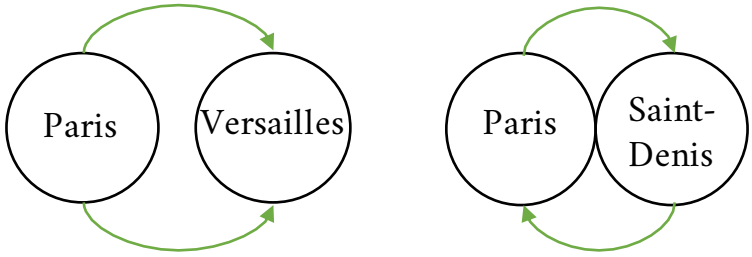


FIGURE 13 – Relation d'adjacence.

Dans la section suivante, nous présentons la Spatial Textual Representation, le modèle de représentation proposé qui agrège les concepts d'entités spatiales et de relations spatiales.

2.1.3 *La Spatial Textual Representation*

Pour représenter l'information spatiale présente dans un document, nous proposons d'associer deux concepts : les entités et les relations spatiales. Dans le domaine de Qualitative Spatial Reasoning (LIGOZAT, 2013), l'intégration des relations qualitatives s'effectue dans une structure graphe appelée *Qualitative Constraint Network* (WALLGRÜN,

1. *terrain features* en anglais

WOLTER et RICHTER, 2010). Un QCN est une structure graphe composée d'entités articulées par des relations qualitatives. Cependant, la détection des entités et de leurs relations est effectuée à l'aide d'un système à base de règles. Dans une représentation spatiale (WALLGRÜN, WOLTER et RICHTER, 2010), cela implique une vérification de la cohérence spatiale des relations extraites. De plus, les entités détectées ne sont pas associées à un identifiant unique, ce qui rend difficile la comparaison des configurations spatiales entre les documents.

Dans nos travaux, nous proposons une structure similaire au QCN, nommée la **Spatial Textual Representation** (STR) composée des entités spatiales (*c.f.* Section 2.1.2) présentes dans un document et des relations spatiales qu'elles entretiennent. Deux relations spatiales sont utilisées : **l'inclusion** et **l'adjacence**. Contrairement au QCN, les entités spatiales extraites sont associées à un identifiant dans une base de données géographiques et les relations spatiales utilisées sont extraites à l'aide des informations associées à ces mêmes entités.

La **Spatial Textual Representation** est représentée par un graphe G_{STR} dirigé et étiqueté défini par :

$$G_{STR} = (V_{ES}, E_{RS}, T_{RS}) \tag{1}$$

Avec :

— V_{ES} correspond à l'ensemble des entités spatiales dans un document ;
— E_{RS} l'ensemble des relations entre les entités $e \in V_{ES}$;
— T_{RS} l'ensemble des types de relation entre les entités spatiales V_{ES}.

La Figure 14 montre une STR générée à partir d'un document, où les arcs verts correspondent aux relations d'adjacence et les arcs rouges correspondent aux relations d'inclusion.

Dans la section suivante, nous présentons le processus de génération automatique d'une STR à partir d'un document.

2.1.4 *Présentation du processus de génération d'une STR*

La génération de la STR, ou *georepresentation*, d'un document se déroule en deux étapes. La première étape, ou **Geoparsing**, consiste à identifier les entités spatiales présentes dans un document (*c.f.* Section 2.2). Par exemple, dans le document de la Figure 15, les entités spatiales suivantes sont identifiées : *Ukraine*, *Europe*, *Russia*. Chaque entité identifiée dans un document est associée à une entrée d'un index géographique. Dans nos travaux, nous utilisons GEODICT (**Fize, Jacques** et SHRIVASTAVA, 2017), un index géographique conçu à partir des informations nécessaires à la génération de la STR (*c.f.* Chapitre 3). La construction de cet index géographique est détaillée dans le

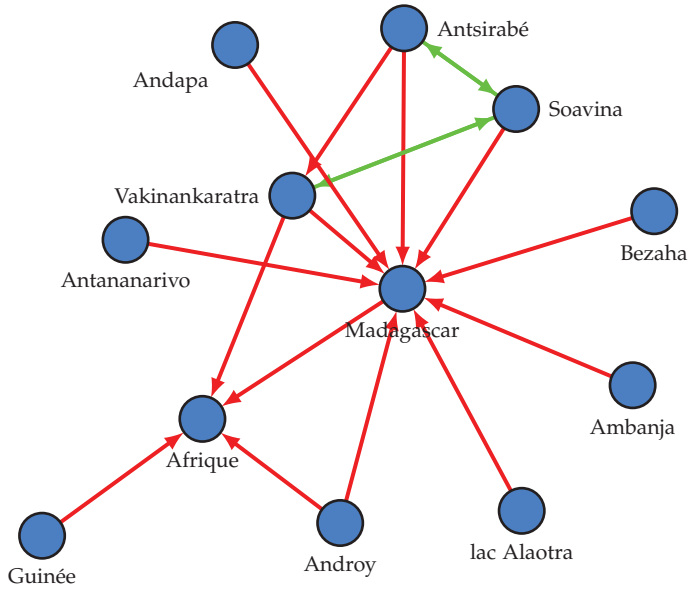


FIGURE 14 – Exemple d'une Spatial Textual Representation générée à partir d'un document du corpus *AgroMada* (*c.f.* Section 5.1.2). Les arcs verts correspondent aux relations d'adjacence et les arcs rouges correspondent aux relations d'inclusion.

Chapitre 3. Dans la seconde étape, la **géocompletion**, les relations spatiales entre les entités sont identifiées à l'aide de leur empreinte spatiale (*c.f.* Section 2.3) et des méta-données présentes dans GEODICT.

Chacune des étapes du processus de **Georepresentation** – le *geoparsing* et la *geocompletion* – est détaillée dans les sections 2.2 et 2.3.

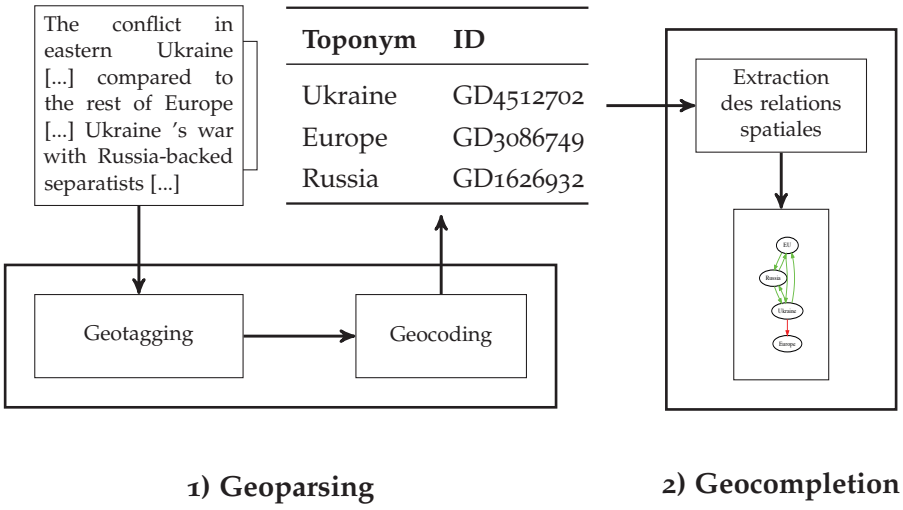


FIGURE 15 – Processus de Georepresentation.

2.2 GEOPARSING : IDENTIFICATION DES ENTITÉS SPATIALES DANS LES DOCUMENTS

Le **geoparsing** (Gritta et al., 2018 ; Moncla, Renteria-Agualimpia et al., 2014) est une famille d'algorithmes dédiés à l'identification et la résolution des toponymes (nom de lieu) présents dans un document. Un algorithme de *geoparsing* est divisé en deux étapes (illustration dans la Figure 16) :

1. **Geotagging** (Section 2.2.1) Le *geotagging* consiste à identifier les toponymes des entités spatiales présentes dans des documents.
2. **Geocoding** (Section 2.2.2) Le *geocoding* consiste à associer une signature géographique pour chaque toponyme identifié durant le *geotagging*. Cette étape se divise en deux phases, la première consiste à résoudre les ambiguïtés levées par certains toponymes. Par exemple, il existe plusieurs lieux nommés *Paris*. La seconde étape consiste à retourner la représentation de l'entité sous forme d'un identifiant ou celle d'une empreinte spatiale (e.g. des coordonnées). Les identifiants ont un rôle d'intermédiaire, ils permettent d'accéder à l'empreinte spatiale stockée dans une base de connaissances.

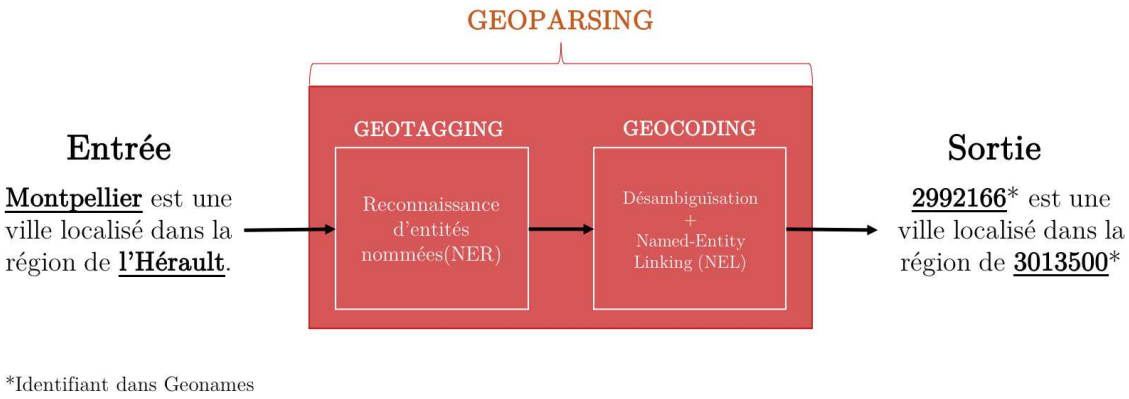


Figure 16 – Fonctionnement global d'un algorithme de Geoparsing. Source : (Gritta et al., 2018).

2.2.1 *Geotagging*

Le *geotagging* consiste à identifier les toponymes dans un document, *e.g. London is the capital of England*→ {England, London}. Le *geotagging* est une sous-catégorie des méthodes de Reconnaissance d'entités nommées (NER) qui sont utilisées dans l'identification de différentes entités présentes dans des documents. Communément, la plupart des entités identifiées sont divisées en trois types : Personne, Organisation et Lieu (Nadeau et Sekine, 2007). D'autres méthodes permettent de détecter différentes entités telles que les dates, les nombres, etc.

Dans le cas du *Geotagging*, le focus se trouve sur la partie responsable de la détection d'entités de type LIEU.

Différentes méthodes de reconnaissance d'entités nommées sont proposées dans la littérature. Elles sont divisées en deux catégories : (i) à base de règles (Section 2.2.1.1), ou (ii) utilisant des algorithmes d'apprentissage automatique (Section 2.2.1.2).

2.2.1.1 *Système à base de règles*

Un système à base de règles est défini selon deux composantes. L'ensemble de règles est la première composante. Une règle est divisée en deux parties : la condition de son activation et l'action qui en dépend. Dans un système de reconnaissance d'entités nommées, un ensemble de *mots* qui respectent une règle est associé à un libellé (*tag* en anglais). La deuxième composante concerne l'ordre de vérification d'une règle. Généralement, il est défini par des experts, mais peut dans certaines propositions être ajusté selon les besoins d'un utilisateur.

Différentes approches de *geotagging* s'appuyant sur des systèmes à base de règles sont proposées. Dans (ZHOU et SU, 2002), les auteurs proposent un modèle à base de règles s'appuyant sur des chaînes de Markov. Dans (CHITICARIU et al., 2010), les auteurs proposent le language NERL dédié à la création de modèle à base de règles pour la détection d'entités nommées.

2.2.1.2 *Apprentissage Automatique*

Contrairement à l'utilisation de systèmes à base de règles, les méthodes d'apprentissage automatique utilisent des données annotées (LEIDNER, 2007 ; Michael D LIEBERMAN, SAMET et SANKARANARAYANAN, 2010 ; NOTHMAN et al., 2013 ; TJONG KIM SANG et DE MEULDER, 2003). Ces approches proposent de transformer le problème de *geotagging* en un problème de classification d'un mot w_i selon n classes en s'appuyant sur une fenêtre donnée $(w_{i-k}, \ldots, w_{i-1}, w_{i+1}, w_{i+k})$. Parmi ces méthodes, l'utilisation *Conditional Random Field* (MANNING et al., 2014) a fait ses preuves. De nouvelles approches dans (AKBIK et al., 2019 ; AL-RFOU et al., 2015 ; HONNIBAL et MONTANI, 2017) proposent de s'affranchir des spécificités des différentes langues.

2.2.1.3 *Méthode hybride*

D'autres contributions proposent une approche hybride qui combine l'expertise apportée par les systèmes à base de règles et la souplesse des modèles d'apprentissage automatique. Dans (GALI et al., 2008), les auteurs utilisent conjointement un modèle CRF (Conditional Random Field) (J. LAFFERTY, MCCALLUM et PEREIRA, 2001) et un ensemble de règles linguistiques propres à l'hindi. Dans (GELERNTER et W. ZHANG, 2013), le système à base de règles est utilisé pour normaliser

le document : uniformiser les noms de rues et de bâtiments ; corriger les erreurs d'orthographe des toponymes.

2.2.1.4 *L'ambiguïté des toponymes.*

La principale difficulté dans l'identification des toponymes dans un document provient de la confusion *"géo/non-géo"*(Buscaldi, 2010), i.e. entre un toponyme et le nom d'une organisation ou entre un toponyme et le nom d'une personne. Par exemple, il existe une ville du Texas [1] ayant un toponyme identique au nom de famille de la chancelière allemande (Angela Merkel).

Une autre source d'ambiguïté provient des métonymies (Gritta et al., 2018), i.e. lorsqu'un toponyme est utilisé pour représenter une idée, un concept particulier. Par exemple, dans la phrase "*Londres est rentré en contact avec Paris*", les toponymes *Paris* et *Londres* représentent les gouvernements des deux capitales.

2.2.2 *Geocoding*

Pour obtenir l'empreinte spatiale (*c.f.* Définition 2.1.2) associée à chaque toponyme d'un document, nous entrons dans la deuxième étape du processus de geoparsing, le *geocoding*. Le **geocoding** associe une signature spatiale à une ressource (textuelle ou non). Dans le cas présent, nous cherchons à associer une empreinte spatiale à chaque toponyme identifié dans un document. Selon les méthodes, cette empreinte peut être limitée à des coordonnées latitude-longitude, *e.g.* *Londres (51,50,-0.12) est la capitale de l'Angleterre (52.16045, -0.70312)*, ou à un identifiant d'une entrée dans un index géographique contenant différentes informations spatiales comme les coordonnées, la classe (ville, région, etc), *e.g.* *Londres (GD269488) est la capitale de l'Angleterre (GD4233692)*.

Comme pour le *geotagging*, le processus de *geocoding* est soumis au problème d'ambiguité. Pour le *geocoding*, la possible [2] ambiguïté repose sur l'ensemble d'entités spatiales partageant un même toponyme. Par exemple, il existe plusieurs villes qui se nomment *Paris* : *e.g.* *Paris, Tennessee ; Paris, Texas ; Paris, Panama*. (Purves, Clough, Jones, Arampatzis et al., 2007) nomment ce phénomène : *referent ambiguity*. L'utilisation d'un algorithme de désambiguïsation est nécessaire pour identifier l'entité spatiale exacte pour chaque toponyme d'un document.

La section suivante présente différents algorithmes de désambiguïsation utilisés dans nos travaux.

1. *Merkel, Texas* https://www.openstreetmap.org/way/33174741
2. *Possible* car il existe certaines entités qui ne partagent pas leur toponyme.

2.2.2.1 *Algorithme de désambiguïsation*

Un algorithme de désambiguïsation consiste à associer une entité spatiale (*c.f.* section 2.1.2) à chaque toponyme présent dans un document. Pour un ensemble de toponymes $t \in T$, un algorithme de désambiguïsation est défini par la formule suivante :

$$f(t, E_t) = \arg\max_{e_i \in E_t} score(e_i)$$

où E_t correspond à l'ensemble des entités spatiales candidates pour le toponyme t et $score(e_i)$ la pertinence d'une entité spatiale pour le toponyme t.

Nous séparons ces algorithmes en trois catégories. La première catégorie d'algorithme s'appuie uniquement sur les informations des entités spatiales candidates. Par exemple, certaines méthodes utilisent la superficie pour choisir l'entité spatiale. La seconde catégorie analyse la proximité spatiale entre les différents candidates des différents toponymes pour sélectionner les entités spatiales exactes. Par exemple, dans (H. Li et al., 2003), les auteurs proposent de désambiguïser les toponymes en s'appuyant sur les similarités spatiales (proximité, appartenant à une région commune, etc.) qu'entretiennent les différentes entités candidates. Sur une approche similaire, la troisième catégorie s'appuie sur la fréquence d'apparition d'entités spatiales dans un même contexte (e.g. corpus de documents).

Dans nos travaux, nous avons choisi de comparer trois algorithmes : MostCommon, SharedProp puis WikiCooc. Pour illustrer les différents algorithmes, nous utilisons le toponyme ambigu Paris (+ de 20 entités avec le toponyme *Paris* [1]).

mostcommon. Plusieurs approches (Rauch, Bukatin et Baker, 2003 ; Smith et Mann, 2003) proposent d'associer un toponyme à une entité spatiale avec une valeur maximale selon une variable. Par exemple, cette variable peut correspondre à la superficie ou la population. Dans nos travaux, nous proposons un algorithme qui s'appuie sur la *commonness*, i.e. l'association la plus courante d'une entité spatiale avec un toponyme. Dans le cas de *Paris*, le toponyme est le plus souvent associé à l'entité *Paris,FR*.

Pour représenter la *commonness*, nous mesurons le *pagerank* (Page et al., 1999) de chaque entité sur le corpus de *Wikipedia*. *Wikipedia* est une encyclopédie composée d'un volume important de données textuelles et de contextes variés qui nous permettent de capturer la *commonness* d'une entité spatiale [2]. En reprenant, l'exemple de *Paris* illustré dans la Figure 17, le rang de *Paris,FR* est égal à 4712 et la

1. Source [Geonames] : https://www.geonames.org/advanced-search.html?q=Paris&country=&featureClass=P&continentCode=

2. Ici, un article *Wikipedia* est considérée comme une représentation d'une entité spatiale.

rang de *Paris,TX* est égal à 3,27. Par conséquent, l'algorithme associera *Paris,FR* avec le toponyme *Paris*.

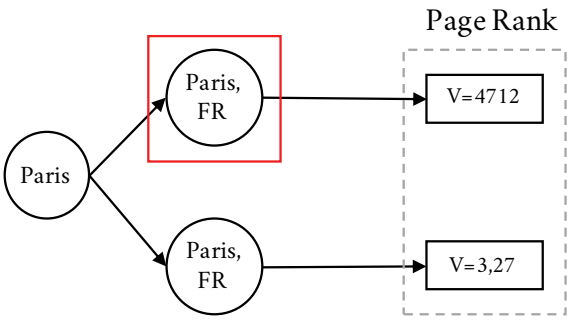


FIGURE 17 – Illustration de l'algorithme MOSTCOMMON. L'entité sélectionnée par l'algorithme est encadrée en rouge.

SHAREDPROP. Dans l'algorithme MOSTCOMMON, l'entité spatiale choisie correspond à la plus couramment associée avec le toponyme. En s'appuyant sur l'exemple de la désambiguïsation du toponyme *Paris*, *Paris,FR* est le résultat le plus courant. Toutefois, il est possible que dans certains documents, *Paris,FR* ne soit pas spatialement cohérente avec le reste des entités d'un document. Par exemple, dans un document contenant les entités spatiales *Brookston,TX*, *Powderly,TX* (villes de l'état du Texas), l'entité spatiale *Paris,TX* est plus cohérente.

Pour mesurer la cohérence spatiale, il est nécessaire de différencier deux types de toponymes : les toponymes ambigus et les toponymes uniques. Les toponymes ambigus sont partagés par différentes entités spatiales comme *Paris*. Les toponymes uniques ne sont associés qu'à une entité spatiale comme le *Port Racine*[1]. Nous définissons comme *entité fixe*, une entité spatiale ayant un toponyme unique. S'appuyant sur les définitions précédentes, nous proposons de désambiguïser les toponymes ambigus en utilisant les informations de proximités avec les entités fixes. Dans l'exemple de la Figure 18, le document est composé du toponyme ambigu *Paris*, de l'entité fixe *France*. L'ensemble {Paris,FR ;Paris,US} correspond aux entités spatiales candidates pour le toponyme de *Paris*. En s'appuyant sur l'appartenance (relation d'inclusion) de *Paris* aux territoires de la *France*, la désambiguïsation de Paris retourne *Paris,FR*.

S'inspirant de l'approche de (H. LI et al., 2003), nous proposons SHAREDPROP, un algorithme de désambiguïsation qui associe une entité à un toponyme selon la somme de ses proximités avec les entités fixes présentes dans un document. La proximité d'une entité candidate à une entité fixe est mesurée en fonction des relations spatiales (*c.f.* Chapitre 2.1.2) existantes entre elles. Le score de proximité $score(e_i)$ pour chaque entité candidate e_i est défini selon la formule suivante :

1. Le Port Racine est le plus petit port de France, situé en Normandie, dans le nord du département de la Manche. [Source]

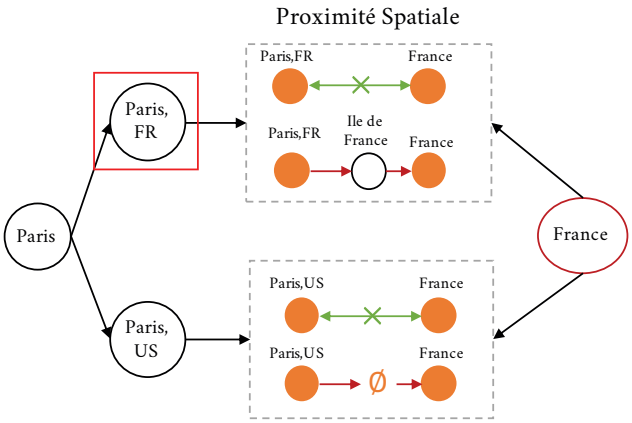


FIGURE 18 – Illustration du fonctionnement de l'algorithme SHAREDPROP sur le toponyme *Paris*. En rouge, le toponyme *France* qui n'est associé qu'à une entité spatiale. L'entité sélectionnée par l'algorithme est encadrée en rouge.

$$\mathrm{score}(e_i, E_f) = \sum_{e_j \in E_f} (\mathrm{scoreInclusion}(e_i, e_j) + \mathrm{scoreAdjacence}(e_i, e_j))$$

À chaque type de relation est associé à un score. Si une relation d'adjacence est identifiée entre e_i et e_k, alors $\mathrm{scoreAdjacence}$ retourne 2. Pour le calcul de $\mathrm{scoreInclusion}$, nous proposons de mesurer le nombre d'entités communes entre la chaîne d'inclusion d'une entité candidate et celle d'une entité fixe. Une **chaîne d'inclusion** est une liste chainée où une entité spatiale es_{n-1} est incluse dans une entité es_n. Par exemple, la chaîne d'inclusion de *Paris* est $\mathrm{Paris} \rightarrow$Ile-de-France$\rightarrow \mathrm{France} \rightarrow \mathrm{Europe} \rightarrow \mathrm{Terre}$. Le rôle des chaînes d'inclusion est de savoir si deux entités sont incluses l'une dans l'autre ou dans une même entité et sur quelle étendue dans la hiérarchie spatiale. Par conséquent, plus deux chaînes d'inclusion sont similaires, plus proches sont les entités dans la hiérarchie spatiale. Dans l'exemple de la Figure 19, nous avons un document composé de deux toponymes : *Paris* et *Boulogne-Billancourt* [1]. Boulogne-Billancourt correspond à l'entité fixe pour mesurer la cohérence de l'entité spatiale qui sera associée au toponyme *Paris*. Dans la Figure 19, le score d'inclusion entre les chaînes d'inclusion de *Paris,FR* et *Boulogne-Billancourt* est égal à 2, tandis que *Paris,US* et *Boulogne-Billancourt* ($\mathrm{scoreInclusion} = 0$). Dans cet exemple, vu qu'aucune relation d'adjacence n'a été identifiée, score(Paris(FR),{Boulogne-Billancourt}= 2. Le fonctionnement de *SharedProp* est détaillé dans l'Algorithme 8.

WIKICOOC. L'algorithme de désambiguïsation SharedProp est fondé sur la nécessité de parvenir à une cohérence spatiale entre les entités

1. Boulogne-Billancourt est une ville au sud-ouest de Paris

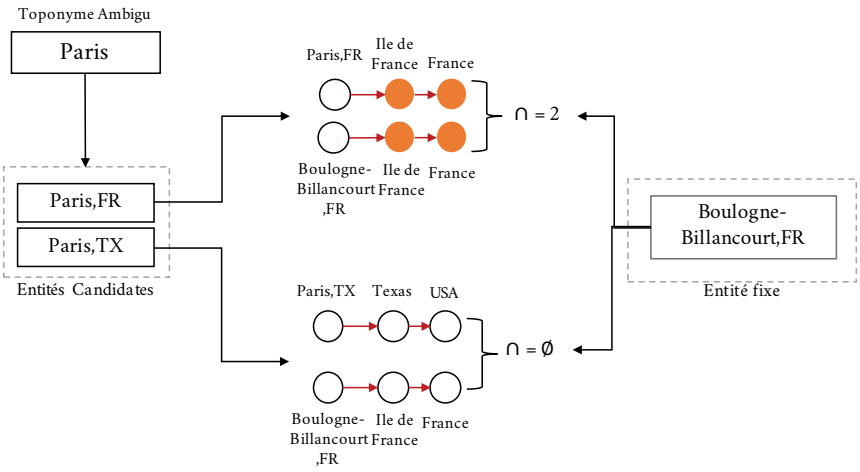


FIGURE 19 – Calcul du score d'inclusion pour chaque entité candidate d'un toponyme ambiguë (Paris).

identifiées dans un document. Cependant, il arrive que les entités spatiales dans un document ne partagent pas nécessairement une relation de proximité spatiale mais une proximité contextuelle. Par exemple, *Paris,FR* et *New York City* sont deux entités régulièrement mentionnées sur des supports différents (encyclopédie, presse, etc.).

Dans (DELOZIER et al., 2015), les auteurs proposent de mesurer la probabilité d'association d'un mot à une région. Ces régions correspondent à un découpage de la surface terrestre par bloc de $50km^2$. Pour mesurer la probabilité d'appartenance d'un mot avec une région, les auteurs utilisent la fréquence d'apparition d'un mot dans les pages *Wikipedia* géolocalisées, i.e. coordonnées dans la zone *info-box*. Au-delà des performances remarquables, l'approche de (DELOZIER et al., 2015) permet de "géocoder" les toponymes sans utiliser d'index géographique. Certains domaines de recherche comme l'analyse de documents historiques bénéficient de ce genre d'approche à cause de l'absence d'index géographiques pour certaines époques. Le principal inconvénient de cette approche concerne le stockage des valeurs de cooccurrence entre mots et régions (64,000 régions * nombre de mots dans le corpus).

Dans nos travaux, nous nous intéressons à des documents contemporains et l'utilisation des données de cooccurrences entre les entités spatiales présentes dans notre index géographique (*c.f.* Section 3) sont suffisantes. Par conséquent, nous proposons une approche similaire, WIKICOOC, qui s'appuie uniquement sur la fréquence des cooccurrences entre les entités spatiales. L'algorithme de désambiguïsation WIKICOOC se déroule en deux étapes. La première étape consiste à construire un *graphe de cooccurrence* intégrant la fréquence de cooccurrences entre chacune des entités spatiales candidates des toponymes d'un document. Par exemple, la Figure 20 illustre le graphe de co-

occurrences des toponymes : *Paris*, *Cherbourg*, *Montpellier*. Dans ce graphe, les entités spatiales sont représentées par les sommets et les arêtes intégrant les valeurs de cooccurrences entre elles.

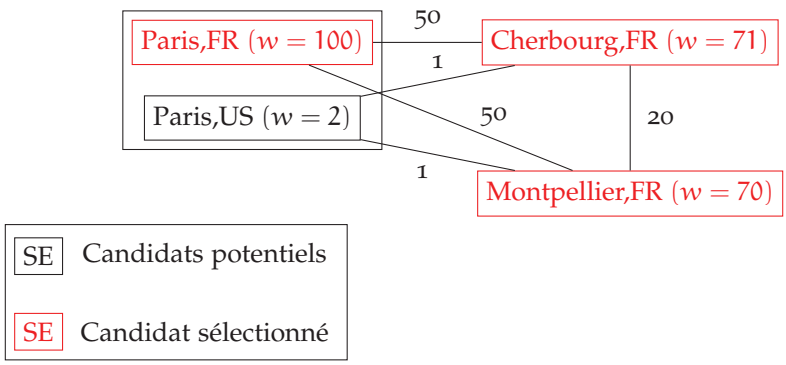


FIGURE 20 – Graphe de cooccurrence utilisée dans l'algorithme de désambiguïsation WIKICOOC.

Une fois le graphe généré, la seconde étape consiste à mesurer le degré pondéré de chacun des sommets, i.e. entités spatiales candidates. En théorie des graphes, le degré d'un sommet correspond au nombre d'arrêtes incidentes. Le degré pondéré d'un sommet correspond au nombre d'arrêtes incidentes pondérées par leur poids. Dans le graphe de la Figure 20, le degré pondéré w du sommet Paris,FR est égal à $w = 50 + 50 = 100$. Enfin, pour chaque toponyme, l'entité spatiale candidate avec le degré pondéré le plus élevé est sélectionnée.

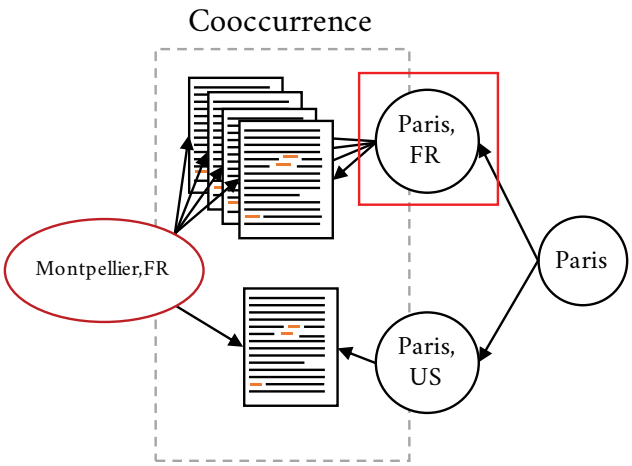


FIGURE 21 – Illustration du fonctionnement de l'algorithme WIKICOOC. L'entité sélectionnée par l'algorithme est encadrée en rouge.

SÉLECTION DES ENTITÉS SPATIALES CANDIDATES. Pour sélectionner les entités spatiales candidates pour chaque toponyme, nous utilisons l'index géographique GEODICT (*c.f.* Chapitre 3). Pour rechercher des entités dans GEODICT, les données sont stockées dans le SGBD

ElasticSearch [1]. La sélection des candidats s'effectue selon les trois requêtes suivantes :

1. La **première requête** retourne les entités ayant le même toponyme dans la langue du document.
2. La **seconde requête** retourne les entités avec un alias égal au toponyme dans la langue du document.
3. La **troisième requête** retourne les entités avec un toponyme similaire, *e.g. "prades-le-Lez" vs "Prades-le-Lez", "Region Alaotra-Mangoro" vs "Alaotra-Mangoro"*.

Nous venons de présenter le processus d'extraction des entités spatiales dans un document. Pour obtenir la forme complète de la STR, la section suivante présente le processus d'extraction des relations spatiales.

2.3 GEOCOMPLETION : EXTRACTION DES RELATIONS SPATIALES

Pour intégrer la notion de topologie entre les entités spatiales dans un document, nous proposons d'utiliser deux relations spatiales : l'adjacence et l'inclusion (*c.f.* Section 2.1.2.2). En s'appuyant sur les informations associées (*c.f.* Chapitre 3) aux entités spatiales extraites dans l'étape précédente, nous proposons deux procédures d'extraction, une pour chaque relation spatiale définie dans la STR.

2.3.1 *Extraction des relations d'inclusion*

Pour identifier une relation d'inclusion entre deux entités spatiales, deux catégories de données peuvent être utilisées. La première catégorie rassemble les méta-données indiquant explicitement les relations spatiales existantes entre certaines entités. La seconde catégorie correspond aux empreintes spatiales disponibles et représentées par un point, polygone ou encore multi-polygone [2]. Dans Geodict (*c.f.* Chapitre 3), nous avons collecté les variables suivantes pour déterminer si deux entités spatiales partagent une relation d'inclusion :

— **La propriété "located in the administrative territorial entity" (P131 [3]).** Cette propriété indique la position de l'entité dans la hiérarchie administrative, e.g. *Le département de l'Hérault fait partie de la région Occitanie.* **[1ère catégorie]**

— **La propriété "located on the specified landform" (P706 [4])**. Elle indique la position d'une entité dans un modelé, *e.g. La statue de Liberté se situe sur Liberty Island.***[1ère catégorie]**

1. https://www.elastic.co/fr/products/elasticsearch
2. Dans OpenStreetMap, la frontière de la France est décomposée en plusieurs polygones, un pour la France Métropolitaine, un pour la Guyane Française, etc.
3. Source :https://www.wikidata.org/wiki/Property_talk:P131
4. Source :https://www.wikidata.org/wiki/Property_talk:P706

— **Les relations hiérarchiques de *Geonames*[1]** Tout comme *P131*, elle indique les relations hiérarchiques administratives entre les différentes entités. **[1ère catégorie]**

— **Géométrie de OpenStreetMap** Contrairement à *Geonames*, OpenStreetMap propose les délimitations (Polygone) de plusieurs entités spatiales.

Pour déterminer si une entité spatiale es_1 est incluse dans une autre entité es_2, la procédure suivante est utilisée. Dans un premier temps, l'existence de la relation dans les variables de première catégorie (P131, P706 et relations hiérarchique dans *Geonames*) est vérifiée. Si aucune relation n'est explicitement indiquée, nous vérifions si la géométrie (Polygone ou Point) de es_1 est contenue dans la géométrie (Polygone) de es_2.

2.3.2 *Extraction des relations d'adjacence*

Tout comme pour la relation d'inclusion, nous identifions les relations d'adjacence entre les entités spatiales d'un document à l'aide des variables suivantes présentes dans GEODICT :

— **La propriété "shares border with" (P47[2]).** Elle indique pour une entité, quelles sont les autres entités avec lesquelles elle partage sa frontière.

— **La géométrie d'une entité.** Chaque entité est associée avec une empreinte spatiale. Celle-ci est représentée sous forme de vecteurs (Point, Polygone), ou plus généralement sa géométrie.

Pour déterminer si deux entités es_1 et es_2 sont adjacentes, nous proposons la procédure suivante :

1. **Vérification de l'existence de relation d'inclusion.** Si une relation d'inclusion existe entre les deux entités, alors aucune relation d'adjacence n'existe.
2. **Test de l'existence de la relation à l'aide de la variable *P47*.**
3. **Test d'intersection entre les signatures spatiales.** Si les deux signatures spatiales de es_1 et es_2 s'entrecoupent alors une relation d'adjacence existe.

Dans les deux sections précédentes, nous avons présenté les deux processus à la base de la génération de la représentation spatiale d'un document. Dans les sections suivantes, nous nous interrogeons sur la complétude de la STR et les solutions à apporter aux problèmes soulevés.

1. Source : https://www.geonames.org/export/place-hierarchy.html
2. Source : https://www.wikidata.org/wiki/Property_talk:P47

2.4 TRANSFORMATION DE LA STR

Au cours du processus de *georepresentation*, certaines entités nécessaires à la représentation de la spatialité d'un document ne sont pas identifiées. Parmi les différentes raisons, nous nous focalisons sur les concepts de **bruit** et de **silence**. Le bruit et le silence sont deux concepts utilisés dans l'évaluation d'un système de Recherche d'Information. Le **silence** représente la proportion d'informations non-reconnues par un système. Le **bruit** correspond aux informations erronées retournées par ce même système. Dans un processus de *geoparsing*, le bruit est associé à la détection d'un toponyme erroné (*geotagging*) ou à une désambiguïsation incorrecte (*geocoding*). Le silence correspond à la non-détection d'un toponyme dans la donnée ou à l'inexistence de l'entité associée dans l'index géographique. Cette perturbation dans la détection des entités spatiales d'un document s'appuie sur différents facteurs. Premièrement, il est possible que certaines entités spatiales soient mentionnées par un alias (e.g. la Ville Lumière), un acronyme (e.g. U.K.), ou une variation orthographique culturelle du toponyme (e.g. Paname). Un deuxième facteur repose sur l'absence de descripteurs dans un document nécessaires à l'identification de toponyme.

En termes de précision et de rappel, le ***silence*** *correspond à* $1 - \mathtt{rappel}$ *et le* ***bruit*** *à* $1 - \mathtt{précision}$

Pour palier ce manque d'information, nous proposons des transformations applicables sur l'ensemble des entités spatiales présentes dans la STR : la généralisation et l'extension.

2.4.1 *Généralisation*

Dans la STR, la configuration spatiale d'un document est articulée par l'ensemble des entités spatiales identifiées. Dans cette section, nous souhaitons poser la question suivante : *Les entités spatiales dans un document sont-elles suffisantes ?* Nous identifions deux facteurs d'incomplétude de l'information intégrée dans la STR dans un contexte de mise en correspondance.

Le premier facteur repose sur la *finesse* des entités spatiales identifiées. La finesse d'une entité spatiale se rapporte à sa position sur l'échelle spatiale, aussi définie comme la **granularité**. Comme illustré par la Figure 22, les entités spatiales telles les rues ou les quartiers possèdent une granularité fine à l'inverse des régions ou des pays. Dans la mise en correspondance de documents, l'existence d'entité spatiale avec une granularité fine (e.g. village, hameau) dans plusieurs documents peut être rare. Toutefois, certaines entités d'un document à l'autre peuvent partager des relations de proximité dans la hiérarchie spatiale. Par exemple, entre un document centré sur *Montpellier* et un document centré sur la *France*, les deux partagent à minima une relation d'inclusion. Par conséquent, la similarité spatiale est non-nulle. Ce

qui nous amène au second facteur, les entités implicites. Dans certains documents comme des articles de journaux régionaux, la mention d'indices spatiaux comme le pays, ou la région dans laquelle se déroule un événement n'est pas nécessaire. Dans ce cas, le lecteur déduit cette information selon le contexte géographique de publication du journal.

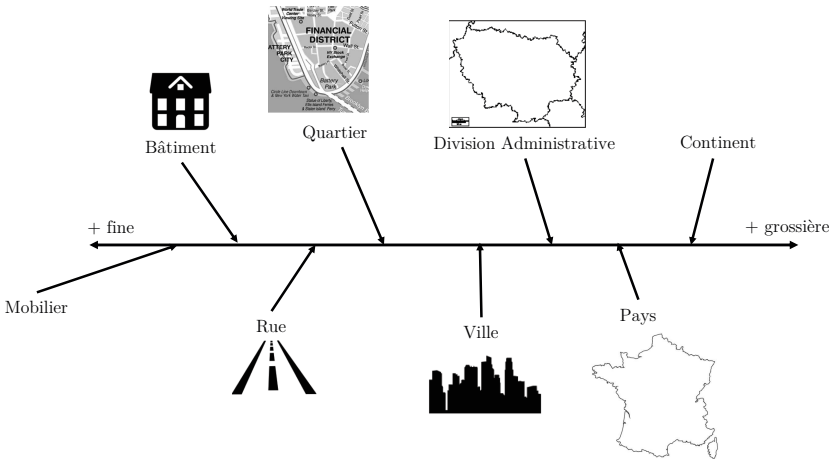


FIGURE 22 – Exemple de positionnement de différentes classes d'entités spatiales selon leur granularité (de la plus fine à la plus grossière).

Pour répondre à ces deux problèmes, nous proposons d'*élever* l'ensemble des entités dans la hiérarchie spatiale jusqu'à une certaine limite. Dans l'exemple illustré dans la Figure 23, cette limite est fixée au niveau départemental. Par conséquent, l'entité *Montpellier* est remplacée par l'entité spatiale du département de *l'Hérault*. Dans nos travaux, nous proposons une première transformation de la STR nommée la généralisation. La **généralisation** d'une STR consiste à remplacer l'ensemble des entités spatiales de la STR par une entité parente dans la hiérarchie spatial. Ces entités parentes sont sélectionnées selon un niveau de granularité défini par une classe. Par exemple, si le niveau de granularité est défini selon la classe Pays [1], l'entité *Paris* est remplacée par *France*.

Le processus de généralisation se déroule de la manière suivante. Nous considérons E, l'ensemble des entités spatiales présentes dans un document et $b \in \{\text{région}, \text{pays}\}$ la limite de la généralisation. Pour chaque entité spatiale $es_j \in E$, sa chaîne d'inclusion (*c.f.* Section 2.2.2) est générée. Par exemple, la chaîne d'inclusion de *Caen* sera $\text{Caen} \rightarrow \text{Calvados} \rightarrow \text{Normandie}\ldots$ Ensuite, chaque entité spatiale est remplacée par une entité de la chaîne ayant la classe b. Dans l'exemple précédent, c'est l'entité Calvados qui sera sélectionnée. Une entité spatiale n'est pas transformée si elle possède une classe avec une

1. Dans la classification de *Geonames*, le code correspondant est A-PCLI.

granularité supérieure ou égale à b. Dans la Figure 23, l'entité spatiale *France* n'est pas généralisée car elle possède une classe supérieure à la limite fixée (i.e. département). Ce processus est repris dans la Procédure 12 et illustré dans la Figure 23.

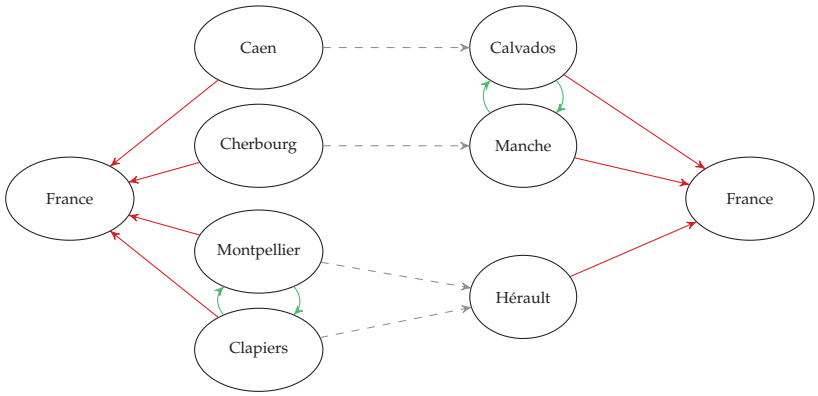


FIGURE 23 – Une STR (à gauche) et sa forme généralisée (à droite). La flèche grise relie une entité à l'entité qui la remplace au cours du processus de généralisation. Ici, la limite de la généralisation est fixée au niveau départemental.

Si la généralisation permet d'obtenir une représentation (i.e. STR généralisée) qui permet d'augmenter le nombre de correspondance entre les STR générées sur un corpus, nous perdons de la finesse dans les correspondances retournées. En s'appuyant sur cette observation, nous proposons une seconde transformation dans la section suivante.

2.4.2 *Extension*

L'hypothèse principale de la généralisation consiste à élever certaines entités d'une STR avec une granularité plus élevée pour obtenir les correspondances implicites comme : les entités spatiales *Paris* et *France* sont différents mais *Paris* partage une relation d'inclusion avec *France*. De plus, en augmentant la granularité moyenne des entités formant la représentation, une entité est remplacée par une entité avec un rayonnement plus important. Cependant, le rayonnement d'une entité n'est pas complètement fondé sur sa position dans l'échelle spatiale. Prenons l'exemple, de *Clapiers* et *Montpellier*, deux villes voisines. Ces deux entités sont associées à une même granularité mais *Montpellier* aura un rayonnement plus important pour des raisons culturelles (musée, salle de concert), politiques (capitale départementale) et physiques (surface).

De ce constat, nous proposons une seconde transformation, **l'extension** qui étend l'ensemble des entités spatiales d'un document avec des entités populaires et proches de celle-ci. En suivant l'exemple illustré dans la Figure 24, du fait de la moindre réputation de l'entité spatiale *Prades-le-Lez*, l'entité *Montpellier* sera rajoutée à la STR. Avant

de présenter la procédure de l'extension, il est nécessaire de définir la notion de popularité d'une entité spatiale. Dans nos travaux, la **popularité** d'une entité spatiale est pondérée par la proportion de personnes en ayant la connaissance. Pour mesurer la popularité d'une entité spatiale, nous proposons de calculer la valeur du PageRank de sa page dans *Wikipedia*. Le PageRank (Page et al., 1999) est un algorithme permettant de mesurer le rang de chaque noeud (page web) dans un réseau (web) à l'aide des connexions qu'il entretient.

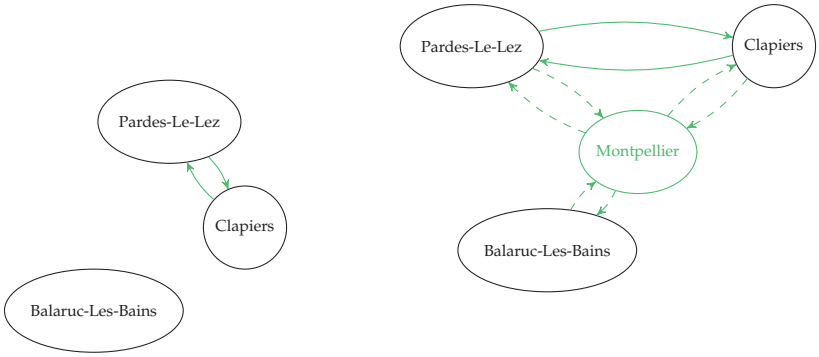


Figure 24 – Une STR (à gauche) et sa forme étendue (à droite) avec les paramètres (n=1 et r=50km).

Le processus d'extension se déroule selon 3 étapes successives. La première étape consiste à **calculer le score de popularité médian** de l'ensemble des entités spatiales présentes dans un document. Dans l'exemple de la Figure 24, le score média de popularité s_{median} est égal à :

$$\begin{aligned} s_{median}(\text{Prades-le-Lez, Clapiers, Balaruc-les-Bains}) &= median(0.41, 0.65, 0.95) \\ &= 0.65 \end{aligned}$$

En s'appuyant sur le score de popularité médian, la deuxième étape consiste à **sélectionner les entités ayant un score inférieur à** s_{median}. Dans notre exemple, seule *Prades-le-Lez* est sélectionnée. Dans la troisième étape, **n entités dans un radius r sont collectées** autour de chaque entité sélectionnée. Pour *Prades-Le-Lez*, l'entité la plus proche qui répond à ce critère s'avère être *Montpellier*. Enfin, les entités collectées sont intégrées dans la STR. La procédure en pseudo-code est définie dans l'Algorithme 13 en Appendix A.3.

Tout au long de ce chapitre, nous avons présenté le processus de représentation de la spatialité d'un document. La majorité de ce processus repose sur un index géographique nommé Geodict conçu dans le cadre de nos travaux. Le chapitre suivant présente Geodict et détaille son processus de création.

3 GEODICT : UN NOUVEL INDEX GÉOGRAPHIQUE

Dans le chapitre 2, nous avons présenté le processus de représentation de la spatialité d'un document. Au travers des deux étapes du processus, l'utilisation d'une base de connaissance, ici géographique, est nécessaire : un **index géographique**. Dans un premier temps, les index géographiques (*gazetteers* en anglais) étaient des livres dans lesquels étaient renseignés des noms de lieux, leurs orthographes autorisées ainsi que des descriptions contextuelles (Aurousseau, 1945 ; Goodchild et Hill, 2008). Avec l'essor de l'informatique dans les années 90, les index géographiques se numérisent et se transforment en **digital gazetteers** (Goodchild et Hill, 2008). Pour pouvoir générer la STR, l'index géographique utilisé doit contenir un ensemble d'informations spatiales nécessaires à différentes phases du processus global :

1. geocoding. Pour identifier le plus d'entités spatiales possibles, chaque entité répertoriée dans un index géographique doit être associée avec plusieurs variations de son toponyme selon la langue, *e.g. New York City (en), New York (fr)*, et les différents alias [1] utilisés (*e.g. NYC, Big Apple (en)*).
2. extraction de relations spatiales. L'identification de relations spatiales entre les entités d'un document s'appuie sur deux catégories de variable. La première catégorie regroupe les variables qui indiquent explicitement les relations spatiales existantes entre certaines entités. Par exemple, il existe une propriété dans Wikidata indiquant l'adjacence entre deux entités, *e.g. P47 (Paris, Seine-Saint-Denis)*. La deuxième catégorie regroupe les variables contenant l'empreinte spatiale (*c.f.* Section 2.1.2) d'une entité, *e.g. latitude-longitude, limites administratives*.

Pour répondre à nos besoins, nous proposons un nouveau *gazetteer* Geodict (**Fize, Jacques** et Shrivastava, 2017), généré à partir des trois sources de données suivantes : *Wikidata* [2], *Geonames* [3] et *OpenStreetMap* [4]. Par construction, Geodict se veut plus riche par sa personnalisation. De part sa nature encyclopédique, *Wikidata* offre un large panel d'informations (spatiales ou non) pour chaque entité. De plus, elle permet d'avoir accès à différentes représentations d'une entité à travers les liens vers une liste exhaustive de sources. Pour cette raison, la construction de Geodict s'appuie sur une extraction

1. toponymes alternatifs
2. https://www.wikidata.org/
3. http://www.geonames.org
4. https://www.openstreetmap.org/

Geonames est composé de ≈ 11 millions d'entités

de *Wikidata*. *Geonames* est un index géographique largement utilisé avec une bonne couverture spatiale, une classification fine [1] et les relations hiérarchiques entre les entités. Dans GEODICT, nous reprenons le système de classification pour catégoriser les entités, ainsi que les relations hiérarchiques pour l'identification de relations d'inclusions. Si *Geonames* assigne des coordonnées (latitude-longitude) pour chaque entrée, aucune représentation (e.g. Polygone) décrivant les frontières des entités n'est disponible. Pour cette raison, nous utilisons *OpenStreetMap* pour extraire les délimitations des surfaces de chacune des entités. L'ensemble des variables collectées dans les différentes sources sont indiquées dans la Table 1.

Variable	Source	Exemple de valeur
Identifiant Unique	Geodict	*GD1286273*
Autres identifiants	Wikidata, Geoname, OSM	*OSM :62578*
Labels	Wikidata	*fr : Cologne, de : Köln, etc.*
Frontières	OpenStreetMap	*Polygon, MultiPolygon*
Coordonnées	Wikidata	*Point(50.942, 6.957)*
Class(es)/Concept(s)	Geonames	*P-PPLA :seat of a first-order...*
Relations spatiales(P47 [2], P131 [3], P706 [4])	Wikidata	*P47 : France shares a border with Belgium*

TABLE 1 – Données disponibles pour chaque entrée de GEODICT. Chaque variable est accompagnée de sa source et d'un exemple de valeur.

Dans la section suivante, nous présentons les sources de données utilisées dans GEODICT.

3.1 SOURCES UTILISÉES POUR CONSTRUIRE GEODICT

WIKIDATA. *Wikidata* est une base de connaissances disponible et modifiable par tous. Chaque entrée de *Wikidata* se divise en deux parties. La première partie détaille les orthographes utilisées pour représenter l'entrée. Les différentes orthographes sont réparties selon la langue ou leur emploi (label officiel ou alias). Dans la deuxième partie, une entrée est associée avec un ensemble d'informations représentées par des *statements*. Semblable à un triplet RDF [5], un *statement* est composé de trois variables :

— une **propriété**, *e.g. pays (P17)*

— la **valeur** ou l'**item** associé, *e.g. France (Q142)*

— et la **source** de l'information *e.g. https ://en.wikipedia.org/wiki/Paris*

OPENSTREETMAP. OpenStreetMap (HAKLAY et WEBER, 2008 ; UNIVERSITY OF NOTTINGHAM, GB et al., 2017) est une source de données

1. https://www.geonames.org/export/codes.html
5. Resource Description Framework https://www.w3.org/RDF/

géographiques structurée selon trois types d'entités : les **noeuds** (*nodes* en anglais), les **chemins** (*ways* en anglais) et les **relations**. Un noeud représente un élément géographique associé avec des coordonnées latitude-longitude. Les chemins sont des ensembles de noeuds pour représenter une *polyline e.g. route, rue*, et des polygones *e.g. frontières*. Une relation est une collection de noeuds et de chemins. Par exemple, la France est représentée par une *relation* contenant les différentes limites administratives (*chemins*) — France Métropolitaine, DOM-TOM— ainsi que leur différents centroïdes (*noeuds*).

Pour intégrer les informations connexes d'une entité dans OSM, un ensemble d'attributs (ou tags) est utilisé. Par exemple, l'entité France possède un attribut *population* associé avec la valeur 63,070,344 [1].

GEONAMES. *Geonames* est un index géographique contenant plus de 11 millions d'entités, classées à l'aide 645 classes [2]. Une classe est composée de deux identifiants, une catégorie générale ou *feature-class* (e.g. *A* = country, state, region, ...) et une sous-catégorie ou *feature code* (e.g. ADM1 = first-order administrative division). Chaque entrée est associée avec différents attributs : *toponyme officiel, alias, coordonnées latitude-longitude, classe, pays, région, population, élévation, fuseau horaire, date de la dernière modification*. De plus, *Geonames* contient les informations de relations hiérarchiques reliant les différentes entrées (Montpellier → Hérault → Occitanie → France).

Dans la section suivante, nous présentons le processus de création de GEODICT.

3.2 PROCESSUS DE CRÉATION DE GEODICT

Pour combiner les différentes sources de données, nous avons mis en place un processus divisé en 6 étapes.

La première étape du processus de création de GEODICT consiste à produire une base sur laquelle de nouvelles informations provenant des différentes sources seront ajoutées. De part les différents liens vers *Geonames* et OpenStreetMap, nous choisissons de fonder GEODICT sur **une première extraction des entités spatiales dans *Wikidata***. *Wikidata* étant une base de connaissances généraliste, nous devons définir quelles sont les propriétés inhérentes à une majorité d'entités spatiales. Pour cette raison, nous choisissons d'utiliser les propriétés *P1566* et *P1402* qui indiquent respectivement l'identité de la représentation dans *Geonames* et dans OpenStreetMap. Cependant, il se peut que certaines entités spatiales dans *Wikidata* ne possèdent pas *statement* associé avec cette propriété. Dans ce cas, les propriétés P706 et P131,

*Les **propriétés P706 et P131** sont détaillées dans la section 2.3.1*

1. La source de cette information est indiquée dans un autre attribut *source :population*
2. Source : https ://www.geonames.org/about.html

respectivement *located in territorial features* et *located in administrative entity* sont utilisées.

Une fois la base de GEODICT générée, chaque entité doit être associée à une classe. Pour **associer une classe à chaque entité**, nous utilisons la classification proposée par la représentation d'une entité dans *Geonames*. Cependant, certaines des entités extraites dans *Wikidata* ne sont pas reliées ou n'existent pas dans *Geonames*. Pour palier ce problème, nous proposons d'utiliser les valeurs associées avec la propriété *P31*, ou *instance of*. La propriété *instance of* permet de caractériser une entité dans *Wikidata*, e.g. France *instance of* Mediterranean country. Pour aligner les différentes valeurs prises par *instance of* avec une classe de *Geonames*, nous proposons d'utiliser les valeurs de cooccurrences lorsque le lien entre les deux représentations existes. Chaque valeur prise par *instance of* est associée à la classe dans *Geonames* avec la plus forte cooccurrence.

Comme pour *Geonames*, le lien entre la représentation entre *Wikidata* et OpenStreetMap peut être absent. Dans cette troisième étape, nous cherchons à **compléter les liens existants entre les différentes représentations d'une entité dans *Wikidata* et OpenStreetMap**. Nous avons remarqué que si le lien vers *Wikidata* n'est indiqué dans OpenStreetMap, il arrive que le lien vers la page *Wikipedia* existe. Un des fondements de la création de *Wikidata* indique que chaque article *Wikipedia* possède une entrée dans *Wikidata* (VRANDEČIĆ et KRÖTZSCH, 2014). Par conséquent, nous récupérons l'identifiant Wikidata de la page *Wikipedia* présente dans chaque entrée d'OpenStreetMap. Chacun des liens identifiés entre les représentations dans *Wikidata* et OpenStreetMap est intégré à GEODICT.

Au delà des propriétés requises pour la construction d'un index géographique (nom, classe, empreinte géographique), il est possible d'**ajouter des informations additionnelles** (taux de fertilité, identifiant de l'entité dans d'autres sources, etc.). Cette quatrième étape consiste à **extraire des propriétés additionnelles** selon le besoin d'un utilisateur. Pour cela, un fichier de configuration permettant d'ajouter de nouvelles propriétés est mis à disposition (*c.f.* Annexe A.2.4).

Certaines entités sont associées avec une empreinte géographique plus complexe que des coordonnées latitude-longitude. Dans le cas d'une ville, cette empreinte correspond à sa démarcation administrative. Pour chaque entité associée avec un identifiant dans OpenStreetMap, **on associe sa représentation vectorielle**, e.g. Polygon ou MultiPolygon.

Si *Wikidata* possède une couverture d'entités spatiales significatives, certaines entités apparaissent uniquement dans *Geonames*. Pour finaliser GEODICT, les entités de *Geonames* absentes de GEODICT sont ajoutées.

L'extraction d'OpenStreetMap *utilisée est disponible à cette adresse :* `https://github.com/missinglink/osm-boundaries`

3.3 STATISTIQUES

Dans la Table 2, différents index géographiques sont comparés avec GEODICT en termes de nombre d'entités, de la présence de frontières, des liens vers d'autres sources de données et de la possibilité de personnalisation. La Table 3 indique le nombre total, puis par classe, d'entités spatiales dans GEODICT.

Gazetteer	**# Entités spatiales**	**Empreintes Spatiales**	**Liens disponibles**	**Personnalisable**
Geonames	11,301,264	Centroïde, frontière pays	*Wikipedia*	-
OpenStreetMap	7,009,806 (relations) + 598,804,573 (ways)	Centroïde, frontière	Wikidata, Wikipedia	✓ (uniquement export)
Geodict	16,162,026	Centroïde, frontières	Geonames, OSM, Wikidata, Wikipedia	✓

TABLE 2 – Comparaison avec d'autres index géographiques.

Classe	**# Geonames**	**# Geodict**	$\frac{\text{Geodict}}{\text{Geonames}}$
T (mountain, hill, rock, ...)	1,539,495	1,665,167	8,16%
H (stream, lake, ...)	2,167,385	2,341,817	8,04%
P (city, village, ...)	4,423,959	6,378,366	44,17%
S (spot, building, farm, ...)	2,315,728	4,065,027	75,53 %
R (road, railroad, ...)	41,148	462,002	1122,78%
L (parks, area,...)	400,561	606,206	51,33 %
A (country, region, ...)	371,782	579,151	55,77 %
V (forest, health, ...)	45,324	52,787	16,46%
U (undersea)	17,338	11,503	-34,66 %
Avec Polygone	0	173,292	-
Total	11,301,264	16,162,026	43,01%

TABLE 3 – Comparaison du nombre d'entités dans GEODICT et *Geonames*. La comparaison est effectuée à l'aide des classes générales définies par *Geonames* et utilisées dans GEODICT.

En termes de couvertures géographiques, les chiffres d'OpenStreetMap sont les plus élevés. Toutefois, OpenStreetMap est une ressource cartographique, ce qui implique qu'elle ne se focalise pas uniquement aux *locations*, e.g. délimitations zones forestières, chemin de randonnées, etc. Concernant *Geonames* et *Getty*, ces deux index géographiques contiennent uniquement[1] le centroïde de chacune des

1. *Geonames* fournit uniquement les frontières de 249 pays

entités spatiales. Inversement, OpenStreetMap et GEODICT propose les représentations vectorielles des surfaces couvertes par certaines entités. Enfin, si OpenStreetMap permet aux utilisateurs de personnaliser les exports, le processus de création GEODICT peut être modifié pour ajouter des informations [1] disponibles dans *Wikidata*.

3.4 CONCLUSION ET PERSPECTIVES

Nous proposons GEODICT, un index géographique généré à l'aide d'un processus de collecte d'informations sur trois sources : *Geonames*, *Wikidata* et OpenStreetMap. Dans le cadre de nos travaux, le processus de génération de GEODICT sélectionne des variables spécifiques dans chacune des représentations d'une entité spatiale : toponyme dans différentes langues (*Wikidata*), classe (*Geonames*), les coordonnées de son centroïde (*Wikidata*, *Geonames*), les entités spatiales adjacentes (*Wikidata*), frontières (OpenStreetMap), etc. La couverture des entités spatiales de GEODICT propose une variété d'informations supérieure à la plupart des index géographiques disponibles, tout comme le nombre d'entités spatiales renseignées. Enfin, le processus de génération de GEODICT est personnalisable selon les besoins de l'utilisateur.

Par exemple, dans l'analyse de documents historiques, les documents étudiés réfèrent à des entités spatiales disparues. Parmi les sources utilisées dans la fabrication de GEODICT, WIKIDATA donne la possibilité d'indiquer une date ou une période pour une valeur d'un *statement*. Par exemple, la propriété *capital* (P36 [2]) contient différentes entités en fonction des périodes historiques. Par exemple, ceci permet de voir que la ville de Bayeux [3] a été la capitale de la France durant la libération de juin à septembre 1944. Dans le domaine d'analyse spatiale de documents anciens, l'intégration d'informations associées à différentes périodes historiques dans GEODICT permettrait l'utilisation de nos méthodes.

Dans le chapitre suivant, nous abordons les méthodes utilisées pour mesurer la similarité entre les représentations générées.

1. Des informations spatiales ou non.
2. Source : https://www.wikidata.org/wiki/Property_talk:P36
3. Ville dans la département du Calvados, France

4 GEOMATCHING : APPARIEMENT DES STR

Dans nos travaux, nous proposons un processus de mise en correspondance spatiale se déroulant sur deux phases. La phase de *Georepresentation* (*c.f.* chapitre 2) extrait puis intègre les informations spatiales d'un document dans une représentation dédiée, la Spatial Textual Representation (*c.f.* section 2.1.3). Dans ce chapitre, nous présentons la deuxième phase du processus, le *Geomatching*, qui consiste à appairer les documents d'un corpus à l'aide de la STR.

Afin de faire correspondre les différentes STRs, des mesures de similarités doivent être étudiées. Étant donnée la nature de la STR, i.e. graphe, nous proposons de mesurer la similarité entre les STR avec des méthodes appartenant au *graph-matching*.

Pour cela, ce chapitre est organisé de la manière suivante. Dans la section 4.1, nous définissons les notions essentielles associées au *graph-matching* et aux algorithmes présentés. Dans la section 4.2, nous proposons une définition générale du *graph matching*. Les différentes mesures de similarité utilisées dans nos expérimentations sont présentées en sections 4.3 et 4.4.

4.1 LA THÉORIE DES GRAPHES

Afin d'aborder la définition du problème de Graph Matching, nous définissons plusieurs notions essentielles de la théorie des graphes.

GRAPHE. Un graphe G est une structure de données définie par $G = (V, E)$; où V représente l'ensemble des sommets (ou noeuds) et E l'ensemble d'arêtes qui relient les sommets de V. Une arête e est définie par $e = (u, v) \in E \subset V \times V$.

SOUS-GRAPHE. Un graphe $G_2 = (V_2, E_2)$ est un sous-graphe de $G_1 = (V_1, E_1)$ si et seulement si $V_2 \subset V_1$ et $E_2 \subset E_1$.

GRAPHE DIRIGÉ ET NON DIRIGÉ. Dans un graphe, l'ensemble des arêtes qui relient les sommets peuvent être dirigées ou non-dirigées. Soit deux arêtes $e_1 = (u, v)$ et $e_2 = (v, u)$ où $e_1, e_2 \in E$, G est un **graphe dirigé** si et seulement si $e_1 \neq e_2$. Inversement, G est un **graphe non-dirigé** si et seulement si $e_1 = e_2$. Par convention, une arête appartenant à un graphe dirigé est appelé *arc*.

TAILLE DE GRAPHE. La **taille** d'un graphe $|G|$ correspond au nombre de sommets qui le compose.

$$|G| = |V|$$

DENSITÉ D'UN GRAPHE. La **densité** D correspond au rapport entre le nombre d'arêtes et le nombre de noeuds dans un graphe.

$$D(G) = \frac{|E|}{|V|^2}$$

GRAPHE ATTRIBUÉ. Un graphe attribué (en anglais *attributed graph*), est un graphe dans lequel les noeuds et/ou les arêtes sont associés à un ou plusieurs attributs V_i et E_i.

$$G = (V, E, V_0, \dots V_n, E_0, \dots, E_m)$$

Parmi les graphes attribués, il existe deux catégories très représentées : les graphes étiquetés ou *labeled graphs*, et les graphes pondérés ou *weighted graph*.

Un **graphe étiqueté** est un graphe avec un label $l_v \in L_V$ (*resp.* $l_e \in L_E$) attribuée à chaque sommet (*resp.* arêtes). Ces labels sont associés aux sommets (*resp. arêtes*) à l'aide d'une fonction α (*resp.* β).

— $G = (V, E, \alpha, \beta, L_V, L_E)$

— $\alpha(v_i) = l_j$ où $l_j \in L_V$, $v_i \in V$

— $\beta(e_i) = l_k$ où $l_k \in L_E$, $e_i \in E$

Un **graphe pondéré** est un graphe où chaque arête est associée à un poids $w \in W_E$.

$$G = (V, E, W_E)$$

PLUS GRAND SOUS-GRAPHE COMMUN. Soit :

— $G_1 = (V_1, E_1)$ et $G_2 = (V_2, E_2)$ deux graphes.

— SG_{G_1,G_2}, l'ensemble des sous-graphes communs à G_1 et G_2.

— MCS, un sous-graphe appartenant à SG_{G_1,G_2}.

MCS est le plus grand sous-graphe commun de G_1 et de G_2 si et seulement si $\nexists\, SG \in SG_{G_1,G_2}$ où $|SG| > |MCS|$.

CLIQUE. Une clique est un ensemble de noeuds $|C|$ d'un graphe $G = (V, E)$ où chaque sommet $v_i, v_j \in C$, $\exists(v_i, v_j) \in E$. Dans la Figure 25, l'ensemble de noeuds A, B, C forme une clique.

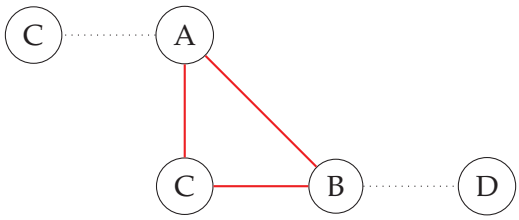


FIGURE 25 – Exemple de clique.

4.2 LE GRAPH MATCHING

Le *graph matching* ou *appariement de graphes* consiste à faire correspondre deux graphes en fonction de leur similarité. Ce sont des méthodes largement utilisées notamment en reconnaissance de formes, en biologie, en chimie moléculaire (Debnath et al., 1991), ou en vision par ordinateur. Parmi les différents méthodes proposées, nous distinguons deux catégories d'algorithme de *graph matching*, ceux qui visent un appariement exact (*exact graph matching*) ou un appariement partiel (*inexact graph matching)*.

Dans un problème d'**appariement exact**, deux graphes $G_1(V_1, E_1)$ et $G_2(V_2, E_2)$, où $|V_1| = |V_2|$ sont appairés si et seulement si pour toute arête $(u, v) \in E_2$, il existe une fonction bijective $f(x)$ telles que $(f(u), f(v)) \in E_1$. De manière générale, la recherche d'un appariement exact entre deux graphes consiste à détecter un isomorphisme de graphe entre G_1 et G_2. Si les deux graphes sont *étiquetés*, il existe un isomorphisme entre $G_1(V_1, E_1, \alpha_1, \beta_1)$ et $G_2(V_2, E_2, \alpha_2, \beta_2)$ si et seulement si :

— $\alpha_1(x) = \alpha_2(f(x))\ \forall x \in V_1$

— $\beta_1((x, y)) = \beta_2((f(x), f(y))\ \forall x, y \in E_1$

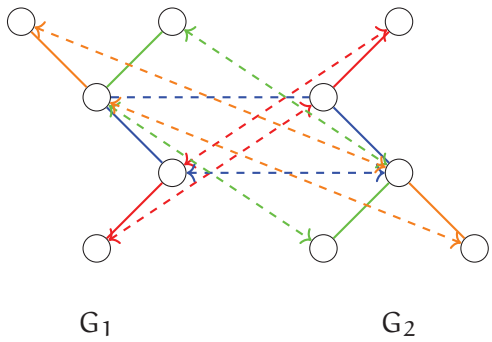


Figure 26 – Appariement exact de graphe.

En reconnaissance de formes, deux graphes sont comparés, un graphe modèle G_m intégrant la forme canonique d'une entité (e.g. plante, visage, lettre manuscrite) et un graphe extrait de la donnée G_d. La plupart du temps, le graphe G_d possède une structure différente du graphe modèle (e.g. nombre de sommets, placement). Dans ce cas, les méthodes d'**appariement partiel** (ou *inexact graph-matching*) sont utilisées. Contrairement à un appariement exact, l'objectif d'un appariement partiel est d'identifier les structures communes entre deux graphes G_1 et G_2.

APPARIEMENT DES STRS : QUELLE APPROCHE CHOISIR ? Dans un premier temps, la STR est une structure graphe s'appuyant sur les entités spatiales présentes dans le document. Etant donné que

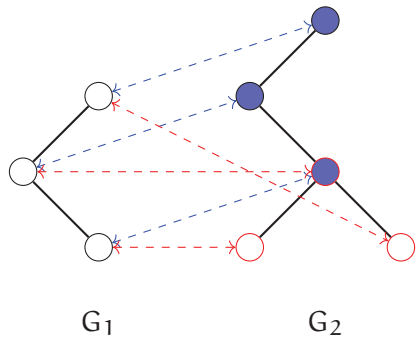


FIGURE 27 – Appariement inexact de graphe.

chaque document possède un nombre d'entités spatiales potentiellement différent du reste, nous utiliserons les mesures de similarité proposées dans le domaine d'appariement partiel de graphe. Dans les sections suivantes, nous présentons différents algorithmes d'appariement partiel de graphe selon deux catégories : *structure-based* et *pattern-based*.

4.3 ALGORITHMES STRUCTURE-BASED

Les algorithmes *Structure-based* s'appuient uniquement sur la comparaison des unités formant les graphes, *i.e.* les sommets dans $v_i \in V$, les arêtes $e_i \in E$ ainsi que leurs attributs. Nous avons sélectionné les algorithmes de référence dans la littérature (RIESEN, JIANG et Horst BUNKE, 2010) : **Maximum Common Subgraph** (BUNKE et ALLERMANN, 1983)(MCS), **Vertex/Edge Overlap** (PAPADIMITRIOU, DASDAN et GARCIA-MOLINA, 2010)(VEO), une mesure dérivée de **l'index de Jaccard** et la **Graph Edit Distance**.

4.3.1 *MCS*

MCS mesure la proportion que représente le plus grand sous-graphe commun entre G_1 et G_2, sur le plus grand graphe.

$$MCS(G_1, G_2) = \frac{|mcs(G_1, G_2)|}{max(G_1, G_2)} \tag{2}$$

4.3.2 *Vertex Edge Overlap*

Vertex Edge Overlap calcule le pourcentage de sommets et d'arrêtes en commun par rapport à la somme des tailles et densités des deux graphes.

$$VEO(G_1, G_2) = \frac{|V_{G_1} \cap V_{G_2}| + |E_{G_1} \cap E_{G_2}|}{|V_{G_1}| + |V_{G_2}| + |E_{G_1}| + |E_{G_2}|} \tag{3}$$

4.3.3 *Jaccard*

L'indice de Jaccard entre deux ensembles A et B est calculé à l'aide de la formule $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$. La mesure proposée calcule le produit des indices de Jaccard de chacune des composantes d'un graphe : les sommets V et les arêtes E.

$$\text{Jaccard}(G_1, G_2) = \frac{|E_{G_1} \cap E_{G_2}|}{|E_{G_1} \cup E_{G_2}|} \times \frac{|V_{G_1} \cap V_{G_2}|}{|V_{G_1} \cup V_{G_2}|} \tag{4}$$

4.3.4 *Graph Edit Distance*

La **Graph Edit Distance** (GED) (Fischer, Riesen et Horst Bunke, 2017) correspond à la séquence d'édition la moins coûteuse pour transformer un graphe $G = (V_1, E_1)$ en un graphe $H = (V_2, E_2)$. Le calcul de la GED est NP-difficile et plus les graphes sont de taille et de densité élevées, plus longue est la recherche de la séquence d'édition (Neuhaus, Riesen et Horst Bunke, 2006). Différents algorithmes proposant des heuristiques permettant d'approximer la valeur de la GED sont proposés (Neuhaus, Riesen et Horst Bunke, 2006 ; Riesen et Horst Bunke, 2009). La GED est définie selon la formule suivante :

$$\text{GED}(G_1, G_2) = \min_{(e_1,\ldots,e_k) \in \mathcal{P}(G_1,G_2)} \sum_{i=1}^{k} c(e_i) \tag{5}$$

Avec G_1, G_2 deux graphes, e_i une édition de graphe et $P(G_1, G_2)$ l'ensemble de séquences d'éditions possibles pour transformer G_1 en G_2.

Parmi les différentes implémentations de la GED, nous utilisons l'approche *Bipartite Graph Matching* proposée dans (Riesen et Horst Bunke, 2009).

4.3.4.1 *Bipartite Graph Matching (BP)*

L'algorithme **Bipartite Graph Matching** présenté dans (Riesen et Horst Bunke, 2009) propose de transformer le calcul de la GED en problème d'affection. **Un problème d'affectation** consiste à attribuer au mieux des tâches à des agents. Chaque agent peut réaliser une unique tâche pour un coût donné (resp. chaque tâche doit être réalisée par un unique agent). Chaque affectation entre un agent et une tâche est pondérée par un coût. L'objectif des algorithmes pour résoudre ce problème est d'identifier l'ensemble des affectations qui minimise le coût total de réalisation de l'ensemble des tâches. Pour résoudre le problème d'affectation, l'algorithme hongrois ou algorithme de *Munkres* (Munkres, 1957) est utilisé.

Ici, l'approche proposée consiste à transposer la recherche d'un chemin d'édition par un problème d'affectation entre les noeuds de

deux graphes G_1 et G_2. Une affectation $(\epsilon \rightarrow A)$ correspond à l'insertion du noeud A, $(B \rightarrow A)$ correspond à la substitution du noeud B avec le noeud A, et $(A \rightarrow \epsilon)$ la suppression du noeud A. Chaque affectation est associée à un *coût* calculé puis renseigné dans une *matrice coût* C_{G_1,G_2} (voir Figure 28). Le calcul de la matrice de coût entre deux graphes est détaillé dans l'algorithme 1. Une fois les coûts d'affectations calculés, l'algorithme hongrois retourne un ensemble de positions dans C_{G_1,G_2} correspondant au chemin d'édition utilisé pour calculer la *graph edit distance*.

Dans l'exemple illustré dans la Figure 28, le coût de toutes les transformations est fixé à 1. Pour passer du graphe G_1 au graphe G_2, deux transformations sont nécessaires : (i) l'ajout du sommet avec le label C et (ii) l'ajout de l'arête $e = (B, C)$. Dans C_{G_1,G_2}, la somme des coûts de ces deux opérations est égale à 2 $(\epsilon \rightarrow C)$[1]. Par conséquent, l'algorithme devrait retourner une distance d'édition égale à 2 entre G_1 et G_2. Sur cet exemple, la valeur retournée par l'algorithme BP est bien égale à 2.

Algorithme 1 : Calcul de la matrice de coût

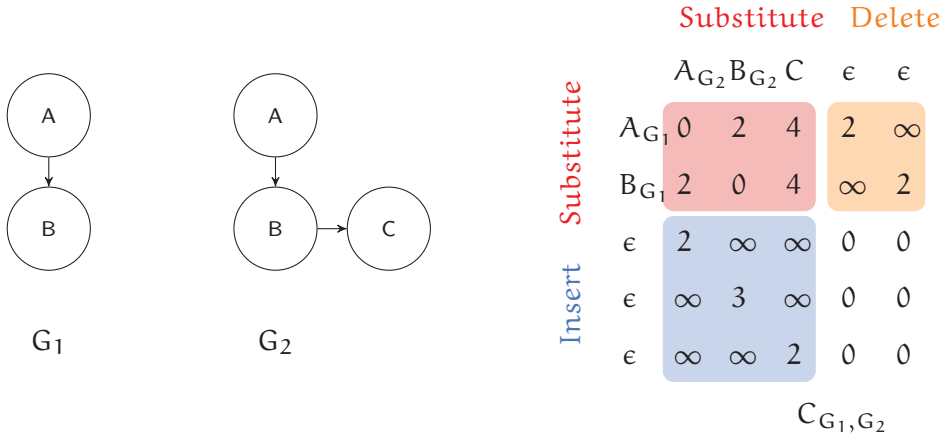
```
Input :
  — G, H deux graphes
  — c(e) avec e ∈ {vertexIns., VertexDel., edgeIns., edgeDel.}
costM_{n,m} ← initialize
for i ∈ 0...n do
    for j ∈ 0...m do
        costM_{i,j} ← substitutionCost(G.node[i], H.node[j], G, H)
for i ∈ 0...m do
    for j ∈ 0...m do
        costM_{i+n,j} ← insertionCost(G.node[i], G)
for i ∈ 0...n do
    for j ∈ 0...n do
        costM_{j,i+m} ← deletionCost(G.node[i], G)
return costM_{n,m}
```

4.4 ALGORITHMES PATTERN-BASED

La plupart des algorithmes *structured-based*[2] permettent de mesurer la similarité entre des graphes avec un temps d'exécution faible. Cependant, si les algorithmes *structure-based* prennent en compte la connectivité (i.e. arêtes) entre les sommets d'un graphe, ils ne comparent pas les motifs présents dans les graphes comparés.

Pour cette raison, nous avons ajouté à notre comparaison un deuxième ensemble d'algorithmes, les algorithmes **pattern-based**. Parmi ces al-

1. Le calcul d'insertion d'un noeud prend en compte l'ajout des arêtes incidentes.
2. A l'exception de la graph edit distance



$$BP(G_1, G_2) = \sum Munkres(C_{G_1,G_2}) = \sum\{0, 0, 0, 0, 2\} = 2$$

FIGURE 28 – Calcul de la GED $BP(G_1, G_2)$ entre deux graphes G_1 et G_2 à l'aide de la matrice de coût C_{G_1,G_2}.

gorithmes, nous comparons différentes méthodes telles que les *graph kernels*, le *graph-embedding*, ainsi qu'une méthode s'appuyant sur une représentation originale : le *Bag of Cliques*.

4.4.1 *Bag of Cliques*

Le modèle de **sac-de-mots** (*bag-of-words* en anglais) est une représentation simplifiée d'un corpus de document, où chaque document est considéré comme un multi-ensemble composé des *mots* présents dans le corpus (RICARDO et BERTHIER, 2011). Dans l'exemple de la Figure 29, l'ensemble des mots dans le corpus $C = \{S1, S2\}$ est {cat, dog, playing, the, is, a}. Soit V, l'ensemble des mots disponibles dans le corpus C, $BW_{C_i}\{(v_0, \ldots, v_n)|n = |V|\}$ est le vecteur représentant le document C_i avec v_i la valeur associée au mot $V[i]V_i$. Généralement, v_i correspond à le nombre d'occurrences (entier) ou la présence (booléen) de V_i dans le document C_i. Dans l'exemple présenté dans la Figure 29, le document S_2 est associé à un vecteur de taille $|V| = 6$ et correspond à l'ensemble des fréquences absolues d'apparitions des mots de V dans S_2. Par exemple, sachant que le mot *a* apparait 2 fois, alors $BW_{S_2}("a") = 2$.

*Un **multi-ensemble** est une collection d'éléments qui peuvent apparaitre plusieurs fois.*

D'autres mesures telle *tf-idf* (SALTON et BUCKLEY, 1991) permettent de pondérer l'apparition d'un mot dans un document avec son apparition dans la globalité du corpus. *tf-idf* est le produit de la fréquence $f(m)$ d'apparition d'un mot m dans un document d_i avec l'inverse de sa fréquence sur l'ensemble des documents du corpus C.

$$tf_i df(t, d_i, C) = f(t) * \log \frac{|C|}{|d \in C : t \in d|}$$

S1 : a cat plays with a dog.
S2 : the dog is with a cat.

	cat	dog	the	plays	is	a	with
S1	1	1	0	1	1	2	1
S2	0	1	1	0	1	0	1

FIGURE 29 – Exemple de représentation sac-de-mots.

Si ce modèle est généralement utilisé sur des données textuelles, d'autres approches proposent d'utiliser ces représentations sur d'autres types de données. Par exemple, en classification d'images, le modèle *bag-of-visual-words* (YANG et al., 2007) propose de représenter une image comme un multi-ensemble de *patch* (région de l'image). Dans nos travaux, nous proposons une représentation intitulée **Bag-of-cliques** s'appuyant sur un modèle de sac-de-mots, où les mots sont les cliques (c.f. Section 4.1) présentes dans un ensemble de graphes.

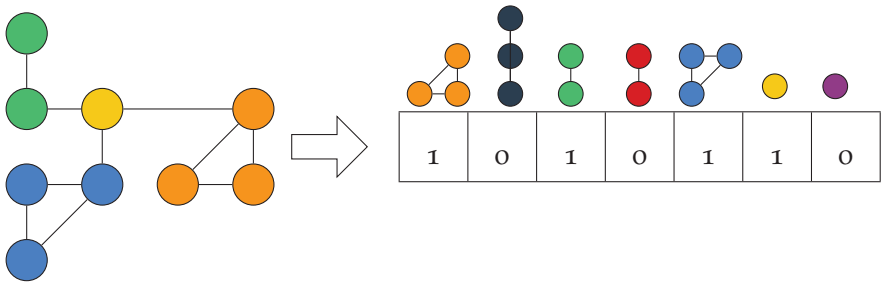


FIGURE 30 – Représentation *Bag of Cliques* (à droite) d'un graphe (à gauche).

Le processus de transformation d'un ensemble de graphe G_s se divise en deux étapes : (i) l'extraction du vocabulaire de cliques et (ii) la construction de la représentation. Dans la première étape, l'ensemble des cliques présentes dans chaque graphe de G_s sont extraites. Pour cela, l'ensemble des cliques maximums sont extraites de manière itérative dans chaque graphe $G \in G_s$. En s'appuyant sur les cliques identifiées dans G_s, une matrice $\mathtt{BagOfCliques}_{n,m}$ est générée, où n correspond au nombre de graphe dans G_s et m le nombre de cliques uniques identifiées. Pour un graphe $G_i \in G_s$, chaque cellule $\mathtt{BagOfCliques}_{i,j}$ est pondérée à 1 si et seulement si la clique C_j apparait dans G_i. L'ensemble de la procédure est détaillée dans la procédure 2 et illustrée dans la Figure 30.

*Une **clique maximum** d'un graphe est une clique dont la taille est la plus grande.*

Une fois la représentation $\mathtt{BagOfCliques}_{n,m}$ initialisée pour tous les graphes de G_s, la similarité entre deux graphes est mesurée avec la similarité cosinus définie par la formule $\cos_{sim}(A, B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$

Algorithme 2 : Génération de la représentation BagOfCliques

```
Input : Gs liste de graphe
V ← []
for G ∈ Gs do
    for c ∈ cliques(G) do
        if !∃c ∈ V then
            append c into V
n ← |Gs| ; m ← |V|
BagOfCliques_{n,m} ← initialisation
for i ∈ 0...n do
    for j ∈ 0...m do
        if v_j ⊂ Gs_i then
            BagOfCliques_{i,j} = 1
        else
            BagOfCliques_{i,j} = 0
return BagOfCliques_{n,m}
```

4.4.2 *Graph Kernels*

Un *kernel* (ou noyau) est une fonction de similarité $k(x, x')$ où x et x' sont les objets comparés. Un kernel doit respecter deux conditions :

— **Symétrie.** $k(x, x') = k(x', x)$

— **Positive semi-défini.** $k(x, x') \geqslant 0, \forall x, x'$

Les *graph kernels* (Kriege, Johansson et Morris, 2019 ; Vishwanathan et al., 2010) forment une famille d'algorithmes d'appariement de graphes performants qui exploite les motifs formés par la topologie présente dans les graphes. Contrairement à plusieurs méthodes de *graph matching*, les *graph kernels* peuvent être utilisés dans des modèles de classification tels que les *Support Vector Machine*. Parmi les différents noyaux de graphes proposés, nous avons sélectionné deux algorithmes majoritairement utilisés dans la litérature : (i) Le *Weisfeiler Lehman Subtree Kernel*, et (ii) le *Shortest Path Kernel*.

4.4.2.1 *Weisfeiler Lehman Subtree kernel*

Le *Weisfeiler Lehman Subtree kernel* est un *graph kernel* présenté dans (Shervashidze et al., 2011) et reposant sur le test d'isomorphisme de Weisefeiler Lehman proposé en 1968 (Weisfeiler et Lehman, 1968).

test d'isomorphisme de graphe de weisfeiler et lehman (Weisfeiler et Lehman, 1968). Le test d'isomorphisme de Weisfeiler et Lehman est fondé sur la comparaison de transformations successives de graphes étiquetés G et H. À chaque fin d'itération,

les ensembles des sommets des deux graphes sont comparés. Si ces ensembles ne sont pas égaux, alors il n'y a pas d'isomorphisme entre les graphes G et H.

WEISFEILEIR LEHMAN SUBTREE KERNEL (SHERVASHIDZE ET AL., 2011). Le Weisfeiler Lehman Subtree Kernel consiste à faire la somme des similarités mesurées sur différentes transformations successives des graphes en entrée, Gs. Les transformations successives des graphes de Gs suivent le protocole proposé dans le test d'isomorphisme de Weisfeiler-Lehman. L'Algorithme 3 prend en entrée une liste de n graphes Gs, un nombre de transformations h et retourne une matrice de similarité $simMatrix_{n,n}$ avec $n = |G_s|$.

Algorithme 3 : Weisfeiler-Lehman Subtree Kernel

```
Input :
  — Gs Liste de graphes
  — L Ensemble des labels
  — h Nombre de transformations
for G ∈ Gs do
    for v ∈ G.sommets do
        append v into L
simMatrix_{n,n} ← init
for i ∈ 0...h do
    for G ∈ Gs do
        if i > 0 then
            L ← init
            G ← WLtransform(G, L, cpt)
        n ← |Gs| ; m ← |L|
        ψ_{n,m} ← init
        for G ∈ Gs do
            for label ∈ L do
                if label ∈ G.sommets then
                    ψ_{i,j} = 1
        simMatrix ← simMatrix+ < ψ_{n,m}, ψ_{n,m} >
return simMatrix
```

4.4.2.2 *Shortest Path kernel*

Le Shortest Path Kernel (BORGWARDT et KRIEGEL, 2005) est un *graph kernel* qui mesure la similarité entre les distributions d'occurrences des longueurs de plus courts chemins extraites dans les deux graphes. Ici, l'algorithme de plus courts chemin utilisé est l'algorithme de

Floyd-Warshall (FLOYD, 1962). Le calcul du kernel est détaillé dans l'Algorithme 4.

Algorithme 4 : Algorithme de Shortest Path Kernel

```
Input : G, H deux graphes
SP_G ← FloydWarshall(G)            // matrice d'adjacence
SP_H ← FloydWarshall(H)            // matrice d'adjacence
V_G ← initVector(max(SP_G))
V_H ← initVector(max(SP_H))
for val ∈ SP_G do
    V_G[val] ++
for val ∈ SP_H do
    V_H[val] ++
return < V_G, V_H >
```

4.4.3 *Graph Embedding*

Depuis plusieurs années, nous assistons à une utilisation croissante des réseaux de neurones profonds (*Deep Learning*) dans la génération de représentation simple (vecteur 1D) intégrant des concepts de haut-niveau. Par exemple, dans le *word-embedding* (MIKOLOV et al., 2013), les modèles proposés permettent de capturer la sémantique des mots en s'appuyant sur des données contextuelles, i.e. sa position par rapport aux autres mots. Depuis, ces modèles suscitent un réel intérêt dans la communauté scientifique travaillant sur des graphes. Le *graph-embedding* vise à encoder dans un vecteur 1D, l'ensemble des aspects structurels et topologiques d'un graphe ou autour de chacun de ces sommets (*node embedding*). Un *graph embedding* est une fonction de *mapping* $f : G \rightarrow (v_0, \ldots, v_d) \in R^d$ où $(v_0, \ldots, v_d)$ est un vecteur de représentation de la topologie du graphe $G = (V, E)$. Une autre catégorie d'algorithme, le *node-embedding* est défini par une fonction de mapping $f : s_i \rightarrow v_{s_i} = (v_0, \ldots, v_d) \in R^d$ où s_i correspond à un sommet de $G = (V, E)$. Ce type d'approche est beaucoup utilisé dans les domaines de détection de communauté (GROVER et LESKOVEC, 2016), de classification de graphes (NARAYANAN et al., 2017) mais aussi de prédiction de lien (M. ZHANG et CHEN, 2018).

Dans la majorité des modèles proposés, l'apprentissage de la topologie d'un graphe ou autour des sommets de celui-ci s'effectue à l'aide de *marches aléatoires*. Une marche aléatoire est une séquence de sommets construite selon un déplacement aléatoire dans un graphe (LOVÁSZ et al., 1993). Dans notre comparaison, nous avons choisi d'évaluer plusieurs modèles s'appuyant sur des marches aléatoires. Parmi ces approches, nous avons sélectionner deux approches de *node-embedding* et une de *graph-embedding* s'appuyant sur des marches aléatoires : *DeepWalk*, *node2vec* et *Graph2Vec*

DEEPWALK. L'approche *DeepWalk* utilise des marches aléatoires pour maximiser la valeur de la probabilité d'observer un sommet en fonction de son entourage. Le modèle génère plusieurs marches aléatoires W_i de longueur $2k+1$. Lors de la création d'une marche aléatoire W_i, le déplacement dans les sommets du graphe suit une loi uniforme, i.e. tous les voisins n d'un sommet ont une probabilité égale d'être visité. Une fois les marches aléatoires générées, une représentation pour chaque sommet du graphe est générée à l'aide du modèle *skipGram* proposé par (MIKOLOV et al., 2013). L'apprentissage du modèle *skipGram* dans (MIKOLOV et al., 2013) s'appuie sur la capacité du modèle à détecter le voisinage autour d'un mot. Ici, l'apprentissage du modèle s'appuient sur sa capacité à prédire le voisinage d'un sommet dans la marche aléatoire.

NODE2VEC. L'approche proposée dans (GROVER et LESKOVEC, 2016) est similaire à celle de DeepWalk. La différence majeure repose sur le parcours du graphe lors de la génération des marches aléatoires. Dans leur approche, (GROVER et LESKOVEC, 2016) proposent une marche aléatoire biaisée qui change de méthodes de parcours (parcours en profondeur ou parcours en largeur) au fur et à mesure de la marche.

GRAPH2VEC. Contrairement aux approches *node2vec* et *DeepWalk*, *Graph2Vec* produit une représentation unique encodant l'ensemble du graphe. Cette approche utilise la fonction de compression des étiquètes de Weisfeiler-Lehman (SHERVASHIDZE et al., 2011) pour encoder l'ensemble des sous-graphes du graphe. Une fois les sous-graphes extraits, les auteurs utilisent le modèle *Doc2vec*(), où le graphe est représenté par un document et les sous-graphes représentent les mots qui le compose.

Pour mesurer la similarité entre les représentations générées par ces modèles, nous utilisons la similarité cosinus sur les vecteurs générés. Dans le cas du *node-embedding*, une moyenne de l'ensemble des vecteurs de sommets est effectué pour obtenir une représentation du graphe.

Dans le chapitre suivant, nous présentons les expérimentations effectuées dans le cadre de l'évaluation de notre processus.

5 PROTOCOLES D'ÉVALUATION ET EXPÉRIMENTATIONS

Pour évaluer le potentiel de notre contribution, nous définissons un protocole d'évaluation pour chaque étape du processus de mise en correspondance. Dans l'évaluation de la phase de *georepresentation* (*c.f.* Chapitre 2), nous évaluons les différentes méthodes d'extraction d'entités spatiales pour obtenir l'information spatiale la plus pertinente. Dans l'évaluation de la phase de geomatching, nous évaluons le potentiel de la STR, de ces transformations et des différentes mesures de similarité (algorithme de Graph Matching). Pour mesurer la qualité des correspondances retournées par les méthodes proposées, nous proposons un ensemble de 6 critères s'appuyant sur la définition de la similarité spatiale (*c.f.* Section 1.4).

Ce chapitre est organisé de la manière suivante. Nous présentons les données utilisées dans nos expérimentations dans la section 5.1. Les protocoles d'évaluations des différentes étapes du processus de mise en correspondance sont détaillés dans les sections 5.2 et 5.3. Enfin, nous présentons les résultats obtenus et les discussions dans la section 5.4.

5.1 DONNÉES UTILISÉES

Dans nos travaux, nous voulons que les méthodes développées puissent être utilisées sur des données hétérogènes avec une composante spatiale importante. À cette fin, nous avons sélectionné deux jeux de données : un jeu de données hétérogène (*AgroMada*) puis un jeu de données plus homogène (*PadiWeb*) pour tester la généricité de notre proposition. *PadiWeb* est un corpus composé de 500 documents extraits d'articles de presses traitant d'épidémies de maladies animales et *AgroMada* regroupe un ensemble de document hétérogènes (rapport de mission, relevé, thèse, article, présentation) produits dans le cadre d'un projet réalisé à Madagascar. Ces deux corpus comportent uniquement des documents en français et en anglais.

5.1.1 *PadiWeb*

Padi-Web (Arsevska, Roche et al., 2016 ; Arsevska, Valentin et al., 2018) est un système de surveillance épidémiologique mis en place par le Cirad en collaboration avec l'Inra. Padi-Web produit une classification et extrait des informations de sources non officielles

(Google News) traitant des épidémies afin de remédier aux délais de publication des arrêtés officiels.

Un corpus de référence (Rabatel, Arsevska et Roche, 2019) a été construit pour évaluer le volume et l'exactitude de l'information extraite. Ce corpus est composé de 442 documents en anglais collectés entre août et septembre 2014 sur la plateforme de Google News. Les maladies, leurs porteurs, les lieux et les dates d'apparitions ainsi que le nombre de cas déclarés sont annotés dans chaque document. Les articles de presse sont collectés selon un ensemble de maladies animales : la maladie Schmallenberg, la fièvre catarrhale ovine, la fièvre aphteuse et la peste porcine africaine. Les lieux mentionnés dans le corpus sont illustrés sur la carte de la Figure 31.

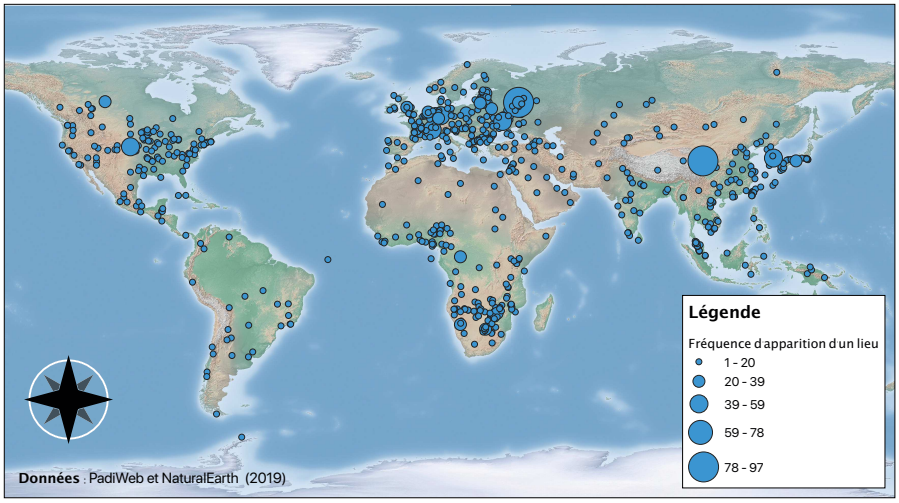


Figure 31 – Carte des lieux mentionnés dans le corpus *PadiWeb*.

5.1.2 *AgroMada*

Un mot sur l'agroécologie ...
Née durant les mouvements sociaux des années 70 au Brésil, l'agroécologie rassemble les pratiques agricoles privilégiant les interactions biologiques et visant à une utilisation optimale des possibilité offertes par les agro-systèmes. Elle s'oppose à l'agriculture intensive qui s'appuie majoritairement sur l'utilisation d'agents artificielles (engrais, pesticides) et d'énergie fossile (véhicule, plastique)[Source]

Dans les années 2000, le Centre de coopération internationale en recherche agronomique pour le développement (CIRAD) s'est impliqué dans le projet des bassins versants du Lac Alaotra (BV-LAC). Les objectifs de ce projet étaient de stabiliser les revenus des producteurs de la région, la protection des bassins versants, ainsi que la promotion des méthodes liées à l'agroécologie. Durant ce projet, un volume important de données a été produit par les agents sur place (relevé de ter-

rain, agenda, rapport, thèse, etc.). Suite à plusieurs discussions avec un expert des données, nous avons choisi de nous appuyer sur un sous-ensemble du corpus traitant du processus d'agroécologie traitée durant le projet BVLAC. Pour extraire les documents associés à cette thématique, nous sélectionnons les documents ayant un ou plusieurs termes dans DicoAE, le dictionnaire de l'agroécologie proposé par l'INRA [1].

Une fois les documentes sélectionnés, nous obtenons *AgroMada*, un corpus composé de 5552 documents en anglais et en français sur le thème de l'agroécologie. La Figure 32 illustre l'ensemble des entités spatiales identifiées dans *AgroMada*.

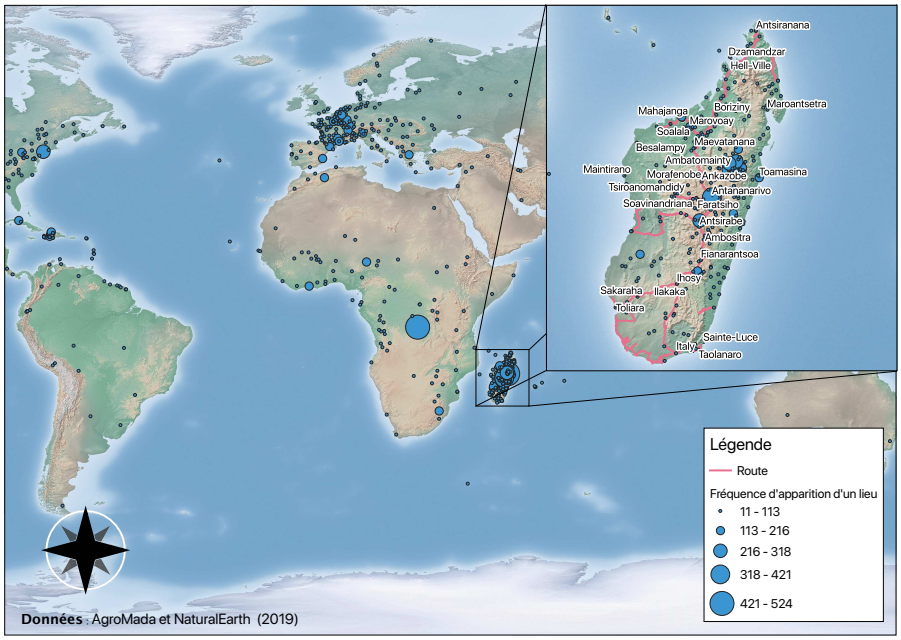


FIGURE 32 – Entités spatiales et leur fréquence d'apparition dans le corpus *AgroMada*.

5.2 ÉVALUATION DE LA GEOREPRESENTATION

Pour obtenir une mise en correspondance pertinente, l'identification de l'information spatiale la plus complète dans un document est cruciale. Dans nos travaux, chaque étape du processus de georepresentation est évaluée sur un échantillon de chaque corpus (*PadiWeb* et *AgroMada*). Pour le corpus de *PadiWeb*, nous utilisons la totalité des documents annotés. Pour le corpus d'*AgroMada*, nous avons collecté et annoté un échantillon de 232 documents (FIZE, 2018).

1. https://dicoagroecologie.fr

5.2.1 *Geotagging*

L'évaluation de l'étape de *geotagging* repose sur la comparaison de différentes méthodes de reconnaissance d'entités nommées. Parmi les différents approches proposées dans la littérature, nous avons choisi d'évaluer les plus utilisées qui proposent des traitements multilingues. Ces outils sont succinctement décrit ci-dessous.

1. STANFORDNER (MANNING ET AL., 2014). StanfordNER est l'extension dédiée à la reconnaissance d'entités nommées de la boîte à outils STANFORDCORENLP développée par l'université de Stanford.
2. GATE (CUNNINGHAM, 2002). Gate, *General Architecture for Text Engineering* est un ensemble d'outils dédié aux traitement automatique des langues, développé à l'université de Sheffield. Parmi les différents outils, nous utilisons le programmeANNIE (A Nearly New Information Extraction System) qui permet de faire de la détection d'entités nommées (NER).
3. NLTK (BIRD ET LOPER, 2004). NLTK, ou Natural Language ToolKit, est une boite à outils pour le traitement automatique des langues développée en Python à l'Université de Pennsylvanie.
4. POLYGLOT (AL-RFOU ET AL., 2015). Polyglot est une boite à outils multilingue supportant jusqu'à 40 langues. Polyglot est utilisée pour réaliser de la détection de langue, la *tokenisation*, l'extraction du *PartOfSpeech*, l'analyse de sentiment sur des données textuelles. Le système de reconnaissance d'entités nommées de Polyglot s'appuie sur un réseau de neurones profonds utilisant les vecteurs des mots (*word-embedding*) pour les classifier selon les *tag* : *Person*, *Organisation*, *Location*.
5. SPACY (HONNIBAL ET MONTANI, 2017). Spacy est une boîte à outils pour le traitement automatique des langues développée en Python. Elle propose plusieurs fonctionnalités comme la détection d'entités nommées, du *PartOfSpeech*, l'extraction d'arbres de dépendances, etc. Comme Polyglot, les modèles de classification s'appuient sur des réseaux de neurones profondes (CNN, RNN).

L'évaluation des performances de chaque système s'appuie sur trois mesures classiques : la précision, le rappel et la F-mesure. La **précision** mesure le nombre d'entités identifiées correctement dans chaque document. Le **rappel** mesure le nombre d'entités identifiées correctement par l'approche par rapport aux nombres exacts d'entités présentes dans le document. Enfin, la **F-Mesure** calcule la moyenne harmonique entre la précision et le rappel. Les formules des trois mesures sont rappelées dans l'Équation 6.

$$\textbf{précision} = \frac{VP}{VP+FP} \qquad \textbf{rappel} = \frac{VP}{VP+FN}$$
$$\textbf{F}-\textbf{mesure} = 2 \cdot \frac{\text{précision} \cdot \text{rappel}}{\text{précision}+\text{rappel}} \tag{6}$$

Avec : VP Vrai positif
FP Faux positif
FN Faux négatif

5.2.2 *Geocoding*

L'évaluation de la résolution de toponyme, ou *geocoding*, consiste à identifier quel algorithme de désambiguïsation (*c.f.* section 2.2.2) permet d'identifier le plus d'entités spatiales pertinentes sur l'ensemble des toponymes d'un document. Pour mesurer les performances des différents algorithmes, nous utilisons les trois mesures suivantes :

1. ACCURACY. La mesure d'exactitude (ou *accuracy*) permet de mesurer la capacité d'un système à donner la sortie attendue pour une donnée en entrée. Dans le cas de la résolution de toponyme, l'objectif est de retourner l'identifiant exact dans l'index géographique.

$$\textbf{Acc}(\textbf{F},\textbf{T}) = \frac{\sum_{i=0}^{|F|} F[i] == T[i]}{|F|} \tag{7}$$

Avec : F Entités spatiales retournées par le système
T Entités spatiales attendues

2. ACCURACY@K (GRITTA ET AL., 2018). Le problème de l'exactitude est qu'elle n'inclut pas la caractéristique spatiale. Dans certains cas, il est possible que plusieurs entités spatiales avec un même toponyme possèdent une signature spatiale proche (courte distance). C'est le cas pour des villes frontalières comme Templeuve (France) et Templeuve (Belgique) ou encore Saint-Gingolph (Suisse) et Saint-Gingolph (France). Dans certains pays, des entités spatiales partageant une relation d'inclusion (e.g. quartier → ville) possèdent le même toponyme comme Sofia (le district) et Sofia (la capitale). Pour évaluer la capacité d'un algorithme de désambiguïsation à retourner une entité spatiale avec une empreinte spatiale proche de l'entité attendue, nous utilisons l'ACCURACY@K.

L'ACCURACY@K évalue la capacité d'un système de geocoding à retourner une entité avec des coordonnées à moins de k degrés de celle attendue.

$$\mathbf{Acc@k}(F, T, k) = \frac{1}{|F|} \sum_{i=0}^{|F|} \begin{cases} \mathrm{dist}(F[i], T[i]) < k & 1 \\ \mathrm{sinon} & 0 \end{cases} \tag{8}$$

Avec :

$\mathrm{dist}(x, y)$	distance euclidienne entre x et y
F	Entités spatiales retournées par le système
T	Entités spatiales attendues
k	seuil de distance

3. DISTANCE MOYENNE D'ERREUR (DME). Dans la continuité de l'*accuracy@k*, la Distance Moyenne d'Erreur permet de mesurer l'erreur moyenne d'un algorithme de désambiguïsation.

$$\mathbf{DME}(F, T) = \frac{\sum_{i=0}^{|F|} \mathrm{dist}(F[i], T[i])}{|F|} \tag{9}$$

Avec :

$\mathrm{dist}(x, y)$	distance euclidienne entre x et y
F	Entités spatiales retournées par le système
T	Entités spatiales attendues

Suivant le processus de mise en correspondance, la prochaine section détaille le protocole d'évaluation du *geomatching*.

5.3 ÉVALUATION DE LA MISE EN CORRESPONDANCE : GEOMATCHING

Dans nos experimentations, nous nous intéressons à évaluer 3 aspects :

1. **La génération de la STR.** Quels types de STR sont utilisés pour faire la mise en correspondance ? (Normal, Généralisé, Étendue, etc).
2. **La mesure de similarité utilisée.** Dans notre approche, nous utilisons des mesures de similarité de graphes provenant de domaine de recherche du Graph Matching (*c.f.* Chapitre 4). Dans notre évaluation, nous voulons évaluer quels algorithmes/mesures répondent le mieux à la problématique de mise en correspondance spatiale.

3. **Les critères utilisés.** Nous avons défini différents critères permettant de vérifier si une mesure de similarité répond bien aux enjeux de la similarité spatiale. Différents regroupements de ces critères peuvent être réalisés afin d'étudier sur certains enjeux (proximité spatiale, entités spatiales partagées, etc.)

5.3.1 *Protocole d'évaluation*

Pour évaluer la mise en correspondance selon les 3 aspects mentionnés (*c.f.* Section 5.3), nous avons mis en place un protocole divisé en 4 étapes illustrées dans la Figure 33 :

1. **Génération de la STR** (normale ou transformée) pour chaque document d'un corpus.
2. **Mesure de la similarité entre les STR** générées en utilisant des algorithmes provenant du Graph Matching (*c.f.* Chapitre 4).
3. **Extraction des** $\text{top} - n$ **correspondances** trouvées pour chaque document d'un **ensemble de 100 documents**.
4. **Mesure des performances** obtenues sur les différents critères.

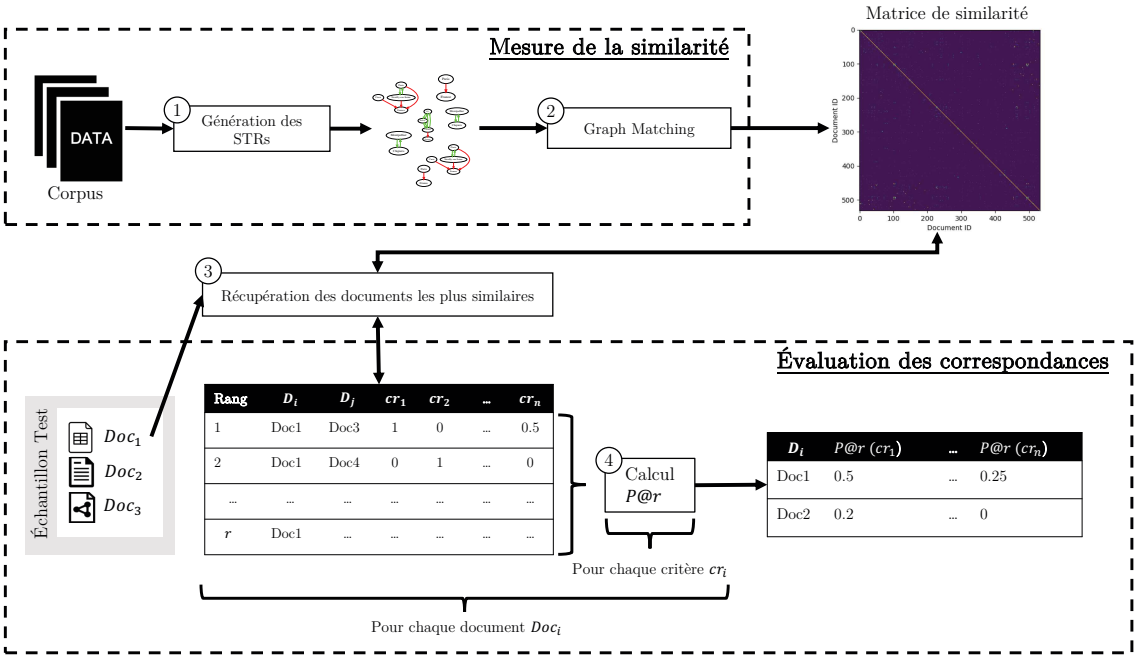


FIGURE 33 – Protocole d'évaluation de correspondances.

Chacune de ces étapes est détaillée dans les sections suivantes.

5.3.1.1 *Évaluation de la génération de la STR*

Pour chaque document d'un corpus, différentes formes de STR peuvent être générées. Pour évaluer le potentiel des transformations, différentes formes de la STR sont générées pour un document :

1. STR normale (aucune transformation)
2. STR Généralisée (Limite = Région)
3. STR Généralisée (Limite = Pays)
4. STR Étendue ($n = 1$) [1]
5. STR Étendue ($n = 2$)

5.3.1.2 *Sous-ensemble des corpus*

Pour évaluer les correspondances retournées par un couple (type de STR, mesure de similarité), nous avons sélectionné un sous-ensemble de documents pour chaque corpus. Pour étudier l'impact du nombre d'entités spatiales sur le comportement des différentes mesures de similarité, les sous-ensembles sont composés de document avec des STRs de différentes tailles. Sur chaque corpus, nous extrayons un sous-ensemble de 100 documents représentés par des STRs divisées en 3 groupes :

— $[0, Q^2]$ (20%)

— $[Q^2, Q^3(G)]$ (40%)

— $[Q^3, max(G)]$ (40%)

où Q^i est le i-quantile calculé sur la totalité des documents selon le nombre d'entités spatiales.

5.3.1.3 *Évaluation des mesures de similarité*

L'ensemble des mesures de similarité présentées dans le Chapitre 4 sont utilisées. Certains algorithmes comme la Graph Edit Distance sont associés à des paramètres. Dans nos tests, voici les paramètres sélectionnés pour chacune de ces mesures :

— **Graph Edit Distance.** Les paramètres de la Graph Edit Distance sont les différents coûts de base des transformations réalisées (*c.f.* Section 4.3.4). Dans le cadre de la comparaison des graphes de la STR, le coût d'insertion/suppression d'un sommet correspond à l'insertion ou la suppression d'une entité spatiale. Dans nos expérimentations, nous considérons que l'ajout et la suppression d'entités pour passer d'une STR à une autre sont tout aussi pénalisant. Par conséquent, l'ensemble des coûts d'édition sont définis à 1.

— **Weisfeiler-Lehman Subtree Kernel.** Le Weisfeiler-Lehman Subtree Kernel (WLK) repose sur l'utilisation successive de transformation d'un graphe simulant un parcours en profondeur. À

1. Nombre d'entités ajoutées par entité sélectionnée

cause de la faible densité générale des graphes des STR générées (*c.f.* Discussion 5.5.1.1), définir un nombre d'itération élevé n'a que peu d'incidence sur l'efficacité de l'algorithme. Pour cette raison, nous limitons le nombre d'itération h à 2.

5.3.2 *Critère d'évaluation d'une correspondance*

Dans nos travaux, nous voulons évaluer les correspondances entre différents documents en s'appuyant sur ces critères. Dans la littérature, plusieurs études comme celles des auteurs de (De Sabbata et Reichenbacher, 2012 ; Purves, Clough, Jones, Hall et al., 2018) s'intéressent à la notion de *geographic relevance* dans le cadre d'un système de *Geographic Information Retrieval* (GIR). Dans (De Sabbata et Reichenbacher, 2012), les auteurs proposent une extension spatiale des critères de pertinence d'un système classique de Recherche d'Information. La *proximité spatiale* (entités spatiales partagées ou proches), l'existence de *cluster* et la proximité dans une hiérarchie spatiale font partie de ces critères.

Dans la section 1.4, nous avons défini trois critères principaux d'une mesure de similarité spatiale en s'appuyant sur différentes définitions de la littérature (D. Lin, 1998 ; Montello et al., 2003 ; Tobler, 1970). Reprenant ces trois critères avec une vision plus précise, nous proposons d'évaluer une correspondance entre deux documents selon les 6 critères de pertinence suivants.

1. ENTITÉS PARTAGÉES (ESP). Ce critère est validé[1] si les deux documents correspondants partagent une ou plusieurs entités spatiales. Par exemple, les deux documents représentés par leurs entités spatiales dans la Figure 34 (différenciés par leur couleur) valident ce critère.

$$\mathbf{ESP}(\mathbf{ES_{D_1}}, \mathbf{ES_{D_2}}) = \begin{cases} 1, & \text{si } |ES_{D_1} \cap ES_{D_2}| \geqslant 1 \\ 0, & \text{sinon} \end{cases} \tag{10}$$

Avec :

ES_{D_i} ensemble des entités spatiales du document D_i

2. ENTITÉS PROCHES (EP). Le critère EP permet d'identifier s'il existe des relations de proximité entre des entités spatiales distinctes appartenant chacune à un document différent. Deux entités spatiales sont proches s'il existe une relation spatiale entre elles (*c.f.* Section 2.1.2.2). Dans la Figure 35, les deux entités **A** et **B** permettent de valider le critère EP entre les deux documents.

1. Prend une valeur entre 0 et 1

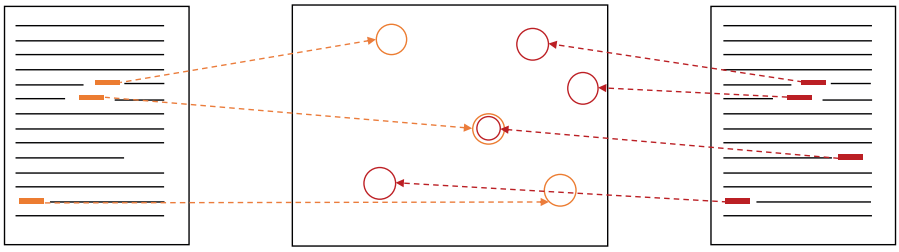


FIGURE 34 – Illustration du critère Entités Partagées (ESP).

$$\mathbf{EP}(\mathbf{ES_{D_1}}, \mathbf{ES_{D_2}}) = \begin{cases} 1, & \text{si} \sum_{i=0}^{|ES_{D_1}|} \sum_{j=0}^{|ES_{D_2}|} p(ES_{D_1}[i], ES_{D_2}[j]) \geqslant 1 \\ 0, & \text{sinon} \end{cases}$$

Avec :

ES_{D_i}	ensemble des entités spatiales du document D_i
$p(es_i, es_j)$	fonction qui retourne 1 si les deux entités sont liées

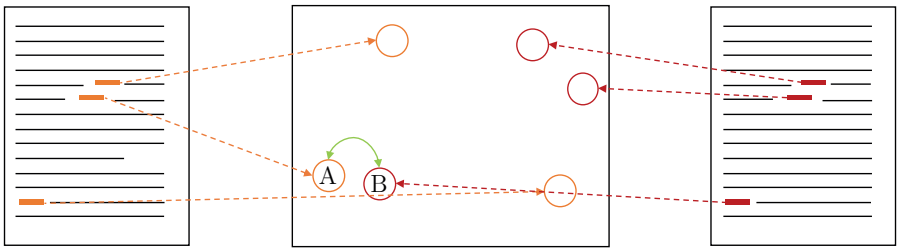


FIGURE 35 – Illustration de l'application du critère d'Entités Proches (EP).

3. EMPRISE SPATIALE SIGNIFICATIVE (ESS). Le critère d'emprise spatiale significative (ESS) consiste à identifier si deux documents partagent au moins un groupe significatif [1] d'entités spatiales proches. Deux groupes d'entités sont similaires, si l'aire d'intersection de leur surface représente plus de 50% du plus petit des deux groupes. Dans l'exemple de la Figure 36, le critère ESS est validé car les deux documents possèdent deux groupes d'entités spatiales proches.

4. EMPRISE SPATIALE STRICTE (ESSC). Le critère d'emprise spatiale stricte (ESSC) vérifie si l'ensemble des entités dans les deux documents sont réparties spatialement de manière identique. Pour cela, nous vérifions si chaque groupe d'entités spatiales présentes dans un document s'entrecoupe avec au moins un autre groupe de l'autre document. Par exemple, chaque groupe d'entités spatiales des documents illustrés dans la Figure 37 s'entrecoupe.

1. Par rapport, à l'ensemble des groupes d'entités spatiales d'un document

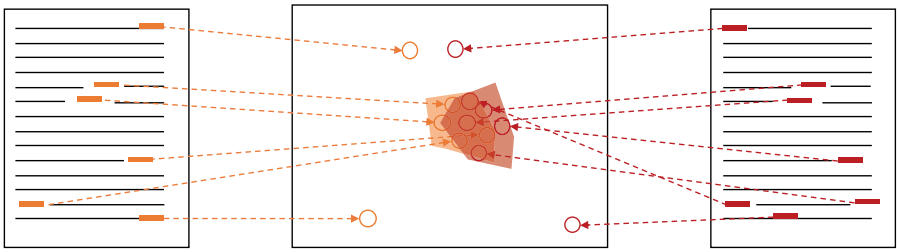


FIGURE 36 – Illustration du critère d'Emprise Spatial Significative (ESS).

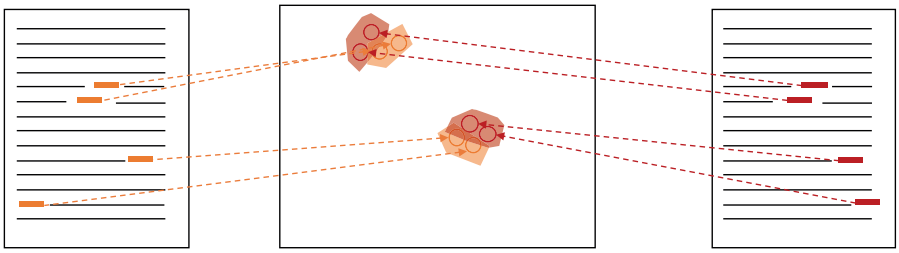


FIGURE 37 – Illustration du critère d'Emprise Spatial Stricte (ESSC).

5. PROXIMITÉ MOYENNE (PM). Dans le critère (EP), nous évaluons s'il existe au moins une relation de proximité entre deux entités, chacune appartenant à un document différent. Pour obtenir plus d'informations concernant la proximité entre les entités présentes dans les documents (Voir Figure 38), nous proposons d'utiliser un cinquième critère : la proximité moyenne (PM). Le calcul du critère PM consiste à calculer l'inverse de la distance moyenne normalisée entre les entités spatiales des deux documents correspondants.

$$\mathbf{PM}(\mathbf{ES_{D_1}}, \mathbf{ES_{D_2}}) = 1 - \left(\frac{1}{(|ES_{D_1}| * |ES_{D_2}|)} \sum_{e_1}^{ES_{D_1}} \sum_{e_2}^{ES_{D_2}} d(e_1, e_2)\right) \tag{11}$$

Avec :

ES_{D_i}	ensemble des entités spatiales du document D_i
$d(es_i, es_j)$	distance entre les entités spatiales es_i et es_j

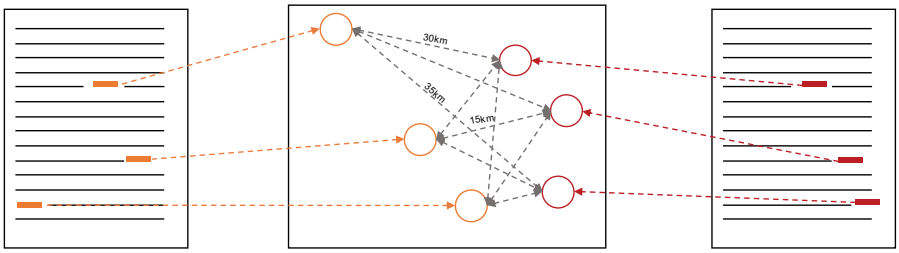


FIGURE 38 – Illustration du critère de Proximité Moyenne (PM).

6. PROPORTION D'ENTITÉS SPATIALES PARTAGÉES (PEP). Pour compléter l'information portée par le critère ESP, le critère Proportion d'entités spatiales partagées (PEP) mesure la proportion d'entités partagées par les deux documents correspondants. Pour mesurer la similarité entre les ensembles d'entités spatiales, nous utilisons le coefficient de DICE (SORENSEN, 1948) défini par la formule $\mathrm{Dice}(A, B) = \frac{2*|A \cap B|}{|A|+|B|}$.

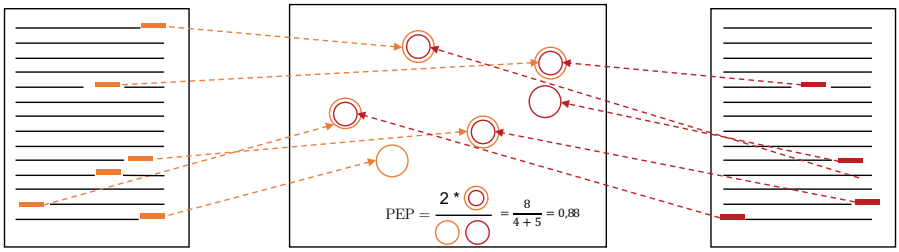


FIGURE 39 – Illustration du critère PEP. Ici la valeur du critère $\mathrm{PEP} = 2 * \frac{8}{4+5} = 0.88$.

Les critères proposés sont utilisés pour mesurer la qualité d'une correspondance sur la dimension spatiale entre deux documents. Les mesures permettant de calculer les performances d'un système de mise en correspondance à l'aide de ces critères sont présentées dans la section suivante.

5.3.3 *Évaluation des top-n correspondances*

Dans le cadre de la mise en correspondance spatiale, l'objectif est de mettre en relation des documents avec la configuration spatiale la plus proche. Contrairement à la mise en correspondance d'enregistrements rapportant à une même personne, une configuration spatiale proche peut être identifiée dans différents documents (pas seulement les doublons). Par conséquent, nous souhaiterions pour chaque document évaluer les n documents correspondants, i.e. les documents les plus similaires.

Dans le domaine de Recherche d'Information, la **précision@n** (CRASWELL, 2009) permet d'estimer la pertinence des top_n résultats d'une requête retournées par un sytème. Concrètement, la precision@n renvoie le pourcentage de résultats pertinents parmi les meilleurs résultats. La formule de la précision@n est la suivante :

$$\mathbf{P@n(R)} = \frac{1}{n}\sum_{i=0}^{n}\begin{cases} \mathtt{estPertinent}(R_i) & 1 \\ \mathtt{sinon} & 0 \end{cases} \tag{12}$$

Avec :

n	le nombre de résultats
R	Sortie du système à évaluer

Dans notre évaluation, nous calculons le précision@n obtenue pour chaque critère sur les $\mathtt{top}-n$ documents similaires pour chaque document d'un échantillon.

$$\mathbf{P@n(D_i, c)} = \frac{1}{n}\sum_{k=1}^{n} c(D_i, D_{ik}) \tag{13}$$

Avec :

$c(D_i, D_{ik})$	valeur du critère **c** entre le document D_i et D_{ik}
D_i	document
D_{ik}	le $k^{\text{ième}}$ document plus similaire avec D_i

Enfin, pour mesurer les performances globales obtenues par une mesure de similarité et un type de STR sur l'échantillon, nous mesurons *Mean Average Precision* ou MAP@n (Beitzel, Jensen et Frieder, 2009) pour chaque critère selon la formule suivante :

$$\mathbf{MAP@n(D, c)} = \frac{1}{N}\sum_{i=1}^{N} P@n(D_i, c) \tag{14}$$

Avec :

D	échantillon de documents
$N = \lvert D \rvert$	nombre de documents dans l'échantillon

Quelles sont les combinaisons dominantes ?

Pour évaluer quelles sont les mesures de similarité et le type de STR qui retournent le plus de correspondances pertinentes, une analyse des valeurs de précision@n des différents critères est nécessaire. À cette fin, nous choisissons d'utiliser le modèle de la Somme Pondérée, ou *Weighted Sum Model* en anglais.

WEIGHTED SUM MODEL. En théorie de la décision, la somme pondérée ou **Weighted Sum Model**, est une des méthodes d'analyse

décisionnelle multi-critères la plus connue. Cette méthode est utilisée pour évaluer un certain nombre d'alternatives — i.e. type de STR, mesure de similarité — à l'aide de plusieurs critères.

Pour un problème donné, m alternatives et n critères de décision sont définis. Dans l'évaluation de correspondances, nous cherchons à maximiser la valeur des critères de similarité spatiale. Soit w_j, la pondération du critère C_j et a_{ij} est la valeur de performance de l'alternative A_i lorsqu'elle est évaluée en fonction du critère C_j. Ensuite, l'importance totale (i.e lorsque tous les critères sont considérés simultanément) de l'alternative A_i, est désignée par le score $A_i^{WSM-score}$ et défini par la formule suivante :

$$A_i^{WSM-score} = \sum_{j=1}^{n} w_j a_{ij} \quad avec\ i \in \{1, \ldots, m\} \tag{15}$$

DIFFÉRENTES PONDÉRATIONS. Lors de la mise en correspondance spatiale entre des documents, nous proposons d'utiliser différents critères pour faire notre évaluation. Les critères ESP (entités spatiales partagées), PEP (pourcentage d'entités spatiales partagées) permettent d'évaluer le potentiel d'une combinaison à associer des documents ayant des entités spatiales communes. Les critères EP (entités proches) et PM (Proximité Moyenne) s'intéressent à la proximité des entités spatiales présentes dans les deux données correspondantes. Enfin, les critères ESS (Emprise Spatial Significative), ESSC (Emprise Spatiale Stricte) se concentrent sur l'évaluation d'une combinaison à retourner des documents partageant une topologie similaire.

Dans nos travaux, nous proposons d'évaluer différents groupes de critères selon différentes exigences attendues lors de de mise en correspondance. Le modèle de la somme pondérée permet d'évaluer différentes combinaisons de type de STR et de mesure de similarité (alternatives) selon des pondérations définies pour chaque critère. Dans nos travaux, nous proposons différentes pondérations appliquées à chaque critère selon différents usages. Ici, la pondération appliquée à un critère correspond à l'importance donnée à celui-ci par rapport aux autres. Par conséquent, la somme des poids assignés à chaque critère est égale 1. Par exemple, une pondération à 0,6 pour un critère **c** veut dire que l'on accorde 60% d'importance au critère c dans l'évaluation des différents systèmes. Dans nos travaux, nous proposons des ensembles de pondération, l'une prend en compte la totalité des critères définis dans la section 5.3.2, l'autre se concentre sur les critères de proximité spatiale.

1. AllCriteria (AC). Les pondérations des critères définies par *AllCriteria* visent à classer les combinaisons en prenant en compte la totalité des valeurs de MAP@n des 6 critères. La somme des pondéra-

tions devant être égale à 1, chaque critère est associé à une pondération égale à $\frac{1}{6}$ (*c.f.* Equation 16).

$$W_{all} = \{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\} \tag{16}$$

2. Spatial relatedness privileged (SRP). Contrairement à la pondération utilisée dans *AllCriteria*, nous proposons un ensemble de pondérations (*c.f.* Équation 17) qui privilégie la proximité existante entre les entités spatiales des deux documents correspondantes (PM, PEP), tout en maximisant le pourcentage d'entités partagées (PEP).

$$W_{sp_rel} = \{0, \frac{1}{3}, 0, 0, \frac{1}{3}, \frac{1}{3}\} \tag{17}$$

5.4 RÉSULTATS ET DISCUSSIONS

5.4.1 *Georepresentation*

Dans cette section, nous présentons les résultats obtenus pour les différentes étapes d'extraction de l'information spatiale dans la Georepresentation : le *geotagging* et le *geocoding*.

5.4.1.1 *Geotagging : Identification des toponymes*

La Table 4 indique les performances des systèmes NER obtenues sur les différents échantillons provenant des corpus *PadiWeb* et *AgroMada*. Chaque sytème est évalué selon trois mesures de performances, la précision, le rappel et la F-mesure.

Sur le corpus de *PadiWeb*, les résultats indiquent que la méthode proposée par le StanfordNER permet d'obtenir les meilleurs performances sur les trois mesures suivi par Polyglot. Sur le corpus d'*AgroMada*, les résultats sont plus disparates. Par exemple, StanfordNER obtient la meilleur précision mais un rappel faible. A l'inverse, *Spacy* obtient un rappel de 84% et une précision très faible. Cependant, nous avons remarqué que le bruit présent dans les résultats de *Spacy* s'exprime par l'identification de chaine de caractères non-associée à un terme mais des nombres, des dates, etc. Par conséquent, les faux-positifs extrait par *Spacy* ont très peu de chance d'être associées à une entité dans Geodict au cours de l'étape de *géocoding*.

Au travers de ces résultats, il existe une différence nette entre les résultats obtenus sur l'échantillon du corpus de *PadiWeb* et de celui d'*AgroMada*. Par exemple, le système StanfordNER ayant obtenu une F-mesure de 0,67 sur *PadiWeb*, se retrouve avec une F-Mesure de 0,22 sur le corpus d'*AgroMada*. Sachant que nous voulons que notre processus de représentation fonctionne sur différents corpus de données (hétérogènes et homogènes), le système Spacy, à travers ces résultats, permet de maintenir un niveau de performance stable entre différents corpus.

	PadiWeb			**AgroMada**		
	précision	*rappel*	*F-mesure*	*précision*	*rappel*	*F-mesure*
StanfordNER(Manning et al., 2014)	**0.59**	**0.77**	**0.67**	**0.31**	0.16	0.22
Polyglot(Al-Rfou et al., 2015)	0.53	0.72	0.61	0.20	0.35	**0.26**
NLTK(Bird et Loper, 2004)	0.42	0.66	0.52	0.13	0.15	0.14
Spacy(Honnibal et Montani, 2017)	0.40	0.65	0.50	0.14	**0.84**	0.25

Table 4 – Performances des différentes méthodes de NER (Reconnaissance d'entités nommées) sur le *geotagging* sur les deux échantillons de corpus.

5.4.1.2 *Résolution des toponymes : Geocoding*

MÉTHODE DE RÉFÉRENCE : MORDECAI. Pour comparer les différents algorithmes de désambiguïsation, nous avons sélectionné MORDECAÏ (HALTERMAN, 2017), un système de *geoparsing* s'appuyant sur SPACY et deux réseaux de neurones. Le premier réseau est utilisé pour prédire le pays dans lequel existe l'entité spatiale pour un toponyme. En s'appuyant sur le résultat précédent et un ensemble de descripteurs (population, altitude, classe), le second réseau de neurones retourne l'entité spatiale la plus probable.

La Table 5 indique l'exactitude (Acc), l'accuracy@k (Acc@k) et la distance d'erreur moyenne (DME) obtenue par chaque algorithme (voir Section 2.2) sur les échantillons des deux corpus.

Au travers des résultats, nous remarquons que les algorithmes de désambiguïsation *MostCommon* et *WikiCooc* obtiennent en moyenne les meilleurs scores sur les deux corpus avec une exactitude à 0,71 pour *PadiWeb* et 0,96 pour *AgroMada*. Contrairement aux approches basées sur la cohérence spatiale : *SharedProp* et *Mordecai*.

	PadiWeb				**AgroMada**			
	Acc	*Acc@0.5*	*Acc@1*	*DME*	*Acc*	*Acc@0.5*	*Acc@1*	*DME*
MostCommon	0.71	**0.84**	**0.85**	**2.43**	0.94	0.97	0.97	**0.49**
SharedProp	0.66	0.78	0.79	4.82	0.88	0.89	0.90	3.36
WikiCooc	0.71	0.83	**0.85**	2.53	**0.96**	**0,97**	**0.97**	0.62
Mordecai	**0.74**	0.76	0.77	2.73	0.55	0.61	0.64	13

TABLE 5 – Performances des différents algorithmes de désambiguïsation (*geocoding*) sur les deux échantillons des corpus.

5.4.2 *Geomatching*

Les résultats de l'évaluation du *geomatching* sont divisés en deux parties. Nous présentons les résultats des différentes mesures selon différents paramètres n de la Mean Average Precision ou **MAP@n**. Puis, nous présentons les combinaisons dominantes — i.e. avec les plus fortes valeurs de précisions— selon différents sous-ensembles et pondérations utilisées.

5.4.2.1 *Méthodes de références*

Pour évaluer l'intérêt de la STR et des mesures de similarité issues du Graph Matching, nous comparons nos résultats avec les représentations et mesures de similarité suivantes.

BAG OF SPATIAL ENTITIES (BOSE). La première approche de référence appelée Bag of Spatial Entities (**BOSE**) est un modèle de sac-de-mots ou chaque élément d'un vecteur de document correspond à un toponyme.

INTERSECTION DE POLYGONE (POLYINTERSECT). S'inspirant de (WOODRUFF et PLAUNT, 1994), la deuxième méthode de référence utilise les propriétés géométriques des entités spatiales pour calculer la similarité spatiale. Pour ce faire, nous fusionnons la géométrie (Polygone, centroïde) de chaque entité spatiale identifiée, renvoyant ainsi une géométrie unique, *i.e. un MultiPolygon*. La similarité correspond à la surface de la zone d'intersection entre les nouvelles géométries des deux documents.

SAC-DE-MOTS (CLASSICBOW). La troisième méthode de référence consiste à utiliser une représentation sac-de-mots (SALTON et BUCKLEY, 1991) produites sur l'ensemble des documents d'un corpus. Les valeurs stockées pour chaque mot correspondent à leur fréquence dans le document correspondant.

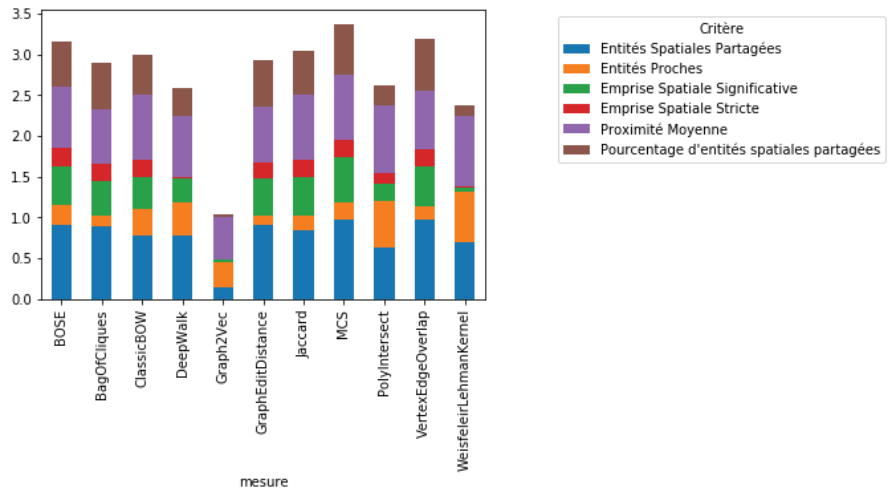
5.4.2.2 *Les mesures de similarité*

Dans cette première partie, nous nous interrogeons sur l'efficacité des différentes mesures de similarités dans le processus de mise en correspondance. Les Figures 40 et 41 présentent les valeurs maximales [1] de MAP@n obtenues pour chaque mesure dans les deux corpus.

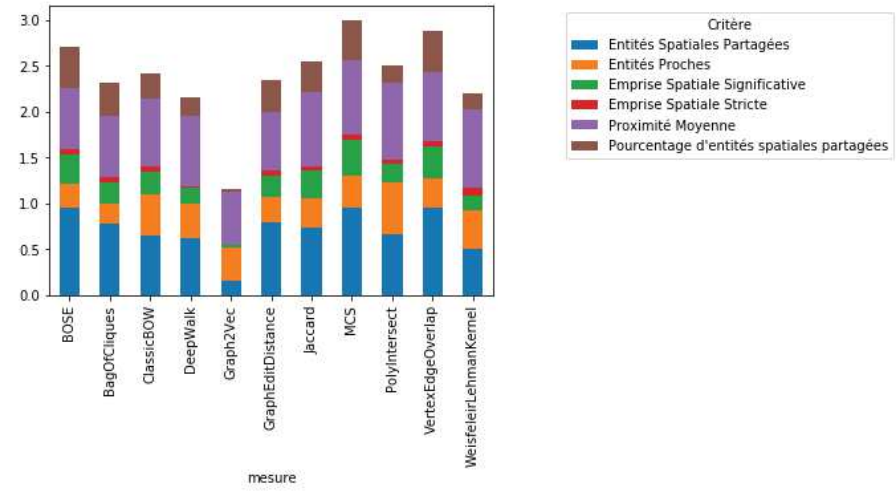
De manière générale, la majorité des algorithmes *pattern-based* obtiennent de plus faibles performances que les algorithmes *structure-based*. Dans l'ensemble des corpus, la mesure MCS (Most Common Subgraph) obtient les meilleurs scores de MAP@1 et MAP@5. Dans *AgroMada*, *VertexEdgeOverlap* et *BagOfCliques* obtiennent des résultats proches de *MCS*. Inversement, les algorithmes de graph-embedding comme *DeepWalk*, ou *Graph2Vec* obtiennent de faibles valeurs de MAP@n. Une des raisons principales repose sur la sparsité des STR générées (voir Discussion 5.5.1.1). Concernant les mesures de références, *BOSE*, *ClassicBOW* et *PolyIntersect* obtiennent des performances inférieures aux mesures de *graph matching* s'appuyant sur la STR.

Enfin, des différences de résultats sont observables entre les résultats de MAP@1 et de MAP@5 sur le corpus *AgroMada*. Par exemple, les performances de *ClassicBOW* chutent entre MAP@1 et MAP@5 (*c.f.* Discussion 5.5.2.2). Le même phénomène est observable sur la *GraphEditDistance*.

1. Extraites de la combinaison dominante sur l'ensemble des combinaisons dans lesquelles apparaissent la mesure de similarité concernée



(a) MAP@1



(b) MAP@5

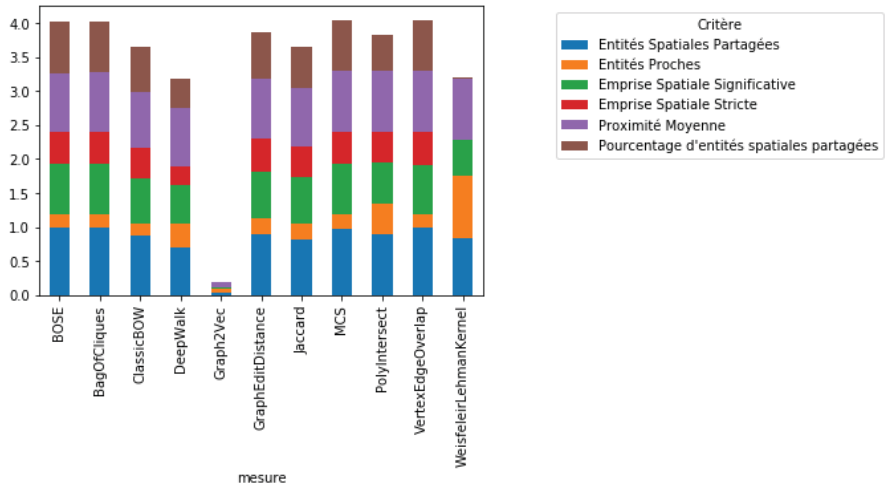
FIGURE 40 – Valeurs cumulées de MAP@n de la combinaison dominante pour chaque mesure de similarité évaluée (*PadiWeb*).

5.4.2.3 *Combinaisons dominantes*

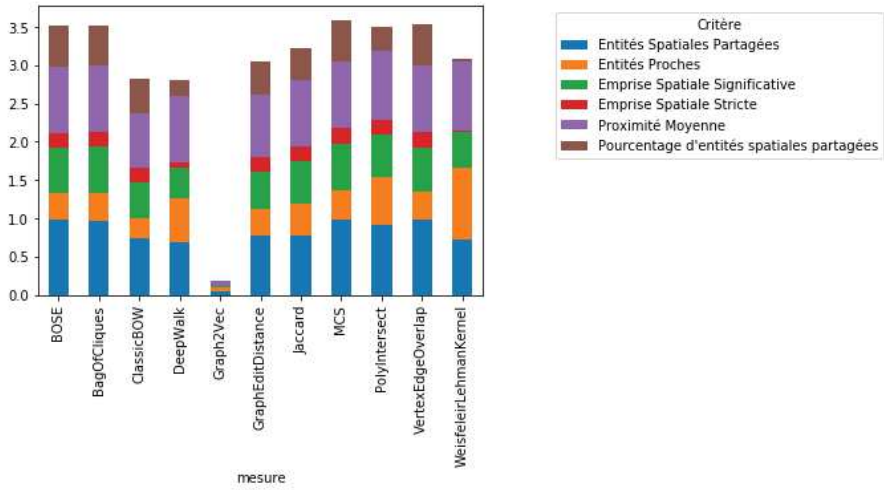
Les Tables 6 et 7 indiquent les résultats de MAP@1 et MAP@5 obtenus par les combinaisons dominantes [1] selon différents filtres — taille de la STR [2] et ensemble de critères (*c.f.* Section 5.3.2)— sur les deux corpus (*PadiWeb* et *AgroMada*).

Les résultats des Tables 6 et 7 indiquent que les combinaisons dominantes sur l'ensemble des critères (AC) sont composées de la mesure de la similarité *MCS* (MostCommonSubgraph) et de la forme de la STR normale, à l'exception des résultats de MAP@1 sur *AgroMada*, qui utilisent une forme généralisée au niveau régionale. Sur les critères SRP (Spatial Relatedness Privileged), les combinaisons dominantes

1. type STR, mesure de similarité
2. nombre d'entités spatiales



(a) MAP@1



(b) MAP@5

FIGURE 41 – Valeurs cumulées de MAP@n de la combinaison dominante pour chaque mesure de similarité évaluée (*AgroMada*).

sont composées de la mesure de similarité *MCS* et de la forme de la STR généralisée au niveau du Pays.

Comme dans l'évaluation des mesures (*c.f.* Section 5.4.2.2), les mesures s'appuyant sur des algorithmes *structure-based* permettent de maximiser la qualité des correspondances (selon les critères) sur l'ensemble des documents. Inversement, les algorithmes *pattern-based* appartiennent aux combinaisons dominantes dans les résultats impliquant uniquement les documents très spatialisés ($|ES| > 46$ et $|ES| > 72$) d'*AgroMada*. Dans l'ensemble des résultats, la mesure de similarité de l'algorithme *MCS* permet de maximiser la totalité des critères dans les différents corpus. Par ailleurs, nous observons qu'une des mesures proposées, *BagOfCliques* obtient de très bonnes perfor-

Mesure	Filtre	Forme de la STR	ESP	EP	ESS	ESSC	PM	PEP	Ensemble de critères
MCS	Aucun	Généralisée (Région)	0,98	0,20	0,76	0,47	0,89	0,75	AC*
WLK	Aucun	Généralisée (Pays)	0,67	0,92	0,46	0,02	0,90	0,05	SRP**
ClassicBOW	$\|ES\| > 46$	text	0,93	0,22	0,85	0,70	0,93	0,75	AC
ClassicBOW	$\|ES\| > 46$	text	0,93	0,22	0,85	0,70	0,93	0,75	SRP
ClassicBOW	$\|ES\| > 72$	text	0,91	0,27	0,91	0,64	0,91	0,69	AC
ClassicBOW	$\|ES\| > 72$	text	0,91	0,27	0,91	0,64	0,91	0,69	SRP

(a) MAP@1

Mesure	Filtre	Forme de la STR	ESP	EP	ESS	ESSC	PM	PEP	Ensemble de critères
MCS	Aucun	Normale	0,98	0,38	0,62	0,20	0,88	0,54	AC
WLK***	Aucun	Généralisée (Pays)	0,73	0,93	0,47	0,01	0,90	0,04	SRP
BagOfCliques	$\|ES\| > 46$	Normale	1,00	0,61	0,90	0,30	0,99	0,57	AC
BagOfCliques	$\|ES\| > 46$	Étendue (n=1)	1,00	0,61	0,89	0,30	0,99	0,57	SRP
BP***	$\|ES\| > 72$	Généralisée (Région)	0,95	0,58	0,75	0,29	0,79	0,50	AC
DeepWalk	$\|ES\| > 72$	Normale	0,87	0,76	0,51	0,13	0,98	0,25	SRP

(b) MAP@5

* All Criteria
** Spatial Relatedness Privileged
*** WLK = Weisfeiler-Lehman Subtree Kernel (Shervashidze et al., 2011) ; BP = Bipartite Graph Matching (Riesen et Horst Bunke, 2009)

Table 6 – Les combinaisons (type de STR, mesure de similarité) les plus performantes selon différentes pondérations de critères (*AgroMada*).

mances (voir Table 6b), particulièrement dans les critères ESS et PM (proximité spatiale et groupes significatifs) (90% de MAP@5).

Concernant les mesures de référence (*BOSE*, *ClassicBOW*, *PolyIntersect*), nous observons deux phénomènes. Dans les résultats de MAP@1 sur *AgroMada* (voir Table 6a), l'approche *ClassicBOW* [1] appartient aux combinaisons dominantes sur la partie des documents ayant plus de 46 entités spatiales. De plus, ces combinaisons obtiennent des valeurs de MAP@1 les plus élevées (supérieure à 64%) sur le critère ESSC [2]. Le second phénomène concerne la présence de *PolyIntersect* sur les documents les plus spatialisés (>7 et >4) du corpus *PadiWeb* (*c.f.* Discussion 5.5.1.4).

1. Utilisation de la représentation sac-de-mots
2. Emprise spatiale stricte

Mesure	Filtre	Forme de la STR	ESP	EP	ESS	ESSC	PM	PEP	Ensemble de critères
MCS	Aucun	Normale	0,98	0,21	0,54	0,22	0,79	0,63	AC*
MCS	Aucun	Généralisée (Pays)	0,89	0,34	0,49	0,22	0,81	0,54	SRP**
MCS	$\|ES\| > 4$	Normale	0,99	0,25	0,66	0,26	0,94	0,62	AC
MCS	$\|ES\| > 4$	Généralisée (Pays)	0,91	0,39	0,60	0,26	0,96	0,54	SRP
VertexEdgeOverlap	$\|ES\| > 7$	Généralisée (Pays)	0,98	0,32	0,60	0,23	0,96	0,54	AC
PolyIntersect	$\|ES\| > 7$	Généralisée (Pays)	0,85	0,70	0,32	0,17	0,98	0,33	SRP

(a) MAP@1

Mesure	Filtre	Forme de la STR	ESP	EP	ESS	ESSC	PM	PEP	Ensemble de critères
MCS	Aucun	Normale	0,96	0,35	0,38	0,07	0,81	0,44	AC
MCS	Aucun	Étendue (n=2)	0,96	0,37	0,35	0,07	0,80	0,44	SRP
MCS	$\|ES\| > 4$	Normale	0,99	0,41	0,46	0,07	0,95	0,43	AC
PolyIntersect	$\|ES\| > 4$	Étendue (n=1)	0,74	0,64	0,24	0,06	0,98	0,21	SRP
MCS	$\|ES\| > 7$	Étendue (n=1)	1,00	0,57	0,45	0,05	1,00	0,41	AC
PolyIntersect	$\|ES\| > 7$	Étendue (n=1)	0,91	0,75	0,32	0,07	0,99	0,25	SRP

(b) MAP@5

* All Criteria
** Spatial Relatedness Privileged

TABLE 7 – Les combinaisons (type de STR, mesure de similarité) les plus performantes selon différentes pondérations de critères (*PadiWeb*).

Concernant les effets des transformations de la STR, nous observons qu'elles permettent d'augmenter la MAP@n sur le critère EP (existence de relation spatiale inter-STR). La généralisation appartient à une majorité des combinaisons dominantes sur *AgroMada*. Inversement, les effets de la généralisation dans *PadiWeb* sont moindres (voir Table 7b) face à l'extension. Cette différence de comportement repose principalement sur la granularité moyenne des entités spatiales dans les corpus (*c.f.* Discussion 5.5.1.4). Par conséquent, les formes généralisées de la STR sont très présentes dans l'ensemble des combinaisons dominantes selon l'ensemble de critère SRP [1] (voir Table 6a).

Concernant les différents critères, les combinaisons dominantes retournent plus de 96% de documents avec une ou plusieurs entités spatiales partagés (ESP). Le critère EP est très sensible à la généralisation (*MCS* MAP@5 +10% sur *PadiWeb* et +55% sur *AgroMada*). Pour les critères ESS et ESSC (emprise spatiale significative et stricte), les

1. Spatial Relatedness Privileged (*c.f.* Section 5.3.3)

valeurs de MAP@1 et MAP@5 dépassent les 62% dans *AgroMada* et de 32% dans *PadiWeb*. Généralement, le critère ESSC (Emprise Spatiale Stricte) est associé à de faibles valeurs de MAP@n. A l'exception des résultats de la Table 6a, où *ClassicBOW* obtient les plus hautes valeurs de ESSC, une conséquence de l'existence de doublons dans le corpus de *AgroMada* (*c.f.* Discussion 5.5.2.2). Tout comme ESP, le pourcentage d'entités partagés (PEP) diminue avec la généralisation (perte de finesse).

5.5 DISCUSSIONS

Nous proposons d'axer les différentes discussions des résultats selon les deux phases générales du processus de mise en correspondance proposé. Dans une première partie, nous discutons les impacts des différentes formes de la STR sur les résultats de mise en correspondance. Dans une seconde partie, nous discutons des différentes catégories d'algorithmes de *graph matching* et de leurs effets dans les correspondances retournées.

5.5.1 *Georepresentation*

5.5.1.1 *Choix des relations spatiales*

Le choix des relations spatiales, i.e. l'inclusion et l'adjacence (*c.f.* Chapitre 2), s'appuie sur l'intégration de la topologie spatiale dans la représentation de la spatialité d'un document. Toutefois, il est possible qu'apparaissent dans certaines configurations spatiales des *clusters* (i.e. sous-graphes) déconnectés. Par exemple, dans la Figure 42, sur les six entités spatiales présentes dans un document, la STR générée est divisée en deux sous-graphes. Avec l'augmentation de la sparsité dans les graphes, nous diminuons le potentiel des algorithmes *pattern-based* comme les approches de *graph-embedding* qui s'appuient sur des marches aléatoires (*c.f.* Section 4.4.3). L'utilisation de nouvelles relations spatiales permet de densifier, de diminuer la sparsité du graphe de certaines STR et ainsi améliorer l'efficacité des algorithmes *pattern-based*. Par exemple, une nouvelle relation pourrait prendre en compte l'enchainement des occurrences des entités spatiales dans le document. Toutefois, si l'utilisation de nouvelles relations permet d'augmenter l'information spatiale existante puis de densifier, il y a un risque de bruitage de l'information présente dans la représentation.

5.5.1.2 *Attributs des entités spatiales de la STR*

Dans la génération d'une STR, les sommets du graphe produit ne sont ni pondérés [1], ni associés à un attribut particulier. Toutefois,

1. Nous considérons que chaque entité est représentative de manière égale dans la configuration spatiale.

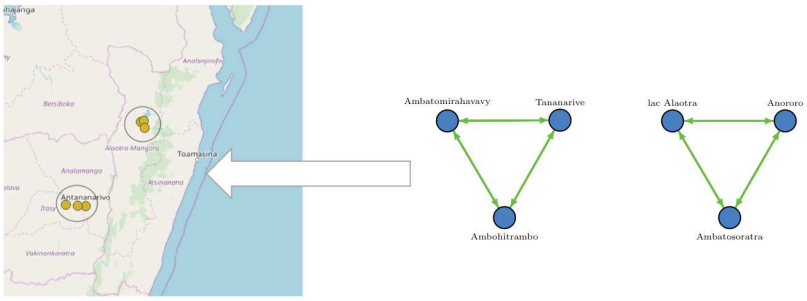


FIGURE 42 – Illustration de la sparsité de certaines STR.

certains attributs pourraient se révéler pertinents dans le cadre d'un processus de mise en correspondance spatiale. Dans la définition des relations spatiales, l'inclusion et l'adjacence permettent d'intégrer une information sur l'échelle spatiale de manière implicite. Une des évolutions possibles de la STR, pourrait être d'utiliser la classe associée pour chaque entité spatiale et construire un graphe multi-couche (*c.f.* Figure 9) ou développer une mesure plus complexe entre deux STR s'appuyant sur la comparaison de deux graphes : un graphe où les noeuds sont identifiés par leur identifiant et l'autre selon leur classe.

5.5.1.3 *Pondération des arêtes*

Dans les algorithmes de *graph matching*, l'intérêt principal de la *Graph Edit Distance* réside dans la définition des coûts de transformation. Par exemple, nous considèrons deux graphes G =(V= $\{A, B\}$, E=$\{(A \rightarrow^1 B\})$ et G' =(V= $\{A, C\}$, E=$\{(A \rightarrow C\})$. Dans le cadre de mesure de similarité ensembliste, la similarité entre G et G' est égale à 0.5. Dans le calcul de la GED[2], la différence repose sur le coût de substitution entre B et C (et inversement). Le coût de substitution s'appuie sur la différence entre les deux sommets et les arêtes incidentes. Par conséquent, le coût de la substitution de B vers C ne prend pas en compte l'ajout d'arête car celle-ci existe déjà dans G. Pour prendre un exemple, soit $STR_1 = \{Paris \rightarrow Ile\text{-}de\text{-}France\}$ et $STR_2 = \{Paris \rightarrow France\}$, les deux STR partagent une même entité *Paris* mais aussi une relation d'inclusion vers la seconde entité.

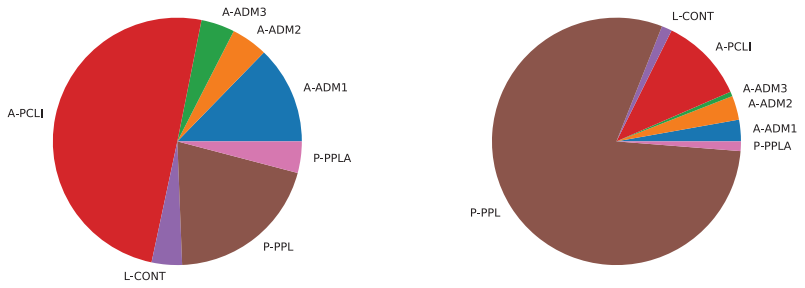
La GED a été utilisée dans la reconnaissance de forme (e.g. lettre manuscrite) et s'appuie sur la topologie mais aussi sur la distance (normalisée) entre chaque sommet. Dans notre évaluation, nous avons ajouté la distance entre les deux entités reliées par une arête dans les graphes de la STR. Les résultats indiquent que les documents correspondants retournés par la GED et les STR avec les arêtes pondérées par la distance, obtiennent une valeur de précision supérieure sur le critère ESS (groupes significatifs d'entités spatiales partagés). Toutefois, la GED n'apparait pas dans les combinaisons dominantes.

1. Relation d'inclusion
2. Graph Edit Distance

5.5.1.4 *Impact de la granularité*

Dans la Figure 43, nous illustrons la distribution des classes des entités présentes dans les différents corpus. Dans la Figure 43a, nous remarquons que le corpus de *PadiWeb* contient beaucoup d'entités associées à la classe *A-PCLI*, i.e. independent political entity, mais aussi *A-ADM1*, i.e. première division administrative (e.g. région française, état aux USA) et P-PPL, i.e. populated places. Dans la Figure 43b, nous remarquons que la grande majorité du corpus *AgroMada* est composée d'entités associées à la classe *P-PPL*. En comparant ces deux graphiques, deux tendances globales se distinguent concernant la granularité des entités employées dans les deux corpus. Dans le corpus de *PadiWeb*, les différentes configurations spatiales se situent au niveau du pays, de la capitale ou de la région, ce qui indique une granularité *"grossière"*. Inversement, dans le corpus d'*AgroMada*, les entités possèdent une granularité plus "fine" avec une présence forte d'entités avec la classe *P-PPL*, associée le plus souvent à des petites villes, villages, etc.

Dans *PadiWeb*, cette forte granularité va avoir un impact sur la réussite de *PolyIntersect* sur l'ensemble des critères *Spatial Relatedness Privileged*, particulièrement par la présence significative d'empreintes spatiales détaillées [1] des entités sur lesquelles s'appuie cet algorithme.



(a) PadiWeb (b) AgroMada

FIGURE 43 – Distributions des classes des entités présentes dans les différents corpus.

5.5.2 *Geomatching*

5.5.2.1 *Faiblesse des algorithmes pattern-based*

Les algorithmes *pattern-based* s'appuient sur la topologie des graphes, et par conséquent nécessitent des graphes avec une densité et/ou un

1. Polygone, Multi-Polygones

volume minimum. Par exemple, les méthodes de *graph-embedding* s'appuient sur plusieurs marches aléatoires pour générer la représentation vectorielle d'un graphe. Par conséquent, plus le graphe est éparse, plus la qualité des *features* extraites diminue. Dans les résultats de l'évaluation, les algorithmes DeepWalk, ou *BagOfCliques* apparaissent uniquement dans le combinaisons dominantes sur l'ensemble des documents très spatialisés.

5.5.2.2 *Détection de doublons dans AgroMada (ClassicBOW)*

Contrairement à *PadiWeb*, le corpus *d'AgroMada* contient plusieurs documents qui existent en plusieurs exemplaires ou en différentes versions. Ce sont des documents longs comme des rapports, des thèses, des articles ou encore des monographies. Généralement, ces documents sont associés avec une configuration spatiale très particulière avec un nombre important d'entités spatiales, ainsi qu'une distribution de mots très proches. De fait, ils sont très similaires avec leurs doublons. Par conséquent, *ClassicBOW* rapprochera ceux-ci en premier. La Figure 44 illustre les correspondances existantes autour d'un rapport sur *La mise en place d'un guichet financier à Madagascar* avec *ClassicBOW*.

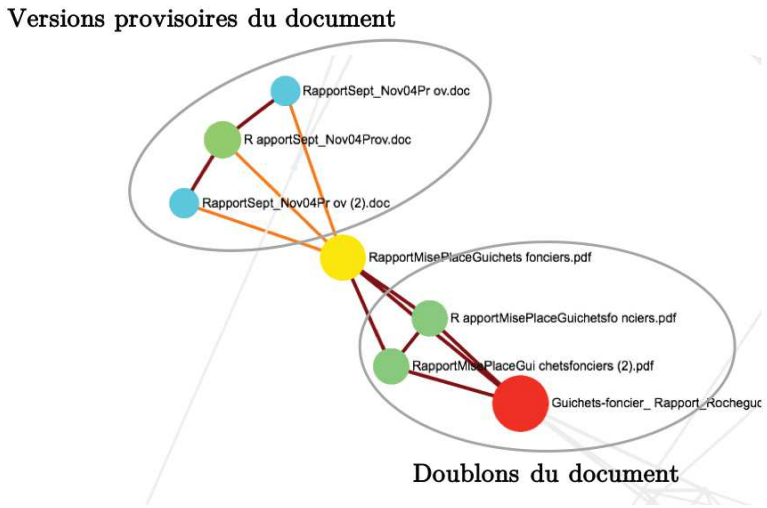


FIGURE 44 – Exemple de correspondances entre des doublons retournées par ClassicBOW.

5.5.2.3 *Effet des transformations*

Le ***rayonnement*** *ou la popularité d'une entité est définie par le nombre d'individus en ayant la connaissance.*

Dans nos travaux, nous proposons différentes transformations pour étendre ou généraliser (*c.f.* Section 2.4) l'information spatiale présente dans les STR. La généralisation (*c.f.* Section 2.4.1) permet de remplacer des entités avec une granularité fine avec une entité plus générale contenant celle-ci. L'extension (*c.f.* Section 2.4.2) permet d'étendre les entités spatiales avec un rayonnement faible pour garantir la finesse de la représentation spatiale en plus d'augmenter son rayonnement.

Dans les Tables 8 et 9, nous indiquons les valeurs de précision moyenne obtenues sur l'ensemble des combinaisons selon la forme de STR choisie. Nous observons une tendance où la généralisation favorise deux documents avec des STR possédant des relations inter-STR, ce qui a pour conséquence d'augmenter la valeur du critère EP (Entité proches). Concernant l'extension, peu de changements sont observés. Une des perspectives envisagées est de fusionner les deux formes issues des transformations de la STR, i.e. généralisation et extension.

Type de STR	**ESP**	**EP**	**ESS**	**ESSC**	**PM**	**PEP**
Étendue (n=1)	0,99	0,20	0,73	0,48	0,89	0,75
Étendue (n=2)	0,99	0,19	0,75	0,47	0,89	0,74
Généralisée (limite = Pays)	0,94	**0,24**	0,74	0,48	0,90	0,70
Généralisée (limite = Région)	0,98	0,20	**0,76**	0,47	0,89	0,75
Normale	**0,99**	0,19	0,74	**0,48**	**0,90**	**0,75**

TABLE 8 – Valeurs moyennes obtenues pour chaque critère selon la forme de la STR (*AgroMada*).

Type de STR	**ESP**	**EP**	**ESS**	**ESSC**	**PM**	**PEP**
Étendue (n=1)	0,98	0,20	0,52	0,22	**0,81**	0,63
Étendue (n=2)	0,98	0,20	0,52	0,22	**0,81**	0,63
Généralisée (limite = Pays)	0,89	**0,34**	0,49	0,22	**0,81**	0,54
Généralisée (limite = Région)	0,96	**0,23**	0,53	0,22	0,79	0,63
Normale	**0,98**	0,21	**0,54**	**0,22**	0,79	**0,63**

TABLE 9 – Valeurs moyennes obtenues pour chaque critère selon la forme de la STR (*PadiWeb*).

5.5.2.4 *Une mise en correspondance personnalisable*

Dans les résultats de mise en correspondance, nous observons différentes combinaisons dominantes selon la taille des graphes de la STR. Si les STR possèdent un nombre d'entités spatiales important, les algorithmes *pattern-based* permettent d'obtenir de meilleures correspondances. Inversement, les algorithmes *structure-based* sont plus efficaces sur les documents moins spatialisés. De plus, différents ensembles de critères (*c.f.* Section 5.3.3) sont proposés pour permettre d'évaluer les résultats de mise en correspondance selon la mise en correspondance souhaitée : stricte ou approximative. Une mise en correspondance stricte garantit l'existence d'entités similaires entre les

STR tandis que l'approximative se focalise sur la proximité géographique. Dans la Table 10, nous répartissons les mesures selon leurs affinités avec la nature du corpus et le type de mise en correspondance (stricte, approximative).

Mise en correspondance	**Nombre d'entités spatiales**	
	Faible	*Grand*
Stricte	MCS	BagOfCliques, GraphEditDistance
Approximative	Weisfeleir-Lehman Subtree Kernel	BagofCliques, DeepWalk

Table 10 – Choix de mesures en fonction des besoins de mise en correspondances.

Deuxième partie

DIMENSION THÉMATIQUE

Dans la continuité de la contribution sur la mise en correspondance de données textuelles hétérogènes, la seconde contribution se focalise sur l'aspect thématique. Dans cette partie, nous présentons notre seconde contribution qui propose une extension de la Spatial Textual Representation en intégrant les thèmes apparaissant dans la même fenêtre contextuelle qu'une ou plusieurs entités spatiales. L'objectif est d'évaluer si des thèmes apparaissant dans des contextes proches permettent d'induire certaines similarités spatiales dans la mise en correspondance de documents. Pour modéliser cette information dans la STR, nous proposons deux concepts, l'entité thématique puis la relation thématique.

INTRODUCTION

Dans la partie précédente, un processus de mise en correspondance sur la dimension spatiale des données textuelles hétérogènes a été présenté. Ce processus se déroule en deux étapes. La première étape (*c.f.* Partie 1 Chapitre 2) consiste à extraire les descripteurs spatiaux d'un document pour les intégrer dans une représentation : la Spatial Textual Representation (STR) (*c.f.* Section 2.1.3). Dans la deuxième étape, différentes mesures de similarité du domaine du *graph matching* sont étudiées (*c.f.* Partie 1 Section 4) pour identifier les correspondances entre les STR générées. Dans le but d'améliorer les résultats de mise en correspondance, différentes transformations de la STR ont été proposées pour étendre ou généraliser la configuration spatiale d'un document (*c.f.* Partie 1 Section 2.4). Dans l'évaluation du processus, différentes combinaisons (type de STR, mesure de similarité) sont évaluées selon un ensemble de critères de similarité spatiale. Dans les deux corpus utilisés (*c.f.* Partie 1 Section 5.1) dans nos expérimentations, la STR ou sa forme généralisée associée à des mesures de *graph matching* de catégorie *structure-based* permettent d'obtenir de meilleurs résultats sur l'ensemble des documents spatialisés.

L'existence et la détection des entités spatiales sont cruciales à la STR. À ces deux problématiques, l'utilisation des transformations est une première solution. Cependant, il existe toujours des représentations spatiales faibles —i.e. avec peu d'entités spatiales—, ou éparses par la présence de sous-graphes déconnectés dans certaines des STR générées (*c.f.* Partie 1 Discussion 5.5.2.1). C'est la raison pour laquelle les algorithmes *pattern-based* peinent à être convaincants dans la mise en correspondance de ces représentations. Pour cela, nous souhaitons explorer la contextualisation des entités spatiales autour de certaines thématiques pour permettre d'améliorer les résultats de mise en correspondance spatiale. À cette fin, nous proposons un processus d'identification des relations thématiques entre les entités spatiales et des entités dites thématiques.

Cette partie est organisée de la manière suivante. Dans le premier chapitre, nous présentons l'extraction et l'intégration de l'information thématique dans la STR. Dans le second chapitre, nous présentons les résultats obtenus sur la mise en correspondance spatiale ainsi que la couverture des différents vocabulaires permettant de détecter les entités thématiques.

1

INTÉGRATION DE L'INFORMATION THÉMATIQUE DANS LA STR

Trois dimensions d'information sont distinguables dans un document : *spatiale, temporelle et thématique.* La **dimension spatiale** d'un document est représentée par ses entités spatiales et les connexions qui existent entre elles. La **dimension temporelle** d'un document s'appuie sur différentes périodes (ou instants) représentées par des dates explicites (e.g. "[. . .] nous sommes arrivés le 29 septembre 2019" ou relative (e.g. "[. . .] nous sommes arrivés le lendemain [. . .]"). Enfin, la **dimension thématique** indique le ou les sujets abordés dans un document (e.g. "[. . .] d'interventions de développement agricole (notamment rizicole) depuis les années 50 [. . .]" selon le vocabulaire adopté. Si chacune de ces dimensions peuvent être analysées séparément, plusieurs applications (analyse épidémiologique (Arsevska, Valentin et al., 2018), politique (Agnew, K. Mitchell et Toal, 2008)) requièrent des analyses combinant ces informations. Par exemple, il est possible de combiner les informations de deux dimensions de la manière suivante :

— **Thématique + Spatiale.** La localisation des thèmes est analysée (géographique ou non).

— **Thématique + Temporelle.** L'utilisation des thèmes sur différentes périodes est analysée.

— **Spatiale + Temporelle.** L'évolution de la configuration spatiale est étudiée sur différentes périodes.

D'autres contributions se focalisent sur une analyse combinant les trois dimensions (J. Li et al., 2018 ; Peuquet, 1994 ; Tao et al., 2018). (Peuquet, 1994) propose la définition des trois combinaisons suivantes :

1. **Où + Quand → Quoi.** L'analyse des thèmes abordés pour une localisation et une période donnée.
2. **Quoi + Quand → Où.** L'analyse de la localisation d'un thème à une période donnée.
3. **Où + Quoi → Quand.** L'analyse de la période relativement à un thème et une localisation donnés.

Dans la partie i, nous avons présenté les difficultés de l'extraction d'entités spatiales, souvent liées aux données traitées (i.e. hétérogénéité), l'index géographique utilisé, l'existence implicite d'informations, . . . Des problèmes similaires sont associés à l'information tem-

porelle, notamment dans sa relativité[1]. Inversement, l'information thématique est inhérente à un document.

Dans nos travaux, l'information thématique se réfère à un ou plusieurs thèmes qui sont abordés dans un document. Un "**thème**" est défini comme : *[. . .] une idée, un sujet développé dans un discours, un écrit, un ouvrage*[2]. Chaque thème d'un document peut être représenté explicitement (e.g. mots-clés) ou implicitement au travers du vocabulaire utilisé. Nous nous appuyons sur la notion d'*entité thématique* représentée par un terme appartenant à un vocabulaire (*c.f.* Sections 1.2.2 et 1.2.3).

Nous proposons d'exploiter l'information thématique pour améliorer la mise en correspondance spatiale. La section suivante présente le lien existant entre les thèmes et les lieux et comment il peut améliorer la mise en correspondance spatiale.

1.1 EXPLOITER L'INFORMATION CONTEXTUELLE ENTRE THÈME ET ESPACE

De manière générale, les informations thématiques et spatiales sont liées, et dans certains cas, la première permet d'induire la seconde. Par exemple, la ville de *Cannes* est régulièrement associée à la thématique du *cinéma*. Ces deux informations sont liées car elles apparaissent fréquemment dans un même contexte. Nous identifions deux catégories de contextes illustrés dans la Figure 45, dans lesquels une information spatiale et une information thématique peuvent être liées.

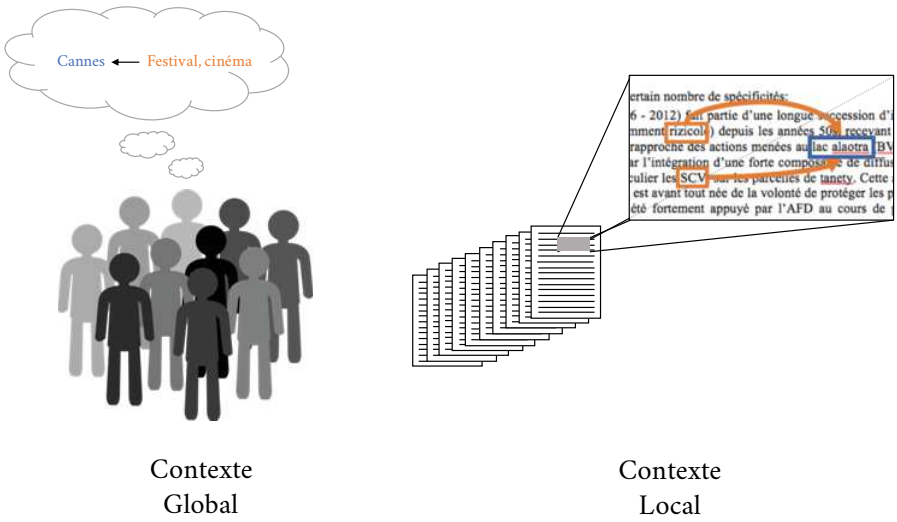


FIGURE 45 – Illustration de la différence entre contexte local et global.

1. Relative à une date ou un moment du récit
2. Définition du dictionnaire de l'Académie Française : https://www.cnrtl.fr/definition/thème

CONTEXTE GLOBAL. Dans un contexte global, les relations entre thème et espace sont construites (resp. disparaissent) dans le temps selon différents facteurs (culture, événement, enseignement). Nous proposons d'illustrer ces relations au travers du système d'auto-complétion implémenté dans les moteurs de recherche. Lors de l'écriture d'une requête, le système propose plusieurs requêtes complètes similaires. Généralement, ces propositions s'appuient sur leur similarité sémantique avec la requête initiale et leur fréquence d'utilisation. La Table 11 indique les 4 premières propositions faites par le système pour les thèmes *corrida*, *festival* et *manifestation* à la date du 24 juin 2019. Nous avons choisi deux moteurs de recherche, l'un prenant en compte la localisation de l'utilisateur (*Google* [1] [2]) et l'autre pas (*Qwant* [3]).

Dans la majorité des résultats, le système propose de compléter l'ensemble des requêtes (i.e. thèmes) par une localisation. Étonnamment, l'auto-complétion du système *Qwant* associe la ville Langueux (Bretagne,FR) avec la corrida, un événement de course sportive des *Côtes d'Armor* et non la course de taureau.

Début de Requête	**Moteur de recherche**	
	Google	*Qwant*
corrida	toulouse	landerneau
	France	langueux
	maugio	–
	nimes	–
festival	de cannes	cannes
	de nimes	cannes 2019
	avignon	avignon
	de lunel	cannes
	montpellier	beauregard
manifestation	montpellier	paris
	toulouse	1er mai
	gilet jaunes montpellier	police
	gilet jaunes toulouse	1er mai paris
	–	–

TABLE 11 – Exemple de requêtes proposés par les systèmes d'auto-complétion des deux moteurs de recherche : GOOGLE et QWANT (date : 24 Juin 2019).

1. https://www.google.fr
2. À notre connaissance, il n'existe pas de moyen de désactiver la prise en compte de la géolocalisation dans les paramètres de Google.
3. https://www.qwant.com

CONTEXTE LOCAL. Le contexte local se limite soit à une configuration spatiale particulière ou une (ou plusieurs) thématique spécifique. Dans nos travaux, cette limite est représentée par l'ensemble de documents du corpus étudié. Parmi les données utilisées (*c.f.* Partie 1 Section 5.1), le corpus d'*AgroMada* contient des documents associés à la thématique de l'agroécologie et centrés sur le pays Madagascar. Le corpus de *PadiWeb* est composé de documents (articles de presse) traitant d'épidémies de maladies animales. Le document de la Figure 46 est composé de différents termes associés à des techniques agricoles telles que le semi sous couvert végétal (SCV) qui apparaissent à côté de toponymes associés aux zones d'études.

Dans nos travaux, nous souhaitons exploiter l'information contextuelle *locale* dans le but d'améliorer la mise en correspondance spatiale dans les corpus étudiés.

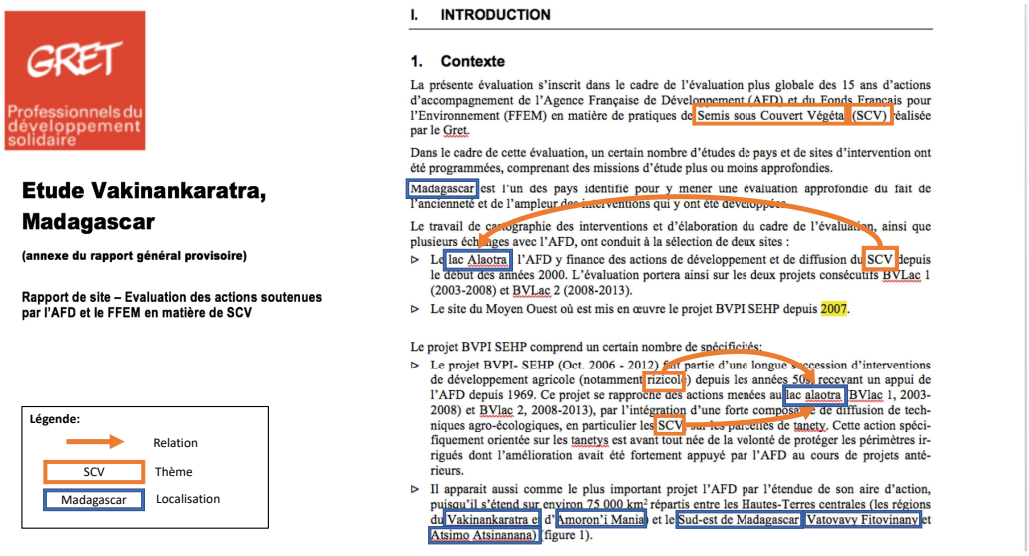


FIGURE 46 – Illustration de relation récurrente entre thème et espace dans un corpus.

Pour représenter l'information thématique associée à une ou plusieurs entités spatiales, nous proposons d'intégrer un nouveau type d'entité à la STR, l'**entité thématique**, reliée à une entité spatiale par une **relation thématique**. Tout comme l'entité spatiale, l'entité thématique est associée à un identifiant d'une ressource terminologique (dictionnaire, ontologie, thesaurus, etc.), ce qui permet de mesurer la portion d'informations thématiques partagées entre deux documents durant le processus de mise en correspondance. Enfin, la relation thématique permet d'indiquer si une entité spatiale appartient à une ou plusieurs fenêtres contextuelles dans lesquelles apparaît une entité thématique.

Pour illustrer le potentiel de ces concepts, la démonstration suivante s'appuie sur les documents représentés dans la Figure 47 selon leurs informations spatiales et thématiques. Les deux documents partagent une même thématique autour de la technique agricole, *SCV*. Inversement, les deux documents mentionnent chacun un lieu différent, le *lac Alaotra* [1] dans le document **1** et la commune de *Ambatondrazaka* [2] dans le document **2**. Dans cet exemple, la similarité entre les entités spatiales présentes dans les deux documents est égale à 0. Pourtant, le *lac Alaotra* et la commune de *Ambatondrazaka* sont deux entités adjacentes. Par conséquent, la similarité spatiale entre les deux documents est non-nulle. Dans ce contexte, la relation qu'entretiennent les entités spatiales avec l'entité thématique *SCV* permet d'induire cette proximité entre les deux entités spatiales. Dans le cadre du processus de mise en correspondance spatiale, nous nous intéressons à ce **transfert de propriété spatiale** d'une entité spatiale vers une entité thématique que permet cette nouvelle relation.

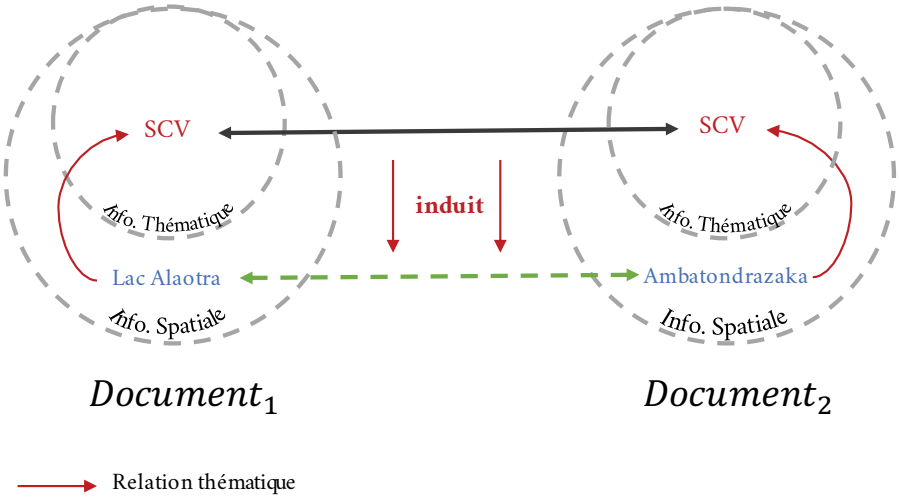


FIGURE 47 – Informations partagées entre deux documents en prenant conjointement les dimensions thématique et spatiale.

Dans la section suivante, nous présentons la représentation choisie de l'information thématique d'un document, l'entité thématique.

1.2 ENTITÉ THÉMATIQUE

Pour représenter les unités thématiques partagées entre différents documents et les liens qu'elles entretiennent avec les entités spatiales, nous proposons de définir un nouveau type d'entité : l'**entité thématique**. Dans cette section, nous définissons l'entité thématique et les vocabulaires utilisés pour l'identifier dans les données (i.e. docu-

1. https://www.openstreetmap.org/relation/1154818
2. https://www.openstreetmap.org/node/1492685514

ments). Ce vocabulaire doit permettre de représenter au mieux les thématiques globales du corpus étudié. À cette fin, nous proposons différents types de vocabulaire pour chaque corpus :

1. **Le vocabulaire de domaine** est produit à l'aide de connaissances expertes d'un domaine particulier (e.g. médicale, agronomique).
2. **Le vocabulaire de corpus** est généré selon une méthode d'extraction supervisée ou semi-supervisée sur un ensemble de documents.

Cette section est organisée de la manière suivante. Dans la section 1.2.1, nous donnons la définition de l'entité thématique. Une fois définie, nous introduisons les deux types de vocabulaire sélectionnés pour identifier les entités thématiques dans les documents. Pour les vocabulaires de domaine (Section 1.2.2), nous présentons les différents formats disponibles, puis les ressources sélectionnées. Pour le vocabulaire de corpus (Section 1.2.3), nous présentons différentes méthodes permettant de générer ce type de vocabulaire, puis les vocabulaires résultant des données étudiées.

1.2.1 *Définition*

Une **entité thématique** est un concept appartenant à un domaine (i.e. médical, agronomique, etc.) ou à un corpus spécifique. Elle est représentée par un ou plusieurs syntagmes définis dans un vocabulaire. Un syntagme correspond à une combinaison de morphèmes ou de mots qui se suivent et produisent un sens acceptable [1]. Deux formats de vocabulaire se distinguent : structuré ou ordonné.

Un **vocabulaire structuré** permet d'intégrer des relations hiérarchiques (thésaurus, ontologie) entre les concepts. Un **vocabulaire ordonné** est une liste de concepts selon un ordre particulier (alphabétique). Pour chaque concept, un vocabulaire peut proposer d'autres représentations : traduction, noms alternatifs, description verbale. La Table 12 présente les données associées au concept de "parcelle" dans la base de connaissances *Wikidata*.

Langue	Label	Alias
English	land lot	lot, plot tract of land, parcel of land, real property, real estate
French	parcelle	

Table 12 – Informations sur l'entité thématique *parcelle* sur *Wikidata*

1. Source : https://www.cnrtl.fr/definition/syntagme

Dans nos travaux, un vocabulaire doit être composé d'entités propres aux thématiques générales d'un corpus. Dans les vocabulaires sélectionnés pour représenter l'information thématique d'un document, nous faisons une division en deux catégories : **un vocabulaire de domaine** et **un vocabulaire de corpus**. Dans les sections suivantes, nous définissons chaque catégorie et les vocabulaires associés. La totalité des vocabulaires utilisés sont indiqués dans la Table 13.

Vocabulaire	Corpus	Type de vocabulaire
DicoAE	AG *	domaine
Maladie Infectieuse	PW **	domaine
Développement Durable	AG	domaine
Biotex PadiWeb	PW	corpus
Biotex AgroMada	AG	corpus
LDA AgroMada	AG	corpus
BiotexLDA AgroMada	AG	corpus

* AgroMada
** *PadiWeb*

TABLE 13 – Ensemble des vocabulaires utilisés pour la détection d'entités thématiques dans les corpus étudiés.

1.2.2 *Vocabulaire de Domaine*

Un vocabulaire de domaine est un ensemble d'entités thématiques, chacune représentant un concept propre à un domaine spécifique (médical, agronomique, etc.). Contrairement au vocabulaire de corpus, ce type de vocabulaire est construit et validé par un ou plusieurs experts du domaine. Il peut être structuré selon différents formats comme :

1. **Terminologie.** Une terminologie [1] est un ensemble des termes, rigoureusement définis, spécifiques d'une science, d'une technique, d'un domaine particulier de l'activité humaine.
2. **Thesaurus.** Dans les années 1960, le thésaurus est une ressource utilisée pour indexer et rechercher des documents (DALBIN, 2007 ; VICKERY, 1960). Un thésaurus est une liste organisée de termes contrôlés et normalisés représentant les concepts d'un domaine de connaissances. Les termes sont reliés entre eux par des relations de synonymie (terme équivalent), de hiérarchie (terme générique et terme spécifique) et d'association (terme associé) ; chaque terme appartient à une catégorie ou domaine.

1. https://www.larousse.fr/dictionnaires/francais/terminologie

3. **Ontologie.** Une ontologie permet de représenter, de nommer et définir un ensemble de propriétés et de relations entre des concepts d'un ou plusieurs domaines [1]. Les termes d'*ontologie* et de *vocabulaire* sont souvent confondus. Selon la définition du W3C (World Wide Web Consortium) d'une ontologie [2] : " *vocabularies define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern.*" ; et : "*There is no clear division between what is referred to as "vocabularies" and "ontologies". The trend is to use the word "ontology" for more complex, and possibly quite formal collection of terms[. . .]*". Si les ontologies sont utilisées pour modéliser des concepts (e.g. ville), leur instances (e.g. Paris) et les relations qu'ils entretiennent (Paris se situe en France) sont généralement utilisées comme ressource terminologique à l'aide des propriétés propres aux normes telle SKOS [3] avec les propriétés `http://www.w3.org/2008/05/skos#altLabel` ou RDF [4] avec `http://www.w3.org/2000/01/rdf-schema#label`.

Les avantages des vocabulaires de domaines sont nombreux. Premièrement, ce sont des ressources construites ou validées par des experts. Deuxièmement, elles proposent de multiples représentations (orthographe). Enfin, elles intègrent des relations entre les différents concepts dans la ressource ou vers une autre.

1.2.2.1 *Vocabulaires sélectionnés*

Dans les corpus de documents utilisés, nous avons identifié deux thématiques générales. Pour le corpus d'*AgroMada*, la thématique principale repose sur **l'agroécologie** et pour *PadiWeb*, les **maladies infectieuses animales**. En conséquence, nous avons sélectionné trois vocabulaires en lien avec ces thématiques : le dictionnaire de l'agroécologie produit par l'INRA (*AgroMada*), le vocabulaire du Développement Durable (*AgroMada*) et un vocabulaire de maladie infectieuse (*PadiWeb*) (*c.f.* Section 5.1).

LE DICTIONNAIRE DE L'AGROÉCOLOGIE (AGROECOINRA). Véronique Batifol-Garandel et Marie-Colette Fauré [5] ont mis en œuvre un dispositif de veille territoriale sur l'agroécologie dans la région Midi-Pyrénées. Tout au long du dispositif, des informations extraites à partir de sources scientifiques, règlementaires, associatives, administratives issues du Web (presse, blogs, flux RSS) sont collectées, analysées et validées. Les informations collectées sont renseignées dans DicoAE, un dictionnaire capitalisant plus de 300 termes liés à l'agroécologie. DicoAE est disponible à cette adresse : `https://dicoagroecologie.fr`

1. Source : `https://en.wikipedia.org/wiki/Ontology_(information_science)`
2. `https://www.w3.org/standards/semanticweb/ontology.html`
3. `https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html`
4. `http://www.w3.org/2000/01/rdf-schema`
5. INRA, Toulouse

Label français	Label anglais
Agro-sylvo-pastoralisme	Agro-forestry-pastoral
Agrobiodiversité	agro-biodiversity
Agroécologie	agroecology
Agroécosystème	agroecosystem
Agroforesterie	agroforestry
Agriculture biologique	Organic Agriculture
Agriculture durable	Sustainable agriculture
Agriculture intégrée	Integrated agriculture
Approche participative	Participatory approach

TABLE 14 – Extrait de la terminologie provenant du dictionnaire de l'agroécologie.

VOCABULAIRE MALADIE INFECTIEUSE (MALADIE INFECT.). Pour analyser les informations sur les différentes maladies animales présentes dans les extractions [1] de *PadiWeb*, un vocabulaire décrivant la terminologie des différentes maladies animales est proposé. Les termes officiels et leurs variations orthographiques sont collectés à partir de ressources tels que *Agrovoc* [2], *UMLS* [3], et *Wikipedia* par une experte des données [4]. Ce vocabulaire est composé de 175 entités associées avec des labels dans différentes langues ainsi que des alias (noms alternatifs).

Label	Alias anglais (Agrovoc)	Alias français (Agrovoc)
Ornithosis	chlamydia psittaci infections, chlamydiosis	Ornithose
Tularemia	tularaemia, tularemia	tularémie
Toxoplasmosis	toxoplasmosis	Toxoplasmose
Borna disease	Borna disease	Maladie de Borna
Feline infectious peritonitis	feline infectious peritonitis	Péritonite infectieuse féline

TABLE 15 – Extrait de la terminologie sur les maladies infectieuse.

1. Ensemble de documents extraits de fils d'actualités comme *Google News*
2. AGROVOC est un vocabulaire contrôlé couvrant tous les domaines d'intérêts de l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO) http://aims.fao.org/fr/agrovoc
3. Unified Medical Language System https://www.nlm.nih.gov/research/umls/index.html
4. Sarah Valentin, CIRAD

VOCABULAIRE DU DÉVELOPPEMENT DURABLE. Le Vocabulaire du développement durable est un dictionnaire publié en 2015 par la Délégation générale à la langue française et aux langues de France. Il comprend 610 termes (et leurs définitions) associés à des concepts du domaine de l'agriculture, la pêche, l'environnement et du développement durable. Ce vocabulaire est disponible à l'adresse suivante : https://frama.link/9McF5W5G

Label en français	**Label en anglais**
accaparement de terres	land grabbing
agriculture biologique	organic farming
agriculture durable	sustainable agriculture
agroalimentaire	agri-food
agro-écologie	agroecology
agroéquipement	agricultural equipment
agroforesterie	agroforestry
agro-industrie	agricultural industry
allélopathie	allelopathy
allomone	allomone

TABLE 16 – Extrait du Dictionnaire du Développement Durable.

Dans la section suivante, nous présentons la deuxième catégorie de vocabulaire spécifique aux documents formant un corpus.

1.2.3 *Vocabulaire de Corpus*

En analyse du discours (GEE, 2004), un ensemble de méthodes se concentre sur l'analyse du vocabulaire dans les corpus de textes. Plusieurs aspects sont étudiés comme la variété (taille du vocabulaire), l'enchainement (expression), le sens (polysémie) et les écritures (alias, langue) des termes. Pour étudier ces vocabulaires, l'utilisation de l'outil informatique permet d'appliquer efficacement des méthodes d'analyse statistique sur de larges corpus. Contrairement aux vocabulaires de domaine, les vocabulaires de corpus ne sont pas consacrés à la formalisation, la modélisation des concepts d'un domaine mais à identifier les termes saillants [1]. Ces méthodes sont utilisées en classification de documents. Pour identifier ces termes saillants, nous proposons d'utiliser deux types d'approches suivantes.

LES APPROCHES DE TOPIC-MODELING. *Le topic modeling* consiste à mesurer la pertinence pour chaque mot d'un corpus selon différents

1. Ils peuvent aider à leur conception.

topics dont le nombre n est défini par l'utilisateur. Un *topic* correspond à un sujet sous-jacent dans un ou plusieurs documents. La pertinence des mots dans chaque *topic* est représentée par différentes distributions de probabilités sur l'ensemble des mots apparaissant dans un document. Par exemple, un *topic* sur la thématique de l'agronomie associera une forte valeur de probabilité à *parcelle*, *engrais*, *culture*, etc. Ces modèles sont largement utilisés dans l'analyse de corpus ainsi que dans l'extraction et l'analyse de terminologie (H. Wang et al., 2011 ; Zhao et al., 2011).

Plusieurs approches ont été proposées dans le domaine du *topic modeling*. Dans un premier temps, la Latent Semantic Analysis (LSA) (Deerwester et al., 1990) et son évolution la Probabilistic-LSA (Hofmann, 1999) sont proposées. La LSA est une décomposition en valeurs singulières (SVD) d'une matrice *terme-document* M pour interpoler la matrice terme-topic et la matrice document-topic, indiquant l'affinité d'un terme et celle d'un document avec les différents *topics*. Pour répondre aux limites de la LSA, le modèle Latent Dirchlet Analysis (LDA) (David M Blei et al., 2003) est proposé. D'autres modèles dérivés de la LDA sont proposés pour faire de la classification d'image (M. Li et Yuan, 2005), l'analyse de réseaux sociaux ou encore de l'analyse de sentiment (C. Lin et He, 2009). Les auteurs de (Alghamdi et Alfalqi, 2015 ; Jelodar et al., 2019) proposent une étude exhaustive des modèles de *topic modeling* proposés ainsi que leurs applications.

L'EXTRACTION AUTOMATIQUE DE MOTS-CLÉS. En Recherche d'Information, une pratique courante consiste à associer manuellement ou automatiquement un ensemble de **mots-clés** pour chaque document. Un mot-clé est défini [1] comme : "*[. . .] essentiel d'une phrase, d'un vers ou d'un document, d'un sujet de rédaction ou de dissertation*". Dans (Hulth, 2003), les auteurs présentent les mots-clés comme un résumé dense pour un document [2]. De manière formelle, un **mot-clé** est représenté par un syntagme — i.e. composé d'un ou plusieurs mots allant jusqu'à la phrase simple — pouvant appartenir à un ensemble de documents ou à un domaine particulier.

Les approches d'extraction automatique de mot-clés consistent à extraire les termes ou phrases saillants d'un document en s'appuyant sur des approches statistiques reposant sur la fréquence (Ramos et al., 2003), la cooccurrence (Matsuo et Ishizuka, 2004) ou les motifs syntaxiques (Bourigault, Gonzalez-Mullier et Gros, 1996 ; Bourigault et Jacquemin, 1999 ; Daille, 1994 ; Lossio-Ventura et al., 2014) dans lesquels apparaissent certains termes. D'autres méthodes utilisent les informations typographiques (Humphreys, 2002) présentes dans des documents enrichis tels que les pages web. Plusieurs études exhaus-

1. Source : https://www.larousse.fr/dictionnaires/francais/mot-clé_mots-clés/52770

2. Citation en anglais : *Keywords may, for example, serve as a dense summary for a document, lead to improved information retrieval, or be the entrance to a document collection.*

tives des contributions dans ce domaine sont proposées dans (Bharti et Babu, 2017 ; Cohen, 1995 ; Hasan et Ng, 2014).

Nous présentons les méthodes choisies pour chacune des catégories ainsi qu'une contribution dans le domaine de l'extraction de vocabulaire de corpus.

1.2.3.1 *Méthodes d'extraction de vocabulaire de corpus*

Dans la section précédente, nous avons présenté deux catégories d'extraction de vocabulaire de corpus : les méthodes de *topic-modeling* et les méthodes d'*extraction automatique de mots-clés*. Dans nos travaux, nous avons sélectionné une approche pour chacune de ces catégories. Pour le modèle de *topic modeling*, nous avons sélectionné le modèle canonique de Latent Dirchlett Analysis (LDA) présenté dans (David M Blei et al., 2003). Pour le modèle d'extraction automatique de mots-clés, nous avons sélectionné Biotex, un système proposé dans (Lossio-Ventura et al., 2014). Bien que la contribution de (Lossio-Ventura et al., 2014) se focalise sur des données biomédicales, l'approche proposée a fait ses preuves sur des données avec des thématiques différentes telles que l'agronomie (Roche et al., 2015).

Dans les paragraphes suivants, nous présentons le modèle de la LDA (David M Blei et al., 2003) ainsi que la méthode développée dans Biotex (Lossio-Ventura et al., 2014). Puis, nous proposons une méthode d'extraction s'appuyant sur la combinaison de l'extraction de mots-clés de Biotex et la division en *topics* de la LDA. Les interactions entre les trois méthodes sont illustrées dans la Figure 48.

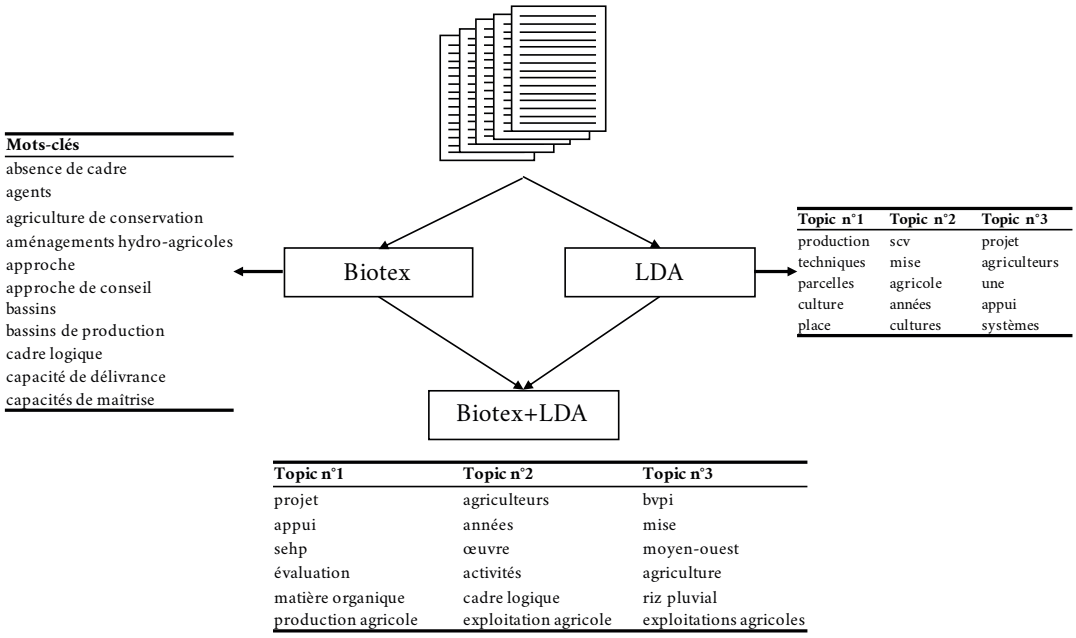


Figure 48 – Interaction entre les trois méthodes d'extraction de vocabulaire de corpus.

1.2.3.2 *Latent Dirchlett Allocation*

La Latent Dirchlet Allocation (LDA) (David M Blei et al., 2003) est un modèle probabiliste génératif de corpus. L'hypothèse de la LDA repose sur une vision d'un ou plusieurs documents comme un mélange aléatoire de *topics* sous-jacents. Chaque *topic* est représenté par une distribution de probabilités sur l'ensemble du vocabulaire (des mots) d'un ou plusieurs documents. Dans l'exemple de la Table 17, trois *topics* sont produits par la LDA. Dans le premier *topic*, des fortes valeurs sont associées aux mots liés à la génétique comme *gene*, *dna*. Dans le deuxième *topic*, les probabilités les plus fortes sont attribuées aux mots *life*, *evolve* qui représentent la thématique de la *biologie*.

Topic 1		**Topic 2**		**Topic 3**	
w_i	$P(w_i\|T_1)$	w_i	$P(w_i\|T_2)$	w_i	$P(w_i\|T_3)$
gene	0,04	life	0,02	brain	0,04
dna	0,02	evolve	0,01	neuron	0,02
genetic	0,01	organism	0,01	nerve	0,01
...	...	...	...	...	...

Table 17 – Exemple de distributions de probabilités sur les mots dans différents *topics* retournés par la LDA.

Étant un modèle non-supervisé, le choix du nombre de *topics* est primordial. Pour déterminer le nombre de *topics*, la cohérence des top-k mots doit être mesurée. Pour cela, plusieurs résultats de la LDA sur différents nombres de *topics* sont évalués pour leur pertinence par des experts ou à l'aide de méthodes automatiques (Musat et al., 2011 ; Röder, Both et Hinneburg, 2015). Ces méthodes automatiques s'appuient sur différents critères comme la cooccurrence de termes d'un même *topic* dans un corpus, ou encore la mesure d'information mutuelle (PMI). Parmi les méthodes automatiques, différentes approches proposent de calculer la cohérence des *topics* extraits s'appuyant sur des mesures dédiées. Parmi les mesures proposées dans (Röder, Both et Hinneburg, 2015), nous avons sélectionné la mesure C_V en nous appuyant sur les résultats des expérimentations menées dans (Röder, Both et Hinneburg, 2015). La mesure C_V calcule la similarité contextuelle entre les top-n mots (mots plus représentatifs d'un *topic*). Pour un mot w_i et w_j avec $\{w_i, w_j\} \in W$, avec $\boldsymbol{W}$ les top-n mots, les vecteurs $\vec{u}$ et $\vec{v}$ contenant les valeurs de similarité de w_i et w_j avec l'ensemble des mots de W sont calculés. La similarité correspond à la valeur normalisée de l'information mutuelle (NPMI). Une fois les deux vecteurs calculés, la cohérence $c(w_i, w_j)$ entre les deux mots w_i et w_j est mesurée à l'aide de la similarité cosinus entre $\vec{u}$ et $\vec{v}$. La cohérence de l'ensemble du *topic* calculée dans C_V correspond à la moyenne des valeurs de $c(w_i, w_j)$ avec $w_i, w_j \in W \times W$.

La Figure 49 illustre la progression du score du cohérence obtenu selon différents nombres de *topics* dans le corpus d'*AgroMada*.

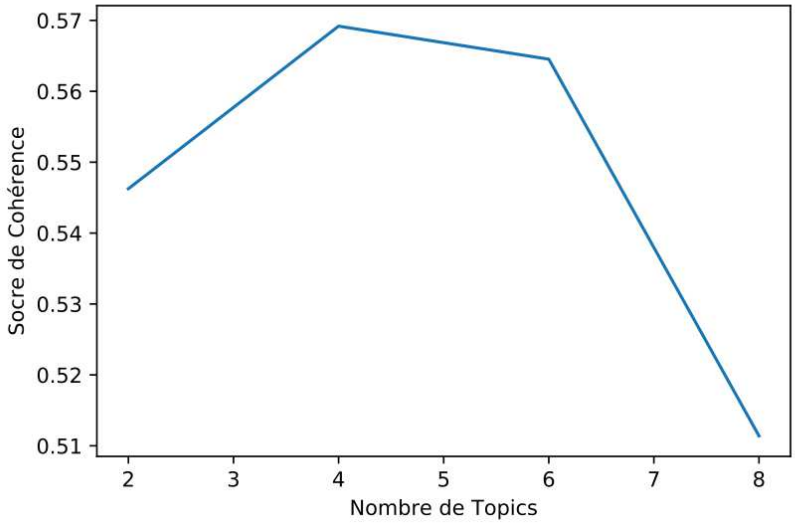


FIGURE 49 – Valeurs de cohérence obtenues sur différents nombres de *topics* extraits par la LDA sur le corpus *AgroMada*.

1.2.3.3 BIOTEX : *Biomedical Terminology Extraction*

BIOTEX (LOSSIO-VENTURA et al., 2014) est un outil d'extraction automatique de mots-clés sur des documents produits dans le domaine biomédical. Les mots-clés extraits sont de tailles variées, i.e. n-grams avec $\mathbf{n} \in \{1, 2, 3\}$. L'approche proposée s'appuie sur une connaissance *a priori* représentée par l'ensemble des motifs syntaxiques fréquents dans lesquels apparaissent les mots-clés. Un motif syntaxique correspond à une séquence de mots ayant une classe grammaticale particulière. Dans l'exemple de la Table 18, le motifs syntaxique correspond à $\text{NOM} + \text{VERBE} + \text{NOM}_2 + \text{ADP}$, où NOM est un mot-clé.

Mot	Viral infections	are	major	drivers	of
Classe grammaticale	**NOM**	**VERBE**	**NOM**		**ADP**

TABLE 18 – Le mot-clé *Viral infections* apparait dans le motif $\text{NOM} + \text{VERBE} + \text{NOM}_2 + \text{ADP}$.

En s'appuyant sur cette connaissance, le processus de détection de mots-clés multilingues [1] de BIOTEX se déroule en deux étapes. Dans une première étape, les n-grams [2] apparaissant dans les motifs syntaxiques *a priori* sont collectés. Dans une seconde étape, les n-grams

1. Ensemble de langues disponibles dans TreeTagger
2. Un n-gram est une sous-séquence de n éléments construite à partir d'une séquence donnée. Ici, la séquence correspond à un document.

extraits sont triés selon une mesure telle que *TF-IDF* (SALTON et BUCKLEY, 1991) ou *C-Value* (FRANTZI, ANANIADOU et MIMA, 2000). Dans les extractions de mots-clés effectuées avec BIOTEX, nous avons sélectionné la mesure $F - TFIDF - C_M$ en nous basant sur les évaluations de (LOSSIO-VENTURA et al., 2014). $F - TFIDF - C_M$ est la moyenne harmonique de la valeur de *TF-IDF* et de la *C-Value* d'un n-gram. La Table 19 présente une extraction des mots-clés de BIOTEX sur le document illustré dans la Figure 46.

1-gram	**2-gram**	**3-gram**
projet	production agricole	zones à risques
agriculteurs	couvert végétal	ouvrages de stabilisation
appui	couverture végétales	cultures en semis
approche	organisations paysannes	systèmes de cultures
nutriments	mode opératoire	rapport de site

TABLE 19 – Mots-clés extraits avec l'outil BIOTEX sur le document illustré dans la Figure 46.

1.2.3.4 *Méthode hybride : BiotexLDA*

Précedement, nous avons présenté le modèle de la LDA (David M BLEI et al., 2003) et l'outil BIOTEX proposé dans (LOSSIO-VENTURA et al., 2014). Nous considérons que les deux méthodes possèdent des caractéristiques combinables pour obtenir une extraction de vocabulaire plus pertinente. Dans le modèle de la LDA (David M BLEI et al., 2003), la caractéristique principale repose sur la quantification de la pertinence des mots selon différents *topics*. Cependant, le modèle ne prend pas en compte des unités de documents telles que les n-grams. Inversement, dans l'approche de BIOTEX, les mots-clés extraits dépassent le 1-gram. Dans notre approche, nous proposons de combiner l'identification de n-grams pertinents proposée par BIOTEX et la distinction des termes dans différents *topics* de la LDA. Plusieurs approches (David M. BLEI et J. D. LAFFERTY, 2009 ; X. WANG, MCCALLUM et WEI, 2007) proposent d'intégrer la notion de n-grams dans la définition du modèle de la LDA. En s'appuyant sur les travaux de (VELCIN, ROCHE et PONCELET, 2016), les auteurs évaluent l'impact de bigrams extrait par BIOTEX sur la qualité des topics construits par la LDA. Nous proposons une extension de cette approche qui intègre l'ensemble des n-grams extraits par BIOTEX dans le calcul de la LDA.

Nous proposons BIOTEXLDA, une approche intégrant la notion de n-grams pertinents dans le calcul de la LDA. Les n-grams sont sélectionnés à l'aide d'un modèle d'extraction automatique de mots-clés — ici BIOTEX. Le processus s'effectue selon les deux étapes suivantes :

1. **Pré-traitement du corpus.** Les termes significatifs sont extraits avec BIOTEX. Les n-grams extraits par BIOTEX et identifiés dans le corpus sont transformés en unigrammes (1-grams). Pour cela, les espaces dans les n-grams extraits sont remplacés par le caractère "_". Par exemple, le terme *Agriculture Durable* est transformé en *Agriculture_Durable*.
2. **Génération de la LDA.** Une fois le corpus transformé, les distributions de probabilités des n-grams sur les documents sont calculées à l'aide du modèle *LDA*. Un fois extraits, les top n-gram pour chaque *topic* sont collectés.

Dans la Table 20, nous présentons les résultats obtenus par BIOTEXLDA et la LDA selon le même nombre de *topics* sur le document illustré dans la Figure 46. Contrairement à la LDA, l'approche proposée permet d'obtenir une extraction de vocabulaire représentant différents *topics* avec une variété de termes et une sémantique plus fine. Par exemple, dans le premier *topic*, la LDA fait ressortir le terme *production* tandis que le pré-traitement à l'aide des mots-clés extraits par BIOTEX permet d'obtenir un syntagme plus précis comme *production agricole*.

	Topic n°1	**Topic n°2**	**Topic n°3**
Avec Biotex	projet	**agriculteurs**	bvpi
	appui	années	mise
	sehp	œuvre	moyen-ouest
	évaluation	activités	agriculture
	matière organique	**cadre logique**	riz pluvial
	production agricole	**exploitation agricole**	exploitations agricoles
	engrais chimiques	**mode opératoire**	**œuvre déléguée**
	marge brute	**organisations paysannes**	**pression foncière**
	temps de travail	**systèmes de production**	systèmes de culture
	aménagements hydro agricoles	plantes de couverture	agriculture de conservation
	rapport de mission	ouvrages de stabilisation	mise en place
	dialogue entre techniciens	**diffusion de connaissances**	**journée de travail**
Sans Biotex	production	scv	projet
	techniques	mise	agriculteurs
	parcelles	agricole	une
	culture	années	appui
	place	cultures	systèmes

TABLE 20 – Comparaison entre les résultats obtenues à l'aide du modèle de la LDA avec et sans l'utilisation des mots-clés extraits par BIOTEX.

1.2.3.5 *Les vocabulaires produits*

POST-TRAITEMENT DES VOCABULAIRES PRODUITS. Dans les différents vocabulaires de corpus générés, l'ensemble des toponymes sont enlevés pour éviter de biaiser les résultats de l'évaluation de la mise en correspondance sur la dimension spatiale. Les doublons correspondant aux formes pluriels sont supprimés.

BIOTEX PADIWEB. Nous avons à disposition deux corpus de données produites par le système de *PadiWeb*. Le premier corpus correspond au *golden-standard* (RABATEL, ARSEVSKA et ROCHE, 2019) mis en place pour évaluer le système *PadiWeb*. Le second corpus correspond à un ensemble de 35 000 documents extraits par *PadiWeb*. En utilisant les données de chaque corpus dans BIOTEX, deux vocabulaires sont produits : BIOTEXPADI500 et BIOTEXPADI35K. Des extraits des deux vocabulaires générés sont présentés dans la Table 21.

Top N-Gram (500 documents)			**Top N-Gram (35,000 documents)**		
1-Gram	*2-Gram*	*3-Gram*	*1-Gram*	*2-Gram*	*3-Gram*
poultry	bird flu	service de santé	poultry	poultry products	politica de confidentialitate
market	avian flu	cas de grippe	virus	wild birds	titularilor de copyrightşi
pigs	wild birds	campagne de vaccination	cattle	spongiform encephalopathy	influenza a virus

TABLE 21 – Extrait du vocabulaire obtenu par BIOTEX sur les corpus de documents de *PadiWeb*.

BIOTEX AGROMADA. Ce vocabulaire contient 1293 mots-clés sélectionnées à partir des 1500 mots-clés retournées par BIOTEX sur l'ensemble du corpus d'*AgroMada*. Un extrait du vocabulaire est présenté dans la Table 22.

BIOTEXLDA AGROMADA Ce vocabulaire est composé de 11902 termes extraits en utilisant la méthode BIOTEXLDA combinant l'extraction de mot-clés de BIOTEX (LOSSIO-VENTURA et al., 2014) et la division en *topics* de la LDA (David M BLEI et al., 2003). La Table 23 présente un extrait du vocabulaire généré par la méthode.

Dans cette section, l'entité thématique et les vocabulaires sélectionnés ou générés ont été présentés. Dans la section suivante, nous présentons la relation thématique et son intégration de la STR ainsi que ces transformations.

1-Gram	2-Gram	3-Gram
projet	sécurisation foncière	projet de mise
production	assemblée générale	membres de bureau
art	crédit solidaire	rapport de campagne
appui	produits ligneux	pièces de rechange
paysans	rapport final	mission de cadrage
riz	gestion foncière	réunions de bilan
article	rapport trimestriel	renouvellement de traitement

Table 22 – Extrait du vocabulaire du corpus *AgroMada* obtenu par Biotex.

1-Gram	2-Gram	3-Gram
Appui	production agricole	accumulation de sédiments
Rapport	activités trimestriel	séances de présentation
Projet	assistance technique	plebeius en milieu
Total	bassins versants	hectares de superficie
valeur	temps partiel	formation remise niveau
travaux	sarclage chimique	fois par ans
TOTAL	engrais chimique	ensablement de rizières

Table 23 – Extrait du vocabulaire obtenu par l'approche BiotexLDA sur *AgroMada*.

1.3 RELATION THÉMATIQUE

Une entité thématique est un concept représenté selon différentes orthographes dépendant de la langue ou de leurs utilisations communes (alias). Dans nos travaux, l'objectif de l'intégration de l'information thématique dans la STR consiste à relier les thèmes partageant un contexte similaire avec une ou plusieurs entités spatiales de telle sorte que ces relations permettent d'induire et d'inclure des similarités spatiales entres des entités spatiales distinctes dans deux documents. À cette fin, nous proposons d'intégrer une nouvelle relation à la STR, **la relation thématique**, reliant une entité spatiale à une entité thématique.

1.3.1 *Etat de l'art*

De manière générale, l'extraction de relation s'effectue entre les entités formant le triptyque (voir Figure 50) : Personne, Lieu, Organisation ainsi que les variables qui leur sont associées. Ces variables sont généralement des quantités, des dates (e.g. date de naissance). De nombreux problèmes d'extraction d'information bénéficient de ces méthodes. Dans (Aprosio, Giuliano et Lavelli, 2013 ; Gardner et

T. MITCHELL, 2015), les auteurs utilisent des méthodes d'extraction de relation pour compléter et étendre l'information contenue dans des bases de connaissances telles que DBPEDIA [1] (APROSIO, GIULIANO et LAVELLI, 2013). Les systèmes de Question Réponse (VOORHEES, 2003) utilisent des approches combinant l'extraction de relations et l'utilisation d'informations présentes dans des bases de connaissances (FADER, ZETTLEMOYER et ETZIONI, 2014 ; XU et al., 2016) ou sur web (RAVICHANDRAN et HOVY, 2002). Enfin, une grande partie de la littérature est consacrée à l'extraction de relations dans des données textuelles du domaine biomédical (FUNDEL, KÜFFNER et ZIMMER, 2007 ; GIULIANO, LAVELLI et ROMANO, 2006 ; GURULINGAPPA, MATEEN-RAJPU et TOLDO, 2012 ; POON, TOUTANOVA et QUIRK, 2014 ; SKOUNAKIS, CRAVEN et RAY, 2003), notamment dans l'extraction de relations entre gènes et protéines, ou encore la détection d'effets négatifs d'une posologie en s'appuyant sur des témoignages de patients.

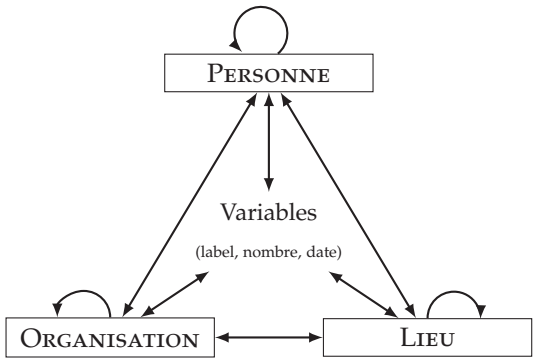


FIGURE 50 – Triptyque commun dans le domaine d'extraction de relation.

Ces méthodes sont divisées en deux catégories : **supervisées** et **semi-supervisées**.

1.3.1.1 *Approche Supervisée*

Les approches supervisées ramènent le problème d'extraction de relations à un problème de classification. Soit S, une fenêtre contextuelle — une fenêtre fixe, une phrase, un paragraphe — dans laquelle apparaissent les deux entités liées e_1 et e_2 ; et $F(S)$ un ensemble de descripteurs associés à la phrase S, *e.g. lemme, arbre de dépendance, type d'entités, PartOfSpeech*. Pour une relation R donnée, une fonction f indique si e_1 et e_2 sont reliées par R dans la formule suivante :

$$f_R(T(S)) = \begin{cases} +1 & e_1 \text{ et } e_2 \text{ sont liées par } R \\ -1 & \text{sinon} \end{cases} \tag{18}$$

1. https://wiki.dbpedia.org

Dans la plupart des approches supervisées, différents descripteurs associés aux entités et aux mots présents dans **S** sont utilisés. Trois descripteurs sont principalement utilisés :

1. **Le type de l'entité.** Chaque terme dans une relation est associé à un *tag*, ou une classe qui définit le type de l'entité. Par exemple, dans la relation (ville, Pays) présente dans la phrase *"Paris est la capitale de la France"*, les termes *Paris* et *France* sont associés aux tags LOCALISATION ou *GPE* (Geopolitical Entity).
2. **La classe grammaticale.** La classe grammaticale, plus connue en anglais par *PartOfSpeech*, indique la nature d'un mot, *e.g. nom, verbe, nom propre, adjectif, déterminant, pronom, ponctuation, etc.*

   Exemple

   > Le chat est endormi. En le réveillant, le chien est griffé par le chat.
   > ↪ Le/**Déterminant** chat/**Nom** est/**Verbe** endormi/**Adjectif**. En/**Préposition** le/**Déterminant** réveillant/**Verbe** , le/**Déterminant** chien/**Nom** est/**Verbe** griffé/**Verbe** par/**Préposition** le/**Déterminant** chat/**Nom**
3. **L'arbre de dépendance** est une structure intégrant les différentes relations grammaticales entretenues par les différents mots d'une phrase. Un exemple d'arbre de dépendance est donné dans la Figure 51.

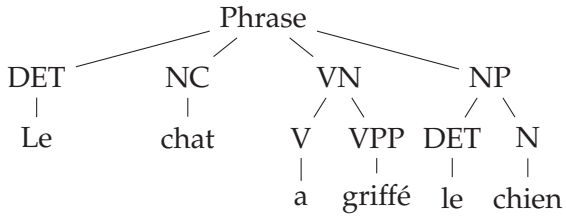


FIGURE 51 – Exemple d'arbre de dépendance généré à partir de la phrase "Le chat a griffé le chien".

1.3.1.2 *Approche Semi-Supervisée*

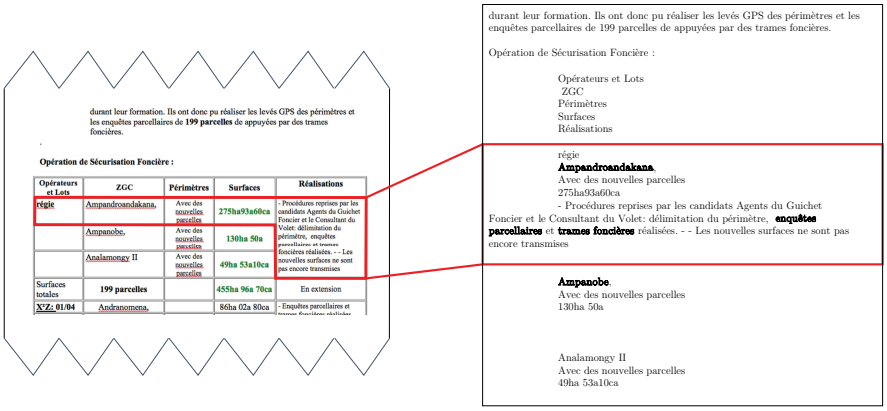
Les approches semi-supervisées d'extraction de relation utilisent des données initiales pour augmenter le nombre de motifs dans lesquels s'exprime une relation entre deux entités. Dans (BRIN, 1999), les auteurs proposent une approche nommée DIPRE qui utilise un ensemble de relations — e.g. *book(title,author)* — construit préalablement. S'appuyant sur le *framework* de DIPRE, (AGICHTEIN et GRAVANO, 2000) proposent différentes mesures de confiance lors de la génération des motifs se basant sur les termes présents à gauche, au milieu et à droite des deux entités reliées par une relation **R**. Dans l'approche de (ETZIONI et al., 2004), les auteurs utilisent un ensemble de motifs

initiaux pour extraire des relations à partir du Web. De manière similaire, les auteurs (FUNDEL, KÜFFNER et ZIMMER, 2007) proposent un ensemble de règles initiales permettant l'identification de la relation R(gène,protéine) sur un ensemble de résumés (*abstracts*) d'articles scientifiques.

1.3.2 *Problématique : Nature hétérogène des documents*

Précédemment, nous avons présenté différentes approches dans le domaine de l'extraction de relation. Dans la majorité de ces méthodes, l'utilisation de descripteurs syntaxiques est cruciale dans la détection ou la validation d'une relation entre deux entités dans un document. Par conséquent, ces méthodes sont fortement dépendantes du bon fonctionnement des méthodes d'extraction de ces descripteurs. Tout comme les méthodes de reconnaissance d'entités nommées, ces méthodes sont très dépendantes de la langue, et par conséquent, elles sont sensibles aux bruits dans un document.

Dans les données textuelles hétérogènes, la structure d'un document est composée de plusieurs blocs : phrase, paragraphe, tableau, énumération, liste. La structure du paragraphe et de la phrase permet d'extraire des descripteurs en utilisant des méthodes de fouille de textes de l'état de l'art. Inversement, cette même extraction s'avère difficile sur des structures moins conventionnelles comme la *table*. Par exemple, la Figure 52 montre un extrait d'un document (à gauche) contenant une table et la donnée texte (à droite) que nous traitons. Dans les deux représentations, l'utilisation d'informations grammaticales est rendue difficile par l'absence de blocs tels que la phrase.



Original Texte

FIGURE 52 – Extrait du document avec une table dans sa forme originale et brute (texte).

Sachant que l'utilisation d'informations syntaxiques (descripteurs) est difficilement identifiable dans les documents étudiés, nous pro-

"You shall know a word by the company it keeps" (FIRTH, JOHN RUPERT, 1957)

posons une extraction de relation s'appuyant sur la cooccurrence entre les entités spatiales et les entités thématiques dans une fenêtre contextuelle.

Dans la prochaine section, nous introduisons la notion de relation thématique.

1.3.3 *Définition de la relation thématique*

Une **relation thématique** R_T relie une entité spatiale es_i à une entité thématique et_k apparaissant dans une même fenêtre contextuelle (*c.f.* Section 1.3.4). Dans le cadre du graphe de la STR, R_T est intégrée sous forme d'arcs avec l'entité spatiale $\mathbf{es_i}$ (la source) et l'entité thématique $\mathbf{et_j}$ (la cible). La Figure 53 illustre la relation thématique entre *Lac Alaotra* et *SCV*.

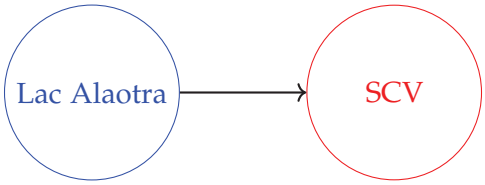


FIGURE 53 – Exemple de relation thématique entre l'entité spatiale *Lac Alaotra* et l'entité thématique *SCV*.

Si une entité thématique et_k est reliée à deux ou plusieurs entités spatiales $\{es_0, \ldots, es_n\} \in E$, alors pour toutes entités $(es_i, es_j) \in E^2$ sont reliées par l'entité thématique et_k (voir Figure 54).

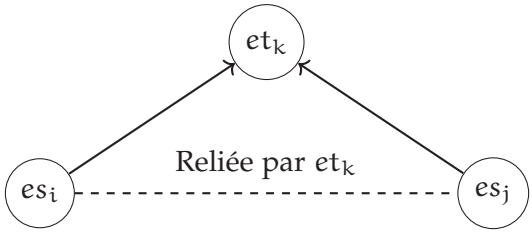


FIGURE 54 – Exemple de relation thématique entre deux entités spatiales

1.3.4 *Différentes fenêtres contextuelles*

Pour extraire une relation thématique $\mathbf{R_T}$ entre une entité spatiale et une entité thématique, la notion de **fenêtre glissante** est choisie. Une fenêtre glissante correspond à une fonction qui capture de manière itérative différentes sections de la donnée (voir Table 24). Différents types de fenêtres sont envisagés :

Itération	Parcours
1	La ville de **Cerbère** propose d'accompagner les **fermiers** dans leur transition vers le BIO.
2	La ville de **Cerbère** propose d'accompagner les **fermiers** dans leur transition vers le BIO.
3	La ville de **Cerbère** propose d'accompagner les **fermiers** dans leur transition vers le BIO.

TABLE 24 – Parcours d'un document à l'aide d'une fenêtre de taille fixe.

TAILLE FIXE. Une fenêtre de taille fixe est définie par un nombre n de mots autour d'une position dans la donnée. Par conséquent, la taille de la fenêtre est de $2n + 1$.

PHRASE. Chaque fenêtre est délimitée par la notion de phrase, i.e. commence par une majuscule et se termine une ponctuation.

PARAGRAPHE. Un paragraphe[1] est une section de document vouée au développement d'un point particulier et composé de plusieurs phrases. Le début d'un paragraphe est marqué par un alinéa ou par un saut de ligne.

Dû à la nature hétérogène des données textuelles traitées, l'identification de la phrase ou paragraphe s'avère difficile. Par conséquent, nous avons choisi d'utiliser une fenêtre glissante comme contexte permettant d'identifier les relations thématiques.

1.3.5 *Détection des entités thématiques*

Pour identifier l'ensemble des entités thématiques $\{et_0, .., et_n\}$ d'un vocabulaire V, nous utilisons plusieurs variations de e_t (i.e. pluriel, singulier, lemme, alias). Pour détecter si une entité thématique appartient à un document, chaque variation est recherchée dans un document D à l'aide de son préfixe. Si un mot $w_i \in D$ est égal au préfixe v_0 d'une variation $v = \{v_0, v_1, \ldots, v_n\}$, les mots $\{w_{i+1}, \ldots, w_{i+n}\}$ sont comparés avec les $n - 1$ mots restants de v. Si $\{w_{i+1}, \ldots, w_{i+n}\} = \{v_1, \ldots, v_n\}$, alors l'entité thématique est identifiée entre la position i et $i + n$ du document D. L'algorithme 5 détaille la recherche des entités thématiques décrites par V dans un document D.

1.3.6 *Extraction des relations thématiques*

Dans nos travaux, l'existence d'une relation thématique existe entre une entité spatiale et une entité thématique si celles-ci appartiennent à une même fenêtre contextuelle dans un document. Précédemment,

1. Définition : https://www.cnrtl.fr/definition/paragraphe

Algorithme 5 : Identification des entités thématiques dans un document

```
Input :
  — D document
  — V Vocabulaire utilisé
  — ET Entités thématiques détectées
ET ← []
foreach e_t ∈ V do
    variations ← récupérerVariations(e_t)
    foreach v ∈ variations do
        while t < |D| do
            i = 0
            start = t
            while w_t == v_i do
                if i == |forme| then
                    append (e_t, start, start + i) into ET
                else
                    t++
                i++
            t ← t++
return ET
```

nous avons présenté l'algorithme utilisé pour détecter une entité thématique et sa position dans un document. La position des entités spatiales dans le document est extraite lors du processus de génération de la STR.

L'extraction de relations thématiques repose sur plusieurs variables : l'ensemble des entités spatiales E présentes dans le document D, le vocabulaire T et la fenêtre contextuelle $f \in \{fixe, phrase, paragraphe\}$ choisie. L'algorithme de détection de relation thématique parcourt le document D, où chaque itération correspond à une fenêtre contextuelle. En s'appuyant sur l'algorithme de détection d'entité thématique (*c.f.* Section 1.3.5), les entités thématiques sont identifiées selon le vocabulaire V. Les entités spatiales qui apparaissent dans la même fenêtre sont collectées. Si une entité thématique e_t et une entité spatiale e_s sont identifiées dans une même itération, i.e. une même fenêtre, alors (e_s, e_t) partage une relation thématique. Le processus d'extraction de relation thématique est indiqué dans l'Algorithme 6.

Algorithme 6 : Extraction de relation thématique

```
Input :
  — E Entités spatiales présentes dans D
  — V Vocabulaire utilisé
  — D Document
  — f Fenêtre choisie
R_T ← {}
for fenetre ∈ extractionFenetre(D, f) do
    entSpatialesDetectees ← IsIn(E, fenetre)
    entThematiquesDetectees ← IsIn(V, fenetre)
    foreach es ∈ entSpatialesDetectees do
        foreach et ∈ entThematiquesDetectees do
            append (es, et) into R_T
return R_T
```

Dans la prochaine section, nous présentons l'intégration des relations thématiques extraites dans la structure de la STR et ses différentes formes (transformations de la STR).

1.3.7 *Intégration à la STR*

La Spatial Textual Representation (STR) est représentée par un graphe G_{STR} dirigé et étiqueté défini par $G_{STR} = (V_{ES}, E_{RS}, T_{RS})$. L'intégration de la thématique dans cette structure implique deux changements. Dans le graphe de la STR, les sommets appartiennent à une même catégorie d'entité, l'entité spatiale. Avec l'ajout de l'entité thématique dans la structure, nous étendons la définition de G_{STR} avec l'ensemble de sommets V_{ET} correspondant aux entités thématiques. Pour la relation thématique, E_{RT}, l'ensemble des arcs représentants les relations thématiques est ajouté à la définition de G_{STR}.

Finalement, la STR intégrant la dimension thématique est représentée par le graphe G_{STR+T} définit par l'expression :

$$G_{STR+T} = (V_{ES}, V_{ET}, E_{RS}, E_{RT}, T_{RS})$$

La Figure 55 illustre le graphe d'une STR intégrant les relations thématiques (arcs bleus) entre les entités spatiales (sommets bleus) et les entités thématiques (sommets rouges). Dans cet exemple, la région de *Vakinankaratra* est liée à l'entité thématique *SCV*.

1.3.7.1 *Propagation sur les transformations*

Dans la partie i, plusieurs transformations de la STR sont proposées (*c.f.* Partie I Section 2.4) pour améliorer les résultats de mise en correspondance spatiale. Par conséquent, la question de l'intégration des relations thématiques dans les différentes formes de la STR est cruciale.

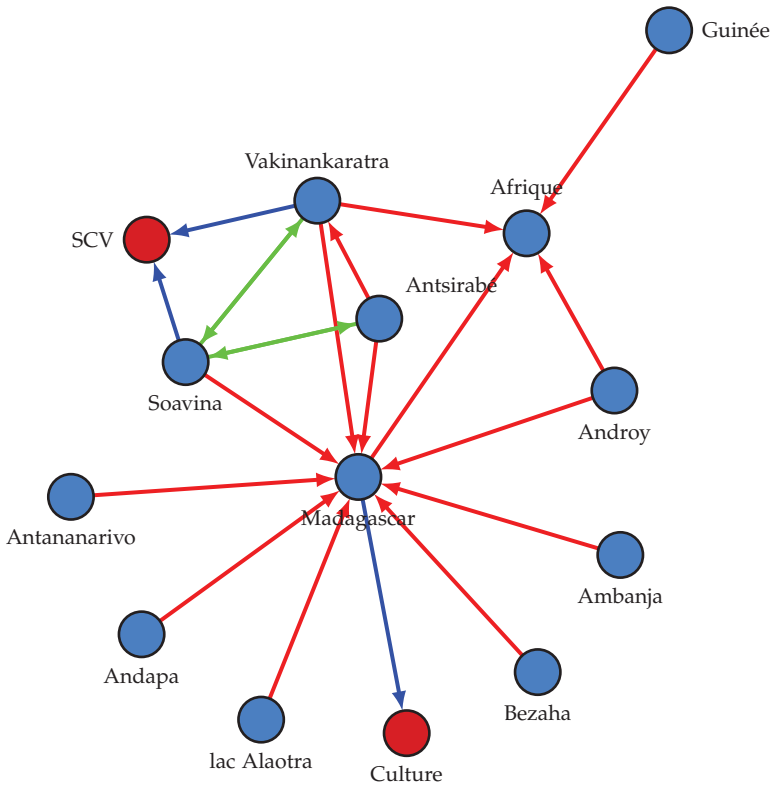


FIGURE 55 – Exemple d'une STR intégrant les relations thématiques.

Dans le contexte de la forme généralisée, l'entité spatiale avec une granularité plus élevée peut-elle bénéficier de la même relation thématique ? Dans la forme étendue, est-ce que la relation thématique peut être propagée à l'entité spatiale étendue ? Pour répondre à ces deux questions, nous nous appuyons sur les relations qu'entretiennent l'entité spatiale et l'entité issue de sa généralisation ou de son voisinage (extension).

Pour répondre à cette question, la notion de partage d'empreinte spatiale est cruciale. En effet, une relation thématique indique une similarité entre une entité thématique et une entité spatiale, et par extension son empreinte spatiale. Dans le cadre de la généralisation, si une entité spatiale es_2 issue de la généralisation d'une entité es_1, alors l'empreinte géographique de es_1 est incluse dans es_2. Si une entité thématique es_t est associée à es_1, alors elle est associée à son empreinte géographique par extension. Par conséquent, si es_t est associée à l'empreinte géographique de es_1 et que es_2 inclue l'empreinte géographique de es_1, alors on considère que es_2 est associée es_t. Inversement, l'extension d'une entité es_1 par l'entité es_2 ne garantit pas l'inclusion de l'empreinte spatiale d'es_2 dans es_1. Par conséquent, es_2 ne peut pas être associée à es_1.

Par conséquent, chaque entité issue de la généralisation hérite des relations thématiques des entités spatiales dont elle est issue. Inversement, les entités spatiales issues de l'extension n'héritent pas des

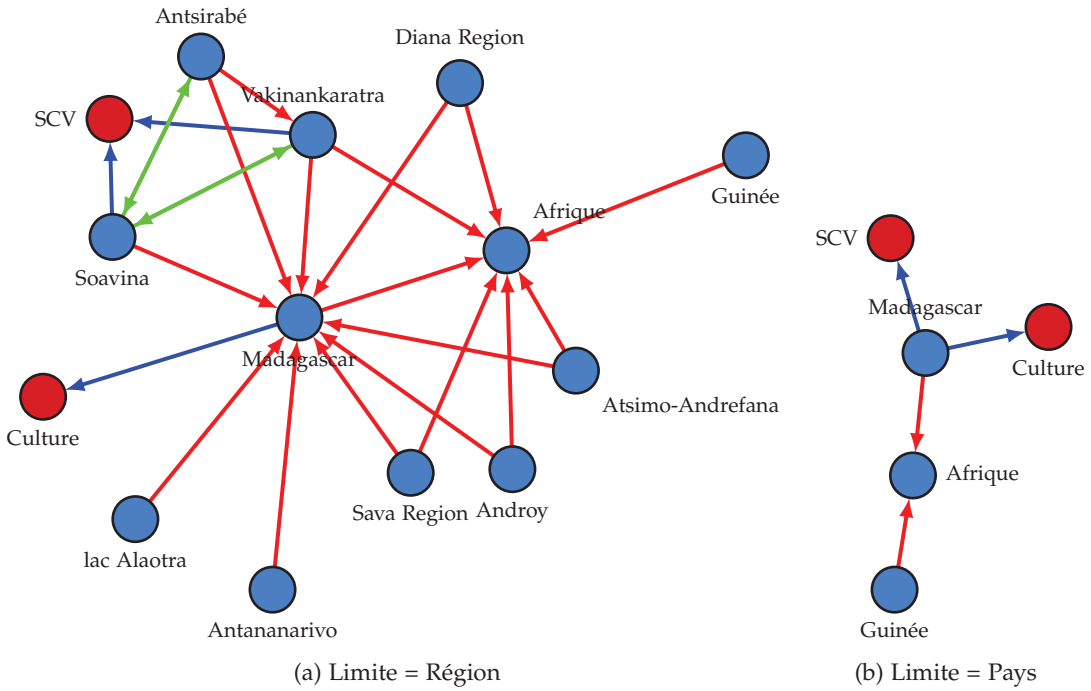
relations thématiques. Dans le cadre de l'intégration de l'information thématique dans la forme généralisée de la STR, les relations thématiques d'une entité spatiale sont propagées à l'entité spatiale remplaçante. Par exemple, l'entité spatiale *Madagascar* hérite des relations thématiques de l'entité *Vakinankaratra*. L'algorithme de propagation est défini dans l'Algorithme 7.

Algorithme 7 : Propagation de l'information thématique sur la forme généralisée de la STR

Input :
- STR_{the} STR avec des relations thématiques
- STR_{gen} STR généralisée

foreach $(es_i, et_i) \in STR_{the}.relationsThematiques$ **do**
  if $es_i \in STR_{gen}$ **then**
    append (es_i, et_i) **into** STR_{gen}
  else
    foreach $es_j \in STR_{gen}$ **do**
      if $es_j.inclus(es_i)$ **then**
        append (es_j, et_i) **into** STR_{gen}

return STR_gen



(a) Limite = Région (b) Limite = Pays

FIGURE 56 – Propogation de l'information thématique sur la forme généralisée de la STR

Dans ce chapitre, nous avons présenté la formalisation et le processus d'extraction de la relation thématique entre entité spatiale et

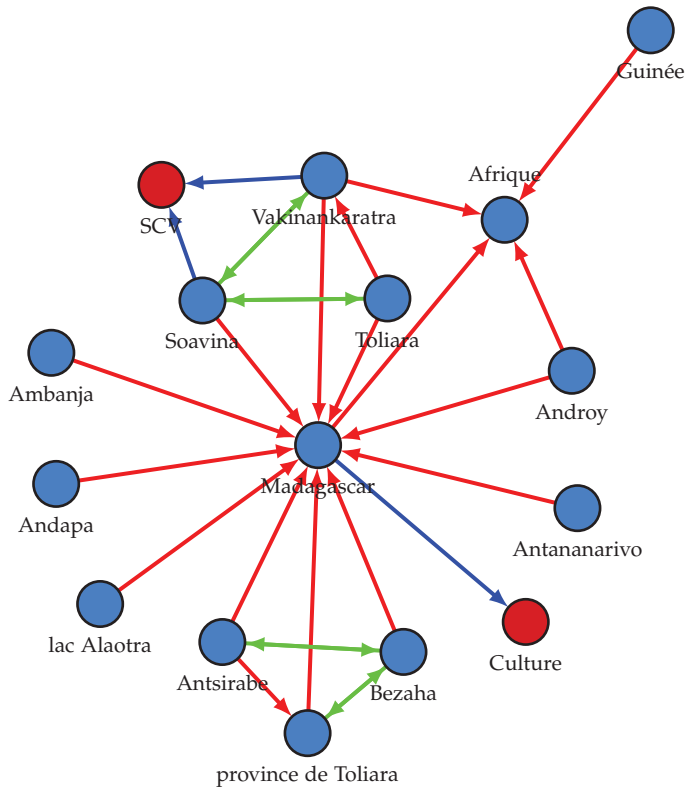


FIGURE 57 – Propagation de l'information thématique sur la forme étendue de la STR.

thématique, ainsi que son intégration dans la structure de la STR. Dans le chapitre suivant, nous présentons les résultats obtenus de la couverture des différents vocabulaires produits et de l'impact des relations thématiques dans la mise en correspondance spatiale.

2 RÉSULTATS ET DISCUSSIONS

Dans le chapitre précédent, nous avons défini la relation thématique, son processus d'extraction, ainsi que son intégration dans la STR. Dans ce chapitre, nous présentons les résultats obtenus. Dans la section 2.1, nous exposons les résultats de couvertures des vocabulaires dans les documents, puis dans les STR des corpus étudiés (*c.f.* Chapitre 5.1). Dans la seconde partie, nous montrons les résultats obtenus sur la mise en correspondance spatiale en prenant en compte des combinaisons comprenant les nouvelles formes de la STR intégrant la dimension thématique.

2.1 COUVERTURE DES TERMINOLOGIES

Dans cette section, nous présentons les résultats de couverture de l'information thématique sur les différents corpus étudiés. Dans un premier temps, les résultats de couverture des vocabulaires proposés sur le texte uniquement sont présentés. Puis, nous présentons les résultats de couverture des entités thématiques dans les STRs à l'issue du processus d'extraction de relation thématique.

2.1.1 *Couverture sur les données textuelles des corpus étudiées*

Dans cette partie, nous étudions les couvertures des différents vocabulaires sur les différents corpus (*PadiWeb* et *AgroMada*). Cela permet d'évaluer si un vocabulaire couvre bien les thématiques présentes dans l'ensemble d'un corpus et avec quelle intensité. Pour cela, nous calculons trois variables :

1. **Pourcentage de documents couverts.** Le nombre de documents dans un corpus incluant une ou plusieurs entités thématiques provenant d'un vocabulaire V.
2. **Variété de termes utilisés.** Le nombre moyen d'entités thématiques différentes identifiées dans un corpus et appartenant à un vocabulaire V.
3. **Nb. de termes détectés.** Nombre d'entités thématiques détectées par document.

Les Table 25 et 26 indiquent les différentes valeurs de couvertures obtenues des vocabulaires sur les corpus de *PadiWeb* et *AgroMada*.

Dans les résultats, nous observons que les vocabulaires de corpus offrent une meilleure couverture (99% des documents dans *AgroMada*

Vocabulaire	Langue	% de document	Variété de termes utilisés (Moy.)	Nb. de termes détectés (Avg.)
BiotexPadi500	en	**94**	12.64	**749.0**
BiotexPadi35k	en	93	**18.02**	418.0
Maladie Infect.	en	59	0.68	32.0

Table 25 – Statistiques de couverture des différents vocabulaires choisis pour le corpus *PadiWeb*.

Vocabulaire	Langue	% de document	Variété de termes utilisés (Moy.)	Nb. de termes detectés (Avg.)
Biotex AgroMada	fr	**99**	39.12	769.0
BiotexLDA AgroMada	fr	**99**	**74.16**	**894.0**
Dev. Durable	fr	73	2.19	87.0
AgroEcoINRA	fr	74	70	23.0
Biotex AgroMada	en	**98**	13.14	346.0
Biotex AgroMada	en	**98**	**46.43**	**542.0**
Dev. Durable	en	61	4.12	104.0
AgroEcoINRA	en	55	1.27	16.0

Table 26 – Statistiques de couverture des différents vocabulaires choisis pour le corpus *AgroMada*.

et 94% dans *PadiWeb*). De plus, ce sont ces vocabulaires qui permettent d'extraire un plus grand nombre de termes variés.

Les vocabulaires avec une "meilleure couverture", i.e. combinant nombre de documents, termes et leur variété, le vocabulaire issu du mélange deBiotex et de la LDA obtient de meilleures performances, plus particulièrement en termes de variété (74% pour *BiotexLDA AgroMada* contre 39% pour *Biotex AgroMada*).

2.1.2 *Couverture sur les STRs intégrant la thématique*

Dans cette partie, nous présentons les résultats obtenus de couverture des différents vocabulaires sur les STRs intégrant la dimension thématique (entité et relation thématique). Pour cela, nous utilisons des mesures présentées précédemment (% de documents, nombre de termes détectés) et la taille moyenne des STRs (sans entités thématiques).

Les Tables 27 et 28 indiquent les statistiques obtenues sur les différents corpus.

Vocabulaire	% de document	Nb. de termes détectés (Avg.)	Taille de la STR (Moy.)
BiotexPadi35k	71	1.86	6.77
BiotexPadi500	74	2.09	6.56
Maladie Infect.	21	0.23	5.59

TABLE 27 – Statistiques de couverture obtenues sur les STRs extraites des documents de *PadiWeb*.

Vocabulaire	% de document	Nb. de termes détectés (Avg.)	Taille de la STR (Moy.)
Biotex AgroMada	98	43.24	94.20
BiotexLDA AgroMada	84	6.31	41.25
Dev Dur.	80	4.00	59.96
AgroEcoINRA	73	2.20	58.00

TABLE 28 – Statistiques de couvertures obtenues sur les STRs extraites des documents d'*AgroMada*.

De manière générale, nous observons une différence significative entre les vocabulaires de corpus et de domaine selon la couverture des entités thématiques sur l'ensemble des STRs générées sur un corpus. Dans *PadiWeb*, le vocabulaire de *Maladie Infect.* couvre 21% contre les 71 et 74 % des vocabulaires *BiotexPadi35k* et *BiotexPadi500*. Dans *AgroMada*, la différence entre les deux types de vocabulaires est présente mais plus faible (4% entre *Dev. Durable* et *BiotexLDA AgroMada*). Enfin, dans *AgroMada*, si la couverture entre *Biotex AgroMada* et *BiotexLDA AgroMada* est similaire, le nombre d'entités thématiques ajoutées est plus faible (-600%).

Dans la section suivante, nous présentons l'impact de l'intégration de la thématique dans la mise en correspondance spatiale.

2.2 MISE EN CORRESPONDANCE SPATIALE

Dans les Tables 29 et 30, nous indiquons les combinaisons dominantes selon le type de STR et la mesure de similarité. À l'inverse des résultats présentés dans la section 5.4, les valeurs indiquent la différence de MAP@n entre la combinaison sans et avec relations thématiques.

AGROMADA. Dans les résultats retournés sur le corpus de *AgroMada*, les combinaisons dominantes sur l'ensemble de l'échantillon (aucun filtre) et l'ensemble de critères AC (All Criteria) (*c.f.* Section 5.3.3)

restent inchangées. Dans les résultats de MAP@1 (Voir Table 29a), l'approche *ClassicBOW* est remplacée par *BagOfCliques* sur les STRs avec un nombre d'entités spatiales $|ES| > 46$, avec une augmentation moyenne de 4% de l'ensemble des critères. Le même phénomène est observable selon l'ensemble de critères SRP (Spatial Relatedness Privileged). Enfin, seuls les résultats des STRs avec un nombre d'entités spatiales $|ES| > 72$ sont améliorés sur la base des performances de MAP@5, avec l'utilisation d'une forme normale de la STR avec des entités thématiques provenant du Vocabulaire du Développement Durable.

PADIWEB. Tout comme les résultats sur *AgroMada*, les mesures *structure-based* comme MCS permettent d'obtenir de meilleures performances sur l'ensemble des documents. À l'inverse d'*AgroMada*, l'intégration de la thématique permet d'augmenter les performances sur les différentes configurations (ensemble de critères et nombre d'entités spatiales) en s'appuyant sur la MAP@1 (voir Table 30a).

De manière générale, l'intégration de relation thématique dans la STR a tendance à améliorer la mise en correspondance spatiale de document. En s'appuyant sur les résultats concernant l'ensemble des STR des différents échantillons (nombre d'entités spatiales $|ES| > 0$), la mesure *MostCommon Subgraph* (MCS) permet d'obtenir de meilleures performances. L'intégration de la thématique dans *AgroMada* bénéficie aux STR de tailles importantes. Inversement, dans *PadiWeb* la quasi totalité des combinaisons dominantes selon la MAP@1 intègrent STR avec des relations thématiques. Concernant les vocabulaires utilisés, l'utilisation de vocabulaire de corpus est majoritaire dans les résultats s'appuyant sur la MAP@1. Inversement les résultats sur la MAP@5 indiquent un gain apporté par les vocabulaires de domaines (*Maladie Infect.* et *Dev. Durable*). Enfin, l'ensemble des combinaisons dominantes utilisant une forme de la STR avec des relations thématiques bénéficient d'une augmentation systématique du critère EM (relation de proximité spatiale inter-STR). Par conséquent, l'intégration des relations thématiques dans la STR permet de mettre en correspondance des documents avec des entités distinctes partageant une proximité spatiale.

2.3 DISCUSSIONS

2.3.1 *Généralisation et extension des entités thématiques*

Dans nos travaux, l'intégration de la thématique dans la forme généralisée et étendue de la STR se concentre sur le transfert de relation depuis la forme normale. Dans les perspectives, l'utilisation des relations sémantiques entre les entités thématiques pourrait être

Mesure	Filtre	Forme de la STR	ESP	EP	ESS	ESSC	PM	PEP	Ensemble de critères
MCS	Aucun	Généralisée (Région)	0,0	0,0	0,0	0,0	0,0	0,0	AC*
WLK***	Aucun	Généralisée (Région) + Biotex AgroMada	-0,08	+0,08	-0,19	-0,02	0,0	-0,03	SRP **
BagOfCliques	$\|ES\| > 46$	Biotex AgroMada	+0,06	+0,04	+0,08	0,0	+0,03	+0,02	AC
DeepWalk	$\|ES\| > 46$	Étendue (n=1) + Biotex AgroMada	+0,03	+0,04	0,0	-0,40	+0,03	-0,34	SRP
ClassicBOW	$\|ES\| > 72$	text	0,0	0,0	0,0	0,0	0,0	0,0	AC
ClassicBOW	$\|ES\| > 72$	text	0,0	0,0	0,0	0,0	0,0	0,0	SRP

(a) MAP@1

Mesure	Filtre	Forme de la STR	ESP	EP	ESS	ESSC	PM	PEP	Ensemble de critères
MCS	Aucun	Normale	0,0	0,0	0,0	0,0	0,0	0,0	AC
WLK	Aucun	Généralisée (Pays)	0,0	0,0	0,0	0,0	0,0	0,0	SRP
BagOfCliques	$\|ES\| > 46$	Normale	0,0	0,0	0,0	0,0	0,0	0,0	AC
BagOfCliques	$\|ES\| > 46$	Étendue (n=1)	0,0	0,0	0,0	0,0	0,0	0,0	SRP
BOSE	$\|ES\| > 72$	Étendue (n=1) + BiotexLDA	-0,05	0,0	-0,16	0,0	-0,31	+0,09	AC
DeepWalk	$\|ES\| > 72$	Normale + Dév. Durable	+0,06	+0,08	+0,04	+0,02	-0,02	-0,01	SRP

(b) MAP@5

* All Criteria
** Spatial Relatedness Privileged
*** WLK = Weisfeiler-Lehman Subtree Kernel (Shervashidze et al., 2011) ; BP = Bipartite Graph Matching (Riesen et Horst Bunke, 2009)

Table 29 – Les combinaisons (type de STR, mesure de similarité) les plus performantes selon différentes pondérations de critères (*AgroMada*).

utilisée dans le transfert des relations thématiques entre la forme normale et transformée de la STR. Dans le cadre de la généralisation, une entité thématique serait remplacée par son hyperonyme.

2.3.2 *Efficacité des différents types de vocabulaires*

Dans notre contribution, différents types de vocabulaires permettant d'identifier les entités thématiques dans un document sont utilisés : les vocabulaires de corpus et de domaine. Dans les résultats précédents, les vocabulaires de corpus apparaissent majoritairement dans les combinaisons dominantes intégrant une forme de la STR thématique.

Mesure	Filtre	Forme de la STR	ESP	EP	ESS	ESSC	PM	PEP	Ensemble de critères
MCS	Aucun	Généralisée (Région)+ BiotexPadi35k	+0,03	+0,09	-0,01	0,0	+0,06	+0,04	AC*
MCS	Aucun	Généralisée (Région)+ BiotexPadi35k	+0,5	-0,4	+0,4	0,0	+0,04	+0,03	SRP**
MCS	$\|ES\| > 4$	Généralisée (Région)+ BiotexPadi35k	-0,02	+0,08	-0,01	0,0	+0,05	-0,03	AC
MCS	$\|ES\| > 4$	Généralisée (Région)+ BiotexPadi35k	+0,06	+0,06	+0,05	0,0	+0,03	+0,05	SRP
Jaccard	$\|ES\| > 7$	Étendue (n=1) + BiotexPadi35k	+0,02	+0,02	+0,08	0,0	+0,02	+0,05	AC
PolyIntersect	$\|ES\| > 7$	Généralisée (Pays)	0,0	0,0	0,0	0,0	0,0	0,0	SRP

(a) MAP@1

Mesure	Filtre	Forme de la STR	ESP	EP	ESS	ESSC	PM	PEP	Ensemble de critères
MCS	Aucun	Maladie Infect.	0,0	0,0	0,0	0,0	0,0	0,0	AC
WLK***	Aucun	Généralisée (Pays)+ BiotexPadi500	-0,53	+0,32	-0,26	-0,06	+0,06	-0,35	SRP
MCS	$\|ES\| > 4$	Maladie Infect.	0,0	0,0	-0,01	0,0	-0,01	0	AC
PolyIntersect	$\|ES\| > 4$	Étendue (n=1)	0,0	0,0	0,0	0,0	0,0	0,0	SRP
MCS	$\|ES\| > 7$	Étendue (n=1)	0,0	0,0	0,0	0,0	0,0	0,0	AC
MCS	$\|ES\| > 7$	Étendue (n=1) + BiotexPadi35k	+0,07	+0,12	-0,11	-0,02	0,0	+0,13	SRP

(b) MAP@5

* All Criteria
** Spatial Relatedness Privileged
*** WLK = Weisfeiler-Lehman Subtree Kernel (Shervashidze et al., 2011) ; BP = Bipartite Graph Matching (Riesen et Horst Bunke, 2009)

Table 30 – Les combinaisons (type de STR, mesure de similarité) les plus performantes selon différentes pondérations de critères (*PadiWeb*).

2.3.3 *Le critère EP : marqueur de l'efficacité de l'intégration thématique*

Dans l'évaluation des correspondances, le critère EP indique le pourcentage de documents correspondants dans lesquels des entités partagent un relation de proximité. L'intérêt de l'intégration de la thématique reposant sur l'induction de proximité spatiale entre des entités distinctes de deux STR, le critère EP permet de quantifier l'efficacité et l'existence de ce phénomène. Tel que nous l'avons défini, EP indique si une ou plusieurs entités partagent une relation de proximité. Par conséquent, il est impossible de savoir quelle proportion de rela-

tions de proximité *inter-STR* existent. Tout comme PEP (pourcentage d'entités spatiales partagées), la définition d'un septième critère mesurant la proportion de relation de proximité entre les entités spatiales de deux documents est envisagée.

Troisième partie

CONCLUSION ET PERSPECTIVES

1 CONCLUSION

La première contribution de cette thèse repose sur un processus de mise en correspondance de données textuelles hétérogènes. Ce processus se déroule en deux étapes : la *georepresentation* et le *geomatching*. Dans la première phase, nous proposons de représenter la dimension spatiale de chaque document d'un corpus à travers une structure dédiée, la Spatial Textual Representation (STR). Cette représentation de type graphe est composée des entités spatiales identifiées dans le document, ainsi que les relations spatiales qu'elles entretiennent. Pour identifier les entités spatiales d'un document et leurs relations spatiales, nous proposons une ressource dédiée, nommée Geodict. Geodict est un index géographique rassemblant des informations spatiales (coordonnées, classe, géométrie) et linguistiques (toponyme, noms alternatifs) sur différentes sources (Geonames, Wikidata, OpenStreetMap) pour chaque entité spatiale. La seconde phase, le *geomatching*, consiste à mesurer la similarité entre les représentations générées (STR). S'appuyant sur la nature de la structure de la STR (i.e. graphe), différents algorithmes de *graph matching* ont été étudiés. Ces algorithmes sont divisés en deux catégories : *structure-based* et *pattern-based*. L'évaluation du processus de correspondance indique une dominance de la forme normale de la STR combinée aux algorithmes *structure-based*.

L'ensemble des étapes du processus de mise en correspondance évaluées sont les suivantes : l'identification de toponymes, la désambiguïsation de toponyme, ainsi que la qualité des correspondances retournées selon les différentes combinaisons. Une combinaison correspond à une forme de la STR et une mesure de similarité. Pour évaluer la pertinence d'une correspondance, nous proposons un ensemble de 6 critères s'appuyant sur une définition de la similarité spatiale entre deux documents.

La seconde contribution présentée dans ce manuscrit repose sur la dimension thématique des données textuelles et sa participation dans le processus de mise en correspondance spatiale. Nous proposons d'identifier les thèmes apparaissant dans la même fenêtre contextuelle que certaines entités spatiales. L'objectif est d'induire certaines des similarités spatiales implicites entre les documents. Pour cela, nous proposons d'étendre la structure de la STR à l'aide de deux concepts : l'*entité thématique* et de la *relation thématique*. L'entité thématique représente un concept propre à un domaine particulier (agronome, médical) et représenté selon différentes orthographes présentes dans

une ressource terminologique, ici un vocabulaire. Nous proposons d'utiliser deux types de vocabulaires. Les vocabulaires de domaine qui sont produits par des experts et les vocabulaires de corpus qui sont générés de manière automatique à l'aide du corpus. Différents vocabulaires de domaine sont sélectionnés en fonction des thématiques globales de corpus étudiés. Pour les vocabulaires de corpus, différentes méthodes d'analyse de documents (*topic modeling* et extraction automatique de mots-clés) sont utilisées. Parmi ces méthodes, nous proposons BiotexLDA, une combinaison des termes extraits par la LDA et l'identification de mots-clés par la méthode développée dans Biotex. Suivant le processus d'évaluation développé dans la première contribution, les résultats révèlent une amélioration de la qualité des correspondances dû à l'intégration des relations thématiques dans la STR.

Une fois identifiées dans un document, ces entités thématiques sont liées à une ou plusieurs entités spatiales par des relations thématiques. Une relation thématique lie une entité spatiale à une entité thématique si celles-ci apparaissent dans une même fenêtre contextuelle. Les vocabulaires choisis ainsi que la nouvelle forme de la STR intégrant la dimension thématique sont évalués selon leur couverture sur les corpus étudiés, ainsi que leurs contributions dans le processus de mise en correspondance spatiale.

2 PERSPECTIVES

2.1 ÉVOLUTION DES REPRÉSENTATIONS POUR LA MISE EN CORRESPONDANCE

Dans un contexte de mise en correspondance, la représentation d'un document est cruciale. Dans nos travaux, la représentation proposée s'appuie sur les entités spatiales puis les entités thématiques et les relations qu'elles entretiennent. Dans l'objectif d'améliorer l'estimation de la similarité spatiale entre deux documents, différentes transformations, i.e. généralisation et extension, sont appliquées aux entités spatiales. Une possibilité serait d'appliquer ces mêmes transformations aux entités thématiques en s'appuyant sur les relations d'hyperonymie, et de synonymie présentes dans des resources ontologiques.

Les travaux menés sur la mise en correspondance de données hétérogènes sont focalisés sur la dimension spatiale et thématique. La suite logique envisagée repose sur l'intégration de la dimension temporelle dans la structure de la STR. Différentes possibilités comme l'identification de dates dans un document et ses relations contextuelles avec les entités thématiques et spatiales sont envisageables. Une autre possibilité consisterait à produire différentes STR, chacune correspondant à une date (ou période) et une partie du document.

Dans notre deuxième contribution, nous exploitons l'information thématique dans le but d'améliorer la mise en correspondance spatiale. Dans l'objectif de proposer une mise en correspondance la plus variée, de nouveaux processus dédiés à la dimension thématique ou temporelle seule sont nécessaires. Comme la STR, ces processus peuvent combiner des informations de différentes dimensions pour améliorer le résultat de mise en correspondance.

Dans notre processus de mise en correspondance, l'unité de comparaison est le document. Le problème de cette approche réside dans la finesse de la mise en correspondance et n'intègre pas la dimension des structures présentes dans un document, e.g. tables, paragraphes, etc. Une évolution possible serait de comparer le document comme un ensemble de blocs, chacun défini par leur structure. En plus d'augmenter la finesse des correspondances entre les documents, cela permet d'appliquer des processus d'extraction d'informations spécifiques (e.g. geoparsing). Si dans nos travaux, nous nous concentrons sur le triptyque (théme, espace, temps), il serait envisageable de mesurer la similarité entre des blocs spécifiques. Par exemple, cela permettrait de mettre en correspondance différentes tables présentant une même observation.

2.2 VISUALISATION DES CORRESPONDANCES À L'ANALYSE DE CORPUS

Dans le contexte d'analyse de corpus de données textuelles, l'analyse des correspondances selon différentes dimensions peut s'avérer cruciale. Dans le milieu journalistique, l'extraction et la mise en correspondance d'informations sont utilisées pour l'analyse de données massives hétérogènes (Michael Hunger et William Lyon, 2016). Dans nos travaux, nous avons commencé le développement d'un outil de visualisation des correspondances entre les documents mais d'autres visualisations profitant de la multi-dimensionnalité du processus de mise en correspondance sont envisageables. Différentes approches (Cho et al., 2016 ; Ferreira et al., 2013 ; J. Li et al., 2018 ; Peuquet, 1994) proposent déjà un ensemble de visualisations de document sur une ou plusieurs dimensions. De nouvelles méthodes de visualisation des correspondances par rapport à une ou plusieurs dimensions de l'information d'un document sont envisagées.

2.3 ÉVALUATION DES CONTRIBUTIONS PAR UN COMITÉ D'EXPERTS

Plusieurs travaux sont menés pour intégrer les producteurs de données dans le processus d'évaluation de nos contributions. Dans une première contribution (**Fize, Jacques**, Roche et Teisseire, 2018 ; **Fize, Jacques**, Teisseire et Roche, 2017), nous avons proposé un système d'annotation de similarités entre les documents d'un corpus, Gemedoc. Le système permet d'annoter les similarités thématique et spatiale entre deux documents à l'aide d'une graduation selon 4 degrés. Une évolution de l'outil intégrant les critères de similarité spatiale proposés dans la thèse est envisagée. Dans une deuxième contribution, un protocole d'évaluation des vocabulaires de corpus générés selon les différentes méthodes (Biotex, LDA, LDA) auprès des producteurs de données a été mis en place avec l'aide de Jean-Philippe Tonneau, géographe au CIRAD et acteur du projet BVLAC. L'objectif est d'évaluer la pertinence et la couverture des termes par rapport aux thématiques considérées comme cruciales par les experts.

APPENDIX

A

APPENDIX

A.1 CLASSIFICATION DE GEONAMES

Dans nos travaux, nous avons produit un index géographique GEODICT dans lequel sont collectés plusieurs entités spatiales provenant de Wikidata, Geonames et OpenStreetMap. Dans chacune des sources, différentes informations spatiales sont sélectionnées. Pour classifier les entités spatiales, nous utilisons la classification offerte par GEONAMES.

GEONAMES est un index géographique qui s'appuie sur une classification structurée en 9 classes générales (*feature-class*) et 645 sous-classes (*feature-code*). La Table 31 présente un extrait des classes de catégorie A (pays, état, région).

TABLE 31 – Extrait de la classification utilisé par Geonames

Classe	**Nom**	**Description**
A-ADM1	first-order administrative division	a primary administrative division of a country, such as a state in the USA
A-ADM2	second-order administrative division	a subdivision of a first-order administrative division
A-ADM3	third-order administrative division	a subdivision of a second-order administrative division
A-ADM4	fourth-order administrative division	a subdivision of a third-order administrative division
P-PPLA	seat of a first-order administrative division	seat of a first-order administrative division (PPLC takes precedence over PPLA)
P-PPL	populated place	a city, town, village, or other agglomeration of buildings where people live and work
L-CONT	continent	continent : Europe, Africa, Asia, North America, South America, Oceania, Antarctica
A-PCLI	independent political entity	-

A.2 OUTILS DÉVELOPPÉES

A.2.1 ***strpython*** *: Librairie Python*

L'implémentation des différentes contributions sont disponible dans un module écrit en *Python 3*. La librairie **strpyton** est composé de différent sous-modules, chacun responsable d'une étape du processus de génération de la STR :

— **nlp** rassemble les différentes méthodes de *geotagging* et de *geocoding* étudiées.

— **models** contient l'ensemble des classes responsable de la modélisation de la STR et de la STR intégrant la thématique. De plus, ce sous-module contient les implémentations des transformations.

— **visualisation** contient les procédures de générations des réseaux de correspondances utilisées dans l'outil de visualisation.

La librairie est disponible à l'adresse suivante : `https://gitlab.irstea.fr/jacques.fize/str-python`.

A.2.2 *Outils de visualisation des correspondances*

Pour visualiser les correspondances selon une forme de la STR et une mesure de similarité, nous avons développé une application en *Python* s'appuyant sur la librairie Flask[1]. Cette application permet de visualiser les correspondances d'un documents ainsi que leur contenu dans une fenêtre d'aperçu. Celle-ci est disponible à l'adresse suivante : `https://gitlab.irstea.fr/jacques.fize/these_visu_matching`

A.2.3 *GMatch4py : une librairie Python pour le graph matching*

Pour pouvoir utiliser les différents algroithmes de graph matching, nous avons développé une libraire dédiée, GMactch4py. GMatch4Py est un module disponible pour le language Python et compatibleWindows, Linux, Mac OS. GMatch4py est disponible sur la plateforme GitHub sous Licence MIT à cette adresse : `https://github.com/Jacobe2169/GMatch4py`. De plus, dans l'optique de produire une librairie offrant des temps d'exécutions optimales, la majorité des algorithmes s'appuient sur Cython. Cython[2] est un compilateur statique optimisant à la fois le langage de programmation Python et le langage de programmation Cython étendu (basé sur Pyrex). L'implémentation des algorithmes en Cython permet de diviser jusqu'à 30× le temps d'exécution (voir Figure 58). Les algorithmes disponibles dans la librairie sont indiqués dans la Table 32.

1. `https://palletsprojects.com/p/flask/`
2. `https://cython.org`

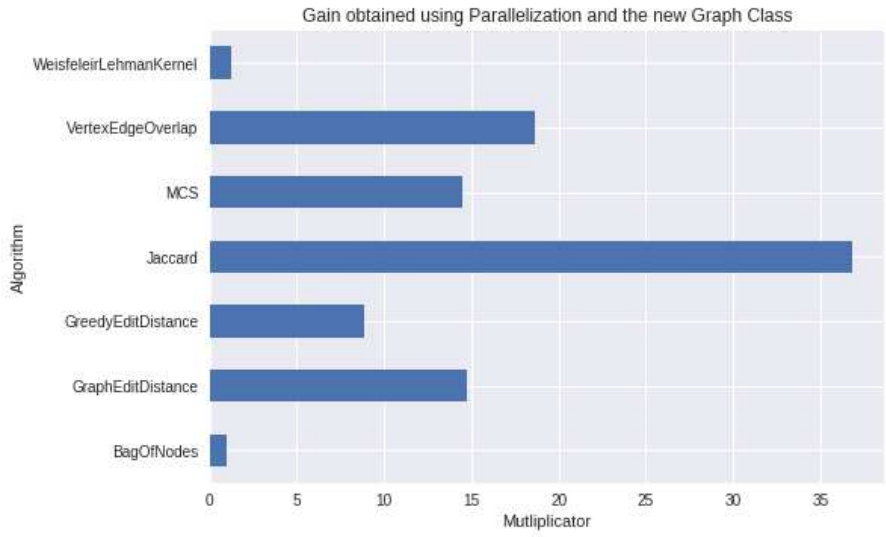


FIGURE 58 – Gains de temps obtenus sur l'implémentation Cython des algorithmes implémentés dans GMatch4py.

Catégorie	Algorithme	Sommet	Arête	Dirigé	Etiqueté	Pondéré	Topologie	Graph Emb.	Node Emb.
Pattern-Based	Graph2vec(Narayanan et al., 2017)	✓	✓		✓	✓	✓	✓	
	Node2vec(Grover et Leskovec, 2016)	✓	✓	✓	✓	✓	✓		✓
	DeepWalk(Perozzi, Al-Rfou et Skiena, 2014)	✓	✓	✓	✓	✓	✓		✓
	Shortest Path Kernel(Borgwardt et Kriegel, 2005)	✓	✓	✓	✓	✓	✓		
	WLK(Shervashidze et al., 2011)	✓	✓	✓	✓		✓		
	Bag of Cliques(**fizematching2018**)	✓	✓		✓		✓		
Structure-Based	GED - BP(Riesen et Horst Bunke, 2009)	✓	✓	✓	✓	(✓)			
	GED - HED(Riesen et Horst Bunke, 2009)	✓	✓	✓	✓				
	GED - BP2(Riesen et Horst Bunke, 2009)	✓	✓	✓	✓				
	GED - Greedy(Riesen et Horst Bunke, 2009)	✓	✓	✓	✓	(✓)			
	Vertex Ranking(Papadimitriou, Dasdan et Garcia-Molina, 2010)	✓	✓	✓	✓				
	VertexEdge Overlap(Papadimitriou, Dasdan et Garcia-Molina, 2010)	✓	✓	✓	✓				
	Maximum Common Subgraph (Horst Bunke et Shearer, 1998)	✓		✓	✓				
	Bag of Nodes	✓	✓		✓				

Table 32 – Algorithmes implémentés dans GMatch4Py

A.2.4 *Geodict*

GEODICT est un index géographique sur lequel s'appuient la majorité des contributions. La version de GEODICT utilisée dans nos travaux est disponible à cette adresse : http://geodict.cirad.fr. Pour obtenir une version personnalisée de GEODICT, c-a-d intégrant d'autres attributs pour chaque entité spatiale, le code source est disponible à cette adresse : https://gitlab.irstea.fr/jacques.fize/geodict. En plus, de la procédure de construction de GEODICT, un outil de visualisation des données de GEODICT est disponible à cette adresse : https://gitlab.irstea.fr/jacques.fize/geodict/tree/master/gui.

Listing 1 – Exemple de fichier de configuration de GEODICT

```
1 {
2     "properties_to_extract":[
3         {"id":"P47",
4         "isMultiple":true,
5         "type":"EntityID"}
6         , ...
7       ]
8 }
```

A.3 ALGORITHMES

Dans cette section, vous retrouverez le pseudo-code de différents algorithmes présentés tout au long du manuscrit.

A.3.1 *Algorithmes de désambiguïsation*

SHAREDPROP

Algorithme 8 : SharedProp

Input : **T** Liste de toponymes
Output : **R** Liste des entités spatiales associées aux toponymes dans T

```
F ← getEntitesFixes(T)
A ← getEntitesAmbigues(T)
for top ∈ A do
    R[T] ← arg max_{c∈candidats(top)} score(c, F)
Function score(c,F) :
    score ← 0
    for es ∈ F do
        if estAdjacent(c,es) then
            score ← score + 2
        score ← score + inclusionScore(c, es)
    return score
Function scoreInclusion(es1,es2) :
    score ← 0
    for p ∈ {P131, P706} do
        inclusion_es1 ← inclusionChain(es1, p)
        inclusion_es2 ← inclusionChain(es2, p)
        score ← score + log(|inclusion_es1 ∩ inclusion_es2|)
    return score
```

WIKICOOC

Algorithme 9 : WikiCooc

Input : **T** Liste de toponymes dans un document
Output : **R** Liste des entités spatiales associées aux toponymes dans T

```
for top ∈ T do
    c ← getCandidats(top)
    append c into listeCandidats
    candidatParToponyme[top] ← c
G ← nouveauGraphe()
for c1 ∈ listeCandidats do
    for c2 ∈ listeCandidats do
        if c1 ≠ c2 or c2 ∉ groupeDeCandidat(c1) then
            ajouterArete(G, c1, c2, coocScore(c1, c2))
for top ∈ T do
    R[top] ← arg max_{c∈candidatParToponyme} degree(G, c)
return R
```

A.3.2 *Identification de relations spatiales*

ADJACENCE

Algorithme 10 : Test d'adjacence

Input : $\mathbf{es_1}, \mathbf{es_2}$ Deux entités spatiales
if $es_1 \in es_2$ ***or*** $es_2 \in es_1$ **then**
  return false
if $es_1 \in es_2$.P47 ***or*** $es_2 \in es_1$.P47 **then**
  return true
if es_1.geometryType $\in$ {Polygon, MultiPolygon} **then**
  return isIntersection(es_1.geometry, es2.geometry)
else if distEuclidienne(es_1.coord, es2.coord) < 1 **then**
  return true
return false

INCLUSION

Algorithme 11 : Test d'inclusion

Input : $\mathbf{es_1}, \mathbf{es_2}$ Deux entités spatiales
$ESQuiInclusEs_1 \leftarrow getData(es_1)$.P131
append $getData(es_1)$.P706 **into** $ESQuiInclusEs_1$
if $es_2 \in ESQuiInclusEs_1$ **then**
  return true
return false

A.3.3 *Transformation de la STR*

GÉNÉRALISATION

Algorithme 12 : Généralisation d'une STR

Input :
- E Entités spatiales dans un texte
- Limite de la généralisation b $\in$ {région, pays}

transformMap $\leftarrow \{\}$
foreach es $\in$ E **do**
  $parent_{e_s} \leftarrow$ es
  f $\leftarrow$ true
  while hasParent($parent_{e_s}$) ***or*** f **do**
    $parent_{e_s} \leftarrow$ récupérerEntitéParente($parent_{e_s}$)
    if $parent_{e_s}$.classe == b **then**
      transformMap[es] = $parent_{e_s}$
      f $\leftarrow$ false
E $\leftarrow$ remplacer(E, transformMap)
return G

EXTENSION

Algorithme 13 : Extension d'une STR

Input :
- **E** Entités spatiales présentes dans un document.
- **n** Nombre d'entités ajoutées à chaque extension.
- **r** Le rayon exprimé en **unit** $\in$ m, km

```
scoreES ← []
for es ∈ E do
    append getScore(es) into scoreES
scoreMedian ← median(scoreES)
selectedEs ← []
for i ∈ 0...|E| do
    if scoreES[i] < scoreMedian then
        append E[i] into selectedES
nouvellesEntité ← []
for es ∈ selectedES do
    append récupérerEntitésProches(n, scoreMedian, r, unit)
     into nouvellesEntité
append entitiesToAdd into E
return E
```

A.4 ENQUÊTE AUPRÈS DES EXPERTS

Projet BVLAC : Que nous disent les données ?

Introduction

Dans le cadre du projet **SONGES**[1] (**S**cience des d**ON**nées hétéro**G**èn**ES**), nous travaillons sur un ensemble de données produites au sein du projet BVLAC. Dans ces données, nous nous intéresserons à deux corpus[2] :

1. **Corpus BVLAC**. Ce corpus est composé de ~15000 documents de natures et de sources variées. On y retrouve des mémoires, des rapports d'agents de terrain, des manuels techniques, des thèses, des articles, des relevés, etc.
2. **Rapport BVLAC.** C'est un rapport publié par l'Agence française de Développement en vue de l'organisation atelier de restitution et d'échanges sur l'évaluation rétrospective du « Projet de mise en valeur et de protection des bassins versants du lac Alaotra « BV LAC II »

Dans cet exercice, nous vous proposons **d'évaluer la qualité** des résultats des **méthodes d'extraction de mots-clés** sur ces deux corpus. Pour cela, nous faisons appel à vos connaissances concernant le Projet BVLAC, pour confirmer si ces méthodes arrivent bien à récupérer l'ensemble des concepts généraux lié à ce projet.

À la fin de cet exercice, les informations récoltées seront analysées et un compte rendu vous sera délivré.

[1] http://textmining.biz/Projects/Songes/
[2] Ensemble de documents

A.5 ENQUÊTE AUPRÈS DES EXPERTS

Formulaire

Nom :
Prénom :
Profession :

Étape 1

Proposez une liste de 4 à 8 mots-clés représentant les concepts généraux autour du projet BVLAC et les termes attendus.

Exemple
Dans cet exemple, nous nous appuierons sur 3 concepts généraux : *Délit, Foncier et Acteur*

Liste de concepts
1. Délit
2. Foncier
3. Acteur
…

À remplir

Liste de concepts
1.
2.
3.
4.
5.
6.
7.
8.

A.6 ENQUÊTE AUPRÈS DES EXPERTS

Étape 2

Indiquez les termes associés à chaque concept. Pour chaque corpus et chaque méthode, une liste des termes extraits est donnée (pages 7 à 10). Pour chaque combinaison, donnez les termes associés à chacun des concepts proposés dans la première étape.

Exemple

Dans cet exemple, nous utiliserons la liste de termes suivantes : *actes de vandalisme, exploitants, exploitations, administrateur de projet, agents , fonciers, agriculteurs, data-mining*.

Liste de concepts	Termes associés
1. Délit	acte de vandalisme
2. Foncier	exploitations, agents fonciers
3. Acteur	exploitants, agriculteurs
…	…

À remplir

Corpus BVLAC - BIOTEX

Liste de concepts	Termes associés
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

Corpus BVLAC - LDA

Liste de concepts	Termes associés
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

A.7 ENQUÊTE AUPRÈS DES EXPERTS

Rapport AFD - BIOTEX

Liste de concepts	Termes associés
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

Rapport AFD - LDA

Liste de concepts	Termes associés
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

A.8 ENQUÊTE AUPRÈS DES EXPERTS

Étape 3

Indiquer les termes pertinents qui n'appartiennent pas aux concepts proposés. Pour chaque corpus et chaque méthode, indiquez les termes qui n'appartiennent pas aux concepts proposés mais qui vous semble pertinents. Si possible, regroupez ces termes dans de nouveaux concepts.

Exemple.

Liste de concepts	Termes associés
1. Informatique	Data-mining
2.	
3.	

À remplir

Corpus BVLAC - BIOTEX

Liste de concepts	Termes associés
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

Corpus BVLAC - LDA

Liste de concepts	Termes associés
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

A.9 ENQUÊTE AUPRÈS DES EXPERTS

Rapport AFD - BIOTEX

Liste de concepts	Termes associés
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

Rapport AFD - LDA

Liste de concepts	Termes associés
1.	
2.	
3.	
4.	
5.	
6.	
7.	
8.	

A.10 ENQUÊTE AUPRÈS DES EXPERTS

Liste de termes extraits sur les différents corpus

Corpus BVLAC

Liste de termes - Corpus BVLAC - BIOTEX		
actes de vandalisme	*édition de rapport*	*rapport de stage*
activité biologique	*enquêtes foncières*	*rapport final*
activités prévues	*exigences de poste*	*rapport foncier*
administrateur de projet	*exploitants*	*rapport trimestriel*
agents fonciers	*exploitations*	*réalisations de saison*
agriculteurs	*façons culturales*	*recensement parcellaire*
amin	*fokontany*	*région*
amin'ny	*fond de carte*	*rendement*
application de champignon	*gestion foncière*	*renouvellement de traitement*
appui	*information mutualisé*	*repiquage*
appui agro-économique	*jour de pluie*	*réseau*
appui individuel	*journée de réflexion*	*réseau hydrographique*
art	*jours de pluie*	*réunion*
article	*lac*	*réunion de travail*
assemblée générale	*lahy sy*	*réunion mensuelle*
assemblée paysanne	*limite communale*	*réunions de bilan*
associations de crédit	*membres*	*riz*
ateliers de production	*membres de bureau*	*riz pluvial*
base de donnée	*mise en page*	*riziculture*
cadre logique	*mise en terre*	*rapport de campagne*
campagne agricole	*mise en valeur*	*saison*
campagne contre saison	*mission*	*saison pluviale*
carte	*mission de cadrage*	*scénario de base*
carte de situation	*mois de mars*	*secteur*
cellule	*mois de novembre*	*sécurisation foncière*
cellule de projet	*nombre de missions*	*services*
changement de règles	*opérateur contracté*	*situation de payement*
chef de division	*opérateurs*	*sol*
chefs de blocs	*organisations professionnelles*	*stage foncier*
comité local	*outils*	*station de pompage*
commission de reconnaissance	*pomme de terre*	*stratégie de services*
compte spécial	*page*	*surface*
conseil	*parc à bois*	*systèmes*
contrat	*parcelle*	*tableaux de bords*
contrat de mission	*parcelles entomopathogènes*	*technicien supérieur*
convention tripartite	*paysans*	*terroirs*
coût total	*personne responsable*	*texture*
couverture morte	*pièces de rechange*	*thèmes de formation*
crédit	*plan de développement*	*trame foncière*
crédit solidaire	*plants*	*type*
culture	*politique foncière*	*type de forêt*
dépenses courantes	*président*	*unité pastorale*
désherbage chimique	*production*	*usages agricoles*
développement régional	*produits ligneux*	
diagnostic foncier	*programme*	*variétés*
données mutualisée	*projet*	*végétation moyenne*
dossiers de crédits	*projet de mise*	*versants*
eau en nature	*qualité supérieure*	*visite de terroir*

A.11 ENQUÊTE AUPRÈS DES EXPERTS

Liste de termes - Corpus BVLAC - LDA		
accusé de réception	fonctionnement	pomme de terre
acte de naissance	fonctions de commercialisation	prise tertiaire
acte de vente	formation	production
activités	frais	production agricole
activités trimestriel	gestion	production de semences
Alaotra	gestion foncière	produits agricoles
amin'ny	guichet foncier	projet
appui	irrigation	rapport
araka ny	itinéraires techniques	rapport de mission
article	izany	règlements en vigueur
ary ny	Jehovah	rehetra
assistance technique	journée de travail	réseaux hydroagricoles
autres	jours	ressources naturelles
banque de données	jours après repiquage	riz pluvial
base de données	koa ny	saison sèche
bases de données	madagascar	santé animale
bassins versants	marge brute	sarclage chimique
campagne agricole	membres de bureau	sarclage manuel
capitalisation en matière	mesuré	sarclage mécanique
cas de force	mettre en place	Secteur
cas échéant	mikasika ny	sécurisation foncière
case de passage	mise à jour	semis direct
conseil de gestion	mise en demeure	sondage de rendement
couverture végétale	mise en meule	station de pompage
date de réception	mise en place	stations de pompage
développement rural	mise en valeur	sy ny
dispositions	montant	système de culture
domaine public	navigué	système de production
données pluviométriques	niveau	systèmes de culture
droit de propriété	œuvre	tableau suivant
élevage	œuvre déléguée	temps partiel
engrais chimique	ordre de service	total
engrais organique	organisations paysannes	toy ny
entrée en vigueur	ouvrage de prise	tranche conditionnelle
entretien	palmier à huile	tranche ferme
études	parcelles	transfert de gérance
exploitants agricoles	paysans	travaux
exploitation	plans fonciers	travaux de recherches
exploitation agricole	plantes de couverture	travaux de réhabilitation
exploitation rizicole	point de vue	tsy maintsy
exploitations agricoles	pois de terre	tsy misy
FIFABE	valeur	unité de production

A.12 ENQUÊTE AUPRÈS DES EXPERTS

RAPPORT BVLAC

Liste de termes - Rapport BVLAC - BIOTEX		
absence de cadre	*cycle de projet*	*millions d'euros*
actes de vente	*défis majeurs*	*mise en vigueur*
agents	*délais de réponse*	*mission de terrain*
agents de terrain	*délivrance*	*modes opératoires*
agriculture de conservation	*demande en dépit*	*niveau local*
aménagements hydro-agricoles	*démarche*	*niveaux de productivité*
animaux	*départements ministériels*	*niveaux de réalisation*
approbation	*développement rural*	*objet*
approche	*différentiel en termes*	*offre de service*
approche de conseil	*diffusion de pratiques*	*pages*
approche holistique	*dimension institutionnelle*	*papam*
appui-conseil	*directeur*	*paradigme*
axes	*dispositif de conseil*	*paysage institutionnel*
axes stratégiques	*diversité*	*phase précédente*
bassins	*documents annuels*	*plans de développement*
bassins de production	*documents de politique*	*populations rurales*
besoin de temps	*dolique*	*programme*
besoins en appui	*droit de propriété*	*programme national*
budget initial	*durabilité*	*programmes de travail*
cadre de résultats	*dynamiques de structuration*	*projet*
cadre logique	*eau avec professionnalisme*	*projet de mise*
capacité de délivrance	*échelle*	*projets fonciers*
capacités de maîtrise	*ensemble de prestataires*	*propriété foncière*
capacités de service	*exploitants agricoles*	*raison*
carte	*exploitations agricoles*	*rapport annuel*
caution solidaire	*exploitations de références*	*rapports annuels*
cellule technique	*financement rural*	*réalités*
changement de culture	*fonctions*	*reboisements*
changement de pratiques	*formulation de projet*	*recommandation*
changements à opérer	*gains de productivité*	*référentiels*
chantiers de réflexion	*gestion durable*	*ressources naturelles*
circonscription	*gestion foncière*	*résultats décevants*
clé de voute	*guichet*	*résultats financiers*
culture	*guichet foncier*	*riz*
com	*impact*	*riz pluvial*
communes	*initiatives individuelles*	*sécheresse*
communes supplémentaires	*insécurité alimentaire*	*secteur financier*
composante	*institutions existantes*	*sécurisation massive*
conseil agricole	*intensification*	*sécurité alimentaire*
conservation	*interventions à caractère*	*service régional*
consultant	*interventions de mise*	*services professionnels*
corps de métiers	*jeunes*	*sigiste*
cours de phase	*micro-crédit*	*sources de revenus*
coût total	*lettre de politique*	*structure pérenne*
couverture sanitaire	*lien entre sécurisation*	*structure régionale*
crise politique	*logique de faire-faire*	*taux de change*
cultures de développement	*maïs*	*techniciens de terrain*
tranche ferme	*mesures correctives*	*techniques*

9

A.13 ENQUÊTE AUPRÈS DES EXPERTS

Liste de termes - Rapport BVLAC - LDA		
accès	ensemble	phase de préparation
activité	évaluation	plan technique
activités agricoles	ex-post	plans topographiques
Agriculture	fiche de performance	plantes de couverture
Ambatondrazaka	fiches techniques	politique
amélioration durable	fin de contrat	politique nationale
aménagement	fiscalité foncière	prise en charge
approche	fonctions de gestion	processus de réforme
aujourd'hui	guichet foncier	producteurs
autonomie financière	hectare de rizière	production agricole
base de vesce	institutions de microfinance	production de semences
bénéficiaires finaux	intervention	production rizicole
cadre	itinéraires techniques	programme
cadre de résultat	lettre de politique	programme national
cadre logique	logique économique	projet
capacités de service	Madagascar	projets de développement
cellule de projet	matériel végétal	provisoire
cellule légère	membres	rapport
cellule régionale	ménages agricoles	recherche appliquée
cette composante	mesure	redevance en nature
collectivités locales	mise à jour	réforme foncière
communautés locales	mise en valeur	renforcement de capacités
compétences formées	modalités de mise	ressources naturelles
conditions	moyenne par rapport	résultats
conditions de durabilité	national	résultats financiers
conditions de sortie	nature qualitative	résultats significatifs
conditions de vie	niveau	résultats techniques
conseiller agricole	niveau de dégradation	revenus agricoles
contrat	niveaux de réalisation	riz pluvial
contrat de partenariat	nombre	sécurité alimentaire
cours de projet	nombre de demandes	sein de filières
curricula de formation	objectif global	semence de vesce
davantage	objectifs spécifiques	services
délivrés	œuvre	situations de référence
développement	œuvre déléguée	sources de revenus
développement agricole	offre de service	stabilisation de lavaka
développement rural	opérateurs	stratégie de sortie
différentiel de précision	organisations professionnelles	taux de recouvrement
difficultés de gestion	outil de type	techniciens de terrain
dispositif de conseil	ouvrage	techniques agro-écologiques
dispositif de mise	pages	temps en temps
dispositif de suivi	particulier	territoire communal
dispositif national	partie intégrante	terroirs de concentration
dispositifs de proximité		titres fonciers
engagement financier		

A.14 ENQUÊTE AUPRÈS DES EXPERTS

Commentaire(s) :

BIBLIOGRAPHIE

Adams, Benjamin (2018). « From spatial representation to processes, relational networks, and thematic roles in geographic information retrieval ». en. In : *Proceedings of the 12th Workshop on Geographic Information Retrieval - GIR'18*. Seattle, WA, USA : ACM Press, p. 1–2.

Agichtein, Eugene et Luis Gravano (2000). « *Snowball* : extracting relations from large plain-text collections ». en. In : *Proceedings of the fifth ACM conference on Digital libraries - DL '00*. San Antonio, Texas, United States : ACM Press, p. 85–94.

Agnew, John A, Katharyne Mitchell et Gerard Toal (2008). *A companion to political geography*. John Wiley & Sons.

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter et Roland Vollgraf (2019). « FLAIR : An Easy-to-Use Framework for State-of-the-Art NLP ». en-us. In : p. 54–59.

Alghamdi, Rubayyi et Khalid Alfalqi (2015). « A survey of topic modeling in text mining ». In : *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6.1.

Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez et Krys Kochut (2017). « Text Summarization Techniques : A Brief Survey ». In : *arXiv :1707.02268 [cs]*. arXiv : 1707.02268.

Al-Rfou, Rami, Vivek Kulkarni, Bryan Perozzi et Steven Skiena (2015). « POLYGLOT-NER : Massive Multilingual Named Entity Recognition ». In : *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, p. 586–594.

Amitay, Einat, Nadav Har'El, Ron Sivan et Aya Soffer (2004). « Web-a-Where : Geotagging Web Content ». In : *Sigir '04*, p. 273–280.

Aprosio, Alessio Palmero, Claudio Giuliano et Alberto Lavelli (2013). « Extending the Coverage of DBpedia Properties using Distant Supervision over Wikipedia ». In : *NLP-DBPEDIA@ISWC*.

Arsevska, Elena, Mathieu Roche, Renaud Lancelot, Pascal Hendrikx, Barbara Dufour, Sylvain Falala et David Chavernac (2016). « Monitoring Disease Outbreak Events on the Web Using Text mining Approach and Domain Expert Knowledge ». In : May, p. 3407–3411.

Arsevska, Elena, Sarah Valentin, Julien Rabatel, Jocelyn de Goër de Hervé, Sylvain Falala, Renaud Lancelot et Mathieu Roche (2018). « Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System ». In : *PLOS ONE* 13.8, p. 1–25.

Aurousseau, M. (1945). « On Lists of Words and Lists of Names ». In : *The Geographical Journal* 105.1/2, p. 61–67.

Beitzel, Steven M., Eric C. Jensen et Ophir Frieder (2009). « MAP ». en. In : *Encyclopedia of Database Systems*. Sous la dir. de LING LIU et M. TAMER ÖZSU. Boston, MA : Springer US, p. 1691–1692.

Belouaer, Lamia, David Brosset et Christophe Claramunt (2016). « From verbal route descriptions to sketch maps in natural environments ». In : *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '16*, p. 1–10.

Bharti, Santosh Kumar et Korra Sathya Babu (2017). « Automatic Keyword Extraction for Text Summarization : A Survey ». In : *arXiv :1704.03242 [cs]*. arXiv : 1704.03242.

Bird, Steven et Edward Loper (2004). « NLTK : the natural language toolkit ». In : *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, p. 31.

Blei, David M, Blei Edu, Andrew Y Ng, Ang Edu, Michael I Jordan et Jordan Edu (2003). « Latent Dirichlet Allocation ». In : *Journal of Machine Learning Research* 3, p. 993–1022.

Blei, David M. et John D. Lafferty (2009). « Visualizing Topics with Multi-Word Expressions ». In : *arXiv :0907.1013 [stat]*. arXiv : 0907.1013.

Borgwardt, K.M. et H. Kriegel (2005). « Shortest-Path Kernels on Graphs ». In : *Fifth IEEE International Conference on Data Mining (ICDM'05)*. Houston, TX, USA : IEEE, p. 74–81.

Bourigault, Didier, Isabelle Gonzalez-Mullier et Cécile Gros (1996). « LEXTER, a Natural Language Processing tool for terminology extraction ». In : *Proceedings of the 7th EURALEX International Congress*, p. 771–779.

Bourigault, Didier et Christian Jacquemin (1999). « Term extraction + term clustering : an integrated platform for computer-aided terminology ». en. In : *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics -*. Bergen, Norway : Association for Computational Linguistics, p. 15.

Brin, Sergey (1999). « Extracting Patterns and Relations from the World Wide Web ». en. In : *The World Wide Web and Databases*. Sous la dir. de Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Paolo Atzeni, Alberto Mendelzon et Giansalvatore Mecca. T. 1590. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 172–183.

Bunke, H et G Allermann (1983). « Inexact graph recognition matching for structural pattern ». In : *Pattern Recognition Letters* 1.May, p. 245–253.

Bunke, Horst et Kim Shearer (1998). « A graph distance metric based on the maximal common subgraph ». In : *Pattern Recognition Letters* 19.3-4, p. 255–259.

Buscaldi, Davide (2010). *Ambiguous Place Names on the Web*. Puebla : Universidad Politecnica de Valencia.

Chiticariu, Laura, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss et Shivakumar Vaithyanathan (2010). « Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks ». In : *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA : Association for Computational Linguistics, p. 1002–1012.

Cho, Isaac, Wewnen Dou, Derek Xiaoyu Wang, Eric Sauda et William Ribarsky (2016). « VAiRoma : A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History ». en. In : *IEEE Transactions on Visualization and Computer Graphics* 22.1, p. 210–219.

Christen, Peter (2012). *Data Matching*. en. Berlin, Heidelberg : Springer Berlin Heidelberg.

Cohen, Jonathan D (1995). « Highlights : Language-and domain-independent automatic indexing terms for abstracting ». In : *Journal of the American society for information science* 46.3, p. 162–174.

Craswell, Nick (2009). « Precision at n ». In : *Encyclopedia of Database Systems*. Sous la dir. de LING LIU et M. TAMER ÖZSU. Boston, MA : Springer US, p. 2127–2128.

Cunningham, Hamish (2002). « GATE, a general architecture for text engineering ». In : *Computers and the Humanities* 36.2, p. 223–254.

Daille, Béatrice (1994). « Study and implementation of combined techniques for automatic extraction of terminology ». In : *The balancing act : Combining symbolic and statistical approaches to language*.

Dalbin, Sylvie (2007). « Thésaurus et informatique documentaires ». fr. In : *Documentaliste-Sciences de l'Information* Vol. 44.1, p. 76–80.

De Sabbata, Stefano et Tumasch Reichenbacher (2012). « Criteria of geographic relevance : an experimental study ». en. In : *International Journal of Geographical Information Science* 26.8, p. 1495–1520.

Debnath, Asim Kumar, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman et Corwin Hansch (1991). « Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity ». en. In : *Journal of Medicinal Chemistry* 34.2, p. 786–797.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer et Richard Harshman (1990). « Indexing by latent semantic analysis ». In : *Journal of the American Society for Information Science* 41.6, p. 391–407.

Delozier, Grant, Jason Baldridge, Loretta London et Austin Tx (2015). « Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles ». In : *Aaai*, p. 2382–2388.

Egenhofer, Max J. et Robert D. Franzosa (1991). « Point-set topological spatial relations ». In : *International Journal of Geographical Information Systems* 5.2, p. 161–174. doi : 10.1080/02693799108927841.

Etzioni, Oren, Stanley Kok, Stephen Soderland, Michael Cafarella, Ana-Maria Popescu, Daniel S Weld, Doug Downey, Tal Shaked et Alexander Yates (2004). « Web-Scale Information Extraction in KnowItAll (Preliminary Results) ». en. In : p. 11.

Fader, Anthony, Luke Zettlemoyer et Oren Etzioni (2014). « Open question answering over curated and extracted knowledge bases ». en. In : *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*. New York, New York, USA : ACM Press, p. 1156–1165.

Ferreira, Nivan, Jorge Poco, Huy T. Vo, Juliana Freire et Claudio T. Silva (2013). « Visual Exploration of Big Spatio-Temporal Urban Data : A Study of New York City Taxi Trips ». In : *IEEE Transactions on Visualization and Computer Graphics* 19.12, p. 2149–2158.

Firth, John Rupert (1957). « A synopsys of Linguistic theory ». In : *Studies in Linguistics*.

Fischer, Andreas, Kaspar Riesen et Horst Bunke (2017). « Improved quadratic time approximation of graph edit distance by combining Hausdorff matching and greedy assignment ». In : *Pattern Recognition Letters* 87, p. 55–62.

Fize, Jacques, Mathieu Roche et Maguelonne Teisseire (2018). « Gemedoc : A Text Similarity Annotation Platform ». In : *Natural Language Processing and Information Systems*. Sous la dir. de Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais et Farid Meziane. Cham : Springer International Publishing, p. 333–336.

Fize, Jacques et Gaurav Shrivastava (2017). « GeoDict : an integrated gazetteer ». In : *Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017) at the Conference on Computational Semantics (IWCS)*. Montpellier, France, p. 11.

Fize, Jacques, Maguelonne Teisseire et Mathieu Roche (2017). « Gemedoc : Un outil pour annoter les correspondances entre les documents ». fr. In : *EXCES - EXtraction de Connaissances à partir de donnéEs Spatialisées, Spatial Analysis and GEOmatics 2017*, p. 6.

Fize, Jacques (2018). *Données pour l'évaluation de méthodes de géocodage*.

Floyd, Robert W. (1962). « Algorithm 97 : Shortest path ». In : *Commun. ACM* 5.6, p. 345.

Frantzi, Katerina, Sophia Ananiadou et Hideki Mima (2000). « Automatic recognition of multi-word terms :. the C-value/NC-value method ». In : *International Journal on Digital Libraries* 3.2, p. 115–130.

Freed, N. et N. Borenstein (1992). *MIME (Multipurpose Internet Mail Extensions) : Mechanisms for Specifying and Describing the Format of Internet Message Bodies*. en.

Fundel, Katrin, Robert Küffner et Ralf Zimmer (2007). « RelEx—Relation extraction using dependency parse trees ». en. In : *Bioinformatics* 23.3, p. 365–371.

Gali, Karthik, Harshit Surana, Ashwini Vaidya, Praneeth Shishtla et Dipti Misra Sharma (2008). « Aggregating Machine Learning and Rule Based Heuristics for Named Entity Recognition ». In : *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

Gandomi, Amir et Murtaza Haider (2015). « Beyond the hype : Big data concepts, methods, and analytics ». In : *International Journal of Information Management* 35.2, p. 137–144.

Gardner, Matt et Tom Mitchell (2015). « Efficient and Expressive Knowledge Base Completion Using Subgraph Feature Extraction ». In : *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal : Association for Computational Linguistics, p. 1488–1498.

Gee, James Paul (2004). « An Introduction to Discourse Analysis - Theory and Method ». en. In : *Routledge*, p. 206.

Gelernter, Judith et Wei Zhang (2013). « Cross-lingual Geo-parsing for Non-structured Data ». In : *Proceedings of the 7th Workshop on Geographic Information Retrieval*. GIR '13. event-place : Orlando, Florida. New York, NY, USA : ACM, p. 64–71. isbn : 978-1-4503-2241-6. doi : 10.1145/2533888.2533943. url : http://doi.acm.org/10.1145/2533888.2533943.

Giuliano, Claudio, Alberto Lavelli et Lorenza Romano (2006). « Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature ». In : *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Goodchild, M. F. et L. L. Hill (2008). « Introduction to digital gazetteer research ». en. In : *International Journal of Geographical Information Science* 22.10, p. 1039–1044.

Gritta, Milan, Mohammad Taher Pilehvar, Nut Limsopatham et Nigel Collier (2018). « What's missing in geographical parsing ? » en. In : *Language Resources and Evaluation* 52.2, p. 603–623.

Grover, Aditya et Jure Leskovec (2016). « node2vec : Scalable Feature Learning for Networks ». en. In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*. San Francisco, California, USA : ACM Press, p. 855–864.

Gurulingappa, Harsha, Abdul Mateen-Rajpu et Luca Toldo (2012). « Extraction of potential adverse drug events from medical case reports ». In : *Journal of Biomedical Semantics* 3.1, p. 15.

Haklay, M. et P. Weber (2008). « OpenStreetMap : User-Generated Street Maps ». In : *IEEE Pervasive Computing* 7.4, p. 12–18.

Halterman, Andrew (2017). « Mordecai : Full Text Geoparsing and Event Geocoding ». In : *The Journal of Open Source Software* 2.9.

Hasan, Kazi Saidul et Vincent Ng (2014). « Automatic keyphrase extraction : A survey of the state of the art ». In : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. T. 1, p. 1262–1273.

Hofmann, Thomas (1999). « Probabilistic Latent Semantic Analysis ». In : *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*. UAI'99. event-place : Stockholm, Sweden. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., p. 289–296.

Honnibal, Matthew et Ines Montani (2017). « spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing ». In : *To appear*.

Hulth, Anette (2003). « Improved Automatic Keyword Extraction Given More Linguistic Knowledge ». In : *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, p. 216–223.

Humphreys, J. B. Keith (2002). « PhraseRate : An HTML Keyphrase Extractor ». In :

Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li et Liang Zhao (2019). « Latent Dirichlet allocation (LDA) and topic modeling : models, applications, a survey ». en. In : *Multimedia Tools and Applications* 78.11, p. 15169–15211.

Jones, Christopher B. et Ross S. Purves (2008). « Geographical information retrieval ». In : *International Journal of Geographical Information Science* 22.3, p. 219–228.

Kho, Abel N et al. (2015). « Design and implementation of a privacy preserving electronic health record linkage tool in Chicago ». In : *Journal of the American Medical Informatics Association* 22.5, p. 1072–1080.

Kriege, Nils M., Fredrik D. Johansson et Christopher Morris (2019). « A Survey on Graph Kernels ». In : *arXiv :1903.11835 [cs, stat]*. arXiv : 1903.11835.

Kuang, Da, Jaegul Choo et Haesun Park (2015). « Nonnegative matrix factorization for interactive topic modeling and document clustering ». In : *Partitional Clustering Algorithms*. Springer, p. 215–243.

Lafferty, John, Andrew McCallum et Fernando CN Pereira (2001). « Conditional random fields : Probabilistic models for segmenting and labeling sequence data ». In :

Leidner, Jochen L (2007). « Toponym resolution in text : annotation, evaluation and applications of spatial grounding ». In : *ACM SIGIR Forum*. T. 41. ACM, p. 124–126.

Li, Huifeng, Rohini K. Srihari, Cheng Niu et Wei Li (2003). *InfoXtract Location Normalization : A Hybrid Approach to Geographic References in Information Extraction :* en. Rapp. tech. Fort Belvoir, VA : Defense Technical Information Center.

Li, Jie, Siming Chen, Wei Chen, Gennady Andrienko et Natalia Andrienko (2018). « Semantics-Space-Time Cube. A Conceptual Framework for Systematic Analysis of Texts in Space and Time ». In : *IEEE Transactions on Visualization and Computer Graphics*, p. 1–1.

Li, Ming et Baozong Yuan (2005). « 2D-LDA : A statistical linear discriminant analysis for image matrix ». In : *Pattern Recognition Letters* 26.5, p. 527–532.

Lieberman, Michael D, Hanan Samet et Jagan Sankaranarayanan (2010). « Geotagging with local lexicons to build indexes for textually-specified spatial data ». In : *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*. IEEE, p. 201–212.

Lieberman, Michael D., Hanan Samet, Jagan Sankaranarayanan et Jon Sperling (2007). « STEWARD : architecture of a spatio-textual search engine ». en. In : ACM Press, p. 1.

Ligozat, Gérard (2013). *Qualitative Spatial and Temporal Reasoning*. en. John Wiley & Sons.

Lin, Chenghua et Yulan He (2009). « Joint sentiment/topic model for sentiment analysis ». In : *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, p. 375–384.

Lin, Dekang (1998). « An Information-Theoretic Definition of Similarity ». In : *In Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann, p. 296–304.

Lossio-Ventura, Juan Antonio, Clement Jonquet, Mathieu Roche et Maguelonne Teisseire (2014). « Biomedical Terminology Extraction : A new combination of Statistical and Web Mining Approaches ». en. In : p. 13.

Lovász, László et al. (1993). « Random walks on graphs : A survey ». In : *Combinatorics, Paul erdos is eighty* 2.1, p. 1–46.

Manning, Christopher, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard et David McClosky (2014). « The Stanford CoreNLP natural language processing toolkit ». In : *Proceedings of 52nd annual meeting of the association for computational linguistics : system demonstrations*, p. 55–60.

Matsuo, Y. et M. Ishizuka (2004). « KEYWORD EXTRACTION FROM A SINGLE DOCUMENT USING WORD CO-OCCURRENCE STATISTICAL INFORMATION ». en. In : *International Journal on Artificial Intelligence Tools* 13.01, p. 157–169.

Mehler, Alexander (2008). « Large text networks as an object of corpus linguistic studies ». In : *Corpus linguistics. An international handbook of the science of language and society*, p. 328–382.

Melo, Fernando et Bruno Martins (2017). « Automated Geocoding of Textual Documents : A Survey of Current Approaches ». In : *Transactions in GIS* 21.1, p. 3–38.

Michael Hunger et William Lyon (2016). *Analyzing the Panama Papers with Neo4j : Data Models, Queries & More*. en.

Mikolov, Tomas, Kai Chen, Greg Corrado et Jeffrey Dean (2013). « Efficient Estimation of Word Representations in Vector Space ». en. In : *arXiv :1301.3781 [cs]*. arXiv : 1301.3781.

Mohammed, Hussam, Nathan Clarke et Fudong Li (2018). « Evidence Identification in Heterogeneous Data Using Clustering ». en. In : *Proceedings of the 13th International Conference on Availability, Reliability and Security - ARES 2018*. Hamburg, Germany : ACM Press, p. 1–8.

Moncla, Ludovic, Mauro Gaio, Javier Nogueras-Iso et Sébastien Mustière (2016). « Reconstruction of itineraries from annotated text with an informed spanning tree algorithm ». In : *International Journal of Geographical Information Science* 30.6, p. 1137–1160.

Moncla, Ludovic, Walter Renteria-Agualimpia, Javier Nogueras-Iso et Mauro Gaio (2014). « Geocoding for texts with fine-grain toponyms : an experiment on a geoparsed hiking descriptions corpus ». In : *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2014)*. Sous la dir. d'ACM. Dallas, Texas, United States.

Montello, Daniel R., Sara Irina Fabrikant, Marco Ruocco et Richard S. Middleton (2003). « Testing the First Law of Cognitive Geography on Point-Display Spatializations ». In : *Spatial Information Theory. Foundations of Geographic Information Science*. Sous la dir. de Walter Kuhn, Michael F. Worboys et Sabine Timpf. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 316–331.

Munkres, James (1957). « Algorithms for the assignment and transportation problems ». In : *Journal of the society for industrial and applied mathematics* 5.1, p. 32–38.

Musat, Claudiu Cristian, Julien Velcin, Stefan Trausan-Matu et Marian-Andrei Rizoiu (2011). « Improving topic evaluation using conceptual knowledge ». In : *Twenty-Second International Joint Conference on Artificial Intelligence*.

Nadeau, David et Satoshi Sekine (2007). « A survey of named entity recognition and classification ». en. In : *Lingvisticae Investigationes* 30.1, p. 3–26.

Narayanan, Annamalai, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu et Shantanu Jaiswal (2017). « graph2vec : Learning Distributed Representations of Graphs ». In : *arXiv :1707.05005 [cs]*. arXiv : 1707.05005.

Neuhaus, Michel, Kaspar Riesen et Horst Bunke (2006). « Fast Suboptimal Algorithms for the Computation of Graph Edit Distance ». In : Im, p. 163–172.

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy et James R. Curran (2013). « Learning multilingual named entity recognition from Wikipedia ». In : *Artificial Intelligence*. Artificial Intelligence, Wikipedia and Semi-Structured Resources 194, p. 151–175.

Nyerges, Timothy L., David M. Mark, Robert Laurini et Max J. Egenhofer, éds. (1995). *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems*. en. Dordrecht : Springer Netherlands.

Page, Lawrence, Sergey Brin, Rajeev Motwani et Terry Winograd (1999). *The PageRank citation ranking : Bringing order to the web.* Rapp. tech. Stanford InfoLab.

Papadimitriou, Panagiotis, Ali Dasdan et Hector Garcia-Molina (2010). « Web graph similarity for anomaly detection ». In : *Journal of Internet Services and Applications* 1.1, p. 19–30.

Perozzi, Bryan, Rami Al-Rfou et Steven Skiena (2014). « DeepWalk : online learning of social representations ». en. In : *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*. New York, New York, USA : ACM Press, p. 701–710.

Peuquet, Donna J. (1994). « It's about Time : A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems ». In : *Annals of the Association of American Geographers* 84.3, p. 441–461.

Phua, Clifton, Kate Smith-Miles, Vincent Lee et Ross Gayler (2012). « Resilient Identity Crime Detection ». en. In : *IEEE Transactions on Knowledge and Data Engineering* 24.3, p. 533–546.

Poon, Hoifung, Kristina Toutanova et Chris Quirk (2014). « DISTANT SUPERVISION FOR CANCER PATHWAY EXTRACTION FROM TEXT ». en. In : *Biocomputing 2015*. Kohala Coast, Hawaii, USA : WORLD SCIENTIFIC, p. 120–131.

Portet, François, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer et Cindy Sykes (2009). « Automatic generation of textual summaries from neonatal intensive care data ». en. In : *Artificial Intelligence* 173.7-8, p. 789–816.

Purves, Ross S., Paul Clough, Christopher B. Jones, Avi Arampatzis et al. (2007). « The design and implementation of SPIRIT : a spatially aware search engine for information retrieval on the Internet ». en. In : *International Journal of Geographical Information Science* 21.7, p. 717–745.

Purves, Ross S., Paul Clough, Christopher B. Jones, Mark H. Hall et Vanessa Murdock (2018). « Geographic Information Retrieval : Progress and Challenges in Spatial Search of Text ». en. In : *Foundations and Trends® in Information Retrieval* 12.2-3, p. 164–318.

Rabatel, Julien, Elena Arsevska et Mathieu Roche (2019). « PADI-web corpus : Labeled textual data in animal health domain ». eng. In : *Data in Brief* 22, p. 643–646.

Ramos, Juan et al. (2003). « Using tf-idf to determine word relevance in document queries ». In : *Proceedings of the first instructional conference on machine learning*. T. 242. Piscataway, NJ, p. 133–142.

Randell, David A, Zhan Cui et Anthony G Cohn (1992). « A Spatial Logic based on Regions and Connection ». en. In : *Principles of Knowledge Representation and Reasoning : Proceedings of the 1st International Conference*, p. 13.

Rauch, Erik, Michael Bukatin et Kenneth Baker (2003). « A confidence-based framework for disambiguating geographic terms ». en. In : *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references -*. T. 1. Not Known : Association for Computational Linguistics, p. 50–54.

Ravichandran, Deepak et Eduard Hovy (2002). « Learning surface text patterns for a Question Answering System ». In : *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA : Association for Computational Linguistics, p. 41–47.

« Qualitative Spatial Representation and Reasoning » (2002). en. In : *Qualitative Spatial Reasoning with Topological Information*. Sous la dir. de Jochen Renz. Lecture Notes in Computer Science. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 31–40.

Ricardo, Baeza-Yates et Ribeiro-Neto Berthier (2011). « Modern information retrieval : the concepts and technology behind search ». In : *New Jersey, USA : Addi-son-Wesley Professional*.

Riesen, Kaspar et Horst Bunke (2009). « Approximate graph edit distance computation by means of bipartite graph matching ». In : *Image and Vision Computing* 27.7, p. 950–959.

Riesen, Kaspar, Xiaoyi Jiang et Horst Bunke (2010). « Exact and Inexact Graph Matching : Methodology and Applications ». In : *Managing and Mining Graph Data*, p. 217–247.

Roche, Mathieu, Sophie Fortuno, Juan Antonio Lossio-Ventura, Amira Akli, Salim Belkebir, Thinhinan Lounis et Serigne Toure (2015). « Automatic extraction of keywords from scientific publications for indexing and open data in agronomy ». fr. In : *Cahiers Agricultures* 5, p. 313–320.

Röder, Michael, Andreas Both et Alexander Hinneburg (2015). « Exploring the Space of Topic Coherence Measures ». en. In : *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*. Shanghai, China : ACM Press, p. 399–408.

Rogot, Eugene, Paul Sorlie et Norman J. Johnson (1986). « Probabilistic methods in matching census samples to the National Death Index ». en. In : *Journal of Chronic Diseases* 39.9, p. 719–734.

Roller, Stephen, Michael Speriosu, Sarat Rallapalli, Benjamin Wing et Jason Baldridge (2012). « Supervised text-based geolocation using language models on an adaptive grid ». In : *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, p. 1500–1510.

Sagiroglu, Seref et Duygu Sinanc (2013). « Big data : A review ». In : *2013 International Conference on Collaboration Technologies and Systems (CTS)*, p. 42–47.

Sallaberry, Christian, Mauro Gaio, Damien Palacio et Julien Lesbegueries (2008). « Fuzzying GIS topological functions for GIR needs ». In : *Proceedings of the 5th Workshop on Geographic Information Retrieval*. ACM, p. 1–8.

Al-Salman, Rami, Frank Dylla et Paolo Fogliaroni (2012). « Matching geo-spatial information by qualitative spatial relations ». In : *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information - GEOCROWD '12*, p. 38.

Salton, G. et C. Buckley (1991). « Global Text Matching for Information Retrieval ». en. In : *Science* 253.5023, p. 1012–1015.

Sanderson, Mark et Janet Kohler (2004). « Analyzing geographic queries ». en. In : 2, p. 8–10.

Shaw, Neal G., Ahmad Mian et Surya B. Yadav (2002). « A comprehensive agent-based architecture for intelligent information retrieval in a distributed heterogeneous environment ». In : *Decision Support Systems* 32.4, p. 401–415.

Shervashidze, Nino, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn et Karsten M Borgwardt (2011). « Weisfeiler-Lehman Graph Kernels ». In : *Journal of Machine Learning Research* 12, p. 2539–2561.

Skounakis, Marios, Mark Craven et Soumya Ray (2003). « Hierarchical Hidden Markov Models for Information Extraction ». In : *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. IJCAI'03. Acapulco, Mexico : Morgan Kaufmann Publishers Inc., p. 427–433.

Smith, David A. et Gideon S. Mann (2003). « Bootstrapping toponym classifiers ». en. In : *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references* -. T. 1. Not Known : Association for Computational Linguistics, p. 45–49.

Sorensen, TA (1948). « Method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish Commons ». In : *Biol. Skr Dan Vid Sel.* 5.

Tao, Fangbo, Chao Zhang, Xiusi Chen, Meng Jiang, Tim Hanratty, Lance Kaplan et Jiawei Han (2018). « Doc2Cube : Allocating

Documents to Text Cube without Labeled Data ». In : *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, p. 1260–1265.

Tjong Kim Sang, Erik F. et Fien De Meulder (2003). « Introduction to the CoNLL-2003 Shared Task : Language-independent Named Entity Recognition ». In : *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*. CONLL '03. event-place : Edmonton, Canada. Stroudsburg, PA, USA : Association for Computational Linguistics, p. 142–147.

Tobler, W. R. (1970). « A Computer Movie Simulating Urban Growth in the Detroit Region ». In : *Economic Geography* 46, p. 234–240.

« A Review of OpenStreetMap Data » (2017). en. In : *Mapping and the Citizen Sensor*. Sous la dir. d'University of Nottingham, GB et al. Ubiquity Press, p. 37–59.

Velcin, Julien, Mathieu Roche et Pascal Poncelet (2016). « Shallow text clustering does not mean weak topics : How topic identification can leverage bigram features ». In : *DMNLP : Data Mining and Natural Language Processing*. T. 1646.

Vickery, B.C. (1960). « THESAURUS — A NEW WORD IN DOCUMENTATION ». en. In : *Journal of Documentation* 16.4, p. 181–189.

Vishwanathan, S.V.N., Nicol Schraudolph, Risi Kondor et K.M. Borgwardt (2010). « Graph Kernels ». In : *Journal of Machine Learning Research* 11, p. 1201–1242.

Voorhees, Ellen M. (1999). « The TREC-8 Question Answering Track Report ». In : *In Proceedings of TREC-8*, p. 77–82.

Voorhees, Ellen M. (2001). « Overview of the TREC-9 Question Answering Track ». In : *In Proceedings of the Ninth Text REtrieval Conference (TREC-9*, p. 71–80.

Voorhees, Ellen M. (2003). « Overview of the TREC 2003 Question Answering Track ». In : p. 54–68.

Vrandečić, Denny et Markus Krötzsch (2014). « Wikidata : A Free Collaborative Knowledgebase ». In : *Communications of the ACM* 57.10, p. 78–85.

Wallgrün, Jan Oliver, Lutz Frommberger, Frank Dylla et Diedrich Wolter (2009). « SparQ user manual v0. 7 ». In : *User manual, University of Bremen*.

Wallgrün, Jan Oliver, Diedrich Wolter et Kai-Florian Richter (2010). « Qualitative matching of spatial information ». en. In : ACM Press, p. 300.

Wang, Huijun, Ying Ding, Jie Tang, Xiao Dong, Bing He, Judy Qiu et David J. Wild (2011). « Finding Complex Biological Relationships in Recent PubMed Articles Using Bio-LDA ». In : *PLoS ONE* 6.3. arXiv : 1103.5181, e17243.

Wang, Xuerui, Andrew McCallum et Xing Wei (2007). « Topical n-grams : Phrase and topic discovery, with an application to information retrieval ». In : *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE, p. 697–702.

Weisfeiler, Boris et Andrei A Lehman (1968). « A reduction of a graph to a canonical form and an algebra arising during this reduction ». In : *Nauchno-Technicheskaya Informatsia* 2.9, p. 12–16.

Wing, Benjamin et Jason Baldridge (2014). « Hierarchical Discriminative Classification for Text-Based Geolocation ». en. In : Association for Computational Linguistics, p. 336–348.

Winkler, William E (2006). « Overview of Record Linkage and Current Research Directions. Tech. Rep. RR2006/02 ». en. In : p. 44.

Woodruff, Allison Gyle et Christian Plaunt (1994). « GIPSY : Automated geographic indexing of text documents ». In : *Journal of the American Society for Information Science* 45.9, p. 645–655.

Wu, Jing (2015). « A qualitative spatio-temporal modelling and reasoning approach for the representation of moving entities ». PhD Thesis. Brest.

Xu, Kun, Siva Reddy, Yansong Feng, Songfang Huang et Dongyan Zhao (2016). « Question Answering on Freebase via Relation Extraction and Textual Evidence ». In : *arXiv :1603.00957 [cs]*. arXiv : 1603.00957.

Yang, Jun, Yu-Gang Jiang, Alexander G. Hauptmann et Chong-Wah Ngo (2007). « Evaluating bag-of-visual-words representations in scene classification ». en. In : *Proceedings of the international workshop on Workshop on multimedia information retrieval - MIR '07*. Augsburg, Bavaria, Germany : ACM Press, p. 197.

Zhang, Muhan et Yixin Chen (2018). « Link Prediction Based on Graph Neural Networks ». In : *arXiv :1802.09691 [cs, stat]*. arXiv : 1802.09691.

Zhao, Wayne Xin, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan et Xiaoming Li (2011). « Comparing Twitter and Traditional Media Using Topic Models ». en. In : *Advances in Information Retrieval*. Sous la dir. de Paul Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee et Vanessa Mudoch. T. 6611. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 338–349.

Zhou, GuoDong et Jian Su (2002). « Named Entity Recognition using an HMM-based Chunk Tagger ». In : *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA : Association for Computational Linguistics, p. 473–480.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX :

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here :

http://postcards.miede.de/

Final Version as of 11 février 2020 (version 1.0).