Intégration et analyse de données hétérogènes massives pour une observation intelligente du territoire

Rodrique Kafando

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

En Informatique

École doctorale I2S

Unité de recherche INRAE-UMR TETIS

Intégration et analyse de données massives et hétérogènes pour une observation intelligente du territoire

Présentée par Rodrique KAFANDO

Le 03 Décembre 2021

Devant le jury composé de

Jérôme DARMONT, Professeur, Université de Lyon 2, ERIC	Rapporteur
Franck RAVAT, Professeur, Université de Toulouse I, IRIT	Rapporteur
Carmen GERVET, Professeur, Université de Montpellier, Espace-Dev	Présidente
Carmen BRANDO, Docteur, Ingénieur de recherche, EHESS, CRH	Examinatrice
Marc APARICIO, Chef du Service SIG Métropolitain, Ville et Métropole de Montpellier	Invité
Rémy DECOUPES, Ingénieur de recherche, INRAE - UMR TETIS	Invité, Encadrant
Maguelonne TEISSEIRE, Directrice de Recherche, INRAE - UMR TETIS	Directrice
Lucile SAUTOT, Maître de conférences, AgroParisTech – UMR TETIS	Encadrante

UNIVERSITÉ DE MONTPELLIER

Résumé

L'avènement des nouvelles technologies de la communication et de l'information (NTIC) accélère la croissance des données produites par les services qu'offrent les grandes villes à leur population dans divers domaines. Parmi ces données, nous pouvons citer : les données textuelles (disponibles sur le Web), les images satellite (données de surveillance), les données issues de capteurs, etc. Ces données sont essentiellement issues des services proposés pour répondre aux besoins quotidiens des habitants comme la mobilité, la communication, la santé, ainsi que les services de gestion des différentes ressources comme l'eau, les exploitations agricoles et urbaines, l'énergie, etc. Cette forte croissance remet en question la complétude et l'efficacité des méthodes et techniques classiquement utilisées en fouille de données. Les difficultés rencontrées sont principalement liées à la volumétrie des données, mais aussi à leur complexité telle que la forte hétérogénéité. Notre sujet de recherche s'inscrit dans le cadre de la démarche ÉcoCité de la Métropole de Montpellier Méditerranée (3M), et vise à développer une démarche d'observation intelligente du territoire à partir des masses de données dont-elle dispose.

Dans cette thèse, nous nous intéressons à l'enrichissement mutuelle entre des données fortement hétérogènes pour le suivi des territoires. Nous limitons notre étude à trois thématiques qui sont, l'urbanisation, l'agriculture et l'hydrologie sur le territoire de la métropole de Montpellier. Pour ce faire, nous déclinons dans cette étude, une chaîne de traitement qui regroupe des approches permettant : 1) de collecter des séries temporelles de données textuelles pour la constitution de corpus thématiques avec un ancrage sur le plan spatial et de les analyser, 2) de stocker ces données massives et hétérogènes afin de les rendre accessibles et analysables par tous, sans pour autant les détériorer, 3) d'extraire des indicateurs permettant de mettre en relation les données des différentes thématiques, tant sur le plan spatial, temporel que thématique, et 4) d'extraire des connaissances à partir de ces données, afin de montrer l'impact de ces thématiques sur l'évolution du territoire de façon globale. À partir de ces différentes approches, nous mettons en évidence l'importance de la mise en relation de données gérées jusqu'ici en silo, en faisant ressortir des connaissances pouvant servir de système d'aide à la décision.

L'ensemble des approches méthodologiques que nous proposons dans cette étude, constitue une chaîne de traitement complète, allant de la collecte des données hétérogènes à leur mise en relation et analyse pour l'extraction des connaissances pour la description d'événements territoriaux sur le plan spatio-temporel.

Abstract

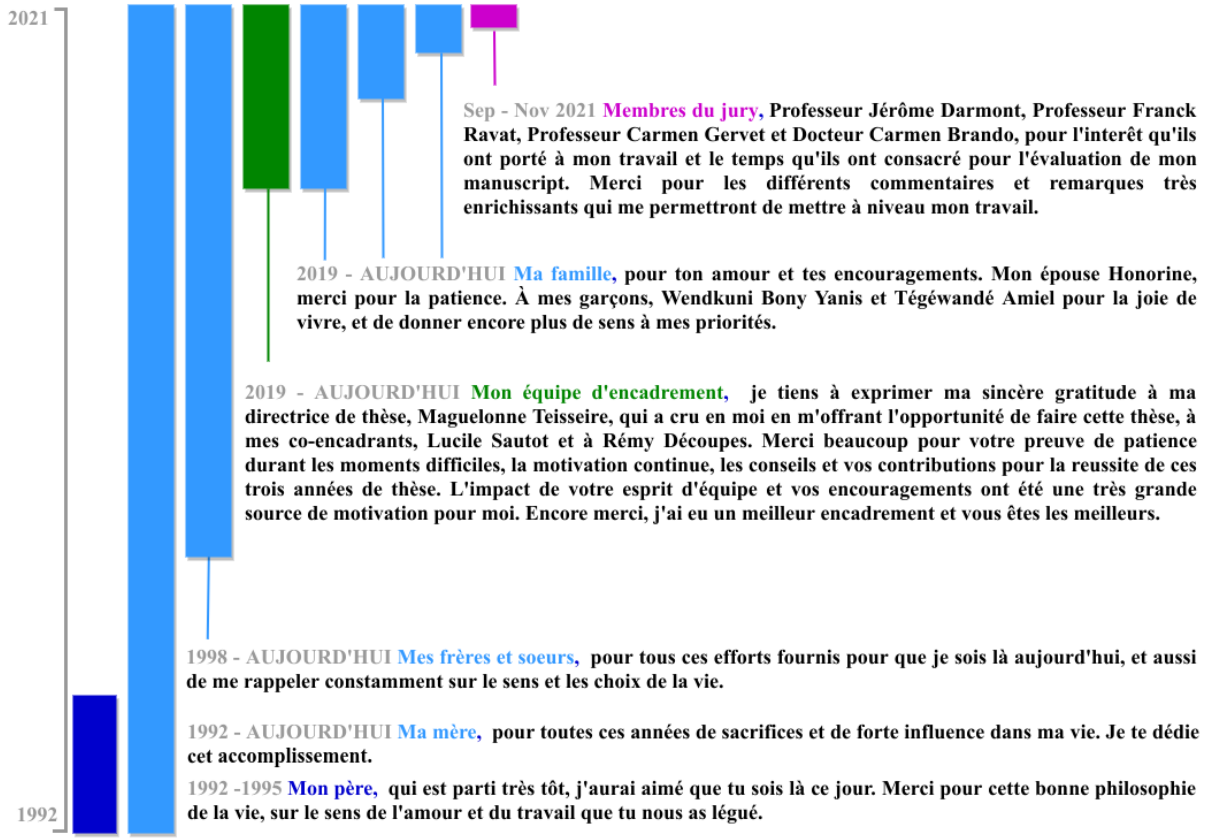
The advent of new communication and information technologies (NICT) accelerates the growth of data produced by services that smart cities offer to their population in various fields. Among these data, we can notice : textual data (available on the Web), satellite images (surveillance data), data from sensors, etc. These data are essentially derived from the services offered by the cities to meet the daily needs of the inhabitants such as mobility, communication, health, as well as services for the management of various resources such as water, agricultural and urban operations, energy, etc. This significant growth of data is challenging the completeness and efficiency of the methods and techniques traditionally used in data mining for knowledge discovery. The difficulties encountered are mainly related to the volume of the data, but also to their complexity such as the high heterogeneity. Our research topic is part of the EcoCity initiative of the Montpellier Mediterranean Metropolis (3M), and aims to develop an intelligent observation approach of the territory from the massive data available.

In this thesis, we are interested in the mutual enrichment between highly heterogeneous data for the monitoring of territories. We limit our study to three themes which are urbanization, agriculture and hydrology on the territory of the metropolis of Montpellier. To do this, we decline in this study, a chain of treatment which gathers approaches allowing : 1) to collect time series of textual data for the constitution of thematic corpora with a spatial anchoring and to analyze them, 2) to store these massive and heterogeneous data in order to make them accessible and analyzable by all, without deteriorating them, 3) to extract indicators allowing to put in relation the data of the different thematics, as well on the spatial, temporal and thematic level, and 4) to extract knowledge from these data, in order to show the impact of these thematics on the evolution of the territory in a global way. From these different approaches, we highlight the importance of linking data that have been managed in silo until now, by bringing out knowledge that can be used as a decision support system.

The set of methodological approaches that we propose in this study constitutes a complete processing chain, from the collection of heterogeneous data to their linking and analysis for the extraction of knowledge for the description of territorial events on the spatio-temporal level.

Remerciements

Je dis merci à...



Sep - Nov 2021 **Membres du jury,** **Professeur Jérôme Darmont, Professeur Franck Ravat, Professeur Carmen Gervet et Docteur Carmen Brando, pour l'interêt qu'ils ont porté à mon travail et le temps qu'ils ont consacré pour l'évaluation de mon manuscript. Merci pour les différents commentaires et remarques très enrichissants qui me permettront de mettre à niveau mon travail.**

2019 - AUJOURD'HUI **Ma famille,** **pour ton amour et tes encouragements. Mon épouse Honorine, merci pour la patience. À mes garçons, Wendkuni Bony Yanis et Tégéwandé Amiel pour la joie de vivre, et de donner encore plus de sens à mes priorités.**

2019 - AUJOURD'HUI **Mon équipe d'encadrement,** **je tiens à exprimer ma sincère gratitude à ma directrice de thèse, Maguelonne Teisseire, qui a cru en moi en m'offrant l'opportunité de faire cette thèse, à mes co-encadrants, Lucile Sautot et à Rémy Découpes. Merci beaucoup pour votre preuve de patience durant les moments difficiles, la motivation continue, les conseils et vos contributions pour la reussite de ces trois années de thèse. L'impact de votre esprit d'équipe et vos encouragements ont été une très grande source de motivation pour moi. Encore merci, j'ai eu un meilleur encadrement et vous êtes les meilleurs.**

1998 - AUJOURD'HUI **Mes frères et soeurs,** **pour tous ces efforts fournis pour que je sois là aujourd'hui, et aussi de me rappeler constamment sur le sens et les choix de la vie.**

1992 - AUJOURD'HUI **Ma mère,** **pour toutes ces années de sacrifices et de forte influence dans ma vie. Je te dédie cet accomplissement.**

1992 -1995 **Mon père,** **qui est parti très tôt, j'aurai aimé que tu sois là ce jour. Merci pour cette bonne philosophie de la vie, sur le sens de l'amour et du travail que tu nous as légué.**

En plus des personnes citées ci-dessus, j'aimerai dire merci à Mathieu Roche pour ses conseils techniques qui m'ont beaucoup aidés. J'aimerai aussi remercier Christiane Weber, mon comité de suivi individuel composé de Phillipe Lemoisson, Eric Barbe pour leurs commentaires et remarques qui m'ont permis d'améliorer mon travail. Merci aux agents de la métropole de Montpellier pour les échanges enrichissants, en particulier Marc Apparicio pour ses remarques très pertinentes.

Pour finir, je suis extrêmement reconnaissant à mes amis et collègues Emmanuel, Maman Sani, Salif, Linda, Jean Eudes, Dav, etc. pour les disscussions enrichissantes et les bons moments passés ensemble. Vous avez fortement contribué à la réalisation de ce travail. Je vous dis merci!

"No one who achieves success does so without acknowledging the help of others. The wise and confident acknowledge this help with gratitude"

Alfred North Whitehead

Sommaire

Table des figures

Liste des tableaux

Nomenclature

3M : Montpellier Méditerranée Métropole

AIDMOIT : Analyse Intégration Données Massive Observation Intelligente Territoire

CSV : Comma-Separated Values

FASTR : FASt Term Recognizer

NDVI : Normalized Difference Vegetation Index

NoSQL : Not Only SQL

NTIC : Nouvelles Technologies de l'Information et de la Communication

OWL : Web Ontology Language

PAC : Pixels Already Covert

RDF : Resource Description Framework

SNE : Spatial Named Entities

STDT : Série Temporelle de Données Textuelles

TALN : Traitement Automatique du Langage Naturel

URL : Uniform Resource Locator

XML : Extensible Markup Language

ObjRefs : Objets de Référence

STIS : Série Temporelle d'Images Satellite

Chapitre 1

Introduction

La croissance du volume de données est une réalité que connaissent toutes les grandes villes de nos jours. Fortement mise en évidence avec l'évolution des nouvelles technologies de l'information et de la communication (NTIC) [Kitchin, 2014], cette croissance accélérée remet en cause les techniques et méthodes précédemment utilisées pour l'exploration et l'analyse des données dans la quête de la découverte de nouvelles connaissances. Face à cette situation, deux principales questions se posent. Il s'agit d'une part, de la gestion et de la bonne gouvernance des données massives, et d'autre part, des techniques et méthodes à mettre en place ou à adapter pour l'analyse de ces masses de données qui sont fortement hétérogènes.

L'hétérogénéité des données se caractérise sous diverses formes. Nous distinguons principalement quatre grands points dans le Chapitre 2 pouvant contribuer à définir de façon générale les données hétérogènes : la source, la sémantique, la structure, et la nature des données. Cette forte hétérogénéité a impacté, au cours de ces dernières décennies, les structures de stockage qui ont connu également des évolutions dans le but de répondre aux contraintes qu'impose cette croissance rapide. Des bases de données relationnelles [Maier, 1983] en passant par les bases de données NoSQL [Han et al., 2011], nous sommes à l'ère des lacs de données [Dixon, 2010]. Une évolution qui marque non seulement la capacité des solutions de stockage à s'adapter aux données, mais aussi ouvre des portes pour d'autres formes d'analyses pouvant contribuer à améliorer les services offerts.

Dans le but de faire émerger des connaissances à priori dissimulées dans ces masses de données, des solutions et méthodes basées sur l'intelligence artificielles, couramment appelées Big Data Analytics Solution [Chen et al., 2012, Russom et al., 2011], sont mises en place pour analyser et extraire des informations à partir de ces données. Ainsi, intervient la notion de mise en relation des données, qui consiste à mettre ensemble plusieurs types de données, de sources différentes dans le but d'exploiter la richesse de chacune d'elles en complément de l'autre pour en extraire d'importantes informations.

À notre connaissance, les solutions de stockage et de mise en relation des données hétérogènes jusque-là proposées, se concentrent sur des domaines ou des thématiques (ex : domaine bio-médical, de transport, d'énergie, etc.) bien précises, ou encore abordent séparément ces deux concepts. Dans le cadre de cette thèse, nous proposons tout d'abord,

un ensemble de solutions liées aux structures de stockage des données massives et hétérogènes adaptées pour les villes intelligentes (ou les grandes villes qui sont des sources de production de données). Nous proposons ensuite des approches de mise en relation entre ces données pour l'extraction des connaissances qu'elles renferment au profit de la Métropole de Montpellier Méditerranée (3M). Aussi, les particularités de notre étude sont les suivantes : 1) elle concerne un cas d'étude concret, 2) elle couvre un processus allant de la collecte des données à leur mise en relation, 3) elle concerne plusieurs thématiques et 4) elle est axée sur une analyse tri-dimensionnelle à savoir la prise en compte des aspects spatiaux, temporels et thématiques.

L'objectif de nos travaux est d'extraire et de mettre à disposition des connaissances qui pourront contribuer à comprendre et à suivre l'évolution du territoire de 3M. Le suivi de cette évolution est fait sur le plan spatial et temporel et cela pour trois grandes thématiques qui sont l'agriculture, l'hydrologie et l'urbanisation. Dans la suite de ce chapitre, nous allons détailler de façon plus précise, les motivations et enjeux associés à cet objectif.

1.1 Motivation

Nous avons introduit précédemment la question sur la forte croissance des données, ainsi que les raisons qui favorisent cette évolution dans les grandes villes. Ces données représentent une source d'information capitale dans laquelle il devient possible de faire ressortir des connaissances décrivant directement ou indirectement le comportement du territoire. Ainsi, il devient nécessaire de fournir des outils d'aide à la décision pour permettre de suivre la dynamique, et de comprendre le comportement et l'évolution des phénomènes territoriaux à partir des données produites par les entités qui le composent. Dans [Al Nuaimi et al., 2015], les auteurs décrivent l'importance et les multiples cas d'utilisation (domaine de l'éducation, de la santé, du transport, etc.) dans lesquels ces données massives peuvent être utilisées pour proposer des services innovants au profit de la population.

En référence au schéma illustré dans la Figure 1, l'exploitation de ces données massives dont dispose un territoire à l'instant présent, devrait permettre d'avoir une vue rétrospective sur l'évolution du territoire ou de la ville, et de pouvoir préconiser des solutions pour l'atteinte d'objectifs dans une vision future.

Dans notre étude, nous assimilons le territoire qui nous intéresse à une ville intelligente, qui dispose d'une masse importante de données. Nous sommes amenés à proposer des techniques et des méthodes, pour extraire des informations pertinentes à partir de ces données qui contribueront à comprendre l'évolution du territoire sur le plan spatio-temporel et thématique.

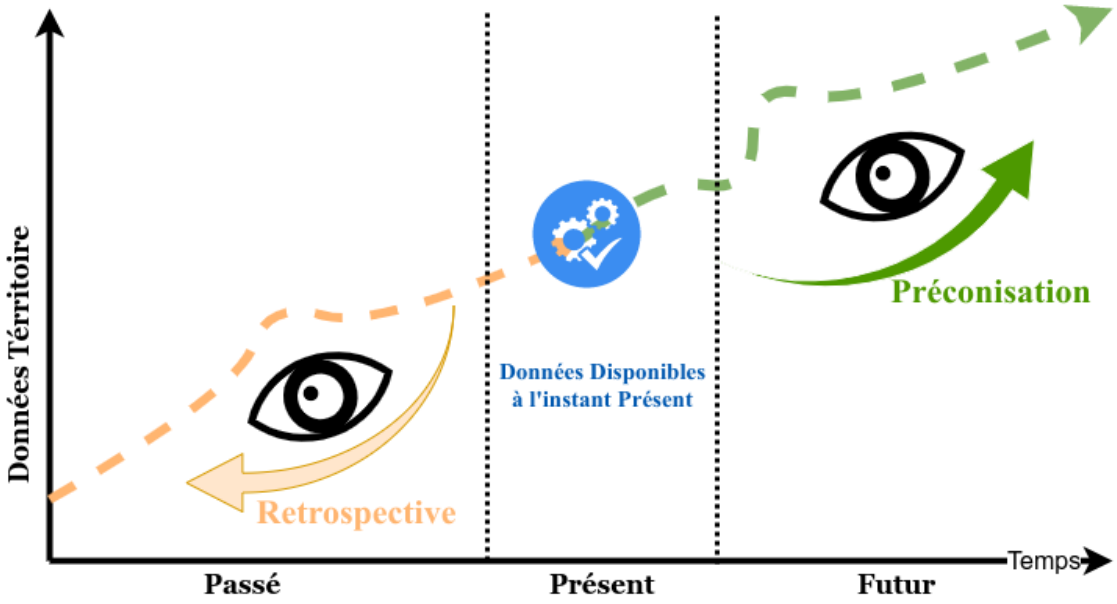


FIGURE 1 – Schéma descriptif de l'évolution des données du territoire en fonction du temps

1.2 Cas d'application

La Métropole de Montpellier Méditerranée (3M) développe depuis plusieurs années une politique qui vise à favoriser les projets de développements numériques innovants. Elle a ainsi distingué un certain nombre de jeunes pousses et d'activités de recherche qui s'intègrent dans son schéma de *ville intelligente*. Elle a, par exemple, financé des capteurs passifs, rechargeables à partir de l'énergie létale des smart-phones, qui collectent des informations sur la qualité de l'air. La métropole tente ainsi de développer à la fois une observation fine de son territoire et de promouvoir des technologies innovantes.

Le travail présenté dans cette thèse participe à ces efforts et vise à développer des compétences autour d'une *observation intelligente du territoire*. Observation à partir de diverses sources d'information (collectées, mesurées, interprétées), intelligente car mettant en œuvre non seulement des technologies innovantes mais aussi des approches fondées sur l'hétérogénéité des sources, l'extraction des connaissances à partir de données massives et enfin un territoire doté d'une gouvernance nourrie par ces observations pour l'aider dans les prises de décision en permettant d'évaluer en continu les actions menées et de les infléchir si nécessaire (logique de gestion adaptative).

De façon générale, l'intégration de données hétérogènes prend en compte l'écosystème *ville intelligente* de la Métropole (laboratoires, start up, collectifs...) et se fonde sur des informations classiques (démographiques, économiques, climatiques...) mais aussi sur des informations venant des projets financés par la Métropole.

Nos travaux dans cette démarche de *ville intelligente*, consistent à apporter des solutions méthodologiques qui vont aider la Métropole de Montpellier à comprendre la dynamique de son territoire à partir des données dont elle dispose. Les différentes parties de notre étude peuvent être résumées en trois points.

- L'une de ses particularités est la prise en compte des trois dimensions majeures qui sont la spatialité, la temporalité et les thématiques associées à ces différentes données ;
- Ensuite, elle doit être basée sur des données hétérogènes. Les données hétérogènes représentent les données de toutes natures dont dispose 3M, pouvant contribuer à décrire au mieux le territoire dans toutes ses dimensions (transport, gestion de risque, aménagement territorial, changement climatique, etc.) ;
- Pour terminer, nos travaux consisteront à trouver des méthodes et techniques permettant la mise en relation des différentes informations dont nous disposons sur le territoire.

Dans le cadre de cette étude, nous avons choisi d'aborder principalement trois thématiques à savoir l'*urbanisation*, l'*agriculture* et l'*hydraulique*. Le choix de ces thématiques est guidé par différents constats sur les mouvements que connaissent les territoires. Ces thématiques ont des liaisons fortes dans le cadre de l'évolution des villes, qu'elle soit

spatiale ou temporelle.

L'urbanisation est un phénomène qui affecte l'agriculture sur l'aspect spatiale. En effet, plus la ville s'étend, plus les surfaces agricoles diminuent. Cette diminution des surfaces agricoles a sans doute des retombées sur la quantité de production.

L'expansion urbaine provoque également une forte augmentation du nombre des ouvrages hydrauliques. À chaque plan d'aménagement, correspond à la construction de nouvelles infrastructures de drainage et d'évacuation. Mais malgré ces nombreuses infrastructures, nous constatons toujours des problèmes liés à la circulation des eaux en milieu urbain, entraînant ainsi des cas d'inondation. Ces mêmes inondations pouvant être causées par la montée des eaux des corridors en cas de grandes pluies. Nous pouvons noter également l'impact que subissent les espaces agricoles en cas d'inondation, affectant ainsi les productions agricoles.

À partir de toutes ces informations thématiques, notre objectif est de montrer l'importance de la mise en relation de données gérées jusqu'ici en silo pour un système d'aide à la décision.

1.3 Objectifs

Nos travaux de recherche visent à apporter des solutions pour la mise en relation des données hétérogènes dont dispose 3M pour un meilleur suivi de son territoire. Indiquées dans l'Annexe A.1, nous avons des données de diverses natures (données structurées, non-structurées, semi-structurées, images satellites) et de diverses sources (web, bases de données locales, services différents, etc.). Le défi majeur de notre étude est d'arriver à enrichir ces données, les unes avec les autres pour faire ressortir des informations jusque là méconnues qui serviront à mieux cerner le comportement du territoire. Pour ce faire, nous avons axé notre étude autour de deux objectifs principaux qui sont :

— montrer comment les données hétérogènes peuvent être mises en lien sémantiquement (documents, médias sociaux, images satellites, etc.) afin de permettre une analyse des phénomènes spatio-temporels complexes sur le territoire de 3M ;

— définir des techniques originales de fouille de données (fouille de données textuelles, fouille de données spatio-temporelles, etc.) adaptées pour l'analyse des données hétérogènes que nous disposons.

Quant aux contributions attendues, elles peuvent être résumées en deux parties :

— Les contributions scientifiques pour la mise en relation et l'analyse des données hétérogènes
  - les solutions à mettre en place pour la collecte et la gestion des données hétérogènes ;

  - les nouvelles méthodes à définir pour l'intégration des données fortement hétérogènes ;
  - le développement de nouvelles approches et méthodes pour la mise en relation et l'analyse de ces données.
- Les contributions thématiques qui prennent en compte les apports de nos travaux vis-à-vis de 3M
  - la proposition d'outils permettant de faire émerger des informations capitales dans le suivi de l'évolution d'un territoire à partir des données dont il dispose ;
  - la mise en évidence de plusieurs indicateurs pouvant contribuer à la prise de décision pour les villes de demain.

1.4 Problématique

Les motivations formulées ci-dessus nous amènent à nous poser un certain nombre de questions. La principale sur le but de notre travail est : comment arriver à retracer l'évolution du territoire à partir des données hétérogènes ? À cette question s'ajoutent les questions secondaires suivantes :

- Quelles stratégies mettre en place pour collecter efficacement les données en lien avec le territoire de Montpellier ?
- Quelle structure de stockage adopter pour le stockage et l'intégration de ces données ?
- Quelles techniques et méthodes mettre en place pour extraire des indicateurs spatio-temporels et thématiques pertinents dans chacune de ces données pour leur mise en relation ?
- Comment analyser efficacement les relations entre ces données ?

La Figure 2 illustre ces différents questionnements dans le contexte de notre étude.

1.5 Organisation du manuscrit

La suite de ce manuscrit est organisée comme suit :

- le Chapitre 2 décrit en premier lieu, la notion de ville intelligente avec les concepts associés ainsi que les principales sources de production de données massives de ces grandes villes. La seconde partie est dédiée à la description de l'architecture générale AIDMOIT que nous avons proposée. Elle fait ressortir les différentes étapes de la chaîne méthodologique ;
- le Chapitre 3 met en exergue le début de la démarche méthodologique. Cette démarche consiste à proposer un protocole pour la collecte de données textuelles des trois thématiques, à savoir les thématiques agriculture, hydrologie, et urbanisation

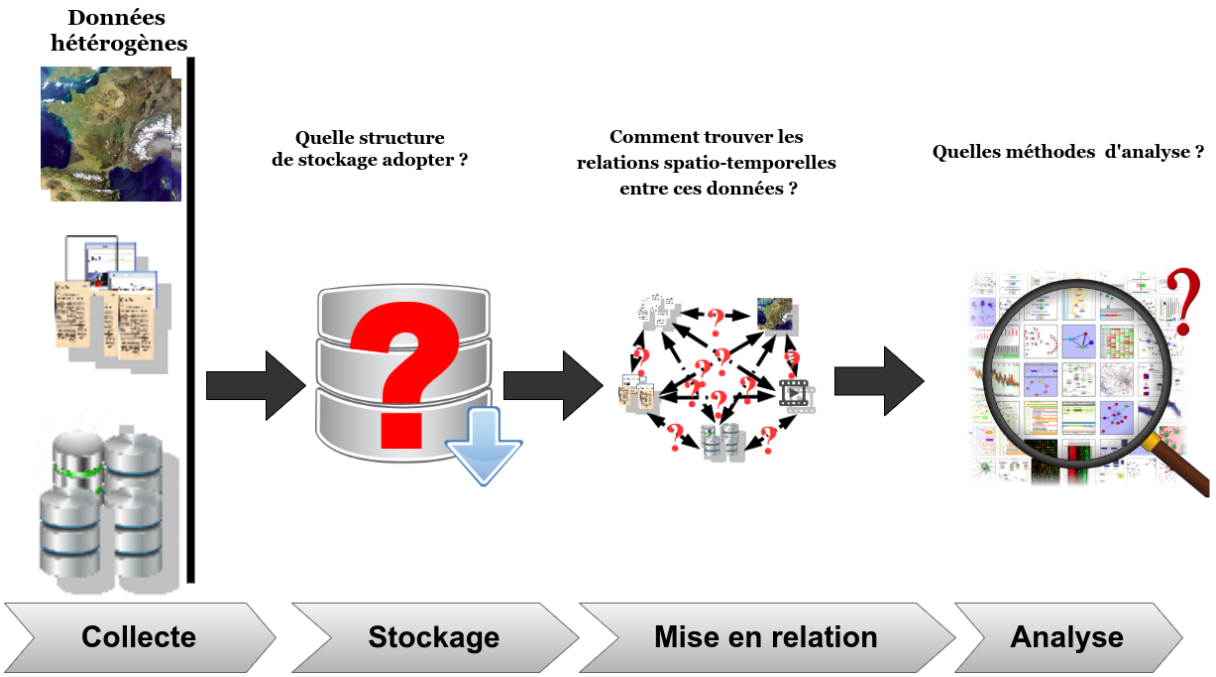


FIGURE 2 – Les différentes problématiques

de 3M (voir Section 1.2), qui constituent un sous-ensemble des données utilisées dans cette étude ;

— le Chapitre 4 concerne toujours le processus de collecte de données, mais prend en compte l'extraction de terminologie avec une illustration thématique dans le domaine bio-médical ;

— le Chapitre 5 aborde les structures de stockages adaptées pour les données massives et hétérogènes. Il détaille essentiellement le lac de données spatiales qui servira pour le stockage des données obtenues dans le Chapitre 3 ;

— le Chapitre 6 décrit les techniques et méthodes pour la mise en relation et l'analyse des données hétérogènes précédemment stockées dans le lac de données ;

— le Chapitre 7 dresse les conclusions issues des chapitres précédents, tout en faisant ressortir les contributions de notre étude ainsi que les perspectives associées à nos travaux ;

1.6 Liste des publications

— Revue internationale

- R. Kafando, R. Decoupes, V. Sarah, L. Sautot, M. Teisseire and M. Roche, "ITEXT-BIO : Intelligent Term EXTraction for BIOmedical Analysis" *Health Information Science and Systems (HISS)*, 2021, pages 1-23

— Actes de conférences internationales

- R. Decoupes, R. Kafando, M. Roche, and M. Teisseire H-TFIDF : What makes areas specific over time in the massive flow of tweets related to the covid pandemic ? *Association of Geographic Information Laboratories in Europe (AGILE) : GIScience Ser.*, Vol 2, 2, https ://doi.org/10.5194/agile-giss-2-2-2021, 2021
- R. Kafando, R. Decoupes, L. Sautot and M. Teisseire Spatial Data Lake for Smart Cities : From Design to Implementation. *Association of Geographic Information Laboratories in Europe (AGILE) : GIScience Series*, Vol. 1, 8, https ://doi.org/10.5194/agile-giss-1-8-2020, 2020

— Actes de conférences nationales

- R. Kafando, R. Decoupes, L. Sautot, M. Teisseire, and C. Weber Constitution de corpus thématique : Pour un meilleur suivi du territoire de la Métropole de Montpellier Méditerranée. *Spatial Analysys and GEOmatics Conference* SAGEO'21 16ème, Mai 2021, pages 165-176.
- R. Kafando, R. Decoupes, L. Sautot, and M. Teisseire Modélisation de la dynamique des territoires : méta-données et lacs de données dédiés à l'information spatiale. *INFormatique des ORganisations et Systèmes d'Information et de Décision INFORSID2020*, Juin 2020, pages 207-222.

Chapitre 2

Vers un système d'information pour les villes intelligentes

Comme nous l'avons précédemment évoqué, les grandes villes produisent aujourd'hui des données hétérogènes, et en très grand volume. Cette croissance continuera à suivre une évolution exponentielle [Gantz and Reinsel, 2012], à partir du moment où les populations sont de plus en plus exigeantes avec un nombre important de services. Nous pouvons noter aussi le fait que certains projets politiques visent à obliger ou à promouvoir l'accès à des services au profit de la population.

Au vu de ce phénomène grandissant, l'une des questions qui se pose est la gestion de ces données. Il devient impératif de trouver des systèmes évolutifs et flexibles, capables de s'adapter à cette évolution, et cela pour deux raisons principales. La première étant que les traitements qui seront appliqués sur ces données ne sont pas connus à l'avance. Les besoins d'exploitations de ces données sont différents selon les utilisateurs. La seconde raison est liée au fait que ces données ne doivent pas être destinées à un type de profil spécifique. Plusieurs profiles (expert territorial, usager, etc.) devraient être capable d'exploiter ces données en fonction de leurs besoins.

Dans ce chapitre, nous abordons, dans un premier temps, la notion des villes intelligentes et leur impact dans la génération massive des données. Dans un second temps, nous présentons l'architecture complète que nous proposons dans le cadre de cette étude, pour la gestion et pour l'exploitation des données massives avec une illustration sur la Métropole de Montpellier Méditerranée (3M).

2.1 Les Villes Intelligentes

2.1.1 Définitions et Concepts

Le concept de ville intelligente ou smart city suscite de plus en plus d'intérêt dans la communauté scientifique, faisant ainsi l'objet de diverses définitions. D'après [Albino et al., 2015], les 'Smart Cities' sont des villes high-tech intensives et avancées qui relient les gens, l'information et les éléments de la ville en utilisant les nouvelles technologies afin de créer une ville durable, plus verte, un commerce compétitif et innovant, et une meilleure qualité de vie. Quant à [Neirotti et al., 2014], le concept de ville intelligente est vu comme un moyen d'améliorer la qualité de vie des citoyens, et prend de plus en plus d'importance dans les agendas des décideurs politiques. Cependant, il n'existe pas de définition partagée de la ville intelligente et il est difficile d'identifier des tendances communes. Plusieurs autres définitions peuvent être trouvées dans [Al Nuaimi et al., 2015] qui a fait une étude sur les cas d'applications des données massives pour les villes intelligentes.

Cependant, les villes intelligentes sont identifiées par un certains nombre de caractéristiques, permettant de les classer en fonction de leur niveau de développement. C'est dans ce sens que des études telles que [Nam and Pardo, 2011, Batty et al., 2012] ont défini trois composantes principales qui permettent de qualifier une ville d'intelligente. C'est trois composantes sont : la technologie, l'humanité, et les institutions. D'après les analyses de [Nam and Pardo, 2011, Batty et al., 2012], la technologie se résume essentiellement aux outils matériels et logiciels, à partir desquels nous pouvons distinguer les concepts tels que :

- ville digitale : qui se définit comme étant un environnement qui regroupe des services innovants permettant une mise en relation étroite (communication, partage d'informations, accès aux services, etc) entre les habitants partout et à tout moment ;
- ville d'information : vue comme un ensemble de plusieurs sources d'information en temps réel et avec des systèmes de collecte automatisés ;
- ville d'intelligence : ou encore cadre de connaissance, d'apprentissage, et de créativité ;
- ville omniprésente : un environnement permettant de mettre en relation les hommes et les services multiples et cela en tout temps et en tout lieu.

Quant à la composante humaine, [Nam and Pardo, 2011, Batty et al., 2012] la déclinent en plusieurs concepts comme la précédente, et nous avons par exemple :

- ville créative : ou une ville qui regroupe toutes formes de créativité dans le cadre social, culturel, artistique, etc.
- ville d'apprentissage : disposant des cadres d'apprentissage continu et compétitif avec une main d'oeuvre abondante et valorisante ;

— ville de connaissance : faisant référence à tous les processus de l'innovation et de l'intelligence économique ;

Pour finir, les auteurs dans [Nam and Pardo, 2011, Batty et al., 2012] abordent la composante institutionnelle, qui fait référence à un suivi de l'évolution des nouvelles technologiques sur le plan juridique, avec la mise en place de lois permettant de réguler les activités au sein de la ville et l'usage des technologies naissantes.

Une ville intelligente est donc une ville qui a pour objectif d'offrir un cadre de vie idéal et des services de qualités à ses habitants, grâce à l'utilisation des nouvelles technologies. Pour atteindre ces objectifs, elle met en place des solutions basées sur des réseaux de capteurs et des services numériques permettant de faciliter l'accès à des services qui entrent dans le quotidien de chaque habitant. Pour ce faire, plusieurs services sont proposés dans différents domaines. Nous pouvons citer par exemple, les dispositifs mis en place pour faciliter l'accès aux moyens de transport, pour une gestion intelligente et efficace des ressources naturelles comme l'eau, pour l'amélioration de la vie de la population dans son ensemble et aussi dans les domaines tels que la santé, la culture, l'éducation, la restauration, la habitation, etc.

C'est dans cette démarche, que la Métropole de Montpellier investit dans des projets innovants dans le but d'offrir une vie de meilleure qualité à ses habitants. À travers le projet ÉcoCité [1], la ville de Montpellier a vu naître un grand nombre de projets dans divers domaines. Le projet ÉcoCité est une initiative ouverte et partenariale [Métropole, 2017] ayant un triple objectif qui est :

— imaginer de nouveaux services urbains ;
— stimuler l'économie par l'innovation ;
— créer de la valeur économique, sociale et sociétale.

L'expérimentation du projet Écocité (voir Figure 3) [2] s'étend sur 2500 ha, dont 40 ha sont dédiés à l'innovation du quartier Eurêka pour favoriser le bien vieillir, 32 000 m2 au sein de l'îlot de la Mantilla pour innover dans la production, la gestion et l'utilisation de l'énergie. Le projet couvre 12 champs d'application qui sont entre autres, l'habitat, la santé, l'énergie, les déchets, la mobilité, la logistique urbaine, le tourisme, l'eau, la culture, la gestion des risques,la participation citoyenne et la cohésion sociale.

Aujourd'hui, nous constatons plus de 45 expérimentations en cours et à venir sur le territoire de la Métropole, avec plus 2,4 M€ investis par les partenaires publics et privés sur les grands projets en 2016-2017. Nous pouvons citer des projets tels que *ville en alerte*, un dispositif qui intègre une prévision météorologique, des "cartes risques" prédictives, un suivi des crues ainsi qu'un outil de collaboration entre les acteurs de la gestion de crise.

1. https ://www.entreprendre-montpellier.com/fr/choisir-montpellier/experimentez-la-cite-intelligente
2. http ://www.ecocites.logement.gouv.fr/ecocite-montpellier-mediterranee-metropole-r83.html

L'application mobile *Tam* qui offre aux usagers des transports des solutions de stationnement, facilite les déplacements dans la métropole, avec la carte d'abonnement unique *Sésame Emma* qui permet l'accès au réseau de mobilité : tramway, bus, parking, Vélomagg', et Auto partage. Dans le secteur énergétique, nous pouvons citer la *Mantilla*, qui est équipée d'un système d'échange de données numériques (Smart network) et qui permet aux résidents de piloter leurs consommations d'eau et d'électricité. Le dispositif facilite aussi la gestion énergétique de l'îlot, alimenté par la centrale biomasse de tri-génération [Métropole, 2017]. Dans la section suivante, nous abordons l'hétérogénéité des données, ainsi que les potentielles sources de données dans les smart cities.

FIGURE 3 – Carte du projet ÉcoCité de Montpellier Méditerranée Métropole (3M)

2.1.2 Données et Villes Intelligentes

Nous avons vu dans le Chapitre 1, que les grandes villes sont des potentielles sources de données massives et hétérogènes. Nous notons également que l'hétérogénéité des données se caractérise sous diverses formes [Wang, 2017], et sa définition est valable dans le contexte d'application où elle est utilisée.

Pour ce faire, nous proposons des définitions plus larges, prenant en compte la majorité des cas des données hétérogènes. Nous distinguons principalement les quatre grandes caractéristiques précédemment mentionnées dans le Chapitre 1, pouvant contribuer à catégoriser les données hétérogènes. Il s'agit essentiellement de la source des données, la sémantique véhiculée par ces données, de leurs structures, et de leurs natures.

— Les données hétérogènes par leur source : dans ce cas précis, nous faisons référence

aux données qui peuvent être de même nature ou de nature différente, mais en provenance de sources diversifiées. Nous prenons par exemple le cas des données d'une entreprise, où nous avons les données clients, les données de vente, de logistiques, etc.

— Les données hétérogènes par la sémantique : dans le cadre des données textuelles, une expression ou un mot peut avoir plusieurs sens selon le contexte, de même qu'une image peut avoir plusieurs interprétations ;

— Les données hétérogènes par leur structure : les données peuvent être stockées différemment et sous plusieurs formes, qu'il s'agisse de données de même nature ou pas. Nous distinguons principalement trois formes de stockage qui sont :
  - les données structurées : ce sont les données stockées généralement dans des bases de données avec des schémas de stockage bien définis ;
  - les données semi-structurées : ce sont les données stockées suivant une certaine ontologie, ou encore des langages de description. Il s'agit généralement des stockages suivant des structures comme XML (Extensible Markup Language), OWL (Web Ontology Language), ou RDF (Resource Description Framework) utilisées pour la gestion et l'échange d'information ;
  - les données non-structurées : qui regroupent l'ensemble des données textuelles libres que nous pouvons rencontrer. Nous avons les rapports, les pages web, des journaux, ou des notes recueillies manuellement. Dans ce cas de figure, des nuances peuvent être faites, suivant que ces données suivent une certaine structure, comme les articles (ex : abstract, introduction, méthodologie, conclusion, tec.), ou qu'il s'agisse des données textuelles comme les tweets, les discussions sur forum, etc.

— Les données hétérogènes par leur nature : nous avons les données textuelles, les données images (photos, images satellites, médicales, etc.) et les vidéos. Chacune d'elles étant une représentation caractéristique qui permet de véhiculer une information. Dans d'autres cas d'étude, on pourrait par exemple s'intéresser uniquement à des données images, mais de formats (png, tif, jpg, jpeg, etc.) différents ou hétérogènes.

2.2 Architecture AIDMOIT

Afin de répondre aux objectifs cités dans le Chapitre 1, Section 1.3, nous proposons l'architecture générale qui est illustrée par la Figure 4. L'acronyme AIDMOIT (Analyse Intégration Données Massive Observation Intelligente Territoire) correspond à l'intitulé de notre projet. L'architecture est composée principalement de quatre (04) parties à savoir : la collecte de données, le stockage de données, la mise en relation et analyse, et enfin la visualisation des résultats. Nous allons successivement décrire ces parties dans les sections

suivantes.

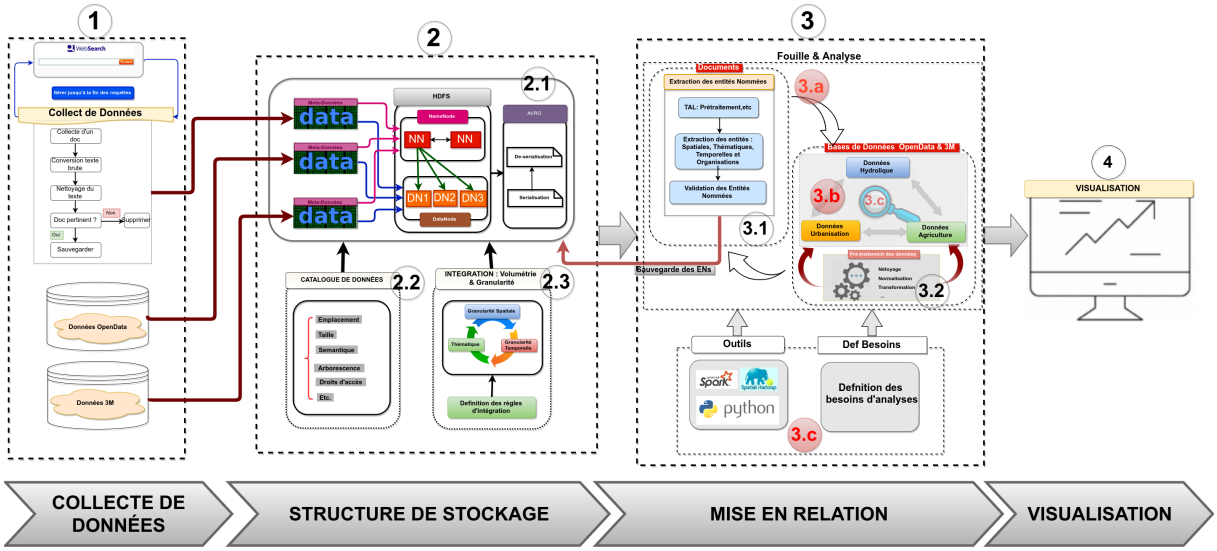


FIGURE 4 – Architecture Générale de AIDMOIT

2.2.1 Collecte de Données

Cette partie représente l'étape de collecte et de regroupement des données de la ville intelligente concernée. En fonction des solutions et services offerts, nous pouvons distinguer des données massives et hétérogènes telles que, des images satellite, des données recueillies par des capteurs (ex : transport, météo, etc.), des données textuelles (ex : pages web, rapports électroniques et/ou manuscrits, des Tweets, etc.), ou des bases de données (publiques ou privées) existantes, etc. En fonction de la source et de la disponibilité des données, nous faisons ressortir principalement deux méthodes de collecte.

La première consiste à identifier les données par un processus d'inventaire. Elle prend en compte, les données dont les sources sont connues d'avance (ex : sites web ou services spécifiques). Nous pouvons citer ici, les bases de données existantes, qu'elles soient publiques ou privées, disponibles en local et/ou sur internet (ex : données en OpenData).

Quant à la seconde, elle concerne les données disponibles sur le Web, et qui doivent faire l'objet de recherches approfondies. Elles sont obtenues à travers des requêtes en utilisant des mots clés spécifiques. Nous pouvons distinguer à ce niveau, les données textuelles, comme les Tweets, des forums, des blogs, des rapports ou mémoires d'études, etc.

Nous illustrons cette partie de l'étude avec des données de 3M, qui sont essentiellement des données textuelles obtenues du Web, des données récupérées auprès de 3M qui ne sont pas ouvertes au grand public, et des données ouvertes et disponibles sur le site de l'OpenData3M [3] ou sur d'autres références.

3. https ://data.montpellier3m.fr/

La première méthode de collecte est traitée dans le Chapitre 5 , Section 5.4.2 , où nous automatisons le processus de collecte à partir des données de 3M recensées dans le Tableau A.1. Elle nous a permis de récupérer et de stocker directement ces données dans le lac de données que nous décrivons dans le Chapitre 5.

Pour ce qui est de la seconde méthode, nous l'avons illustrée à partir de la collecte de données textuelles sur les différentes thématiques de notre étude, l'agriculture, l'hydrologie et l'urbanisation, mentionnées dans le Chapitre 1 , Section 1.2, en proposant un protocole de collecte de données textuelles dans le Chapitre 3. Ce protocole vise à récupérer des données sur le Web, de sources et de types divers et qui puissent être propre à une thématique donnée, avec le moins de bruits (données non pertinentes) possible.

2.2.2 Structure de Stockage

Nous proposons une structure de stockage adaptée pour la gestion des données massives qui sont de nature hétérogène. Parmi ces données, nous avons celles citées dans la Section 2.2.1, qui sont composées de données textuelles, d'images satellite, et de données vectorielles ou shapfiles. Cette structure doit non seulement permettre de stocker les données de grand volume, mais aussi doit permettre de les stocker sans les détériorer. Dans cette partie, trois défis majeurs se présentent à nous. Tout d'abord, il s'agit de pouvoir stocker l'ensemble des données indépendamment de leur hétérogénéité. Ensuite, nous devons tenir compte, du fait que plusieurs utilisateurs doivent pouvoir accéder à ces données, quelque soit leurs besoins, et enfin, chaque utilisateur devra être libre de choisir ou de définir le type de traitement qu'il souhaite réaliser sur ces données.

Nous proposons dans le Chapitre 5 , un lac de données spatiales, accompagné d'un système de gestion de méta-données. En effet, il s'agit d'une structure de stockage qui permet de répondre aux trois défis que nous venons d'évoquer dans le paragraphe précédent. Les données (textes, images, vecteurs, etc.) stockées dans le lac feront l'objet d'une analyse conjointe, une problématique que nous abordons dans la section suivante.

2.2.3 Intégration et Mise en Relation

Cette partie constitue notre troisième contribution. Composée de plusieurs sous parties, elle consiste à mettre en relation l'ensemble des données collectées et stockées dans le lac de données, Chapitre 5, dans le but de faire émerger des connaissances. Nous identifions trois défis majeurs, qui sont : 1) extraire des indicateurs pertinents dans chacune des données, 2) trouver les relations spatio-temporelles et thématiques entre ces différents indicateurs, et 3) montrer que des informations capitales peuvent être obtenues en mettant en relation toutes ces données.

Nous traitons les solutions proposées pour cette partie dans le Chapitre 6. Les méthodes développées consistent à enrichir les informations extraites d'une part des données

textuelles, et d'autre part, des séries d'images satellite. Les approches proposées permettent de décrire et/ou d'enrichir des évènements spatio-temporels qui sont observables à partir des séries d'images temporelles, en s'appuyant sur des corpus de données textuelles obtenus du Web à l'aide du protocole de collecte proposé dans le Chapitre 3.

2.2.4 Visualisation

Elle consistera à proposer tout d'abord, des représentations visuelles accessibles à tous, sur les analyses obtenues dans l'étape précédente, Section 2.2.3. Et aussi à proposer des stratégies de restitution ou de visualisation d'analyses sur les lacs de données, afin de faciliter l'interprétation et l'accès des connaissances.

Cette dernière section n'est pas abordée dans sa globalité dans notre étude, mais garde toute son importance. Néanmoins, nous proposons un cas d'illustration sur le protocole de collecte de données textuelles, Chapitre 3, pour la Métropole de Montpellier.

2.3 Conclusion

Dans ce chapitre, nous faisons ressortir l'impact des villes intelligentes sur l'évolution des masses de données et la complexité liée à leur gestion et exploitation. Nous avons aussi abordé le concept de ville intelligente dans le contexte de la métropole de Montpellier qui est le projet ÉcoCité.

Pour terminer, nous avons présenté dans les grandes lignes, la solution architecturale que nous avons mis en place pour répondre aux défis de notre étude. Le protocole de collecte pour répondre aux défis liés à la collecte de corpus thématiques sur le plan spatio-temporel avec le plus de précision possible, le lac de données spatiales pour répondre aux défis liés au stockage des données fortement hétérogènes sans les détériorer, et les méthodes et techniques de mise en relation pour répondre aux défis liés à l'extraction et à la mise en relation des données sur le plan spatio-temporel pour faire émerger les connaissances. Le prochain chapitre sera consacré à l'étude du protocole de collecte proposé pour la constitution de corpus thématiques.

Chapitre 3

Protocole de collecte de données textuelles

Pour avoir un processus de suivi d'évènements d'un territoire, il est nécessaire de disposer des données en quantités et en qualités sur ce dernier. Dans ce chapitre, nous abordons la problématique de collecte de données textuelles qui constituent une partie des données que nous avons déjà abordées dans le Chapitre 2, Section 2.2.

Nous proposons un protocole de collecte automatique qui a pour but, de constituer des corpus sur des thématiques (voir Chapitre 1, Section 1.2) bien précises à partir du Web. Ces corpus de données thématiques sont enrichis et mis à jour de façon régulière grâce à l'approche développée.

Ce chapitre est organisé en quatre sections. La Section 3.1 est une introduction générale du chapitre. La Section 3.2 propose un état de l'art connexe aux problématiques de notre étude. Nous décrivons, ensuite, l'approche méthodologique de la solution que nous proposons en Section 3.3. Cette approche est ensuite appliquée sur une étude de cas en Section 3.4. Puis une conclusion générale clôturera ce chapitre en Section 3.5.

3.1 Introduction

Introduit dans les années 2001 avec CorpusBuilder [Ghani et al., 2001], la constitution de corpus à partir du Web consiste à effectuer des requêtes sur le Web en utilisant des mots-clés afin de construire des collections de documents en grande quantité.

Par définition, nous considérons un corpus thématique comme étant une collection de données textuelles recueillies à partir de pages Web et relatives à une thématique bien précise. Par données textuelles, nous incluons les pages Web comme les articles de presse, les forums, ou encore des données provenant des réseaux sociaux comme Twitter. La thématique quant à elle ne prend en compte que des données qui abordent un sujet bien précis, comme la santé, le transport, l'urbanisation, l'éducation, etc.

Inspirée des approches proposées dans [Baroni and Bernardini, 2004, Sharoff, 2006], la solution mise en place dans le cadre de cette thèse est basée sur le fait qu'en partant d'un petit ensemble de données, il est possible de l'enrichir en ayant recours à d'autres sources de données.

Ces approches peuvent être résumées en deux points. Le premier consiste à utiliser un ensemble réduit de mots-clés permettant de caractériser une thématique, pour requêter des documents sur le Web. Les documents obtenus sont utilisés pour constituer un corpus. Le second point, consiste à extraire des mots-clés de ce corpus, pour avoir un ensemble plus grand que le premier, et qui servira à son tour à récupérer des documents sur le Web pour constituer un corpus plus grand, et ainsi de suite. En plus de cette approche, nous avons proposé une chaîne de traitement complète qui prend en compte la pertinence des termes ou mots-clés utilisés pour les requêtes, et celle des documents obtenus lors de la phase de collecte, permettant de limiter le bruit.

3.2 État de l'art

Les genres ou catégories textuelles, des ressources indexées par les moteurs de recherche sont très variés : blog, publicité, rapport officiel, articles scientifiques etc. Cette diversité offre des points de vue et des terminologies différents sur des mêmes sujets d'étude. Plusieurs travaux proposent différentes hiérarchies de genre [Vidulin et al., 2009, Santini, 2011] afin de faciliter la classification de nouvelles ressources. Dans [Madjarov et al., 2019], les auteurs exploitent ces hiérarchies pour renforcer les capacités d'un moteur de recherche à proposer des résultats pertinents selon les genres souhaités. L'évaluation de la pertinence des ressources collectées sur le web est bien entendu un enjeu majeur. Une approche classique vise à calculer les co-occurrences de termes entre la requête et les documents rapatriés pour ensuite classer les résultats en fonction du poids du terme dans le document, en utilisant, notamment des mesures tels que TF-IDF [Aizawa, 2003]. Cette

méthode a cependant des limites. Cette approche par mot clé ne peut retranscrire la proximité sémantique entre deux termes comme par exemple, vélo et mobilité douce.

Pour remédier ce problème, [Schaeffer. et al., 2020] proposent de filtrer les ressources pertinentes en calculant la proximité sémantique, via le modèle de prolongement lexical word2vec [Mikolov et al., 2013], entre des termes qui composent le document et un thésaurus thématique. Dans un contexte de surveillance en épidémiologie animale, [Arsevska. et al., 2016] montrent que la construction de requête avec deux mots clés, symptôme et animal, permet de récolter des ressources avec davantage de pertinence. Il est possible, enfin, de combiner ces deux dernières approches, prolongement lexical et requête multi-termes, pour enrichir la sémantique du terme par l'ajout d'un contexte apporté par la phrase, ou le groupe de mots, qui le contient en utilisant le modèle BERT [Devlin et al., 2018].

Dans l'ensemble, nous notons trois techniques utilisées pour la constitution de corpus. La première est une constitution manuelle, la deuxième est basée sur une collecte à partir du web et la dernière est une combinaison des deux. Dans [Wu et al., 2012], les auteurs ont travaillé sur un corpus construit manuellement à partir des notes cliniques de la Clinique Mayo entre le 1er janvier 2001 et le 31 décembre 2010. Pour ce qui est de la constitution de corpus thématique, nous pouvons distinguer deux principales approches. La première consiste à utiliser un petit ensemble de mots clés caractéristiques du domaine pour des requêtes automatisées, les résultats sont ensuite utilisés pour étendre le corpus et ainsi de suite. Nous pouvons citer les travaux comme BootCaT [Baroni and Bernardini, 2004] et [Sharoff, 2006] qui ont également utilisé la même approche avec une liste de 500 mots pour former un corpus généralisé. Dans [Kilgarriff et al., 2010], les auteurs ont également utilisé la même méthode pour construire un corpus des principales langues du monde. La seconde approche repose sur la structure des pages Web qui est le DOM (Document Object Model). Elle consiste à se focaliser sur certaines parties de la page qui sont sensées contenir l'essentiel de l'information. À titre d'exemple, [Lin and Ho, 2002] divisent chaque page Web en blocs et identifient la pertinence du contenu de chacun d'eux.

L'un des inconvénients que nous avons pu noter sur les approches existantes est la généralisation. En effet, chaque approche aborde le problème en prenant en compte des documents spécifiques. Dans le cas où nous souhaitons construire des corpus sans tenir compte du type de document, c'est-à-dire de pages Web de type HTML, des documents PDF, ou des données de Twitter, il devient difficile de trouver une solution unique pour cela.

Nous proposons dans ce chapitre, la mise en place d'un ensemble de techniques qui reposent sur des stratégies d'extraction de concepts (officiels et non officiels) et de constitution automatique de corpus thématiques ancrés sur le plan spatio-temporel, provenant de sources diversifiées, validées par des mesures de similarité sémantique.

3.3 Approche méthodologique

L'approche méthodologique que nous proposons pour la constitution des corpus thématiques porte sur deux grands points. Le premier est la constitution d'un vocabulaire de concepts thématiques. Il consiste à former automatiquement un ensemble de termes relatifs à une thématique bien précise. Le second concerne la collecte de documents en utilisant le vocabulaire de concepts thématiques obtenu dans la première étape pour former les corpus.

3.3.1 Constitution des vocabulaires de concepts thématiques

Le processus pour la constitution de vocabulaire de concepts thématiques est basé sur le principe de construction de corpus proposé dans [Kilgarriff and Grefenstette, 2003, Kilgarriff et al., 2010, Sharoff, 2006]. Il consiste à utiliser un ensemble de graines ou une liste de termes d'un domaine caractéristique pour requêter des documents sur le web. Les résultats sont ensuite utilisés pour étendre la liste des termes et donc le corpus, et ainsi de suite. L'approche proposée se décline en deux phases : la constitution de l'ensemble de termes graines noté TG, et la constitution du vocabulaire de concepts noté VC.

3.3.1.1 Choix des termes graines *TG*

La proposition de l'ensemble TG comprend deux étapes (voir Figure 5). La première consiste à extraire un ensemble de termes à partir d'un mini corpus obtenu de Wikipédia[1] en utilisant l'outil BioTex de [Lossio-Ventura et al., 2014a]. Le choix de BioTex se justifie par ses nombreux avantages que nous décrivons dans le Chapitre 4, Section 4.3.1. Ce corpus est construit en utilisant un mot clé thématique (ex : urbanisation), et les vingt premières pages renvoyées par Wikipédia sont utilisées pour former le corpus. Les ressources de Wikipédia sont riches et proposent des définitions étymologiques sur les thématiques, notre terminologie initiale se nourrit de cette quantité d'information. Bien que paramétrable, nous avons fixé au nombre de 20, les pages récoltées de Wikipédia, puisque, d'après nos différentes expérimentations, le nombre de termes extrait étaient suffisant.

Pour BioTex, nous avons principalement utilisé deux mesures statistiques qui sont C_Value et F-TFIDF-C_M. C_Value donne de l'importance aux termes apparaissant plusieurs fois dans le même document et aux termes composés permettant de valoriser les expressions. F-TFIDF-C_M est une extension de C_Value et de TF-IDF proposée dans BioTex. Elle représente la moyenne harmonique de C_Value et TF-IDF, et permet de classer les termes en fonction de leur pertinence vis-à-vis du document en prenant en compte l'ensemble du corpus. F-TFIDF-C_M présente l'avantage d'utiliser toutes les valeurs de la distribution et de réduire le bruit tout au long du processus d'extraction.

1. `https://fr.wikipedia.org/`

C-Value et F-TFIDF-C_M sont complémentaires, car la première favorise l'extraction des termes composés pertinents et la seconde privilégie les termes discriminants. Discriminants dans le sens ou elle permet de capturer les termes pertinents qui majoritairement ne sont pas reconnus par C-Value et TF-IDF prise individuellement.

La seconde étape consiste à faire choisir un ensemble de termes par des experts du domaine parmi les termes récupérés à partir de Wikipédia. L'ensemble de termes ainsi constitué forme la liste de termes experts que nous appelons ici *TG*.

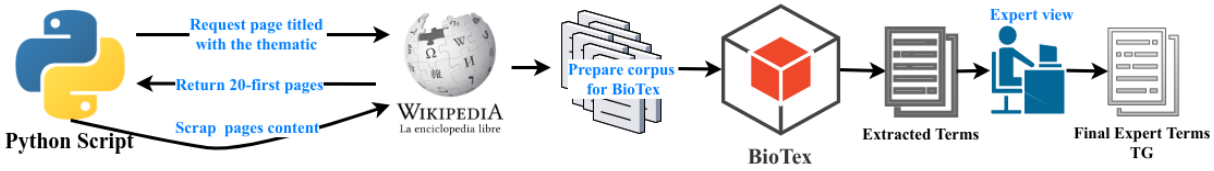


FIGURE 5 – Choix de termes initiaux - TG

3.3.1.2 Constitution du vocabulaire de concepts *VC*

La constitution du vocabulaire de concepts est basée sur une première étape qui consiste à utiliser la liste *TG* comme une liste de graines permettant de collecter un grand nombre de documents via un moteur de recherche comme par exemple Google. L'objectif est d'obtenir des variantes de termes pour faire évoluer la terminologie de TG, afin de récolter des documents parlant de la thématique mais avec des mots et des termes qui ne seraient pas présents dans Wikipédia. Ensuite, grâce à BioTex, le corpus obtenu est utilisé pour extraire un grand ensemble de termes noté TB (termes BioTex), à partir duquel, nous allons réduire le bruit éventuel pour former le vocabulaire *VC*. Pour éviter de prendre en considération les termes généraux ou les termes qui ne permettent pas de définir la thématique étudiée, nous avons mis en place un filtre basé sur une approche sémantique en utilisant DistilBERT [Sanh et al., 2019], reconnue pour l'évaluation de la Similarité Sémantique Textuelle (STS). La valeur de similarité varie entre -1 pour de très faible similarité et 1 lorsque les termes sont très proches ou identiques.

La mesure sémantique est évaluée à partir 1) des termes extraits TB qui sont les nouveaux termes, et 2) les termes graines ou experts TG. Pour chaque terme de TB, nous calculons sa similarité sémantique avec chacun des termes de l'ensemble TG. Ce qui nous donne un nombre de valeurs de similarité égal au nombre de termes dans l'ensemble TG pour chaque terme t de TB. Pour obtenir la valeur de similarité finale pour un terme, nous faisons la moyenne des TOP@n (n correspond aux n termes les plus pertinents) en tenant compte des plus grandes valeurs de similarités pour garder une certaine exhaustivité vis-à-vis de l'ensemble des termes experts. À la fin du processus, nous obtenons un nouvel ensemble ordonné de termes, constitué des mêmes termes de l'ensemble TB, mais avec des scores de pertinence différents. Les vocabulaires de concepts seront par la suite sélectionnés

à partir de ces ensembles de termes pour chaque thématique, en partant de ceux disposant d'un plus grand score. La Figure 6 décrit les différentes étapes pour la construction de vocabulaire de concepts thématiques dans son ensemble et le calcul de similarité entre termes est illustré sur la Figure 7.

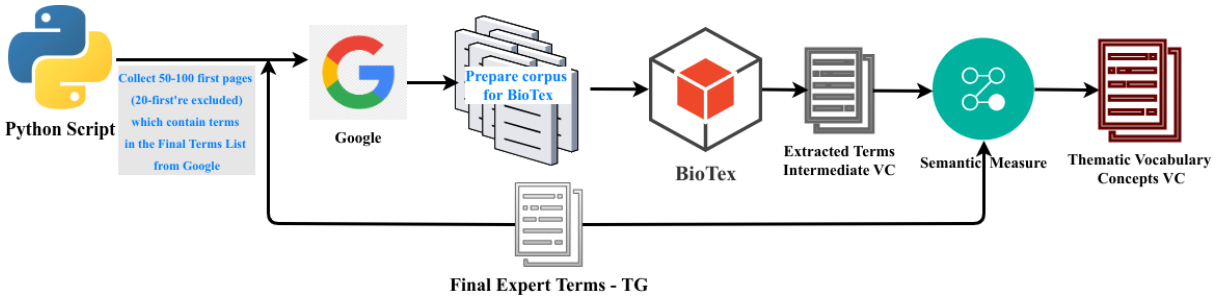


FIGURE 6 – Constitution de vocabulaire de concepts thématiques - VC

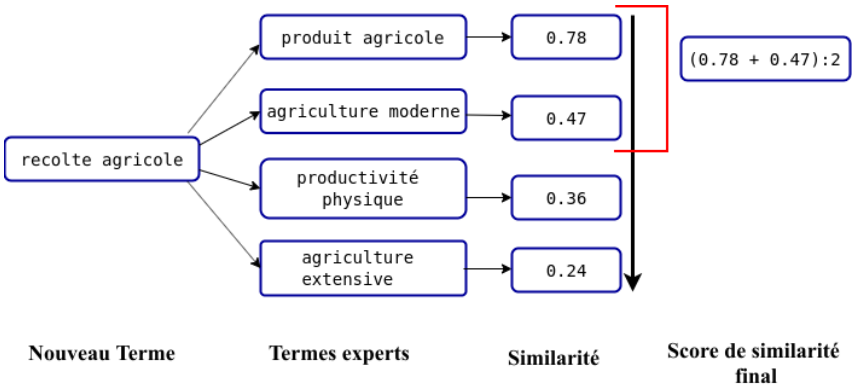


FIGURE 7 – Exemple de calcul de similarité sémantique - score final calculé en faisant la moyenne des 2 plus grandes valeurs

3.3.2 Constitution des corpus thématiques

3.3.2.1 Collecte de documents

Une fois le vocabulaire de concepts thématiques construit, nous avons pour chaque thématique, un ensemble de concepts permettant de la décrire au mieux. La phase de collecte (voir Figure 9) consiste à utiliser un module, qui prend en entrée un vocabulaire de concepts *VC* et une emprise spatiale, pour retourner des documents en relation avec la thématique concernée pour former un corpus de taille importante. Nous pouvons ainsi constituer pour chaque thématique, un corpus le plus en cohérence possible avec son contenu. La prochaine étape permet de filtrer les documents les moins pertinents pour chaque corpus thématique.

3.3.2.2 Évaluation automatique des documents par mesure sémantique

Dans le but de ne considérer que les documents pertinents pour chaque thématique, nous avons introduit une évaluation par mesure de similarité. Identique à la précédente, l'évaluation des documents d'une thématique se fait entre un document collecté par rapport à l'ensemble des termes de son vocabulaire étendu de concepts thématiques VEC (voir Figure 8) obtenu avec WordNet [Miller, 1998]. L'objectif principal de cette étape est de disposer 1) d'une couverture plus large sur les termes en introduisant leur synonyme grâce à WordNet, et 2) d'obtenir des termes moins experts et/ou administratifs qui sont importants pour l'évaluation des documents de la sphère sociétale tels que les blogs, les annonces, etc. Cette technique permet de prendre en compte ces synonymes lors de l'évaluation des documents. Comme décrite précédemment, la valeur de similarité avec DistilBERT varie entre -1 et 1. Un document de la thématique urbanisation sera évalué avec le vocabulaire étendu de concepts de la thématique urbanisation. Plus la valeur de similarité est forte, plus le document sera considéré comme pertinent vis-à-vis de la thématique, et vice-versa.

Le processus pour l'évaluation des documents est illustré sur la Figure 9. Afin d'éliminer les documents qui ne sont pas pertinents, nous définissons lors des expérimentations, une valeur seuil à partir de laquelle, nous décidons de garder le document ou de l'exclure du corpus. Lorsque le document obtenu est pertinent, une fiche de méta-données contenant les informations sur le document est automatiquement générée avant d'être stockée dans un lac de données [Kafando et al., 2020b]. Les différentes étapes, illustrées sur la Figure 10, constituent le protocole de collecte proposé.

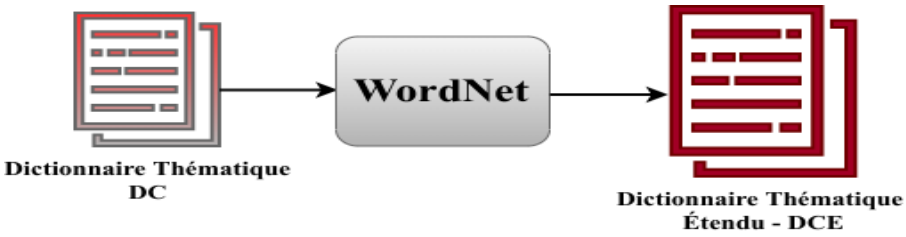


FIGURE 8 – Constitution du vocabulaire étendu de concepts thématiques avec WordNet

La fiche de méta-données d'un document est obtenue lors de la sauvegarde de ce dernier. En effet, un pré-traitement est appliqué à chaque document, lors duquel les méta-données sont extraites. Celles-ci sont de deux types : celles qui sont propres au document, tels que le nom, le titre, la date de publication, etc. et celles obtenues en effectuant des analyses sur le document. Nous pouvons citer, la pertinence du document vis-à-vis de la thématique spécifiée pour constituer le corpus, et la liste des entités nommées spatiales extraites du titre. D'après [Kergosien et al., 2015], les entités nommées spatiales (ENS ou SNE) peuvent être classées selon les entités nommées spatiales Absolues (ENS_A) ou Relatives (ENS_R). Une ENS est dite absolue, lorsqu'elle désigne un emplacement,

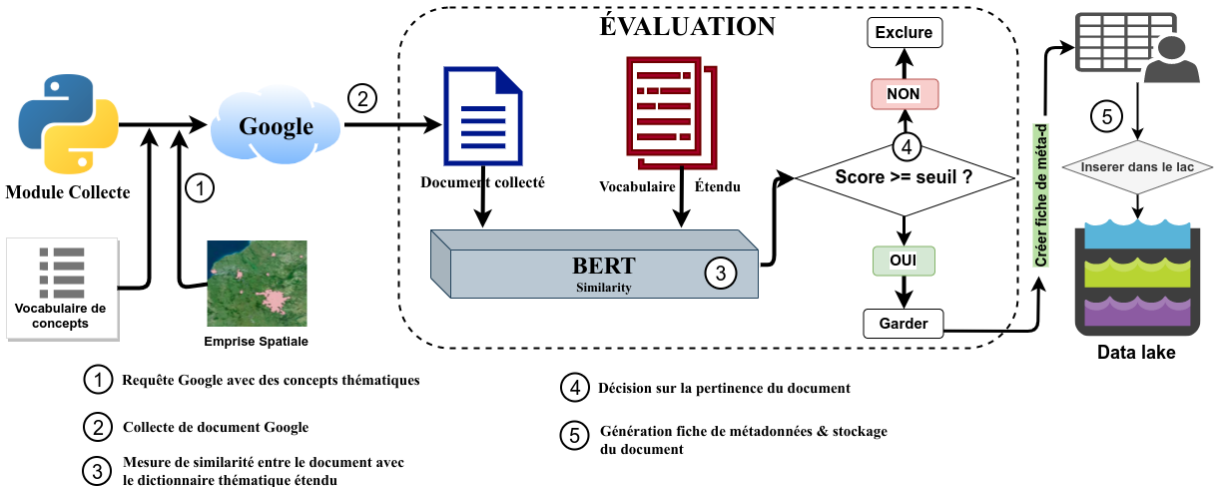


FIGURE 9 – Collecte et évaluation des documents

pouvant être identifié avec des coordonnées bien précises (ex : Montpellier-France), et relative lorsqu'elle fait appel à des indicateurs de type orientation avec des estimations telles que Centre de Montpellier, Nord-Ouest de Paris, aux alentours de la gare St Roch, etc. De façon plus précise, une fiche de méta-données d'un document est constituée des éléments suivants :

- *Name* : nom de base utilisé pour le document au niveau de l'Url ;
- *Type* : type du document, à savoir s'il s'agit d'un document html, pdf ou autre ;
- *Source* : Url complète du document, permettant de le retrouver sur internet ;
- *Request* : combinaison de mots clés utilisée lors de la recherche sur Google ;
- *Text* : contenu ou le corps du document à proprement dit ;
- *Title* : titre ou intitulé du document ;
- *Post_date* : date de création par défaut, si elle n'est pas disponible, la date de publication est récupérée à la place ;
- *City* : nom de la ville ou de l'entité spatiale définit lors de la requête sur Google ;
- *SNE* : ou encore spatial named entities, regroupe l'ensemble des entités spatiales identifiées dans le titre du document ;
- *Pertinence* : pertinence du document vis-à-vis de sa thématique ;
- *Thématique* : thématique à laquelle appartient le document.

3.3.3 Protocole d'évaluation

Nous proposons dans cette partie, un processus d'évaluation permettant de vérifier la pertinence du protocole de collecte. Le processus d'évaluation est construit à deux niveaux et est basé sur des avis, recueillis auprès des experts de chaque thématique. La première évaluation est faite sur la liste de termes experts constituée selon le protocole décrit, Section 3.3.1.1 , et la seconde sur les corpus thématiques obtenus selon le protocole décrit, la Section 3.3.2.

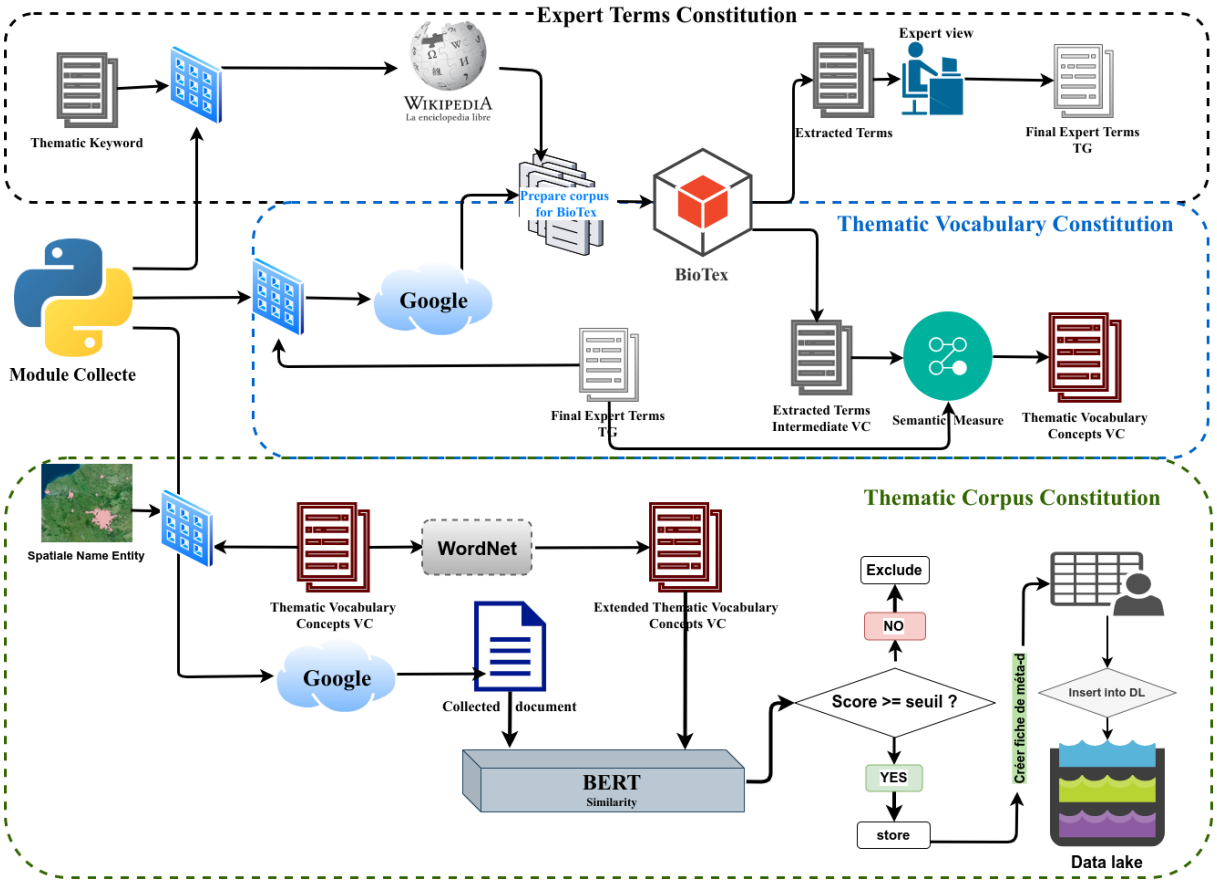


FIGURE 10 – Les étapes du protocole de collecte

3.3.3.1 Évaluation des termes experts

Pour chaque thématique, l'analyse des termes experts consiste à attribuer une valeur ou un score de pertinence à chacun des termes initialement proposés. Ce score varie entre 0 et 5. La valeur 0 indique que le terme n'est pas pertinent ou n'est pas adapté à la thématique, tandis que 5 indique qu'il est parfaitement adapté au sens sémantique.

Chaque liste de termes est soumise à deux évaluateurs experts. La valeur du score final d'un terme est obtenue en faisant la moyenne des deux valeurs. Ensuite, nous catégorisons pour chaque thématique, les valeurs de scores en trois intervalles. Le premier **Très_Adapté** = [**4,5**], le deuxième **Adapté** =]**2,3**], et le troisième **Pas_Adapté** = [**0,2**]. Ces intervalles permettent de classer les termes respectivement, à savoir s'ils sont très adaptés, moyennement adaptés ou pas du tout adaptés vis-à-vis de la thématique. À la fin de cette étape de regroupement, seuls les termes présent dans les intervalles **Très_Adapté** et **Adapté** seront pris en compte et considérés pour la collecte de documents qui serviront pour la constitution des vocabulaires de concepts.

3.3.3.2 Évaluation des corpus

Comme celle effectuée avec les termes experts, cette étape consiste à avoir une vue globale sur la qualité des documents, tout d'abord de leur contenu, et dans un second temps, de leur richesse en diversité en les regroupant par catégorie et par couverture spatiale et temporelle. Les évaluations sont faites par thématiques, et chaque document est évalué par deux experts du domaine ou de la thématique. Le score final de chaque document est obtenu en effectuant la moyenne des notes comme pour les termes experts. Les documents de chaque thématique sont respectivement regroupés en fonction des trois intervalles précédemment définis.

Une fois cette étape réalisée, nous soumettons à nouveau ces documents aux experts, afin qu'ils puissent les catégoriser par type, et d'évaluer s'ils sont bien définis spatialement et/ou temporellement. Pour ce faire, nous avons défini au préalable 3 types de document, à savoir, les documents de type **académique, administratif**, et **sociétal**. En plus de cela, nous avons défini une catégorie **hors sujet**, prenant en compte, les documents ne traitant pas la thématique correspondante et ceux parlant de publicités, d'offres d'emploi, et les annonces immobilières. Pour chacune des catégories nous faisons ressortir leur emprise spatiale et temporelle. La Table 3.1 présente les critères d'évaluation que nous avons proposés.

Type	- Académique (thèse, rapport, article) - Administratif (rapport d'étude, plan, etc.) - Société (article de presse, blog, etc.)
Spatialité	- Concerne la Métropole de Montpellier - Concerne la région de Occitanie en générale - Ne concerne ni la Métropole de Montpellier, ni Occitanie
Temporalité	- date de production inférieur à 1 an - date de production comprise entre 1 an et 5 ans - date de production supérieur à 5 ans
Hors sujet	- Publicités (offres de formations, d'emplois, annonces immobilières, etc.) - Hors contexte

TABLE 3.1 – Critères d'évaluation établis pour les corpus

Pour évaluer l'accord inter-évaluateurs, nous avons utilisé le coefficient de Pearson (voir Equation 3.1) [Benesty et al., 2009] d'une part, et d'autre part une représentation linéaire dans le but d'illustrer la distribution des valeurs attribuées par chaque paire lors de l'évaluation, 1) sur les termes experts, et 2) sur les corpus thématiques.

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \tag{3.1}$$

Le coefficient de Pearson varie entre -1 et +1, et détermine l'intensité de la relation linéaire entre les différentes évaluations.

— Une valeur de ρ égale à -1 ou à +1 indique l'existence d'une relation linéaire parfaite (fonctionnelle), négative ou positive entre les deux variables ;
— Le coefficient est nul ($\rho = 0$) lorsqu'il n'y a pas de relation linéaire entre les variables.

Dans la section suivante, nous présentons l'application de l'approche méthodologique proposée pour la collecte de données textuelles sur le territoire de la Métropole de Montpellier.

3.4 Cas d'application : 3M

L'objectif recherché de cette étude est de récupérer des documents sur le Web qui abordent des problématiques en relation avec le territoire de la Métropole de Montpellier Méditerranée (3M). Comme mentionné précédemment, nous avons choisi d'aborder les trois thématiques qui sont, l'**agriculture**, l'**urbanisation** et l'**hydrologie**.

3.4.1 Constitution de vocabulaires de concepts thématiques

La première étape est le choix de termes graines *TG* décrit dans la Section 3.3.1.1. La seconde étape consiste à soumettre cet ensemble de termes à un groupe d'experts dont le rôle est de sélectionner un sous-ensemble (liste finale de TG) de termes permettant de décrire au mieux la thématique.

Afin de constituer les mini-corpus via Wikipédia qui serviront pour les choix des listes TG, nous utilisons le libellé de la thématique comme mot clé principal de la recherche sur Wikipédia, soit 'urbanisation', 'agriculture', ou 'hydrologie'. Le mini-corpus ainsi constitué de chaque thématique est ensuite utilisé pour l'extraction des termes parmi lesquels seront sélectionnés les termes graines.

Termes requêtes	**Urbanisation**	**Agriculture**	**Hydrologie**
Taille Corpus (Nb docs)	20	20	20
Nombre de termes retenu	100	100	100

TABLE 3.2 – Constitution de la liste de termes graines TG

Urbanisation TG-1	**Agriculture TG-2**	**Hydrologie TG-3**
Politiques urbaines	Récolte agricole	Hydrogéologie
Aménagement territoriale	Agriculture	Eau pluviale
Aménagement foncier	Aménagement agricole	Mesure hydrologique
Aménagement touristique	Production agricole	Cycle de l'eau
Architectonique	Exploitation agricole	Réseau hydrologique
Territorialisation	Agriculture de Banlieue	Ouvrage hydraulique
Code de l'urbanisme	Biologie agricole	Alimentation en eau
Composition urbaine	Cartographie agricole	Bilan hydrique
Déplacement urbain	Climatologie	Cours d'eau
Développement durable	Développement agricole	Cours d'eau urbain

TABLE 3.3 – Extrait de termes experts par thématique

Les tableaux 3.2 et 3.3 présentent respectivement, les détails sur les données utilisées pour la constitution de la liste de termes graines TG, et les TOP@10 de la liste de termes graines TG de chaque thématique.

Une fois les termes graines obtenus, nous procédons à la constitution des vocabulaires de concepts *VC*. Pour chaque thématique, la liste TG est utilisée pour servir de mots clés aux requêtes pour constituer de corpus plus importants, sur le web, sans distinction de site. Ceci nous permet d'avoir une forte diversité des données tant dans leurs contenus que dans leurs sources. Nous avons récolté respectivement 105, 120 et 117 documents pour former les corpus Urbanisation, Agriculture et Hydrologie à partir desquels seront extraits les ensembles de termes que nous nommons termes Biotex (TB).

Afin de pouvoir constituer les vocabulaires de concepts thématiques finaux, nous avons utilisé la mesure sémantique proposée, Section 3.3.1.2 qui permet de limiter le bruit, en donnant des poids très faibles aux termes dont la sémantique est éloignée de la thématique.

La mesure sémantique est calculée à partir 1) des termes extraits TB et 2) les termes graines ou experts TG. Pour chaque terme de TB, nous calculons sa similarité sémantique avec chacun des termes de l'ensemble TG. Ceci nous donne un nombre de valeurs de similarité égale au nombre de termes dans l'ensemble TG pour chaque terme t de TB. Pour obtenir la valeur de similarité finale pour un terme, nous faisons la moyenne des TOP@5 plus grandes valeurs de similarités. Chaque terme est associé à une valeur de similarité unique, représentant son poids en terme de pertinence par rapport à la thématique, sur l'ensemble des termes. À la fin du processus, nous obtenons un nouvel ensemble TS (termes sémantiques), constitué des mêmes termes de l'ensemble TB, mais avec des scores ou rangs de pertinence différents.

Les dictionnaires de concepts seront par la suite sélectionnés à partir de ces ensembles de termes TS pour chaque thématique, en partant de ceux disposant d'un plus grand score.

Dans le Tableau 3.4, nous présentons d'abord les TOP@10 termes de l'ensemble TB avant l'application de la mesure sémantique. Les Tableaux 3.5, 3.6 et 3.7, donnent respectivement les TOP@10 et LAST@10 termes après l'application de mesure sémantique pour les thématiques agriculture, hydrologie et urbanisation.

Tout d'abord, nous constatons que l'ordre de pertinence des termes a changé entre les deux étapes. Dans le Tableau 3.4 de la thématique *Agriculture*, nous remarquons la présence de termes comme *google scholar*, *field sizes*, qui sont des termes qui n'ont visiblement pas un sens sémantique proche de l'agriculture, mais qui occupent des rangs importants dans les classements de BioTex (c'est-à-dire dans l'ensemble TB). Après avoir appliqué le calcul de similarité sémantique entre ces termes avec les termes experts (voir Tableau 3.3), nous constatons par exemple dans le Tableau 3.5, l'apparition de nouveaux termes dans les TOP@10. Cela signifie que les nouveaux termes présents dans la deuxième étape occupaient un rang de pertinence moins fort avant le calcul de similarité. Nous avons par exemple *production agricole* qui passe du 9ème rang au 1er rang avant et après. Cette remarque est aussi valable pour les thématiques Urbanisation et Hydrologie.

Nous constatons aussi l'amélioration des termes suivant leur score par rapport aux termes présents dans les LAST@10 termes, qui sémantiquement parlant n'ont aucune relation avec les thématiques étudiées.

Thématique Agriculture	
Multi-termes	**Simple Termes**
dégradation des terres	agriculture
agriculture urbaine	terres
etats membres	sols
changement climatique	désertification
google scholar	production
occupation du sol	développement
field sizes	food
lutte contre la desertification	scholar
production agricole	espaces
matiere de dégradation des terres	parcelles
Thématique Urbanisation	
Multi-termes	**Simple Termes**
étalement urbain	urbanisation
taux d'urbanisation	population
aménagement du territoire	species
plan local d'urbanisme	espaces
développement durable	ville
communauté urbaine de bordeaux	aménagement
local d'urbanisme	développement
population urbaine	villes
gradient d'urbanisation	urban
seed bank	territoire
Thématique Hydrologie	
Multi-termes	**Simple Termes**
étude hydraulique complémentaire de la vallée	eau
vallée de la leze	étude
oh00 pont	débit
étude du risque statistique	crue
zonage du risque inondation	débits
période de retour	modèle
commune de montfaucon	données
étude hydraulique complémentaire	projet
étude du risque	ans
zonage du risque	tableau

TABLE 3.4 – TOP@10 termes thématique de l'ensemble TB avant la mesure sémantique

Thématique Agriculture			
Multi-termes	rank	**Simple Termes**	rank
production agricole	0.946	agroenvironnement	0.862
exploitation agricole	0.932	agroenvironnementaux	0.857
agricole production	0.905	agronomiquement	0.851
d'oeuvre agricole	0.903	agrochimie	0.845
entreprise agricole	0.810	agroecologie	0.840
service agricole	0.895	l'agroecologie	0.838
techniques de production agricole	0.895	agroecologique	0.838
production agricole perdu	0.891	agricuture	0.831
agricole perdu	0.891	agroecologue	0.825
gestion de la production agricole	0.891	agroforesterie	0.816
—	—	—	—
final models	0.101	relatives	0.184
top two models figures	0.096	password	0.179
other authors such	0.092	gudmundsson	0.179
political weekly 5021	0.091	lindangelund	0.177
political weekly 5221	0.089	wiley	0.177
their models although these	0.088	stocks	0.176
forgot your password	0.087	cookies	0.168
their models although	0.085	dating	0.155
login to your account	0.081	acebook	0.152
online library requires cookies	0.059	twitter	0.139

TABLE 3.5 – TOP@10 et LAST@10 termes de la thématique agriculture après la mesure sémantique

Thématique Hydrologie			
Multi-termes	rank	**Simple Termes**	rank
situation hydrologique	0.894	hydrologiquement	0.885
regime hydrologique	0.891	hydrogeologie	0.874
station hydrologique	0.883	hydrometrique	0.864
scenario hydrologique	0.880	hydrometeorologie	0.862
hydrologique station	0.877	hydrodynamique	0.861
description hydrologique	0.876	hydrologik	0.852
cours d'eau	0.875	hydrodynamisme	0.846
information hydrologique	0.875	hydrometriques	0.845
cycle hydrologique	0.873	eau	0.837
etude hydrologique	0.866	hydrometeorology	0.836
—	—	—	—
seizures in fetal sheep	0.146	adherents	0.198
positions relatives	0.145	relationships	0.198
cookies to make	0.144	reader	0.193
questions relatives	0.141	ranch	0.192
with other grains whether	0.140	relationship	0.182
type d'informations relatives	0.139	topmodel	0.178
other prediction methods	0.136	pronostic	0.173
website's visitors including	0.136	couples	0.167
with other grains	0.118	couple	0.160
pickup are preferable	0.116	yearbooks	0.152
website's visitors including their	0.115	relatives	0.133

TABLE 3.6 – TOP@10 et LAST@10 termes de la thématique hydrologie après la mesure sémantique

Thématique Urbanisation			
Multi-termes	rank	**Simple Termes**	rank
planification urbaine	0.897	urbaine	0.862
configuration urbaine	0.885	urbaines	0.841
urbaine planification	0.883	urbanisme	0.838
approche d'urbanisme	0.882	urbanisee	0.805
enjeux urbains	0.882	urbanismes	0.802
plan d'urbanisme	0.882	territorialisations	0.801
planification urbaine classique	0.882	urbanisation's	0.780
enjeux de la planification urbaine	0.881	urbanisation	0.797
notions de l'urbanisme	0.877	urbanisees	0.789
composition urbaine	0.877	territorialisation	0.783
—	—	—	—
logements en hausse	0.472	crush	0.181
modifications engendrees	0.472	parents	0.177
temps etre	0.472	terrorism	0.176
secteur en zone villa	0.472	facebook	0.172
modifications d'habitats	0.472	ninja	0.166
centimetres de l'horizon	0.472	wife	0.163
travail de couture	0.472	dating	0.162
puissante force motrice	0.47214	twitter	0.153
immatriculees au nom de l'etat	0.47214	demons	0.138
inegal acces	0.472	cookies	0.130

TABLE 3.7 – TOP@10 et LAST@10 termes de la thématique urbanisation après la mesure sémantique

3.4.2 Constitution des corpus thématiques

3.4.2.1 Vocabulaire de concepts thématiques final

Le choix du vocabulaire de concepts thématiques final est effectué en fixant un seuil minimal spécifique à chaque thématique. Pour ce faire, nous regardons donc à partir de quelle valeur du score, les termes commencent à être éloignés de la thématique. Pour la suite de notre étude, nous avons considéré les TOP@1000 termes de chaque thématique. Le tableau 3.8 indique le seuil correspondant pour chaque thématique.

Une fois le seuil fixé, la liste de termes est étendue en utilisant WordNet. Les synonymes obtenus permettent de sortir du cadre de langage officiel ou académique, et d'utiliser des expressions moins techniques, qui permettront par exemple d'évaluer des documents en relation avec des réseaux sociaux, des blogs ou encore des annonces telles que les offres

Thématique	**Urbanisation**	**Agriculture**	**Hydrologie**
Taille Corpus (Nb documents)	105	120	117
Taille du VC	TOP@1000	TOP@1000	TOP@1000
Valeur du seuil de similarité	0.80	0.75	0.77

TABLE 3.8 – Statistiques sur la constitution du vocabulaire de concepts thématiques VC

d'emplois. L'exemple présenté dans la Figure 11 correspond à un exemple de synonymes obtenus avec le terme *zone d'aménagement concerté* de la thématique urbanisation.

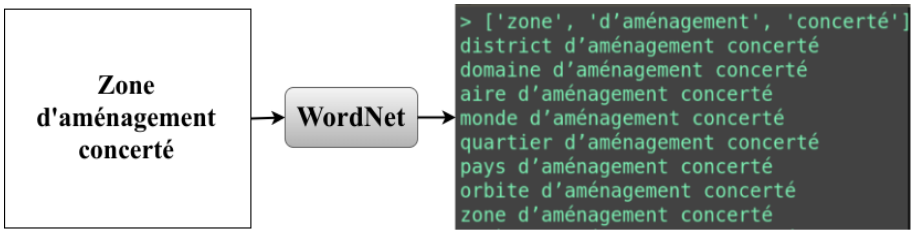


FIGURE 11 – Principe de constitution du vocabulaire étendu de concepts thématiques

3.4.2.2 Évaluation automatique des documents par mesure sémantique

Cette évaluation est faite par calcul de similarité sémantique entre chaque document collecté avec le vocabulaire étendu de concepts thématiques de la thématique concernée.

Pour chaque thématique, nous avons considéré les TOP@1000 termes de son vocabulaire étendu de concepts thématiques pour l'évaluation des documents la concernant. Ce choix est fixé suite aux différentes expérimentations qui montrent une très faible variation, voir nulle au delà des TOP@1000. À la fin du processus, pour chaque thématique, un score est affecté à chaque document, représentant sa proximité sémantique avec le vocabulaire étendu de concepts thématiques précédemment obtenu. Nous avons défini une valeur seuil de 0.5 comme valeur minimale. Un document dont la valeur de la proximité sémantique est inférieure à 0.5 ne sera donc pas considéré dans le corpus, et ce pour l'ensemble des thématiques. Cette valeur est fixée de façon expérimentale, après une vérification des résultats obtenus. Le tableau 3.9 indique les caractéristiques pour chaque corpus thématique.

Thématique	**Urbanisation**	**Agriculture**	**Hydrologie**
Taille du VC utilisé	TOP@500	TOP@500	TOP@500
Taille du VEC utilisé	TOP@1000	TOP@1000	TOP@1000
Seuil de similarité	0.5	0.5	0.5
Taille Corpus	867	1.400	1380

TABLE 3.9 – Statistiques sur les corpus constitués par thématique

3.4.3 Évaluation & Discussion

Dans le Tableau 3.10, nous décrivons le nombre de termes et la taille des corpus des différentes thématiques que nous avons évaluées.

Thématique	**Urbanisation**	**Agriculture**	**Hydrologie**
Liste TG	114	100	100
Taille Corpus	125	200	155
Nb évaluateurs	2	2	2

TABLE 3.10 – Statistiques sur les données évaluées par les experts

3.4.3.1 Évaluation des termes experts

Le regroupement des termes dans les trois intervalles que nous avons prédéfinis est décrit dans la Section 3.3.3.1. Le premier intervalle correspond, **Très_Adapté** = [4,5], le deuxième, **Adapté** =]2,3], et le troisième, **Pas_Adapté** = [0,2]. Dans la Table 3.11, nous présentons la répartition des termes sur les trois intervalles issus des évaluations expertes.

Qualité termes	**Urbanisation**	**Agriculture**	**Hydrologie**
Très_Adapté	35	68	54
Adapté	75	68	31
Pas_Adapté	10	6	13

TABLE 3.11 – Répartition des termes en fonction de leur évaluation qualitative

À l'issu de cette évaluation qualitative, nous remarquons que pour chaque thématique, environ les 1/10ème des termes ont été évalués comme n'étant pas adaptés à la thématique. Au moins 90% des termes proposés par les experts pourront être utilisés lors de la phase de constitution de vocabulaire de termes thématiques.

Les Figures 12 et 13 illustrent respectivement la matrice de corrélation représentant les accords inter-évaluateurs des différentes thématiques, et les représentations linéaires montrant la distribution des valeurs données par chaque couple d'évaluateurs.

Nous avons respectivement 0.71, 0.86 et 0.32 pour les valeurs des coefficients de Pearson entre évaluateurs de la thématique urbanisation, agriculture et hydrologie. Nous notons un accord assez important, illustré dans la Figure 13, entre les évaluateurs de la thématique agriculture, avec la majorité des points qui sont proches de la courbe linéaire. Contrairement à la thématique hydrologie, où nous constatons une forte distribution (désaccord) entre les évaluateurs.

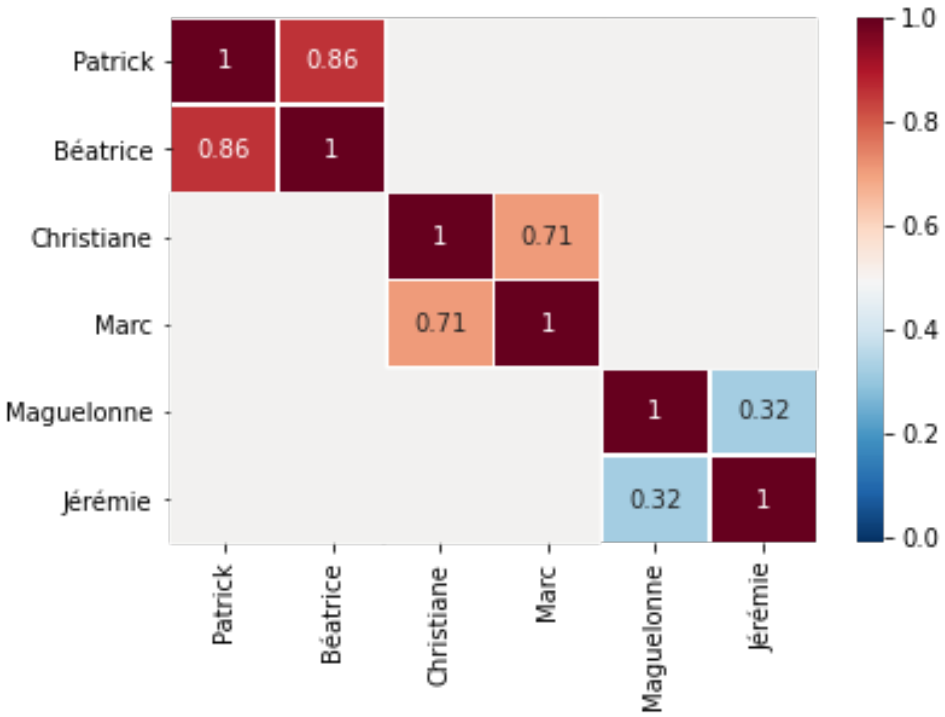


FIGURE 12 – Accords inter-annotateurs - Corrélation de Pearson

3.4.3.2 Évaluation des corpus

Le Tableau 3.12 présente les résultats issus des évaluations expertes sur la qualité des documents pour chaque corpus. Contrairement à l'évaluation des termes, celle des documents montre une forte distribution des valeurs pour chaque intervalle.

Qualité termes	**Urbanisation**	**Agriculture**	**Hydrologie**
Très_Adapté	17	88	47
Adapté	58	52	47
Pas_Adapté	50	60	60

TABLE 3.12 – Répartition des documents en fonction de leur évaluation qualitative

Les Figures 14 et 15 nous montrent respectivement les accords inter-évaluateurs et les représentations linéaires avec la distribution des valeurs données par chaque couple d'évaluateurs.

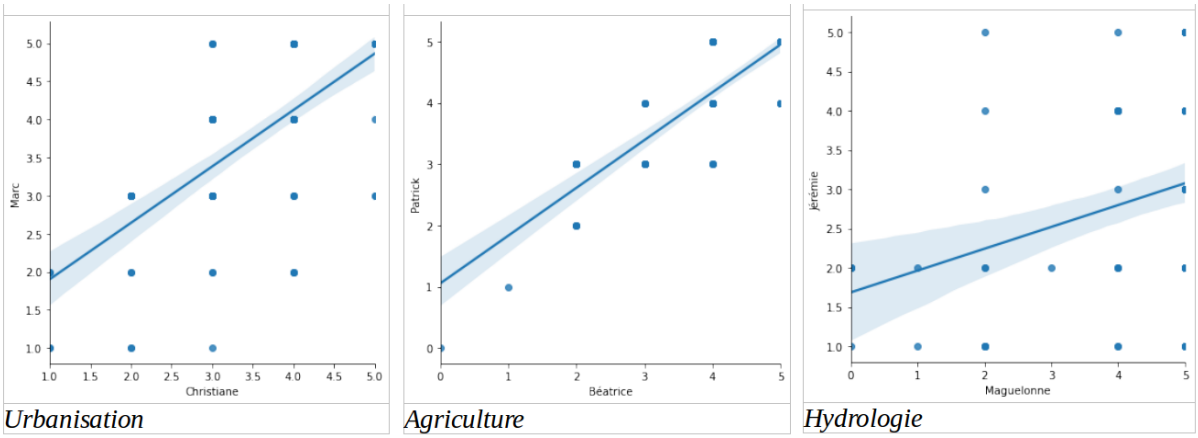


FIGURE 13 – Distribution des scores inter-évaluateurs

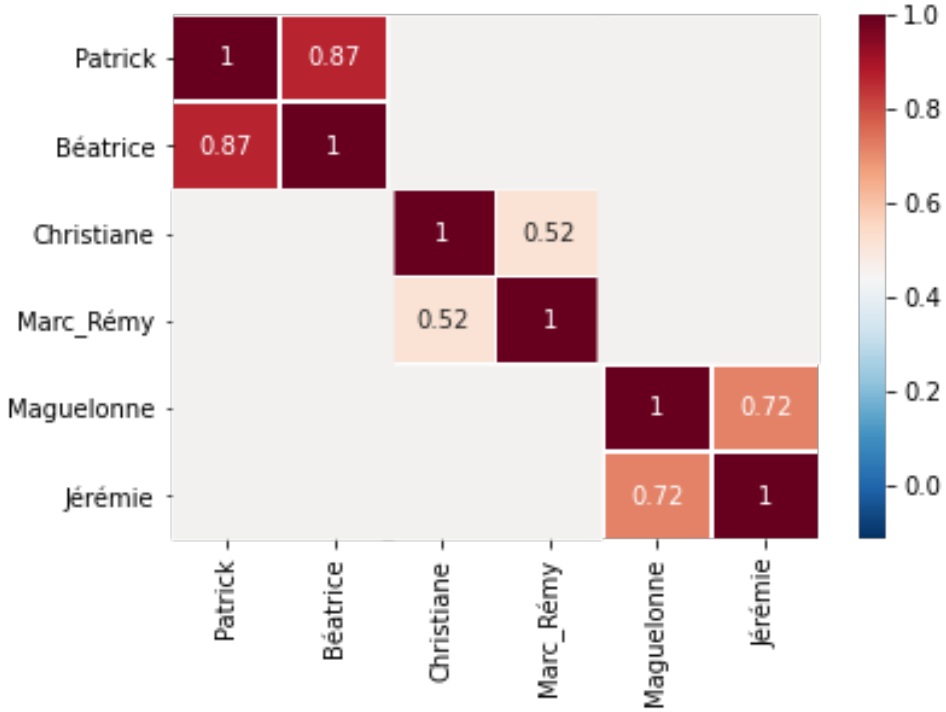


FIGURE 14 – Accords inter-annotateurs - Corrélation de Pearson

Nous remarquons un accord inter-évaluateur assez important pour la thématique agriculture et hydrologie, soit 0.87 et 0.72 pour les valeurs de Pearson. Ces accords sont aussi illustrés par les courbes représentatives de chaque thématique dans la Figure 15.

Le regroupement par intervalles, les évaluations des accords inter-évaluateurs avec le coefficient de Pearson ou les courbes représentatives ne nous permettent pas d'évaluer efficacement la qualité des documents comme celle des termes experts. En effet, la qualité d'un document est mesurée en utilisant plusieurs critères et non uniquement sur sa sémantiques. C'est pour cela que nous avons défini des critères d'évaluations dans le Tableau

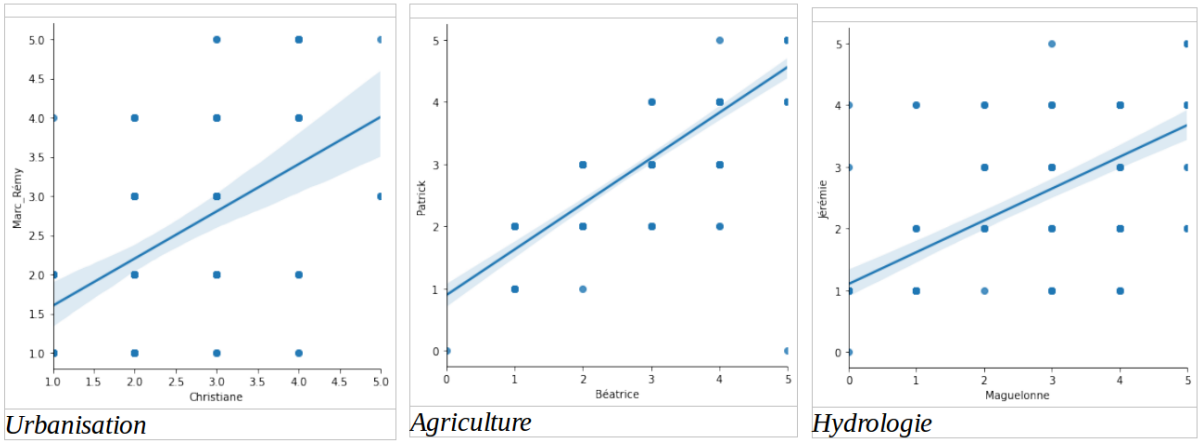


FIGURE 15 – Distribution des scores inter-évaluateurs

3.1 nous permettant de classer les documents par type ou catégorie, et pour chaque type sa pertinence sur la couverture spatiale, temporelle et sémantique.

Tout d'abord, pour chaque thématique, nous faisons ressortir les données réelles, et les redondances. Lors de la phase de collecte, il arrive qu'une même donnée soit récupérée sur des sources (urls) différentes. Il est donc impératif de les identifier et de les éliminer. Ensuite, à partir des données réelles, nous faisons ressortir les données contextuelles, les données dont les contenues parlent du contexte ou de la thématique abordée, et celles que nous avons classées hors sujets (essentiellement les publicités et les hors contextes). La Figure 16 illustre les statistiques pour les corpus des différentes thématiques, respectivement pour les données agriculture, hydrologie et urbanisation.

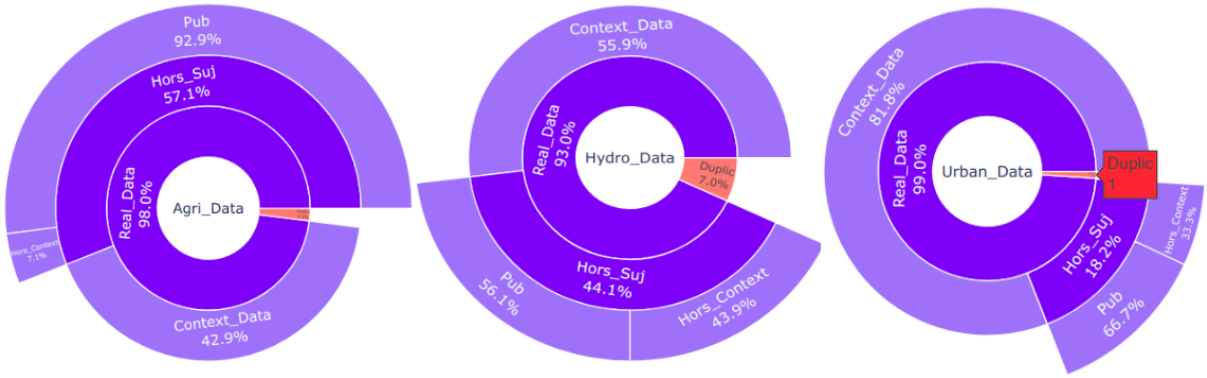


FIGURE 16 – Statistiques sur les corpus évalués

Nous enregistrons respectivement 2%, 7% et 1% de redondances entre les documents des thématiques agriculture, hydrologie et urbanisation. Parmi les données réelles, nous constatons un pourcentage élevé des données de la catégorie hors sujet pour les corpus agriculture et hydrologie (57.1%, 44.1%), contrairement à l'urbanisation où nous enregistrons un faible pourcentage de 18.2%. Les pourcentages élevés pour les thématiques

agriculture et hydrologie sont essentiellement dûs au fait que, beaucoup d'offres d'emplois, de publicités et de formations traitent de ces thématiques par rapport à celle de l'urbanisation.

Les analyses suivantes, sont consacrées uniquement aux données que nous avons identifiées comme étant contextuelles (celles non classifiées en "hors sujet"), prenant en compte, celles de type académique, administratif et sociétal. Ainsi, les analyses prennent en compte, les 42.9% de données de l'agriculture, 55.9% des données de l'hydrologie, et les 81.8% des données de l'urbanisation.

Nous catégorisons les données de chaque thématique en fonction des trois types précédemment cités. Pour chaque type, nous faisons ressortir la couverture spatiale et temporelle des documents qui le composent. La couverture spatiale comme temporelle est évaluée en fonction des critères définis dans le Tableau 3.1.

Les Figures 17, 18, et 19 illustrent respectivement les répartitions pour chaque catégorie, la couverture spatiale et temporelle des données qui la composent.

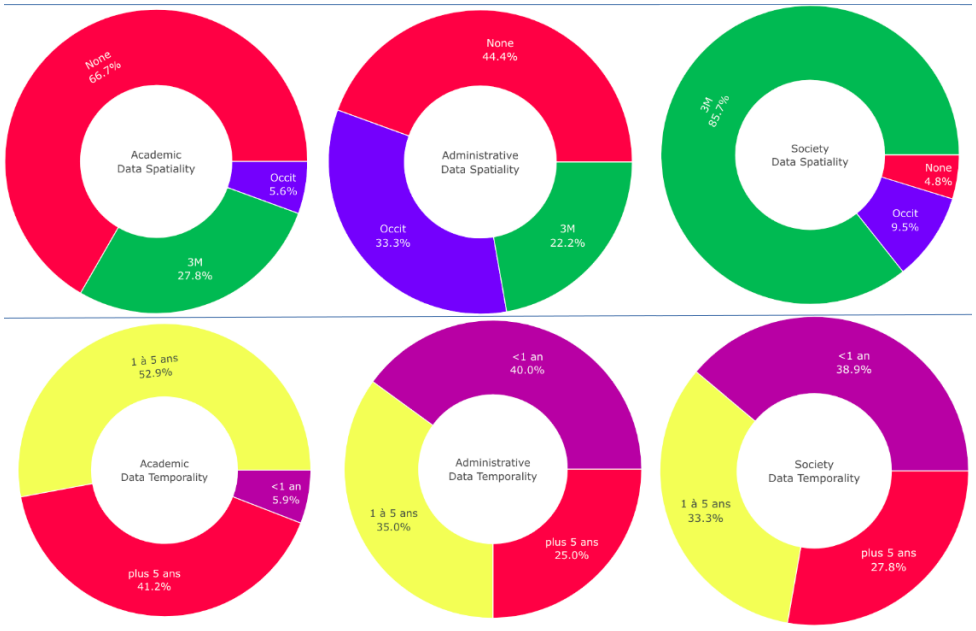


FIGURE 17 – Représentation par catégorie de document en fonction de la couverture spatiale (Data Spatility) et temporelle (Data Temporality) : thématique Agriculture

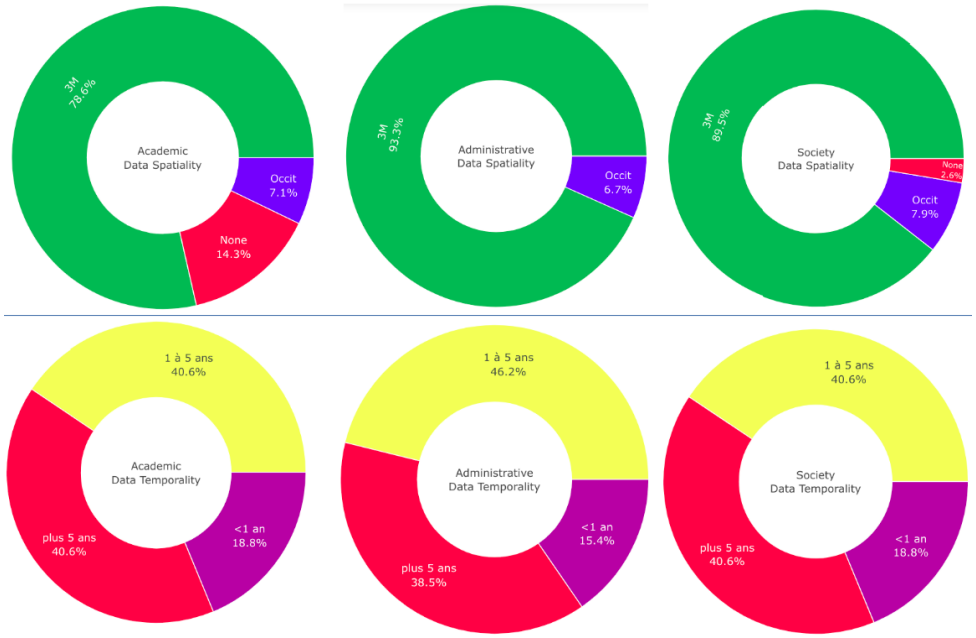


FIGURE 18 – Représentation par catégorie de document en fonction de la couverture spatiale (Data Spatility) et temporelle (Data Temporality) : thématique Urbanisation

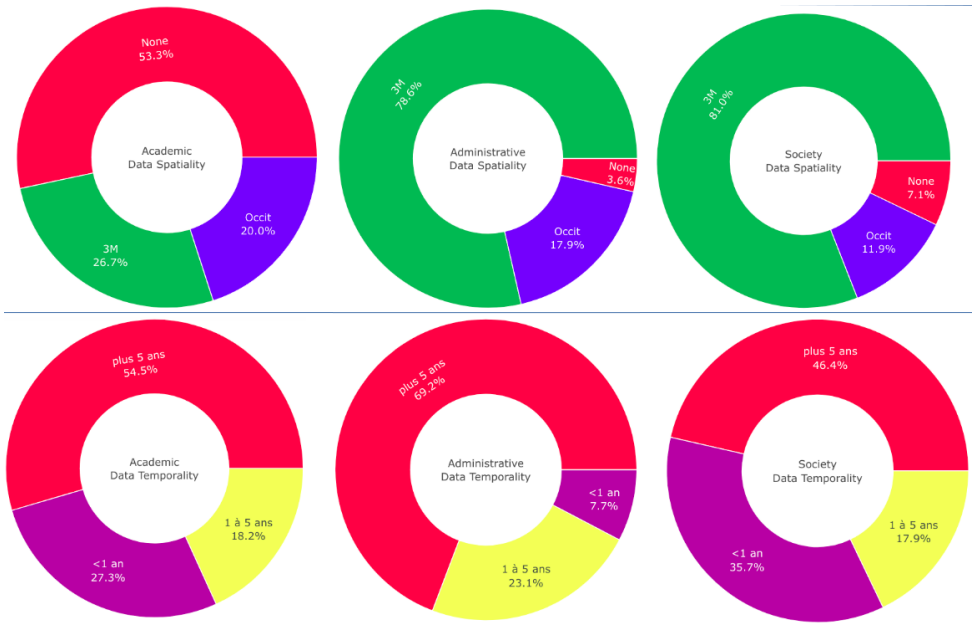


FIGURE 19 – Représentation par catégorie de document en fonction de la couverture spatiale (Data Spatility) et temporelle (Data Temporality) : thématique Hydrologie

Pour chacune des thématiques, nous remarquons une disponibilité conséquente sur les données de la catégorie société qui est d'une grande importance dans notre étude, tant sur la spatialité que sur la temporalité. Nous notons respectivement que 85.7%, 89.5%, et 81.0% des documents de type société de la thématique agriculture, urbanisation et

hydrologie traitent des problèmes qui sont directement liés à la Métropole de Montpellier. En ce qui concerne la couverture temporelle, nous remarquons respectivement que 72.8%, 59.4%, et 53.6% des données de la thématique agriculture, urbanisation et d'hydrologie ont une durée maximum de 5 ans.

Ces différentes analyses nous permettent d'illustrer l'efficacité du protocole proposé pour la collecte de données de type sociétés qui viennent en complément de celles qui sont fortement de type administratives ou éducatives.

3.4.4 Prototype d'une application Web

Dans le but de permettre aux utilisateurs d'effectuer des requêtes et d'explorer la base de données des corpus thématiques, nous avons proposé un prototype d'application Web. À partir de cette application, l'utilisateur a la possibilité de renseigner les critères de sa recherche, visualiser et télécharger des données s'il le souhaite.

L'application est subdivisée en deux parties : l'indexation de la base de données qui constitue la partie back-end et l'interface utilisateur qui constitue le front-end. L'indexation consiste à indexer les corpus thématiques contenus dans le lac de données pour faciliter leur exploitation. Nous avons opté ElasticSearch [2] pour indexer la base de données et son remplissage a été automatisé grâce à logstash [3] (voir Figure 20). Lors du processus d'automatisation, les entités nommées spatiales absolues présentent dans les méta-données des documents sont géocodées, c'est-à-dire qu'elles sont associées aux coordonnées de géoréférencement de leur centroïde (longitude, latitude) et des bounding box respectifs. Plusieurs raisons ont guidé le choix d'ElasticSearch :

- un moteur de recherche distribué ;
- il offre une interface REST ;
- il permet de faire des recherches avancées ;
- il offre diverses types de traitement et d'analyse, avec des formes de visualisation pour les données à travers Kibana [4].

Les codes sources du protocole de collecte sont disponibles dans le dépôt Github Protocole Collecte. Les fichiers de configuration pour l'indexation des données sont disponibles dans le dépôt Github Indexation Elastic Search et le protype dans le dépôt Github Prototype.

2. https ://www.elastic.co/fr/elasticsearch/
3. https ://www.elastic.co/fr/logstash/
4. https ://www.elastic.co/fr/kibana/

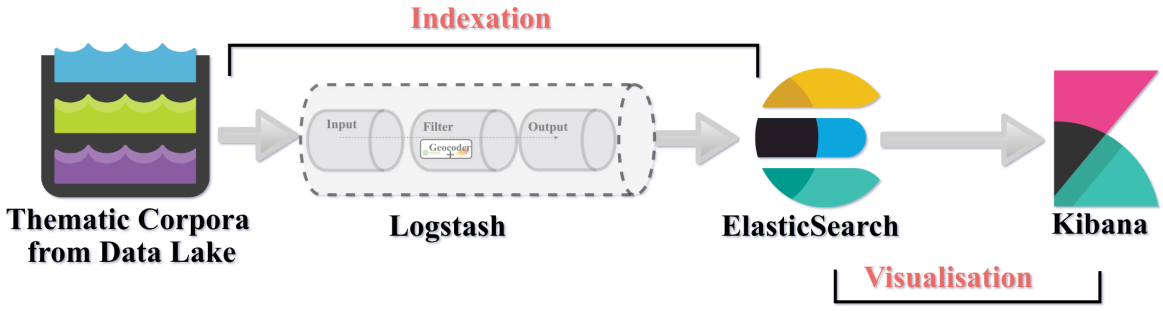


FIGURE 20 – Processus d'indexation dans ElasticSearch

Pour ce qui concerne l'interface Web, nous l'avons développée en utilisant Streamlit [5], un framework open-source dédié à la mise en production des solutions en science de données. L'application Web prend en entrée principalement trois paramètres (voir Figure 21) qui sont :

- une thématique : nom de la thématique sur laquelle l'utilisateur souhaite interroger les données. Dans le cadre de ce prototype, nous proposons les thématiques de 3M sur lesquelles nous travaillons, à savoir l'agriculture, l'hydrologie et l'urbanisation ;
- un intervalle temporel : il permet de définir l'intervalle de temps sur lequel l'utilisateur souhaite récupérer les données. Un intervalle de temps compris entre 2015 et 2020 permet de récupérer les documents dont les dates de publications sont comprises entre 2015 et 2020 incluses ;
- une base de données : correspond à la base de données sur laquelle l'utilisateur veut effectuer sa requête. Dans ce cas d'étude, il s'agit de la base de données de 3M (précédemment indexée dans ElasticSearch) constituée de corpus thématiques.

Une fois les paramètres d'entrée définis, l'utilisateur peut cliquer sur le bouton *Start Retrieving* pour lancer sa requête. À la fin de la requête, deux sorties sont proposées à l'utilisateur. Illustrée dans la Figure 22, la première présente un extrait des données obtenu sous forme de tableau avec des informations telles que le nombre de documents que comprend le corpus obtenu, la taille du corpus en Mb, et un lien permettant de télécharger les résultats de sa requête sous la forme d'un fichier CSV.

La seconde sortie, illustrée dans la Figure 23, donne des informations sous forme de graphiques, montrant les répartitions spatio-temporelles sur l'ensemble du corpus. Dans la première colonne, nous illustrons tout d'abord la répartition sur le plan spatial. *With_SNE* donne le pourcentage des documents pour lesquels des entités spatiales sont bien identifiées, et *WithOut_SNE* le pourcentage de ceux sur lesquels elles n'ont pas pu être identifiées lors de la collecte. Les nuages de mots ou wordcloud permettent de constater les entités nommées qui dominent dans le corpus. Pour ce qui est de la seconde colonne, elle contient les illustrations sur la répartition temporelle des documents. *With_TNE* indique

5. https ://streamlit.io/

le pourcentage des documents sur lequel les dates de productions et/ou de publications sont indiquées ou disponibles, et sont conformément subdivisés suivant les intervalles que nous avons définis dans la Section 3.4.3.2. Nous avons aussi *WithOut_TNE* qui indique le pourcentage de ceux n'ayant pas de date explicitement défini. De même que pour les SNE, nous faisons ressortir les dates les plus fréquentes ou dominantes à partir du wordcloud.

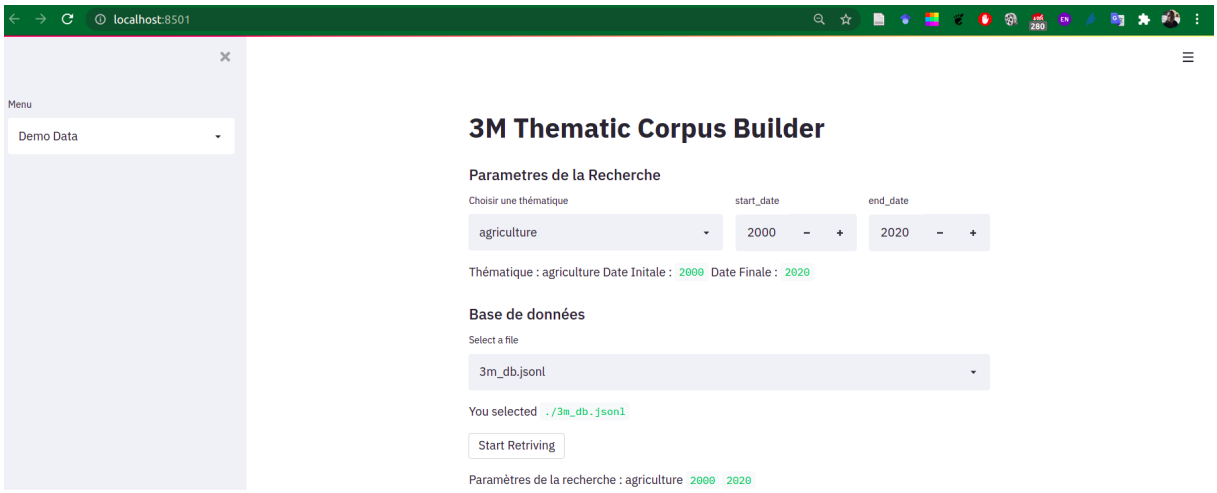


FIGURE 21 – Interface Web : Inputs

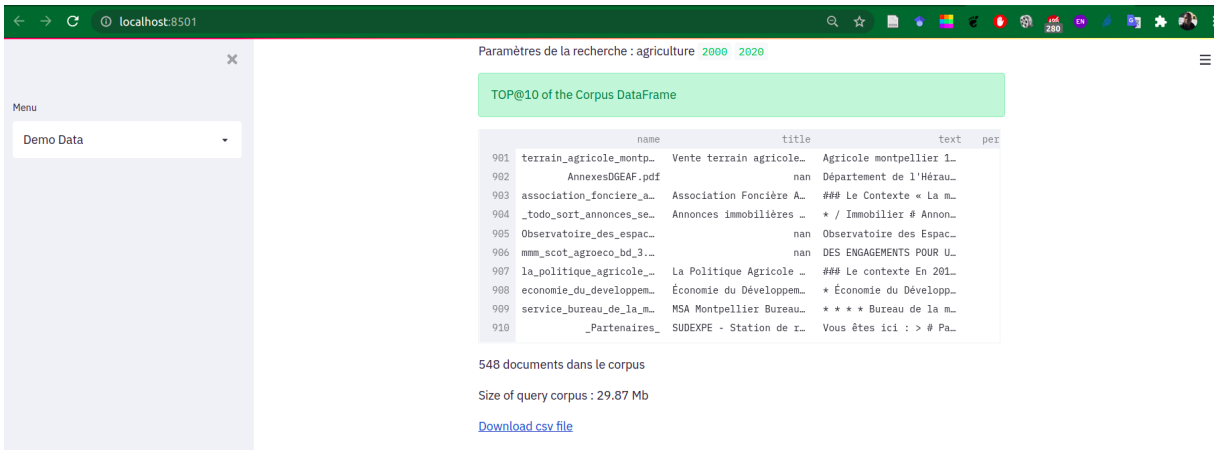


FIGURE 22 – Interface Web : Output 1

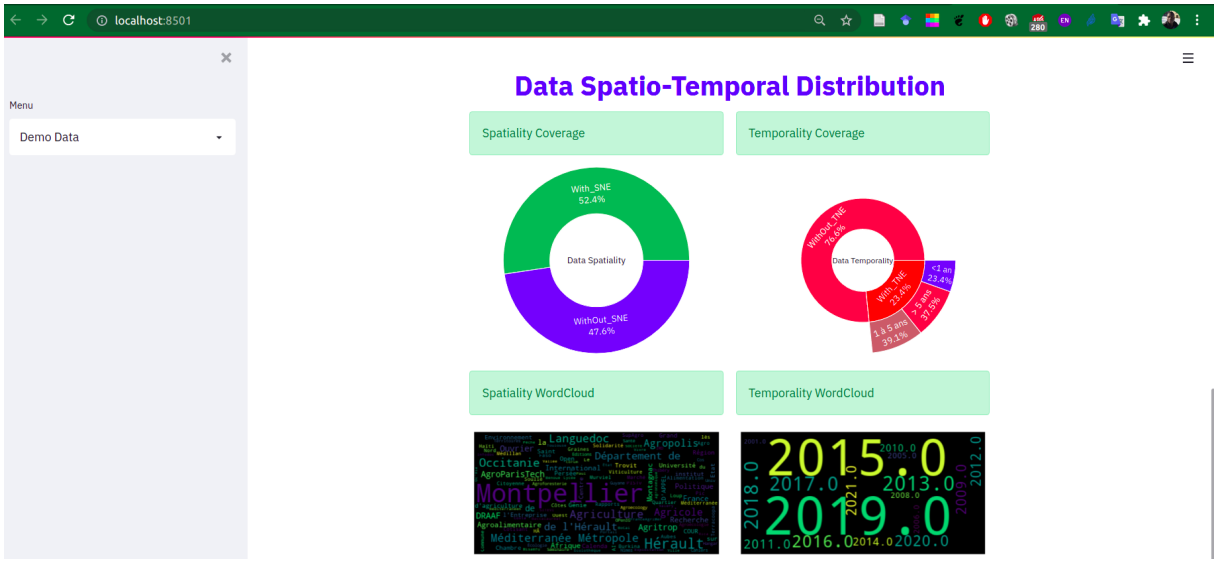


FIGURE 23 – Interface Web : Output 2

3.5 Conclusion

Le protocole que nous avons élaboré vise à proposer une démarche générique permettant de constituer 1) des vocabulaires thématiques et 2) des corpus thématiques spécifiques. Il s'appuie principalement sur le principe de faire évoluer une terminologie allant de la constitution de graines à la formation d'un vocabulaire de concepts thématiques bien spécifique. Quant à la constitution des corpus, elle se fait en utilisant chaque vocabulaire thématique comme ensemble de mots clés de recherche et une spécification de l'emprise spatiale d'intérêt.

Chacune des étapes est tout d'abord soumise à une évaluation sémantique automatique, dans le but de limiter le bruit aussi bien dans le vocabulaire de concepts que dans les corpus thématiques. Après la constitution de termes graines, ainsi que les corpus thématiques, une évaluation experte est soumise à ces deux étapes, afin de valider qualitativement la liste de termes graines initialement constituée, et les corpus thématiques obtenues à partir de ces listes.

Illustrés sur les thématiques, agriculture, urbanisation et hydrologie de la métropole de Montpellier, les résultats des différentes évaluations nous ont permis de tirer quelques conclusions. Pour chaque thématique, plus de 90% des termes graines étaient adaptés ou étaient sémantiquement valides. Quant à l'évaluation des corpus thématiques, nous avons pu conclure que les données de type société obtenues sont bien ancrées sur le plan spatial (respectivement 85.7%, 89.5%, et 81.0% pour agriculture, urbanisation, hydrologie), à savoir le territoire de Montpellier et aussi sur le plan temporel avec une forte quantité (respectivement 72.8%, 59.4%, et 53.6% pour agriculture, urbanisation et hydrologie) de

données dont la date de production était inférieure ou égale à 5 ans.

Bien que la protocole de collecte soit une approche générique, nous notons le fait que la pertinence de ce protocole reste fortement liée à celle des termes graines filtrés par les experts dans la Section 3.3.1.1. Cette étape peut être un handicap lorsque des experts ne sont pas disponibles pour proposer ces termes.

Pour ce faire, nous proposons dans le chapitre suivant, un ensemble de stratégies permettant d'obtenir des termes (pouvant servir de termes experts) sur une thématique bien précise à partir d'un corpus ou d'un corpora (ensemble de corpus).

Le protocole de collecte proposé dans ce chapitre est utilisé pour la constitution de corpus thématiques composés de différents types de documents (administratives, académiques, sociétales, etc.) comme indiqué dans la Section 2.2.1 du Chapitre 2. Ces corpus thématiques constituent en partie les données de 3M, qui seront stockées dans le lac de données du Chapitre 5. L'ensemble de ces données seront utilisées dans le Chapitre 6 pour les analyses et la mise en relation.

Chapitre 4

Extraction & analyse de terminologies : ITEXT-BIO

Dans ce chapitre, nous abordons un ensemble d'approches méthodologiques permettant de suivre l'émergence d'une thématique donnée dans le temps. Cette émergence se distingue par l'apparition d'une nouvelle terminologie décrivant le plus souvent un évènement qui impacte la thématique. Cette approche s'inscrit d'une part, dans le long terme pour la Métropole de Montpellier Méditerranée (3M) pour le suivi des thématiques, et d'autre part se concentre sur l'extraction générique des terminologies pouvant servir de termes experts (voir chapitre précédent) pour la collecte de données textuelles. En effet, comme mentionné dans le Chapitre 1, Section 1.2, nous travaillons sur des données de plusieurs thématiques choisies avec 3M. Il est donc impératif de pouvoir mettre à jour les données en capturant les nouvelles informations relatives à des évènements nouveaux que pourraient connaître ces thématiques dans le futur (voir Figure 1).

ITEXT-BIO qui se définit par Intelligent Term EXTraction for BIOmedical Analysis est composé de : ITEXT pour désigner le processus d'extraction intelligent utilisé pour extraire et analyser les termes, et BIO pour dire que le processus a été appliqué sur des données du domaine biomédicale, à savoir la Covid-19. Cette spécificité permet de mentionner le fait que la méthode proposée peut-être appliquée sur des données d'un autre domaine, notamment sur les thématiques de la Métropole de Montpellier. Nous pouvons avoir par exemple ITEXT-AGRI, ITEXT-URBAN, ou encore ITEXT-HYDRO, respectivement pour les thématiques agriculture, urbanisation et hydrologie.

ITEXT-BIO est illustré avec des données de la pandémie du coronavirus. Le choix de ces données se justifie par 1) la forte disponibilité de données sur la thématique, et 2) du fait que l'émergence de la pandémie a un impact majeur dans le domaine de la santé.

Ce chapitre est organisé comme suit : nous proposons tout d'abord une introduction générale dans la Section 4.1, la Section 4.2 porte sur l'état de l'art pour dresser un panorama des travaux existants. Dans la Section 4.3, nous abordons l'ensemble des stratégies que nous avons proposées pour l'extraction et l'enrichissement des termes. Nous terminerons avec la Section 4.4 qui décrit le cas d'étude associé à l'approche méthodologique sur les données de la COVID-19.

4.1 Introduction

L'extraction de connaissances à partir de données textuelles est une problématique qui a suscité de nombreuses recherches et discussions. Parmi les nombreuses tâches, nous avons l'extraction de connaissances sous forme de terminologies, qui consiste à extraire des termes significatifs ou les plus informatifs à partir d'un grand ensemble de texte. Ce processus bien établi est utilisé dans le traitement du langage naturel et a conduit au développement de plusieurs outils tels que TBXTools [Oliver and Vàzquez, 2015], TermSuite [Cram and Daille, 2016], BioTex [Lossio-Ventura et al., 2014a], etc.

Basée sur [Lossio-Ventura et al., 2014a], notre proposition porte sur l'extraction de terminologies dans un domaine spécifique à partir de corpus hétérogènes, et illustre comment procéder pour une analyse quantitative et qualitative plus efficace. À cette fin, nous proposons une démarche générique reposant sur une combinaison de stratégies d'extraction et d'analyse. Les stratégies d'extraction de termes sont basées sur des combinaisons d'approches linguistiques, de mesures statistiques et de segmentation de corpus, tandis que les stratégies d'analyse sont basées sur des combinaisons de termes extraits de ces corpus.

Sur la base de cette combinaison de stratégies, ITEXT-BIO vise à extraire : 1) des termes représentatifs, 2) des termes discriminants, et 3) de nouveaux termes pertinents, à partir d'un ou plusieurs corpus. Ces stratégies sont particulièrement utiles et montrent leur importance pour des tâches dédiées, telles que l'analyse de corpus, le suivi de l'évolution des données pour un domaine ou un sujet de recherche spécifique (ex : l'épidémiologie, le Covid-19), ou le suivi (monitoring) de la recherche scientifique.

4.2 État de l'art

L'extraction de terminologie sur des données de domaines spécifiques est un sujet d'intérêt et de discussion majeur qui anime la recherche sur le traitement automatique du langage naturel (TALN). Ce sujet a suscité de multiples propositions sur le plan méthodologique tels que [Kageura and Umino, 1996, Pais and Ion, 2020, Pazienza et al., 2005, Rigouts Terryn et al., 2020] qui sont des méthodes orientées vers l'extraction de termes à partir de corpus. Également connue sous le nom d'extraction automatique de termes ou automatic terms extraction (ATE), cette tâche est prise en compte dans diverses applications de TALN, comme la recherche d'information [Bracewell et al., 2005, Azarafza et al., 2020, Shah et al., 2019, Duari and Bhatnagar, 2020], le topic modeling [Habibi and Popescu-Belis, 2015, Wang et al., 2020b], la surveillance d'évènements [Maynard et al., 2005, Joung and Kim, 2017, Arsevska et al., 2018], l'extraction de mots-clés [Campos et al., 2020] et les résumés de document [Azarafza et al., 2020], ou encore l'acquisition

d'ontologies, la construction de thésaurus, etc.

D'après [Lossio-Ventura et al., 2014c], les techniques d'extraction de termes peuvent être classées en quatre approches : les approches linguistiques, les approches statistiques, les approches basées sur des méthodes d'apprentissage automatique et les approches hybrides.

Dans l'ensemble, les approches linguistiques prennent en compte des règles morphosyntaxiques des parties du discours ou part-of-speech (POS) pour décrire les termes ayant des structures communes [Brill, 1992]. Les approches statistiques utilisent des mesures statistiques telles que la fréquence des termes [Ramos et al., 2003, Whissell and Clarke, 2011], ou la co-occurrence des termes entre les mots et les phrases comme le Chi-carré [Matsuo and Ishizuka, 2004]. Quant aux approches d'apprentissage automatique, elles utilisent des mesures statistiques et sont principalement axées conjointement sur l'extraction de termes [Conrado et al., 2013, Foo, 2009, Campos et al., 2020], la classification [Wang et al., 2016] et le clustering [Azarafza et al., 2020]. Elles combinent des approches linguistiques et statistiques pour extraire les termes des données textuelles afin de construire des modèles d'apprentissage automatique qui serviront à faire des prédictions. Dans [Campos et al., 2020], les auteurs ont souligné que la plupart de ces tâches sont abordées avec des algorithmes d'apprentissage non supervisés. Les approches hybrides comprennent, par exemple, les méthodes C_Value [Pazienza et al., 2005], C/NC_Value [Frantzi et al., 2000] qui combinent des mesures statistiques et des règles linguistiques pour extraire les termes multi-mots et les termes imbriqués. Dans [Campillos Llanos et al., 2013, Neifar et al., 2016], les auteurs combinent des méthodes basées sur des règles linguistiques et des dictionnaires pour extraire respectivement des terminologies dans des données textuelles du domaine biomédical en espagnol et de textes spécialisés en arabe. Des études relatives aux approches hybrides telles que celles réalisées dans [Ji et al., 2007, Lossio-Ventura et al., 2014b] ont révélé leur efficacité et leur performance pour l'extraction des termes composés.

Dans [Lossio-Ventura et al., 2014a], les auteurs ont proposé BioTex, un outil basé sur des mesures linguistiques et statistiques pour extraire des termes liés au domaine biomédical. La même approche a été utilisée pour détecter des termes ou des signaux pour la surveillance des maladies infectieuses sur le web [Arsevska et al., 2018]. Dans [Meystre et al., 2008], une méthodologie hybride a été proposée pour extraire la terminologie dans des versions électroniques de dossiers médicaux. Cette approche hybride a également été adaptée par [Yao et al., 2017] pour extraire des concepts liés à la culture chinoise.

L'ensemble des études connexes précédemment citées se concentrent en majorité sur les techniques et les méthodes d'extraction de termes, principalement à partir de corpus. Sur la base des approches méthodologiques existantes, nous avons orienté notre étude vers le développement d'une approche générique pour l'extraction de termes à partir de corpus hétérogènes, c'est-à-dire une approche qui ne se limite pas à un seul domaine d'application, et aussi un ensemble de stratégies combinées pour analyser ces termes. Notre

approche combine et adapte des critères linguistiques et statistiques associés à des informations structurelles dans les textes afin de mettre en évidence les informations pertinentes qu'ils contiennent sous forme de terminologie. Les stratégies présentées visent également à surmonter les problèmes de temps de mise en place que connaissent les méthodes d'apprentissage automatiques qui nécessitent très souvent des données manuellement annotées en totalité ou partiellement.

4.3 Approches méthodologiques

Nous présentons ici deux approches complémentaires pour l'extraction et l'analyse de termes : l'approche non-guidée (ou Free Extraction Approch) et l'approche guidée (ou Driven Extraction Approch) pour l'extraction de termes. La première est basée sur une combinaison du type de corpus et des mesures statistiques, tandis que la seconde est basée sur une combinaison de corpus avec des règles basées sur des variations morpho-syntaxiques.

4.3.1 Approche non-guidée pour l'extraction de terminologie

L'approche non-guidée que nous proposons pour l'extraction des termes vise à offrir un outil permettant d'extraire des termes significatifs d'un domaine spécifique à partir d'un corpus donné. Comme nous l'avons mentionné dans la Section 4.2, des outils ont été déjà proposés pour l'extraction de termes et de concepts. Nous avons opté pour l'outil BioTex [Lossio-Ventura et al., 2014a] pour supporter l'extraction de termes, et cela pour plusieurs raisons :

- Les règles d'extraction qu'utilise BioTex ne tiennent pas compte de la sémantique d'un domaine spécifique. Il extrait les termes en se basant sur des patrons linguistiques comme *NN NN, JJ NP NP, NN NP NP, etc.* et aussi des patrons pour les verbes, tels que : *NN VBD NN NN, NP NN VBD NN NP, etc.*
- BioTex utilise des mesures hybrides (linguistiques et plusieurs mesures statistiques) pour le processus d'extraction des termes. Une approche qui s'avère être efficace pour l'extraction des termes simples et composés (multi-mots) ;
- La plupart des outils existants (ex : *Maui-indexer*[1], *Topia Termextract*[2], *KEA*[3], etc.) sont conçus pour l'extraction de mots-clés dans des documents uniques et ne prennent pas en compte des documents multi-langues, tandis que BioTex est conçu pour l'extraction termes (simple/multi-mots) et prend en charge des ensembles de documents (corpus) et aussi, une utilisation multi-langues (français, anglais, espagnol).

1. `https://code.google.com/p/maui-indexer/`
2. `https://pypi.python.org/pypi/topia.termextract/1.1.0`
3. `http://www.nzdl.org/Kea/`

L'outil BioTex prend en entrée essentiellement trois paramètres qui sont :

— un corpus : il s'agit de la source de données à partir de laquelle les termes sont extraits. Généralement une compilation de documents ;

— une mesure statistique : comme mentionné ci-dessus, l'approche de BioTex est basée sur des mesures linguistiques et statistiques. Le paramètre linguistique est défini par défaut, mais l'utilisateur doit définir le paramètre statistique, car il en existe plusieurs, afin d'exécuter le processus d'extraction des termes ;

— le nombre de mots à extraire par concept : appelé n-grammes, il concerne la longueur des termes extraits et varie de 1 à 4_grammes. L'utilisateur pourrait donc définir le nombre maximum de mots par terme qu'il aimerait obtenir.

La Figure 24 présente le processus global en trois étapes (1 à 3) pour l'extraction des termes avec l'approche non-guidée.

À la fin du processus d'extraction avec BioTex, les termes extraits sont classés en deux ensembles : TermSet, qui contient uniquement des termes à un seul mot ou single-word term (SWT), et MultiTermSet, qui contient des termes à plusieurs mots ou multi-word term (MWT). En utilisant le processus d'extraction guidé, nous pouvons capturer ou extraire la composition complète d'un terme à partir d'un terme incomplet obtenu lors de la première étape (extraction non-guidée). L'étape du processus d'extraction guidé utilise les termes incomplets pour capturer les termes entiers dans le document. Par exemple, si les termes *higher risk acute* ou *higher risk area* sont extraits lors de l'étape du processus d'extraction non-guidé, un terme entier qui pourrait être *higher risk acute care area* peut-être obtenu lors du processus d'extraction guidé.

4.3.2 Approche guidée pour l'extraction de terminologie

Cette approche d'extraction vise à garantir que les termes extraits à l'aide de BioTex puissent être utilisés pour améliorer la terminologie du domaine concerné. À partir d'un terme donné, le processus vise à extraire certaines variations de ce terme qui existent dans le corpus.

Le processus général dans le cadre de cette approche est traité avec FASTR [Jacquemin, 1994]. FASTR est un outil linguistique basé sur des règles qui génèrent des variantes morpho-syntaxiques des termes. On note respectivement *NN, NNS, NNP, NNPS* pour les patrons qui designent les noms, *VB, VBD, VBG, VBN, VBP, VBZ* pour les verbes, *RB, RBR, RBS* pour les adverbes et enfin *JJ, JJR, JJS* pour les adjectifs. Il permet d'extraire les variantes d'un terme donné dans les documents en texte intégral. Pour un terme donné, FASTR permet d'extraire les termes réduits ou longs qui contiennent le terme initial. La Figure 24 illustre les deux étapes (4 et 5) de l'approche guidée pour l'extraction de termes. Pour un terme donné, FASTR aide à extraire les termes proches

ou longs qui contiennent le terme initial, comme indiqué dans l'exemple précédent avec le terme *higher risk acute*.

Le processus guidé a l'avantage d'extraire de nouveaux termes pertinents que BioTex ne peut pas extraire du corpus lors de la phase d'extraction avec le processus non-guidé.

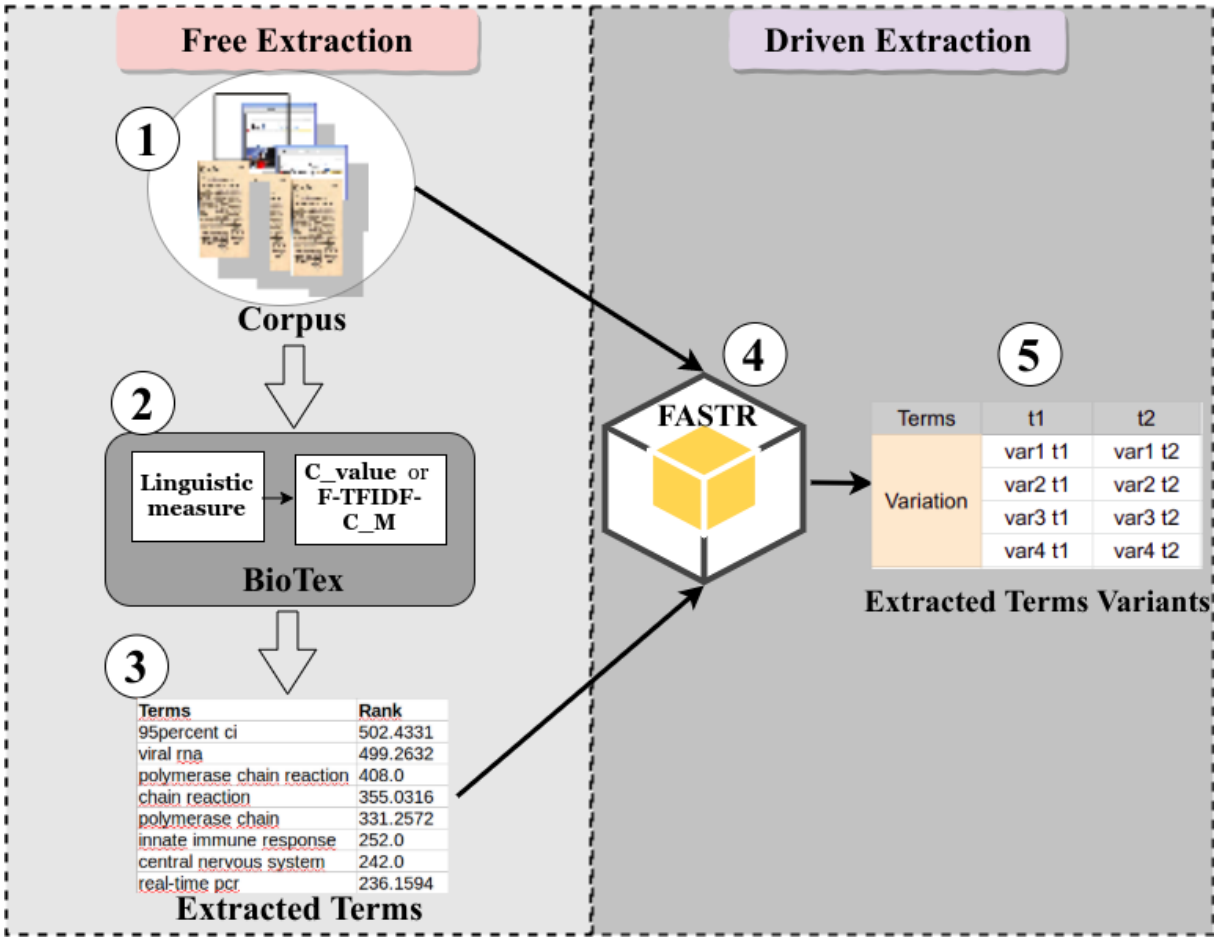


FIGURE 24 – Processus non-guidé et guidé, respectivement pour l'extraction de termes avec BioTex et FASTR

4.3.3 Stratégies de combinaisons proposées pour l'extraction des termes

Sur la base des éléments donnés dans les sections 4.3.1 et 4.3.2, nous proposons un workflow pour l'extraction et l'analyse de termes, dédié aux articles scientifiques, illustré par la Figure 25 . Nous décrivons ce processus en fonction du type de corpus, de la mesure statistique utilisée, et de l'approche (guidée ou non-guidée) :

— *Le type de corpus* : pour un article donné, nous avons principalement considéré trois parties qui sont, le Titre (T), le Résumé ou l'Abstract (A), et le Corps ou le Contenu (C) de l'article. Chaque partie de l'article est considérée pour la constitution des corpus. Nous avons donc un corpus uniquement constitué des Titres (T), un corpus

constitué de Résumés (A) et un dernier composé des Contenus (C) ;

— *Les mesures* : BioTex intègre plusieurs mesures statistiques, chacune d'elles utilisant une stratégie spécifique pour évaluer la pertinence des termes extraits. Dans notre cas, nous avons utilisé deux mesures qui sont C_Value et F-TFIDF-C_M. C_Value indique l'importance des termes qui apparaissent le plus fréquemment dans un document, en partant de l'idée que la fréquence d'apparition d'un terme dans le document reflète son importance dans ce document. De plus, sur la base du critère de fréquence, C_Value favorise l'extraction de termes multi-mots en prenant en compte les termes imbriqués par exemple *virus* dans le terme multi-mots *virus de la grippe* [Frantzi et al., 2000]. F-TFIDF-C_M représente la moyenne harmonique de deux valeurs, C_Value et TF-IDF, et classe les termes en fonction de leur pertinence dans le document tout en prenant en compte l'ensemble du corpus [Lossio-Ventura et al., 2016]. C_Value et F-TFIDF-C_M sont complémentaires, car la première favorise l'extraction de MWT pertinents tandis que la seconde donne du poids aux termes discriminants. Pour chaque mesure, l'objectif est d'organiser les termes extraits en cinq ensembles : 1) les termes correspondant à l'ensemble de corpus Titre (T), 2) les termes correspondant à l'ensemble de corpus Abstract (A), 3) les termes correspondant à l'ensemble de corpus Contenu (C), 4) les termes qui se croisent dans l'ensemble de corpus Titre et Abstract (TA), et 5) les termes qui se croisent dans l'ensemble de corpus Titre et Contenu (TC).

— *L'approche* : les termes peuvent être extraits en utilisant à la fois un corpus donné et une mesure statistique spécifique dans une approche d'extraction non-guidée. De plus, pour le processus guidé, les variations de termes sont extraites en utilisant à la fois un corpus donné et un ensemble spécifique de termes. L'ensemble de termes peut être défini à partir des résultats de l'approche précédente (approche non-guidée).

4.4 Cas d'application : COVID-19

4.4.1 Description des données

Notre étude s'est portée sur le COVID-19 Open Research Dataset [4] [Wang et al., 2020a] qui est constitué d'articles scientifiques sur le COVID-19 et des recherches historiques sur les coronavirus. Tout au long de cette étude, nous nous référons à cet ensemble de données sous le nom de COVID19-MOOD-data.

Le jeu de données COVID19-MOOD-data est divisé en deux corpus principaux, nommés respectivement Papers1 et Papers2. Papers1 contient le sous-ensemble de données *commercial use*, tandis que Papers2 contient les sous-ensembles de données *commercial*

4. https ://www.semanticscholar.org/cord19/download

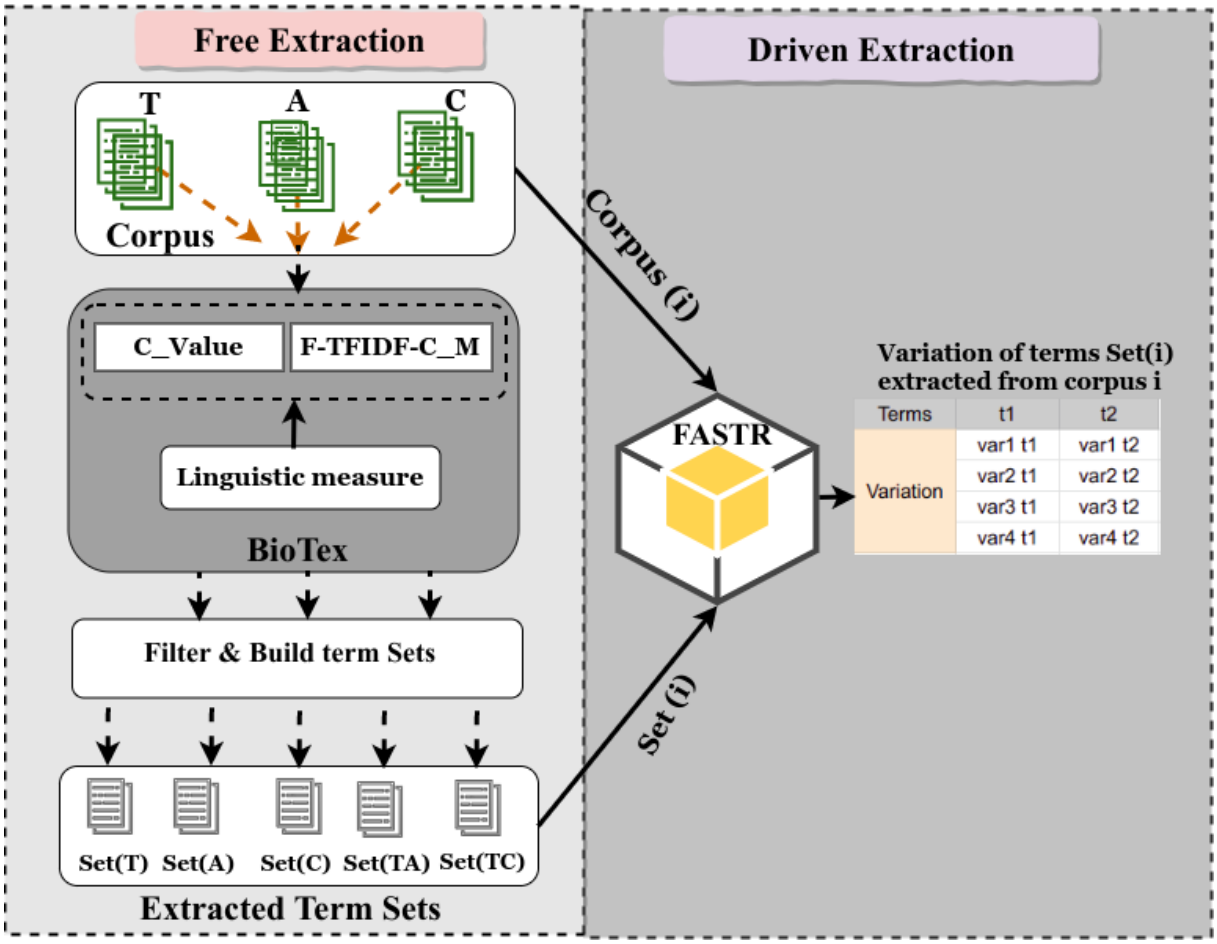


FIGURE 25 – Combinaison proposée pour l'extraction de termes

use, non-commercial use et le custom license.

Nous avons appliqué principalement trois étapes de pré-traitement sur les données de chaque corpus (Papers1, Papers2) afin de créer trois sous-corpus selon leur titre, leur résumé et leur contenu :

— Title : représente le corpus qui contient uniquement les titres des articles ;
— Abstract : représente le corpus qui ne contient que les résumés des articles ;
— Content : représente le corpus qui contient uniquement les contenus des articles.

Nous les avons nommés respectivement PapersX-title, PapersX-abstract et PapersX-content, voir le Tableau 4.1 pour plus de détails et le Tableau 4.2 pour les définitions des acronymes que nous avons utilisés.

Dans la Section 4.4.2, les données présentées ci-dessus sont utilisées pour l'application de l'approche méthodologique proposée.

4.4.2 Paramètres d'expérimentations

Tout au long des expérimentations, nous avons utilisé les paramètres suivants pour l'extraction des termes avec BioTex. Nous avons utilisé C_Value et F-TFIDF-C_M comme

	$NB_d(C)$	$NB_M(d)$	$std(c)$
Papers1			
Papers1-title	9315	15	± 8
Papers1-abstract	9315	180	± 94
Papers1-content	9315	4639	± 359
Papers2			
Papers2-title	32322	13	± 10
Papers2-abstract	32322	168	± 88
Papers2-content	32322	4913	± 720

TABLE 4.1 – Statistiques sur le jeu de données COVID19-MOOD-data

Abréviations	Description
$NB_d(C)$	nombre de document dans le corpus
$NB_M(d)$	nombre moyen de mots d'un document dans le corpus
$std(c)$	écart-type entre documents d'un corpus
NN	nom
$NNNN$	patrons pour les noms singuliers et pluriels
JJ	adjectif
NP	nom propre

TABLE 4.2 – Définition des acronymes

mesures statistiques, 50 patrons différents ou règles d'extraction de termes, et un nombre de mots allant de 1 à 4 grammes ($n = 1, 2, 3, 4$) par terme. Ces paramètres sont appliqués aux corpus décrits dans la Section 4.4.1. Le choix de C_Value et de F-TFIDF-C_M est basé sur les résultats d'études précédentes [Lossio-Ventura et al., 2016, Frantzi et al., 2000] qui ont montré que les deux mesures permettent d'extraire efficacement des termes simples (ou Single Word Terms SWT) et les termes composés ou multi-termes (ou Multi-Word Terms MWT).

Bien avant d'appliquer BioTex, certains pré-traitements spécifiques ont été réalisés sur les corpus Papers1-content et Papers2-content en raison de leur taille. Papers1-content a été sub-divisé en 09 sous-corpus (8 sous-corpus de 1000 documents chacun et 1 sous-corpus de 1315 documents) et Papers2-content en 32 sous-corpus (31 sous-corpus de 1000 documents chacun et 1 sous-corpus de 1332 documents). Chaque corpus a été divisé en unités plus petites pour faciliter le processus d'extraction. Les résultats obtenus à partir des unités plus petites ont ensuite été combinés en calculant la moyenne des valeurs de classement ou ranking. Le ranking final d'un terme donné sera donc égal à la moyenne de ses valeurs de ranking dans tous les sous-corpus dans lesquels il était présent. Le résultat final donne un ensemble de termes, classés par ordre croissant selon les valeurs de classement.

Le Tableau 4.3 présente un exemple de multi-termes ou MWT obtenus à l'aide de BioTex. La colonne *Terms* contient les termes extraits, la colonne *in_umls* indique si le terme correspondant est disponible dans le métathésaurus [Bodenreider, 2004] du système de langage médical unifié (UMLS) ou non, et la colonne *rank* indique l'importance du terme sur la base de la mesure statistique par rapport à la liste complète des termes issus du corpus. Dans notre étude, nous avons utilisé UMLS comme référence pour les termes extraits car notre étude est liée à une analyse de terminologies du domaine biomédicale. Cette comparaison vise à séparer les nouvelles terminologies ou les terminologies qui n'étaient pas encore répertoriées dans UMLS.

Terms	**in_umls**	**rank**
public health	1	1602.3971
respiratory syndrome	0	1481.9399
infectious disease	1	1198.2317
virus infection	1	1126.9083
influenza virus	1	1023.8858
immune response	1	1008.0362

TABLE 4.3 – Exemple de termes extraits avec BioTex

4.4.3 Approche non-guidée pour l'extraction de terminologie

Nous avons utilisé BioTex, comme indiqué dans la Section 4.3.1, pour extraire les termes des corpus en utilisant l'approche non-guidée. Plusieurs analyses sont effectuées ci-dessous sur les résultats obtenus. À cette fin, nous avons mené les expérimentations pour répondre à trois questions principales : 1) pour chaque corpus, quels sont les termes ou concepts thématiques les plus représentatifs (termes qui résument le contenu principal du corpus) par mesure statistique ? 2) pour chaque corpus, quels sont les concepts les plus représentatifs pour les deux mesures (C_Value et F-TFIDF-C_M) ? et 3) quels sont les concepts discriminants et communs dans l'ensemble des corpus ?

Pour chaque cas, nous déterminons si les termes extraits existent ou non dans le thésaurus UMLS.

4.4.3.1 Extraction de termes représentatifs dans un corpus

Dans cette section, nous illustrons comment des termes représentatifs peuvent être extraits de différents ensembles de données. Sur la base des mesures de BioTex, un terme est plus important qu'un autre dans un corpus donné s'il a un classement plus élevé que l'autre terme.

La Figure 26 montre des termes représentatifs pour les corpus Titre, Résumé et Contenu avec les mesures statistiques correspondantes (voir Tableau A.2 et A.3 pour plus de détails). Elle met en évidence les termes qui sont importants dans chaque partie des

articles. Notons que les termes extraits sont différents pour chaque mesure et sous-corpus, mais certains d'entre eux sont commun aux deux mesures. Par exemple, des termes comme *public health, immune responses* sont extraits à la fois par C_Value et F-TFIDF-C_M à partir des corpus obtenus avec les résumés.

Afin de montrer quantitativement le nombre de termes représentatifs qui sont à la fois extraits des différents corpus, nous montrons les termes communs entre les corpus Titre vs Résumé, et Titre vs Contenu pour le corpus Papers2 dans la Figure 27. Pour les deux mesures, les termes du titre sont plus représentatifs dans le résumé que dans le contenu des articles, c'est-à-dire 57% et 27% par rapport à 28% et 5%, respectivement, pour Titre vs Résumé et Titre vs Contenu. Cependant, nous avons remarqué que les termes extraits avec C_Value généraient plus de termes communs que ceux extraits avec F-TFIDF-C_M. Les termes communs représentent les termes extraits à la fois dans les corpus Titre, Résumé et Contenu pour chaque mesure.

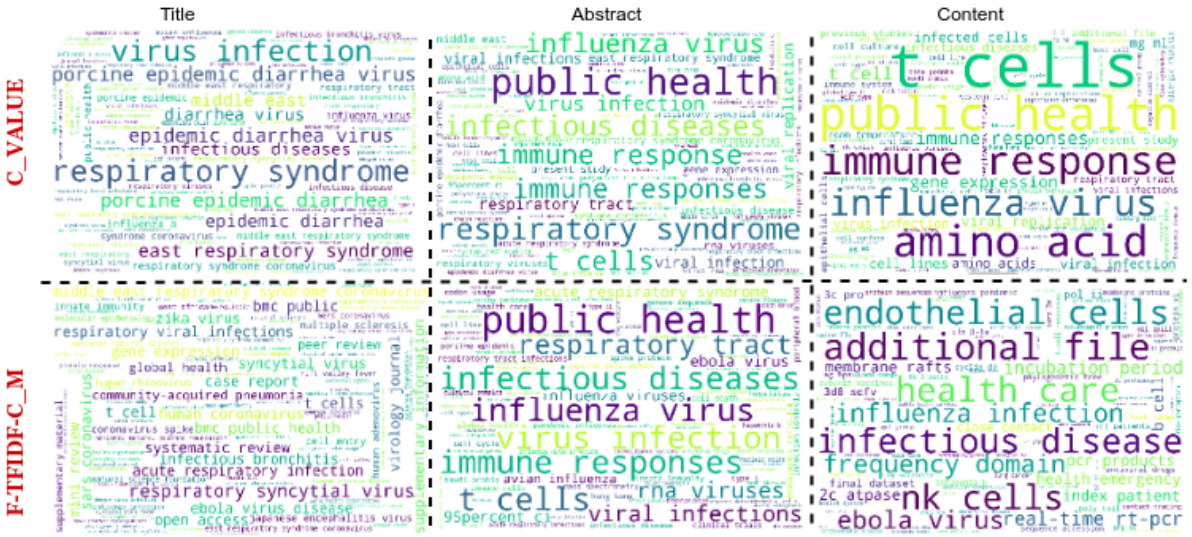


FIGURE 26 – Termes représentatifs extraits du corpus Papers1

Comme déjà indiqué, les termes extraits ont été comparés au thésaurus UMLS. Le Tableau 4.4 montre le TOP@20 des termes extraits pour le corpus de contenu Papers1 en utilisant les mesures C_Value et F-TFIDF-C_M. Les termes en gras sont ceux qui ne figurent pas dans UMLS.

Selon ces TOP@20 termes, nous pouvons noter que :

— la majorité des SWTs ou termes simples sont dans UMLS pour les deux mesures statistiques C_Value et F-TFIDF-C_M ;

— pour les MWTs ou multi-termes, plusieurs ne sont pas dans UMLS. Ces termes peuvent être classés comme suit :

  — *Sous-termes UMLS :* il s'agit de termes qui ne correspondent pas exactement à ceux présents dans le thésaurus d'UMLS mais qui pourraient constituer des sous-parties. Par exemple, *health emergency* fait partir du terme comme *Emer-*

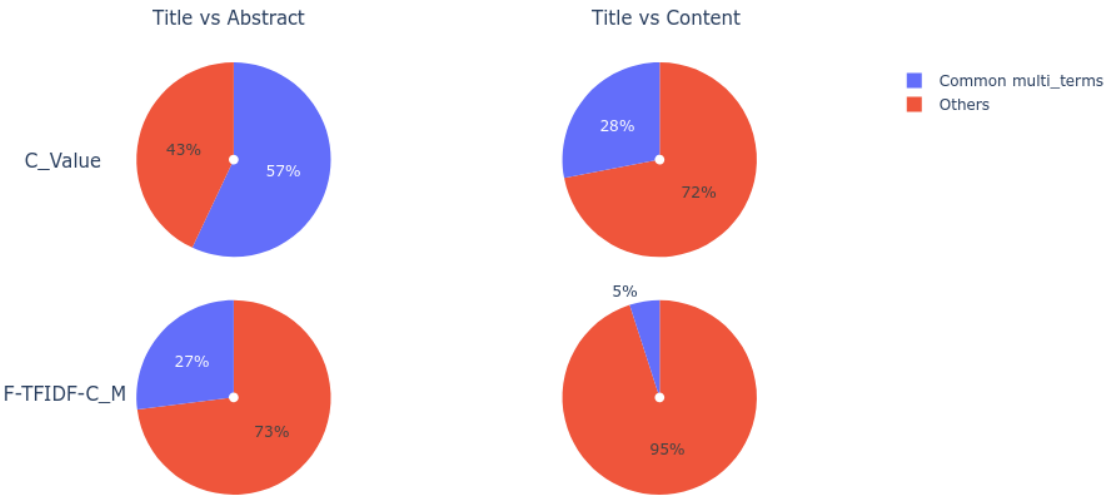


FIGURE 27 – Termes communs extraits du corpus Papers2

gency Health Services dans UMLS ;

— *Nouveaux termes :* ces termes ne figurent pas dans UMLS, mais sont significatifs (ou non) dans le contexte de la COVID-19. Par exemple, des termes comme *close contact* se rapportent au mode de transmission du virus COVID-19.

Les figures 28 et 29 illustrent en pourcentage, le nombre de termes parmi les TOP@100 pour chaque mesure (C_Value, F-TFIDF-C_M) et pour chaque dataset (Papers1-title, Papers2-title) :

— In_UMLS : nombre de termes dans UMLS ;

— Not_In_UMLS_V : nombre de termes qui ne sont pas parfaitement identiques aux termes UMLS, mais qui ont des variantes ou font partir des termes UMLS ;

— Not_In_UMLS : nombre de termes qui ne correspondent pas du tout aux termes UMLS. Nous les indiquons comme de nouveaux termes. Termes qui ne sont pas dans UMLS mais qui pourraient avoir une plus grande signification dans le contexte de l'étude ou qui pourraient être ajoutés à UMLS.

D'après ces statistiques obtenues à partir des corpus de titres Papers1_Title et Papers2_Title, nous constatons tout d'abord que les mesures C_Value et F-TFIDF-C_M permettent l'extraction de termes plus conventionnels ou de termes du thésaurus UMLS quel que soit le corpus. Ensuite, nous constatons que F-TFIDF-C_M génère plus de nouveaux termes (Not In UMLS) que C_Value, quel que soit le corpus. Enfin, le nombre de nouveaux termes est plus important avec les MWTs (Figure 29) que les SWTs (Figure 28) et ce, quelle que soit la mesure.

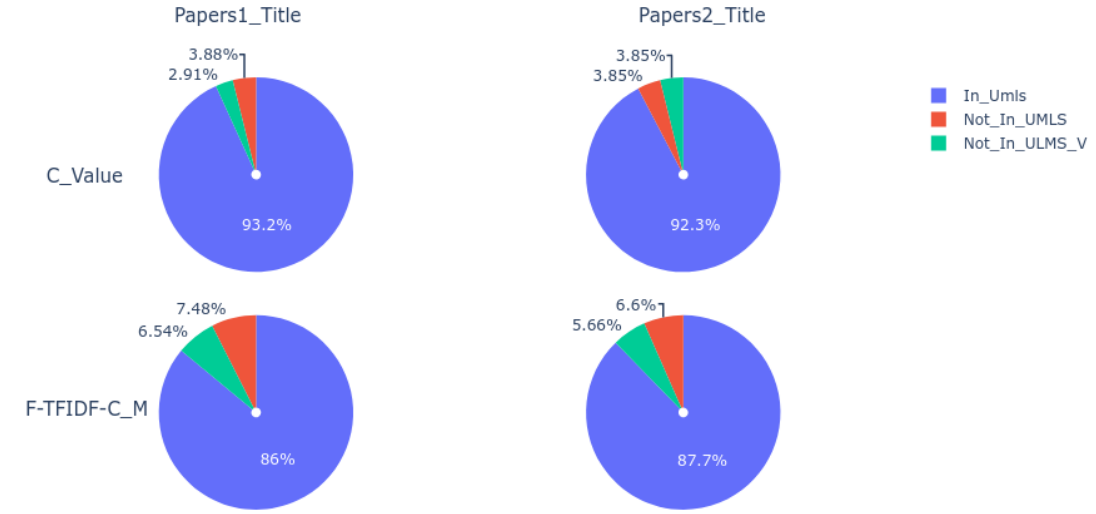


FIGURE 28 – C_Value vs F-TFIDF-C_M SWTs

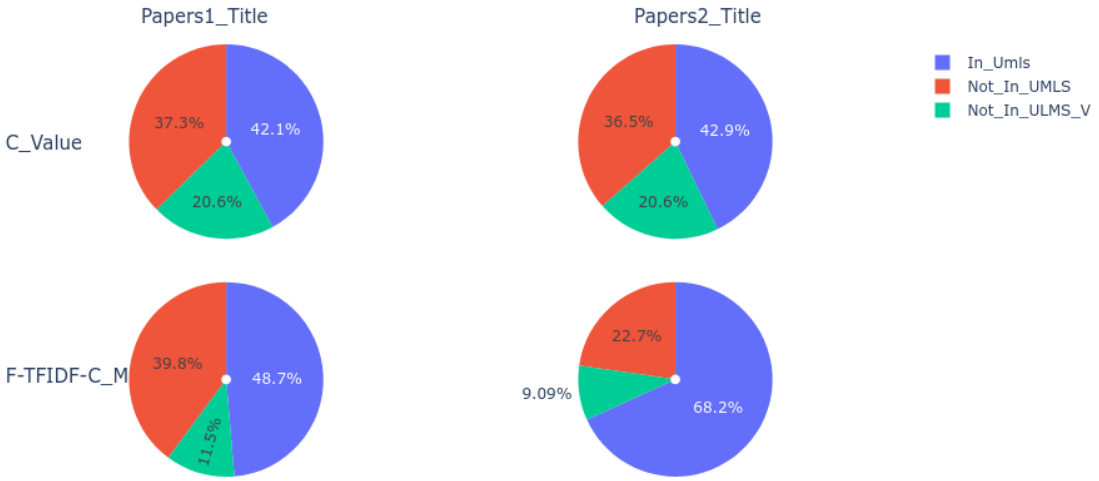


FIGURE 29 – C_Value vs F-TFIDF-C_M MWTs

4.4.3.2 Extraction de termes pertinents à partir de corpus avec les mesures C_Value et F-TFIDF-C_M

Il s'agit d'une analyse quantitative et qualitative des termes extraits de chaque corpus en tenant compte des deux mesures (C_Value et F-TFIDF-C_M). En d'autres termes, elle consiste à analyser et comparer les termes obtenus pour les deux mesures, c'est-à-dire les termes détectés en même temps, ainsi que les termes spécifiques à chacune d'elles.

L'analyse quantitative vise à mettre en évidence, pour chaque jeu de données (corpus), le nombre de termes obtenus pour chaque mesure, le nombre de termes obtenus pour les deux mesures, et qui sont disponibles ou non dans le thésaurus UMLS. Tandis que l'analyse qualitative vise à mettre en évidence, dans chaque cas, l'importance ou la pertinence des termes obtenus par rapport au domaine étudié.

Pour les différentes illustrations, nous tirons profit du diagramme de Venn [Ho et al., 2021] pour faire ressortir la distribution des termes du corpus Papers2-title, illustrée dans la Figure 30. Les termes sont organisés en différentes sections. Par exemple, *gene expression, human coronavirus, case report, public health, respiratory syncytial virus, etc.* sont disponibles dans UMLS et sont extraits par les deux mesures (C_Value et F-TFIDF-C_M). Selon le domaine étudié, ces termes auront tendance à être plus représentatifs et importants dans l'ensemble du corpus. De plus, pour chaque mesure, il y a de nouveaux termes qui ne sont pas dans UMLS.

4.4.3.3 Extraction de termes discriminants et communs à partir de plusieurs corpus

Dans cette partie, l'analyse des termes est effectuée en considérant conjointement plusieurs corpus ou corpora, c'est-à-dire entre les corpus de Titre, Résumé et Contenu. La Figure 31 correspond à l'extraction de termes qui sont discriminants et en commun à partir des corpus Papers1-title, Papers1-abstract et Papers1-content.

Il existe des termes communs dans l'ensemble du corpus, tels que *gene expression, virus replication, influenza virus, etc.*. Ces termes ont tendance à être pertinents dans les corpus Title, Content et Abstract pris conjointement.

De plus, les ensembles de termes, comme *[respiratory infection, acute respiratory infection, etc.]*, *[innate immune response, endoplasmic reticulum, etc.]*, et *[nucleotide sequences, room temperature, etc.]* sont des termes discriminants dans les corpus Titre, Résumé et Contenu pris conjointement.

4.4.4 Approche guidée pour l'extraction de terminologie

Nous avons élaboré une approche guidée (ou driven approach) pour l'extraction de termes et cela grâce à l'outil FASTR [Jacquemin, 1994]. Notre proposition répond à deux questions principales : 1) Pour un ensemble de termes, comment peut-on extraire de nouvelles variantes de ces termes à partir d'un corpus donné ? 2) Certains de ces nouveaux termes ou variantes, existent-ils dans le thésaurus UMLS ? Lors de nos expérimentations, nous avons utilisé les termes communs conjointement obtenus à partir des corpus Title, Abstract et Contenu, dans la Section 4.4.3.3, du fait que ces termes étaient plus représentatifs et pertinents dans l'ensemble des corpus. La Figure 32 montre un exemple de

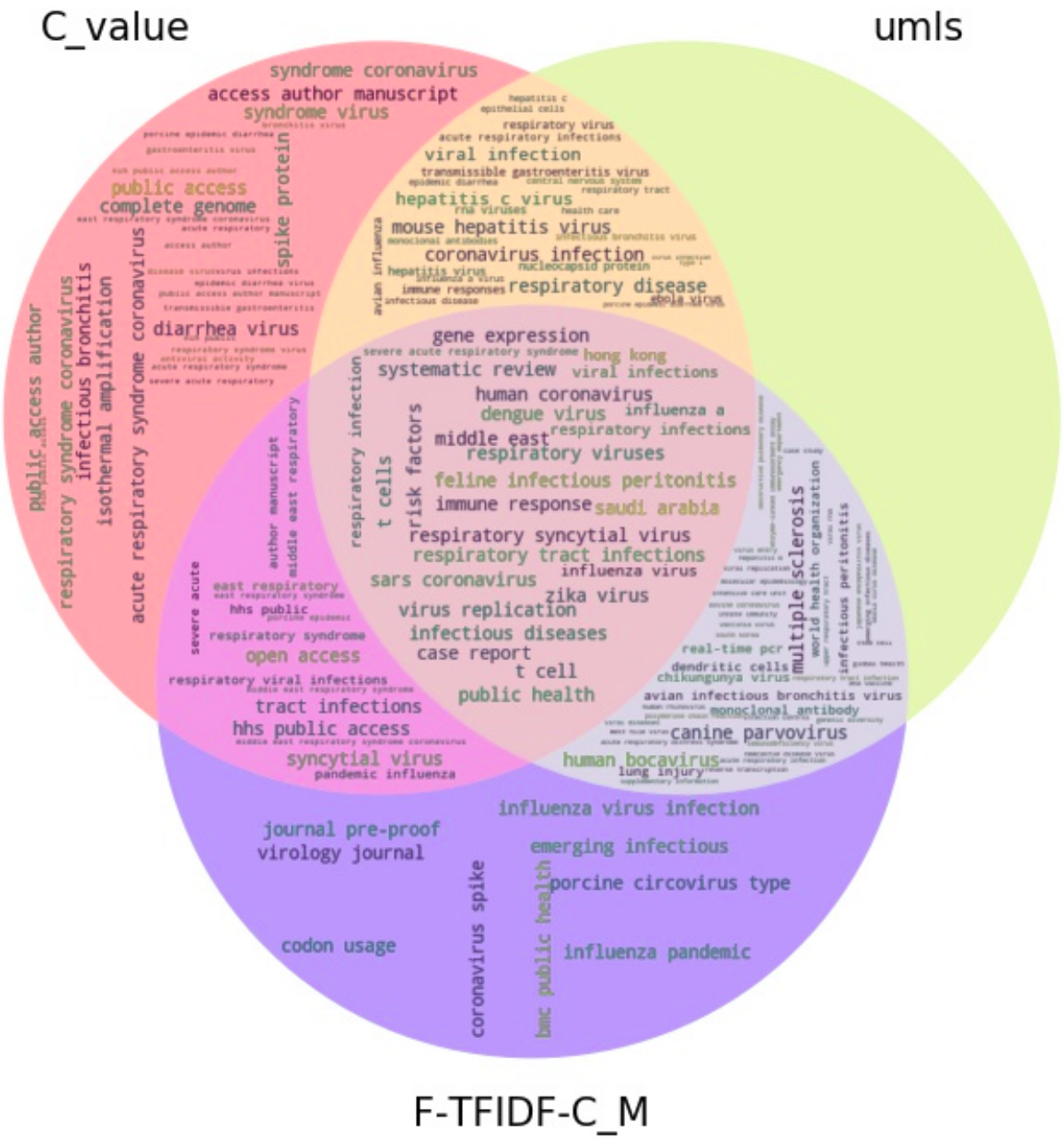


FIGURE 30 – Distribution des termes selon les mesures et leur présence dans UMLS. : Corpus Papers2-title

variantes extraites avec le terme *infectious disease*.

Le Tableau 4.5 contient une liste des TOP@5 des variantes extraites avec six termes initiaux. Parmi eux, nous avons mis en évidence (en gras) les termes qui sont dans UMLS. Tout comme l'extraction avec l'approche non-guidée, cette approche peut être utilisée pour extraire ou enrichir un ensemble de termes.

FIGURE 31 – Distribution des termes représentatifs en prenant en compte plusieurs corpus avec C_Value. : Corpora Papers1

4.4.5 Stratégies de combinaisons proposées pour l'analyse des termes

Les stratégies d'analyses des termes proposées prend en compte deux niveaux : 1) l'extraction de termes en intra-corpus, c'est-à-dire à partir d'un seul corpus, et 2) l'extraction de termes inter-corpus ou à partir de plusieurs corpus.

— Stratégies d'extraction de termes en intra-corpus : elles visent à extraire les termes communs ou discriminants d'un corpus donné. À cette fin, les termes extraits à partir des deux mesures sont comparés. Nous montrons le processus dans la Figure

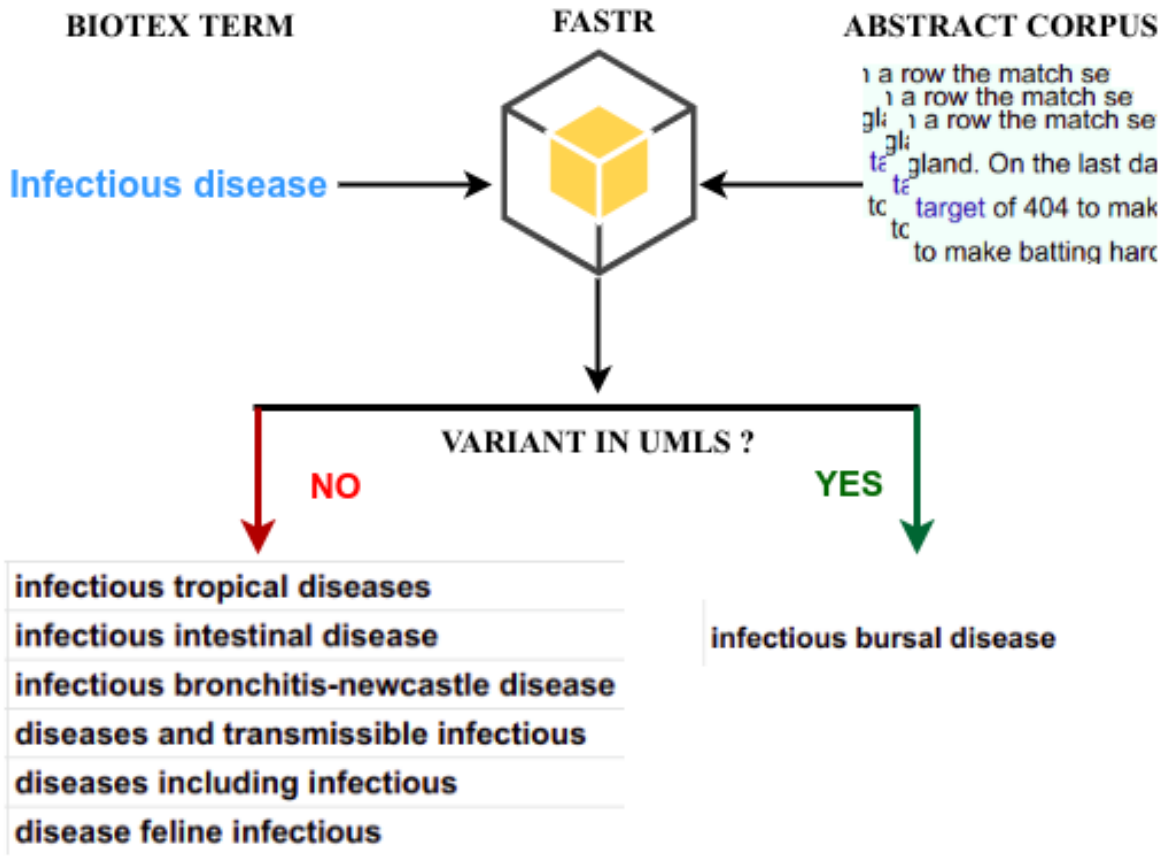


FIGURE 32 – Exemple de variantes du termes *infectious disease* obtenues avec le corpus Abstract

33, où l'ensemble des termes Set(Cp) extraits du corpus Cp (Titre, Résumé ou Contenu) en utilisant chaque mesure (C_Value, F-TFIDF-C_M) sont comparés conjointement avec les termes présents dans UMLS. L'ensemble **A** représente les termes du corpus extraits spécifiquement avec C_Value, l'ensemble **B** représente les termes spécifiques à F-TFIDF-C_M, tandis que l'ensemble **C** représente les termes communs aux deux mesures et qui sont présents dans UMLS. Nous considérons que les ensembles **A** et **B** sont des termes discriminants du corpus selon les mesures, et l'ensemble **C** est considéré comme contenant des termes communs ou les termes les plus représentatifs du corpus.

Le processus d'extraction de nouveaux termes avec FASTR est fait avec l'un des ensembles combinés (discriminant ou commun) et le corpus Cp concerné.

— Stratégies d'extraction de termes en inter-corpus : elles visent à extraire les termes communs et discriminants, tout en prenant en compte plusieurs corpus pour une mesure donnée. Comme l'illustre la Figure 34, pour chaque mesure (C_Value ou F-TFIDF-C_M), les ensembles de termes Set(Cp1), Set(Cp2), Set(Cp3) sont extraits respectivement des corpus Cp1 (Titre), Cp2 (Résumé), et Cp3 (Contenu). Ces ensembles sont comparés afin de calculer l'ensemble de termes communs **D** pour l'ensemble des corpus, et les ensembles de termes discriminants **A**, **B**, **C** res-

pectivement pour les corpus Cp2, Cp1 et Cp3. Dans ce contexte, les nouveaux ou les variantes des termes sont extraits en utilisant l'un des ensembles combinés avec un corpus (Cpx).

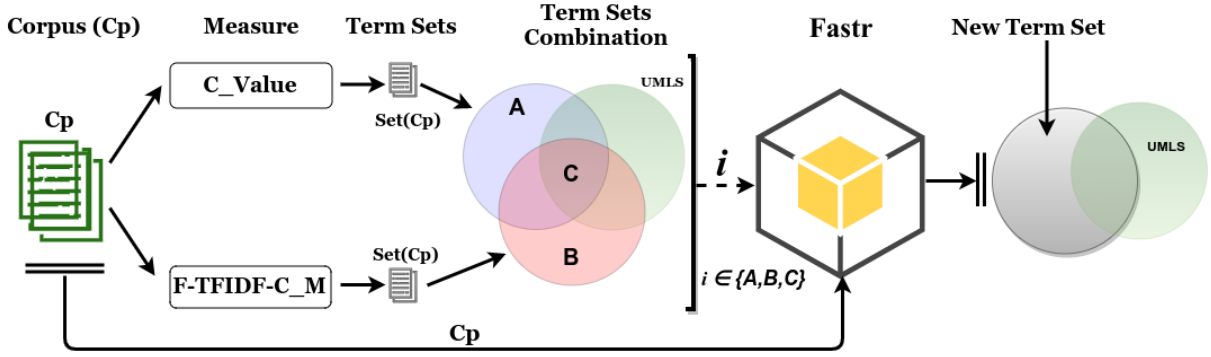


FIGURE 33 – Stratégies d'extraction de termes en intra-corpus

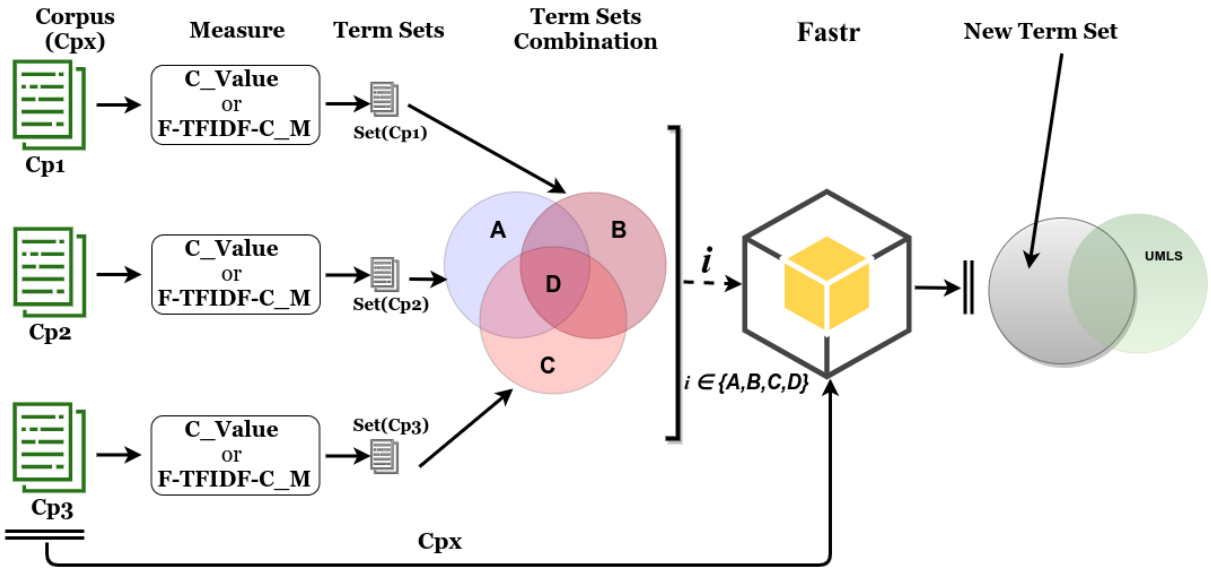


FIGURE 34 – Stratégies d'extraction de termes en inter-corpus

4.4.6 Évaluation : Intelligence épidémique

L'intelligence épidémique (IE) a pour but de détecter, de rechercher et de surveiller les menaces potentielles pour la santé en temps opportun [Paquet et al., 2006]. Outre les systèmes de surveillance classiques, tels que les avertissements d'épidémie de l'Organisation mondiale de la santé animale (OIE), le processus d'IE intègre de plus en plus de données non structurées provenant de sources informelles telles que les informations en ligne. Plusieurs systèmes de surveillance basés sur le Web ont été développés et utilisés pour soutenir la surveillance de la santé publique et de la santé animale (ProMED [Madoff, 2004], HealthMap [Freifeld et al., 2008], GPHIN [Mykhalovskiy and Weir, 2006], PADI-web [Valentin et al., 2020], etc.) Dans cette étude de cas, nous nous sommes concentrés

sur le choix des mots-clés avec le système PADI-Web pour la surveillance du COVID-19 (c'est-à-dire la surveillance guidée par des outils automatiques) et pour la surveillance des maladies inconnues (c'est-à-dire la surveillance syndromique).

La plateforme d'extraction automatisée d'informations sur les maladies à partir du web (PADI-web [5]) est un système de surveillance automatisé permettant de suivre l'émergence de maladies infectieuses animales, y compris les zoonoses [Arsevska et al., 2018, Valentin et al., 2020]. PADI-web surveille les actualités de Google News par le biais de flux RSS (really simple syndication) spécifiques, en ciblant les maladies d'intérêt (par exemple, la peste porcine africaine, la grippe aviaire, etc.). PADI-web utilise également des flux RSS non spécifiques, composés de combinaisons de symptômes et d'hôtes (c'est-à-dire d'espèces), permettant ainsi une surveillance syndromique et la détection d'événements pathologiques inhabituels. Les flux RSS sont constitués de combinaisons de différentes catégories de termes (c'est-à-dire de mots-clés), notamment des symptômes, des noms de maladies et des espèces.

PADI-Web a été utilisé pour la surveillance de la maladie de la COVID-19. Dans ce contexte, le choix des termes de surveillance du COVID-19 est crucial.

Dans les sous-sections suivantes, nous discutons du choix des termes donnés par ITEXT-BIO à utiliser dans le système PADI-Web [Valentin et al., 2020] et d'autres systèmes de surveillance basés sur le web pour la surveillance syndromique ainsi que de la COVID-19 [Madoff, 2004, Freifeld et al., 2008, Mykhalovskiy and Weir, 2006]. Cela permet d'évaluer la pertinence des termes générés par notre approche pour une tâche dédiée, c'est-à-dire la surveillance sanitaire basée sur le web.

4.4.6.1 Extraction de termes pertinents

Nous avons comparé la pertinence des TOP@10 des termes extraits des corpus Papers2 avec C_Value ou F-TFIDF-C (Tableau 4.6). Le Tableau A.4 donne plus de détails sur ces termes. La pertinence a été évaluée en classant les termes dans une ou plusieurs des catégories suivantes :

— Surveillance ou veille de la COVID-19 : termes épidémiologiques spécifiques au COVID-19 (par exemple *coronavirus spike*).
— Surveillance ou veille syndromique : termes épidémiologiques non spécifiques à une maladie particulière (par exemple, *infectious bronchitis*).
— Pertinence au domaine d'étude : termes liés à la santé, c'est-à-dire soit à des maladies spécifiques (par exemple, *porcine epidemic diarrhoea*), soit non spécifiques (par exemple, *virus infections*). La catégorie Pertinence au domaine comprend donc les deux catégories précédentes, en d'autres maladies que la COVID-19.

5. `https://padi-web.cirad.fr/en/`

— Partie d'une expression multi-mots ou multiword expression (MWE) désignant une maladie : partie d'une expression multi-mots correspondant à un nom de maladie (par exemple, *East respiratory syndrome* pour *Middle East syndrome coronavirus*).

Parmi les termes extraits avec C_Value des corpus Titres, Résumés ou conjointement du corpus Titres et Résumés, six à sept font partie d'expressions de maladie multi-mots ou MWE. Seul un terme extrait avec F-TFIDF-C_M fait partie des MWE. C_Value pourrait donc présenter un intérêt particulier pour l'extraction de variantes de noms de maladies, même s'ils sont incomplets.

Pour les termes pertinents sur la surveillance de la COVID-19 et syndromique spécifique au domaine d'étude, F-TFIDF-C_M a obtenu de meilleurs résultats que C_Value, même lorsque la fréquence des termes pertinents était faible (de un à cinq termes sur dix). Aucun terme commun n'a été extrait de (Titre + Résumé) ou de (Titre + Contenu) en utilisant F-TFIDF-C_M. En utilisant C_Value, seuls trois termes communs ont été extraits du Titre + Contenu. Parmi les 10 premiers termes extraits du Titre + Résumé avec ces métriques, sept étaient des parties des expressions multi-mots désignant des maladies. Sans tenir compte de la catégorie de termes, nous avons extrait plus de termes pertinents des Titres et des Résumés que dans les Contenus. Cela correspond au fait que les titres et les résumés sont plus riches en informations clés et en termes pertinents en raison de leur longueur limitée.

4.4.6.2 Extraction de termes par approche guidée ou driven term extraction

Nous avons sélectionné six termes, extraits dans la Section 4.4.6.1 : *respiratory tract, viral infections, SARS coronavirus, incubation period, influenza virus, respiratory infections and infectious diseases*. Nous avons extrait les variantes de manière aléatoire avec FASTR (voir Section 4.3.2). Un épidémiologiste a évalué manuellement la pertinence de 10 variants sélectionnés aléatoirement parmi les variantes de chacun des six termes . Parmi les 60 termes (10*6 termes) évalués (voir Tableau 4.7), 72% (43/60) sont pertinents et 7% (4/60) sont non pertinents. Pour 13 variantes (22%), la pertinence n'a pas pu être évaluée car l'expression était tronquée et ambiguë, telle que *disease has an infectious* pour le terme *infectious diseases*.

FASTR se révèle donc être un outil efficace pour générer des variantes de termes. Cependant, nous avons remarqué que FASTR générait jusqu'à 774 variantes pour un seul terme. Ainsi, pour éviter une sélection aléatoire des termes, il serait intéressant de calculer un indice de pertinence qui pourrait être utilisé pour classer les variantes proposées, afin de pouvoir mettre en avant ceux ayant une forte pertinence. Par ailleurs, plusieurs variants extraites étaient des fragments d'expressions qui n'ont pas pu être évalués. Ce problème pourrait être résolu en affichant le contexte de la variante (c'est-à-dire la phrase dans laquelle la variante apparaît).

4.5 Conclusion

Dans cet chapitre, nous décrivons ITEXT-BIO, une approche méthodologique générique appliquée à l'extraction de termes dans le domaine biomédical. Nous montrons comment elle permet aux utilisateurs d'extraire des termes (ou des concepts) à partir de différents types de données textuelles en combinant plusieurs stratégies. L'approche d'extraction de termes non-guidée extrait des termes à partir d'un corpus, tandis que l'approche guidée extrait à partir d'un corpus et d'un ensemble de termes, des variantes de cet ensemble de termes.

Nous illustrons que les combinaisons de stratégies proposées, basées sur des mesures statistiques et des segments textuels (Titre, Abstract, Contenu), permettent d'extraire et de catégoriser efficacement les termes (termes représentatifs, discriminants et nouveaux) à partir d'un ou de plusieurs corpus. Nous avons également analysé quantitativement et qualitativement les termes extraits pour déterminer ceux qui sont liés au domaine d'application et ceux qui pourraient être considérés comme une terminologie émergente pour le suivi des maladies.

Nos perspectives se concentreront sur l'extraction et l'analyse des termes en : (i) en prenant en compte différentes sections d'articles et en appliquant la méthodologie à différents types de corpus dérivés de journaux ou des réseaux sociaux tels que Twitter, (ii) en considérant des combinaisons d'outils autres que BioTex, et (iii) en introduisant des techniques de word embedding comme BERT [Devlin et al., 2018] pour capturer les aspects sémantiques des termes extraits afin de réduire l'ambiguïté du contexte.

L'approche méthodologique proposée dans ce chapitre vient en complément du protocole de collecte proposé dans le Chapitre 3. En effet, le but de ITEX-X (X, étant la dénomination du domaine) est de permettre l'extraction des termes les plus importants et/ou émergents d'une thématique donnée. Dans le processus de collecte et de mise à jour des données de la Métropole de Montpellier, les termes émergents obtenus avec ITEX-X seront utilisés pour mettre à jour et/ou constituer des vocabulaires de concepts thématiques évoqués dans la Section 3.3.1, qui sont utilisés pour la constitution de corpus thématiques.

	C_Value Measure			
	SWTs			
TOP 20	cells	virus	protein	study
	data	**figure**	patients	expression
	rna	analysis	disease	p
	mice	c	influenza	number
	MWTs			
TOP 20	t cells	public health	immune response	gene expression
	viral replication	**infected cells**	viral infection	virus infection
	mg ml	infectious diseases	respiratory tract	epithelial cells
	previous studies	room temperature	**additional file**	viral infection

	F-TFIDF-C_M Measure			
	SWTs			
TOP 20	mice	patients	influenza	proteins	health
	dna	vaccine	transmission	research	model
	children	outbreak	vaccination	e	china
	peptide	fusion	network	percent	mers-cov
	MWTs			
TOP 20	**additional file**	infectious disease	nk cells	health care	endothelial cells
	frequency domain	ebola virus	**influenza infection**	**real-time rt-pcr**	incubation period
	health emergency	**index patient**	**membrane rafts**	**pcr products**	**2c atpase**
	b cell	**close contact**	**final dataset**	**3d8 scfv**	pol ii

*Les termes en **gras** ne sont pas dans UMLS*

TABLE 4.4 – TOP@20 des termes extraits du corpus Paper1-content avec C_Value et F-TFIDF-C_M - SWTs vs MWTs

Terms	infectious disease	virus replication	laboratory tests	respiratory syndrome	preventive measure	syndrome coronavirus
Variants	diseases including infectious	**replication competent viruses**	laboratory confirmation tests	**respiratory distress syndrome**	**preventive measures**	syndrome coronavirus-related coronavirus
	infectious pulmonary diseases	replication of N1347A virus	**laboratory testing**	respiratory acute syndrome	preventive hygienic measures	Syndrome human coronavirus
	infectious bursal disease	virus optimal replication	Testing presents isolation laboratories	syndrome coronavirus and respiratory	prevention community-engaged measures	Syndromic Surveillance Coronavirus
	infectious lung diseases	replicating influenza viruses	laboratory diagnostic testing	respiratory tract syndromic	preventive health measures	syndrome virus coronavirus
	infectious acute disease	replication of human viruses	laboratory genomic testing	respiratory insufficiency syndrome	preventive behavioral measures	Coronavirus Associated Syndromes

*Termes en **gras** dans UMLS*

TABLE 4.5 – Variantes des termes extraites en utilsant FASTR

Corpus (Papers2)	**Mesure**	**n**	**Pertinence au domaine**	**Veille COVID-19**	**Veille mique**	**Syndro-**	**Partie d'une MWE**
title	C_Value	10	3	0	2		6
title	F-TFIDF-C_M	9	4	1	1		1
abstract	C_Value	10	1	0	0		6
abstract	F-TFIDF-C_M	10	5	1	2		1
content	C_Value	10	0	0	0		1
content	F-TFIDF-C_M	10	2	0	2		0
title + abstract	C_Value	10	3	0	0		7
title + abstract	F-TFIDF-C_M	0	-	-	-		-
title + content	C_Value	3	1	0	0		2
title + content	F-TFIDF-C_M	0	-	-	-		-

TABLE 4.6 – Termes pertinents extraits du corpus Papers2 en fonction des mesures statistiques (C_Value or F-TFIDF-C_M)

influenza virus	**evaluation**	**respiratory infections**	**evaluation**	**infectious diseases**	**evaluation**
influenza a/wsn/33 virus	not relevant	respiratory virus infections	relevant	diseases relates to infectious	relevant
viruses and conventional influenza	relevant	respiratory viral infection	relevant	disease called feline infectious	relevant
virus remains the influenza	lack of context	infection by respiratory	relevant	infectious animal diseases	relevant
influenza by virus	not relevant	infections of the respiratory	relevant	infectious enteric diseases	relevant
influenza vaccine virus	relevant	infect respiratory	relevant	disease without being infectious	not relevant
virus and canine influenza	relevant	infections are respiratory	relevant	disease has an infectious	lack of context
virus influenza	relevant	infected with respiratory	relevant	infectious disease	relevant
viruses such as influenza	relevant	infection with other respiratory	relevant	disease named it infectious	lack of context
influenza b viruses	relevant	respiratory virus infection	relevant	infectious swine diseases	relevant
viruses and emerging influenza	relevant	infection transmitted via respiratory	relevant	disease models for infectious	lack of context
viral infections	**evaluation**	**sars coronavirus**	**evaluation**	**incubation period**	**evaluation**
viral bronchopulmonary infection	relevant	coronavirus is urbani sars	not relevant	incubating period	relevant
virally infected	relevant	coronavirus of 18 sars	not relevant	periods of incubation	relevant
viral respiratory infections	relevant	coronavirus that causes sars	relevant	incubation periods	relevant
infection and encounter virally	lack of context	coronavirus named sars	relevant	period of incubation	relevant
infection or viral	lack of context	coronavirus related to sars	relevant	period than incubation	lack of context
viral skin infection	relevant	coronavirus isolated from sars	relevant	period and incubation	lack of context
virals infection	relevant	coronavirus responsable du sars	relevant	incubation for period	lack of context
infection with one viral	relevant	sars -associated coronavirus	relevant	period and incubating	lack of context
viral opportunistic infections	relevant	sars human coronavirus	relevant	period covering an incubation	relevant
infection at high viral	lack of context	sars and coronavirus	relevant	period of extrinsic incubation	relevant

TABLE 4.7 – 60 termes extraits aléatoirement des variantes obtenues avec FASTR (Section 4.2)

Chapitre 5

Stockage de données hétérogènes : Lac de données spatiales

Nous avons évoqué dans le Chapitre 2, Section 2.1.2, la forte hétérogénéité et l'abondance des données produites par les grandes villes. Il est nécessaire de penser à des solutions de stockage évolutives, qui s'adaptent à cette forte croissance, et aussi, à la diversité de ces données.

Dans ce chapitre, nous proposons une solution de stockage et de gestion de données basée sur les lacs de données : un lac de données spatiales. Cette solution de stockage est illustrée avec les données de la Métropole de Montpellier Méditerranée (3M).

Ce chapitre est structuré comme suit. La Section 5.1 correspond à l'introduction générale. Nous présentons les concepts et les travaux relatifs aux lacs de données et à la gestion des informations spatiales dans la Section 5.2. Dans la Section 5.3, nous proposons une solution de stockage et de gestion pour les données massives et hétérogènes de façon générale. Nous proposons ensuite un cas d'application de la solution proposée pour la Métropole de Montpellier dans la Section 5.4. La Section 5.5 conclut ce chapitre, tout en dressant les contributions et en les positionnant dans le contexte général de notre étude.

5.1 Introduction

Comme déjà mentionnée dans le Chapitre 2, la forte croissance des données que connaissent les villes intelligentes nous amène à nous poser un certain nombre de questions sur la gestion de ces données, et aussi sur leur exploitation.

Dans le cadre de notre étude, pour la métropole de Montpellier, le principal besoin exprimé par les utilisateurs est d'arriver à explorer sémantiquement de grandes quantités de données disponibles au sein de leur organisation. Parmi ces données (voir Chapitre 2, Section 2.2.1), certaines sont produites par les citoyens, d'autres par les différents services de la métropole et des municipalités associées (transport, tourisme, etc.). Il est donc difficile d'avoir une vue d'ensemble sur les informations à disposition.

Le principal inconvénient des outils existants est la difficulté pour les utilisateurs d'explorer de manière flexible un ensemble de données massives et hétérogènes. Plus précisément, les entrepôts de données [Devlin and Cote, 1996] s'avèrent moins flexibles pour permettre aux utilisateurs de construire de nouvelles analyses qui n'auraient pas été prévues [Madera and Laurent, 2016]. Pour résoudre ce problème, les lacs de données [Dixon, 2010] représentent un nouveau mode de gestion des données, avec un stockage total ou partiel des éléments associés (données et méta-données). Pour ces nouveaux systèmes, dont la théorisation est récente, il y a eu peu de travaux méthodologiques sur la conception de ces infrastructures de données, considérant qu'elles requièrent essentiellement des compétences techniques. Confrontés à la mise en place d'un lac de données en conditions réelles dans le contexte d'une ville intelligente, nous sommes amenés à contredire cette hypothèse.

Dans ce chapitre, nous nous intéresserons ainsi à la conception et à l'implémentation d'un lac de données, en partant d'une masse de données hétérogènes avec une forte composante spatiale. En nous inspirant de travaux précédents sur la normalisation de données spatiales [ISO/TC 211, 2019] et sur les lacs de données [Ravat and Zhao, 2019, Madera and Laurent, 2016, Sawadogo et al., 2019], nous développons une approche de conception, avec une illustration sur les données (voir Chapitre 3 et Tableau A.1) de la Métropole de Montpellier. Le code développé de notre proposition est mis à disposition de la communauté. Nous démontrons aussi que les infrastructures type lac de données ne sont pas réservées aux experts mais peuvent être proposées à d'autres utilisateurs à la condition de leur fournir une interface adaptée.

5.2 État de l'art

Plusieurs systèmes de gestion et de stockage de données ont émergé pour supporter le Big Data [McAfee et al., 2012]. Parmi eux, nous pouvons citer les bases de données NoSQL (Not Only SQL) [Bruchez, 2015], les entrepôts de données [Kimball and Ross, 2011, Phipps and Davis, 2002] et les lacs de données [Russom, 2017, Hai et al., 2016].

5.2.1 Entrepôts de données et lac de données

Les entrepôts de données ont été conçus comme une optimisation des bases de données relationnelles pour l'exécution de requêtes analytiques et sont utilisés comme support à la prise de décision dans les organisations. Les modèles conceptuels des entrepôts de données sont basés sur les concepts suivants : les faits (et mesures), les dimensions, les hiérarchies et les membres [Kimball and Ross, 2013]. De ce fait, concevoir un entrepôt de données revient à définir l'espace des tableaux croisés possibles, qui vont être construits par les utilisateurs pour explorer les données. Les entrepôts de données permettent une exploration facilitée de volumineux jeux de données par les utilisateurs. Mais la mise en place d'un entrepôt de données implique une normalisation des données entrantes issues de sources variées (cette normalisation pouvant être automatisée via un ETL). Malgré quelques propositions intéressantes (voir par exemple [Oukid et al., 2016] et [Minati et al., 2006]), l'intégration de documents et d'images satellitaires dans le même entrepôt de données reste une tâche complexe. Ainsi, la mise en place d'un entrepôt de données nécessite un long processus de préparation des données.

La définition d'un lac de données a été proposée par [Dixon, 2010]. Une comparaison détaillée avec les entrepôts de données a été réalisée dans [Madera and Laurent, 2016] puis reprise dans [Sawadogo et al., 2019]. Les lacs de données sont une solution récente qui a été développée pour répondre à la gestion des Big Data pour lesquelles les entrepôts de données montraient quelques faiblesses. Le principal problème rencontré avec les entrepôts de données est la gestion de données de natures hétérogènes. Un lac de données est une structure de stockage de données massives, qui intègre les données en provenance de différentes sources dans leur format natif, sans qu'il soit nécessaire de réaliser un traitement [Russom, 2017, Hai et al., 2016]. Selon [Sawadogo et al., 2019]. Un lac de données est un système évolutif de stockage et d'analyse de données, stockées dans leur format natif, destiné à des spécialistes tels que des statisticiens, des analystes et des "data scientists". Les principales composantes et caractéristiques des lacs de données sont :

- un catalogue de méta-données qui facilite l'accès aux données et en assure la qualité ;
- des outils de gestion des données ;
- l'accessibilité aux utilisateurs ;

— l'évolution possible des données ;
— l'ingestion des données de toute nature ;
— une organisation logique et physique.

Etant une nouvelle technologie Big Data, les lacs de données ont été étudiés dans de nombreux articles. En raison de leur capacité à gérer de larges volumes de données structurées et non structurées, une étude exploratoire a été réalisée pour mieux comprendre l'utilisation des lacs de données dans le contexte industriel [Llave, 2018]. Dans [Giudice et al., 2019, Mehmood et al., 2019], de nouvelles architectures de lac de données ont été conçues afin d'extraire des informations pertinentes d'un ensemble de données hétérogènes, en se basant sur les sources de ces données. Dans [Quix et al., 2016], un système de gestion de méta-données générique et extensible pour les lacs de données (Generic and Extensible Metadata Management System for Data Lakes, GEMMS) a été développé, en premier lieu pour extraire des méta-données des sources et en second lieu, pour enrichir les sources de données en utilisant des informations sémantiques venant à la fois des données et des méta-données. De nombreux systèmes de gestion de méta-données ont été ainsi proposés par la communauté, mais il reste encore des défis à relever dont en particulier la mise en lien sémantique des données [Nargesian et al., 2019].

En nous basant sur ces travaux, nous définissons un lac de données comme une structure de stockage composée de jeux de données, ayant les caractéristiques précédemment citées et celles décrites dans la Section 5.3.

5.2.2 Information géographique et normalisation

Plusieurs définitions ont été proposées pour le concept de territoire selon le domaine étudié. Dans [Moine, 2006], le territoire est considéré comme étant un système complexe et évolutif qui associe un ensemble d'acteurs, d'une part, et d'autre part, l'espace géographique que ces acteurs utilisent, développent et gouvernent. Dans [Simone et al., 2018], les auteurs quant à eux, considèrent que le territoire est un ensemble composé de trois dimensions : l'espace géographique, le temps et les relations sociales. Ils définissent le territoire comme étant un système complexe situé dans un espace géographique spécifique émergeant de la co-évolution d'un ensemble de processus hétérogènes (anthropologico-culturel, relationnel, cognitif et économico-productif) qui caractérise cet espace d'une manière unique et non répétitive.

Tout en prenant en compte les définitions proposées dans l'état de l'art, nous considérons que le territoire est :

— un ensemble d'acteurs physiques et/ou juridiques. Physique dans le sens où il est habité par un ou plusieurs groupes de personnes interagissant les uns avec les autres, et juridique au sens où il est composé de plusieurs organisations politiques, économiques, etc. ;

— décrit par un ensemble d'informations géographiques, à savoir des entités spatiales, thématiques et temporelles qui interagissent entre elles. Ces informations évoluant dans le temps et dans l'espace.

Dans cette étude, nous nous focalisons sur les données ayant un fort caractère spatial ou contenant des informations géographiques produites et gérées par la Métropole de Montpellier, notre zone d'étude. La solution proposée est basée sur la norme ISO 19115 [ISO/TC 211, 2019]. Cette norme est dédiée à la gestion des données spatiales, en fournissant un certain nombre d'informations permettant de constituer un système de gestion de méta-données efficient. Parmi les méta-données qu'elle contient, nous pouvons citer :

— identification : contient l'ensemble d'informations permettant d'identifier une donnée ;
— étendue : regroupe les informations sur l'étendue d'une donnée spatiale ;
— référence géographique : permet d'identifier les informations de géo-référencement d'une donnée spatiale ;
— qualité : regroupe un certain nombre d'indicateurs, permettant de définir un niveau de qualité pour la donnée ;
— Etc.

5.3 Solution proposée

5.3.1 Modèle conceptuel du lac de données spatiales

Dans cette section, nous présentons une vue générale du modèle conceptuel du lac de données spatiales que nous proposons comme solution de stockage adaptée pour les données massives et hétérogènes dans notre étude. L'un des objectifs visé est de fournir un guide pour sa reproductibilité. La conception proposée implique un certain nombres de contraintes, permettant ainsi de prendre en compte la généricité de la solution proposée, ainsi que des caractéristiques des données hétérogènes :

— la spatialité des données : les données spatiales présentent des attributs et des caractéristiques qui les distinguent des autres types de données. Ceci est valable pour les systèmes de stockage qui les supportent, que sur les méthodes de traitement qui leur sont appliquées. Les villes intelligentes disposent des données telles que des images satellite, des données vectorielles comme les plans d'urbanisation, etc. Dans le but de proposer une solution générique, il est nécessaire de tenir compte de tout type de données que peut disposer une ville intelligente ;
— l'inter-opérabilité : le système proposé doit être inter-opérable avec d'autres systèmes d'information, au niveau local, national et européen ;
— l'exploitation : le système doit permettre à des utilisateurs de profils différents, de pouvoir réaliser des tâches qui leur sont propres, en fonction de leur besoin.

Nous distinguons principalement trois sections dans le modèle conceptuel : la section data, la section metadata et la section interdata-metadata. La section data, le coeur de la structure de stockage, est basée sur Hadoop Distributed File System (HDFS) [Shvachko et al., 2010]. Le choix de HDFS est motivé d'une part, par le fait qu'il permet de stocker les données dans leur format natif (contrairement au système de stockage clé-valeur), et d'autre part, de sa distributivité. Avec HDFS, il est possible détendre (parallèlement) plus facilement la capacité de stockage en cas de besoin et aussi d'effectuer des calculs distribués.

La section metadata est un catalogue de données [Lamb and Larson, 2016], qui décrit les données stockées dans le lac de données. La section interdata-metadata est une partie de la section metadata. Elle permet le stockage de relations entre jeux de données riches sémantiquement.

HDFS est un système performant pour le stockage de données massives et hétérogènes, mais ne peut pas être utilisé tel quel par nos utilisateurs. Les utilisateurs du lac de données ont besoin d'explorer le lac de données afin de trouver les jeux de données les plus pertinents vus leur requête, et éventuellement de découvrir de nouveaux jeux de données. Ces fonctionnalités (exploration, requêtage, découverte) sont supportées par le catalogue de données, qui offre une interface graphique simple pour accéder aux méta-données descriptives du contenu du lac de données.

Le modèle conceptuel que nous proposons est une extension de la norme ISO 19115 [ISO/TC 211, 2019]. Cette norme inclut une description spatiale des données et sert de base à plusieurs profiles de méta-données (INSPIRE, Dublin Core) utilisés habituellement par les institutions publiques.

La Figure 35 est une vue générale du modèle conceptuel étendu proposé. Dans cette figure, la classe représentée en blanc est directement issues de la norme ISO 19115 (voir Figure 35) et les classes représentées en jaune constituent nos ajouts. Afin que les modèles restent lisibles, nous avons fait le choix de ne représenter que la classe principale de chaque package.

Dans la section data (voir Figure 36), nous définissons un lac de données comme un ensemble de ressources. Une ressource peut être un service (voir la norme ISO 19115) ou une série de données. Une série de données est composée d'un ou plusieurs jeux de données, qui partagent une caractéristique. Un jeu de données est une collection de données identifiables. Trois types de jeux de données ont été définis : document, vecteur et raster.

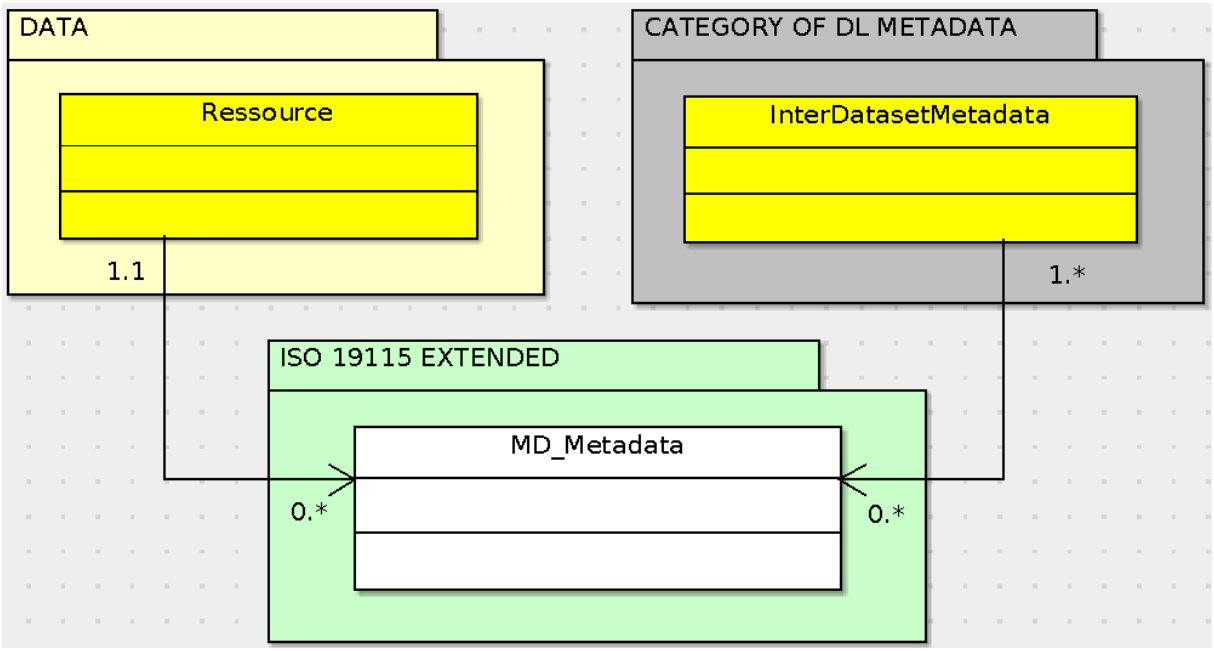
La section metadata décrit les fiches de méta-données associées à chaque ressource (voir Figure 37). Une fiche de méta-données est composée de :

— une identification (obligatoire) qui permet la différenciation des ressources par l'utilisateur ;

— une représentation spatiale (optionnelle), un système de coordonnées de référence (optionnel) et une emprise spatiale et/ou temporelle (optionnelle). Ces trois élé-

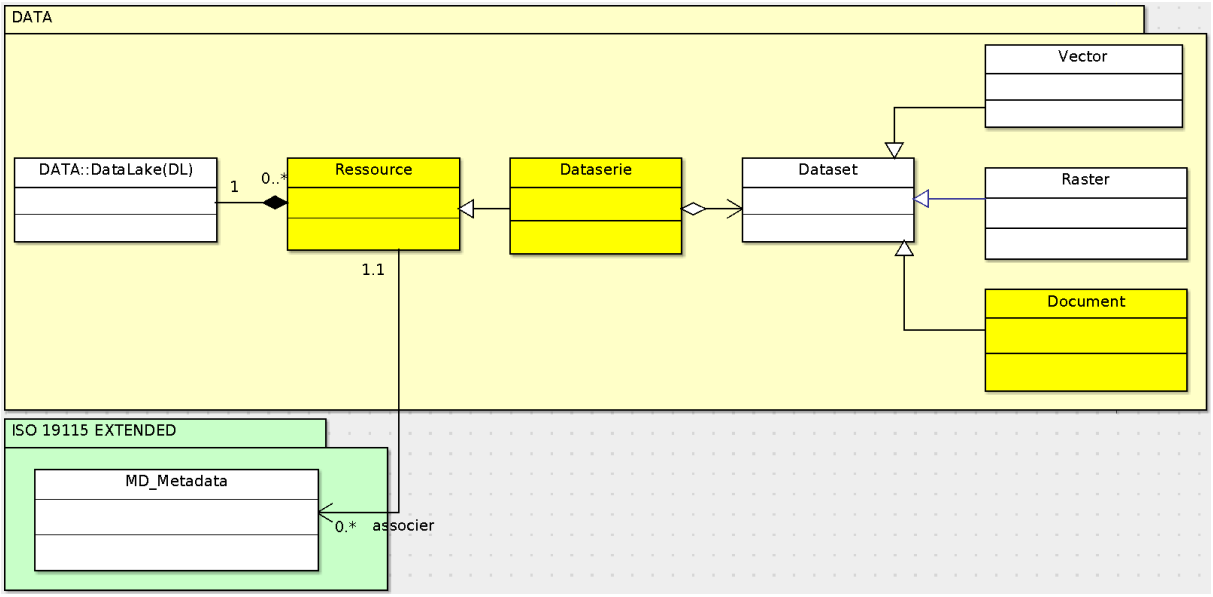
ments permettent de décrire la spatialité de la ressource ;
— une description du contenu de la ressource (optionnelle) ;
— une généalogie (optionnelle), qui explique comment la ressource a été obtenue ;
— un ou plusieurs liens vers des ressources associées ;
— un système de référence (optionnel), qui identifie les systèmes de références spatiaux, temporels et paramétriques utilisés par cette ressource ;
— une emprise, qui décrit l'emprise temporelle et spatiale de la ressource.

Enfin, la section interdata-metadata (voir Figure 38) décrit les relations entre les jeux de données et permet à l'utilisateur d'avoir une visibilité sur les données liées à sa requête initiale. Quatre types de relations ont été proposés, basés sur [Sawadogo et al., 2019] : parenté, inclusion, similarité et regroupement thématique. Le modèle conceptuel proposé dans son ensemble permet de prendre en compte non seulement l'intégration des métadonnées de données spatiales, mais aussi tout type de données stockées dans le lac.



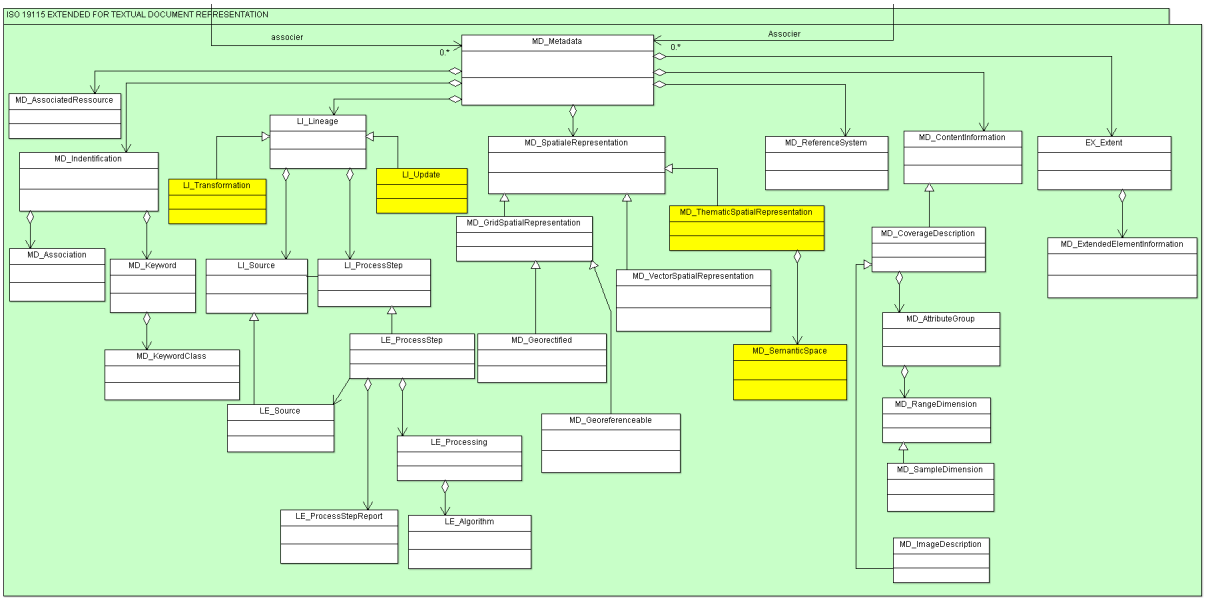
Les classes blanches viennent de [ISO/TC 211, 2019], les classes jaunes ont été ajoutées à la norme.

FIGURE 35 – Vue générale du modèle conceptuel proposé



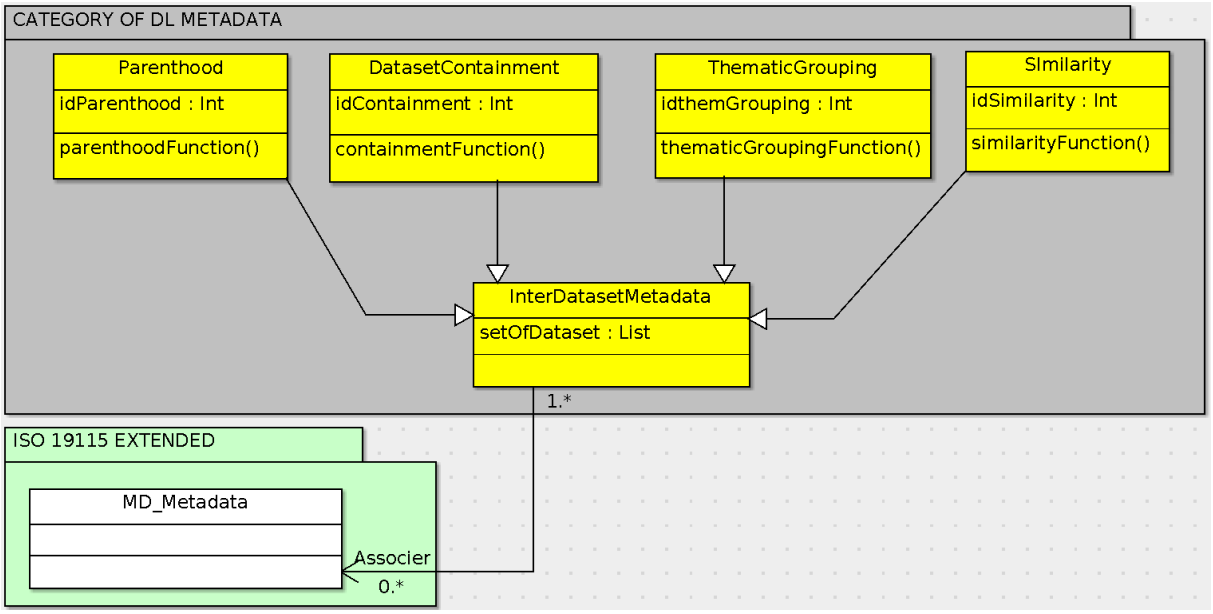
Les classes blanches viennent de [ISO/TC 211, 2019], les classes jaunes ont été ajoutées à la norme.

FIGURE 36 – Section Data



Les classes blanches viennent de [ISO/TC 211, 2019], les classes jaunes ont été ajoutées à la norme.

FIGURE 37 – Section metadata



Les classes blanches viennent de [ISO/TC 211, 2019], les classes jaunes ont été ajoutées à la norme.

FIGURE 38 – Section Interdata-Metadata

5.3.2 Infrastructure système

Comme présenté dans la précédente section, un lac de données est composé de deux sous-systèmes. La partie donnée qui assure le stockage des données physiques, est assurée par la mise en place du système de fichiers distribués. Cette partie est enrichie par un gestionnaire de méta-données qui constitue le deuxième composant du lac de données, assurée par le système de gestion de méta-données.

La partie stockage de données physiques repose sur le système de fichiers distribués HDFS (Hadoop Distributed File System), utilisant la technologie du projet Hadoop [1]. Un ensemble de serveurs HDFS, appelé cluster HDFS, est constitué de trois machines ayant des rôles différents : deux nœuds de stockage (appelé datanode dans la terminologie Hadoop) et un nœud maître (appelé namenode). Ce dernier gère les interactions avec les utilisateurs ainsi que la distribution et la réplication des données sur les datanodes.

Malgré les différentes fonctionnalités offertes à l'utilisateur, le namenode HDFS ne propose ni de système d'indexation de données ni de moteur de recherche. Pour répondre à ce besoin, l'administrateur d'un lac de données doit avoir recourt à un outil de gestion de méta-données de type ElasticSearch [2], construit sur le projet Apache Lucene [Chen et al., 2017] et [John and Misra, 2017]. Nous proposons d'utiliser l'outil GeoNetwork [3] qui est application web de catalogue de données spatialisées. Cet outil open-source embarque un moteur de recherche Apache Lucene et a l'avantage d'implémenter le modèle de la norme

1. Hadoop : http ://hadoop.apache.org/
2. https ://www.elastic.co/fr/elasticsearch/
3. https ://geonetwork-opensource.org/

ISO 19115. Ainsi, le serveur GeoNetwork sauvegarde les méta-données obligatoires et optionnelles, telles que décrites dans la précédente section, et conserve les liens permettant de télécharger les données stockées dans le cluster Hadoop via le namenode. Le moteur de recherche de GeoNetwork permet à l'utilisateur de faire des recherches croisées sur les trois dimensions : spatiale, temporelle et thématique. Le résultat de la recherche est une collection de jeux de données répondant à l'intersection des critères de la requête.

5.3.2.1 Insertion et indexation des données dans le lac de données

Comme illustrée dans la Figure 39, l'insertion de jeux de données dans le lac de données se déroule principalement en cinq étapes. Les deux premières étapes sont réalisées manuellement, les trois dernières étapes sont automatisées. Cette tâche est réalisée grâce à la première méthode de collecte abordée dans le Chapitre 2, Section 2.2.1.

En effet, l'administrateur doit remplir un tableur au format CSV [étape 1]. Ses colonnes sont les méta-données telles que décrites dans la section précédente ainsi qu'un lien HDFS pour indiquer l'emplacement du jeu de données. Chaque ligne représente un jeu de données pour lequel l'administrateur doit compléter les méta-données.
Puis l'administrateur lance un programme, depuis sa machine (ou un serveur du lac de données) [étape 2], qui va lire [étape 3] et extraire les informations du fichier CSV. Le programme télécharge les jeux de données et les insère dans le cluster HDFS [étape 4]. Enfin, le programme crée une fiche de méta-données de type ISO 19115 et l'insère dans GeoNetwork afin de bénéficier de son indexation et de son moteur de recherche [étape 5]. Les étapes 6 et 7 sont réalisées respectivement après chaque insertion de données. Elles consistent à stocker de façon distribuée la donnée dans les datanodes, et à indexer les méta-données correspondantes dans HDFS à partir de GéoNetwork qui représente l'accès utilisateur.

5.3.2.2 Les profils utilisateurs

Le lac de données peut être exploité par deux types de profils utilisateurs. Le premier profil utilisateur concerne le *grand public* ou *tout utilisateur*, le système leur permet d'explorer et de récupérer des données présentes dans le lac à travers l'interface web offert par GeoNetwork sans avoir besoin de compétences sur l'exploitation des lac de données. Le second profil concerne les *utilisateurs expérimentés*, qui en plus de l'exploration peuvent effectuer des traitements et des analyses directement sur le lac de données en utilisant des outils disponibles dans l'écosystème d'Hadoop, comme Apache Spark. Nous illustrons les différentes parties par la Figure 40.

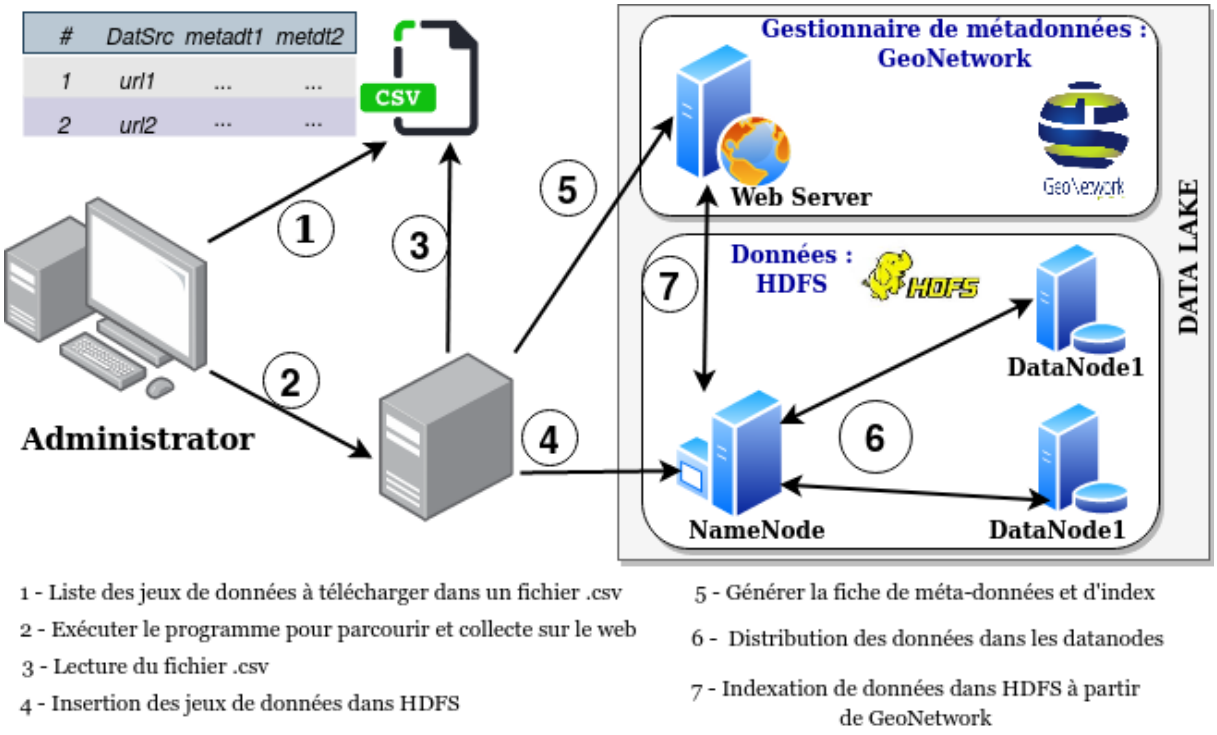


FIGURE 39 – Insertion et indexation de jeux de données dans le lac de données

5.3.2.3 Découverte et accès aux jeux de données

L'utilisateur peut parcourir, découvrir, faire une requête et accéder aux jeux de données en utilisant le moteur de recherche de GeoNetwork (profil grand public). Les recherches peuvent être une combinaison de critères sur les trois dimensions :

- sémantique : basée sur les mots clés ou une recherche en texte plein sur le titre, le résumé ou la généalogie de la fiche de méta-données ;
- spatialisée : en dessinant une emprise spatiale directement sur la carte afin de filtrer les jeux de données qui intersectent l'étendue géographique voulue ;
- temporelle : permet de filtrer sur les années, mois et jours.

GeoNetwork retourne une collection de fiches de méta-données décrivant les jeux de données qui respectent les critères de recherche. En parcourant les jeux de données, l'utilisateur peut accéder à tous les fichiers de données stockés dans le cluster HDFS sans avoir besoin de connaître la syntaxe d'interrogation d'Hadoop (voir Figure 41).

5.4 Cas d'application : 3M

5.4.1 Données

Les jeux de données utilisés dans le cadre de notre étude sont constitués entre autres d'images satellites, des données textuelles, de couches vectorielles et autres données telles que les données de transports, d'urbanisation, d'agriculture, de commerce, etc. Elles pro-

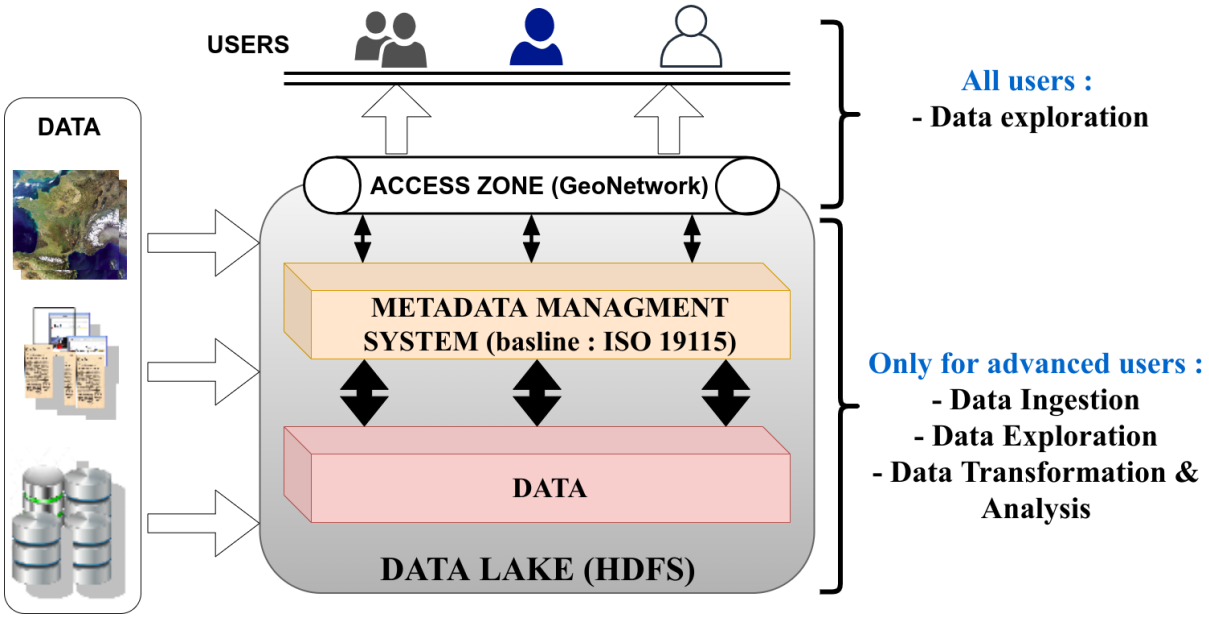


FIGURE 40 – Illustration des profils d'utilisateurs du lac de données

viennent de sources différentes :

- la plate-forme opendata de la Métropole de Montpellier [4]. Elle regroupe un ensemble de données produites par la Métropole de Montpellier et qui sont mises à la disposition du grand public. La liste exhaustive des liens se retrouve dans le fichier datasources.csv présent dans le dépôt logiciel de notre implémentation (https ://github.com/aidmoit/collect/blob/master/input/datasources.csv), accédé le 2020-02-19. Ces jeux de données sont publiées sous licence "Open Data Commons Open Database License" (ODbL) ;
- le web : nous avons constitué des corpus thématiques de données textuelles à partir du web en utilisant le protocole de collecte proposé dans le Chapitre 3 ;
- OpenStreetMap : nous a permis d'obtenir les étendues spatiales des lieux de notre cas d'étude (les communes de la métropole de Montpellier).

Dans les Sections 5.4.2 et 5.4.3 qui suivent, nous présentons respectivement le processus d'automatisation mis en place pour l'exploitation du lac de données au profit de la Métropole de Montpellier [5] et un exemple de recherche de données du lac via GéoNetwork.

5.4.2 Automatisation du déploiement du lac de données

L'installation d'un cluster Hadoop peut s'avérer complexe en fonction du cas d'étude. À des fins de reproductibilité, nous avons automatisé son installation, via les outils Va-

4. Open Data 3M : http ://data.montpellier3m.fr/
5. Montpellier Méditerranée Métropole (3M) : https ://www.montpellier3m.fr/

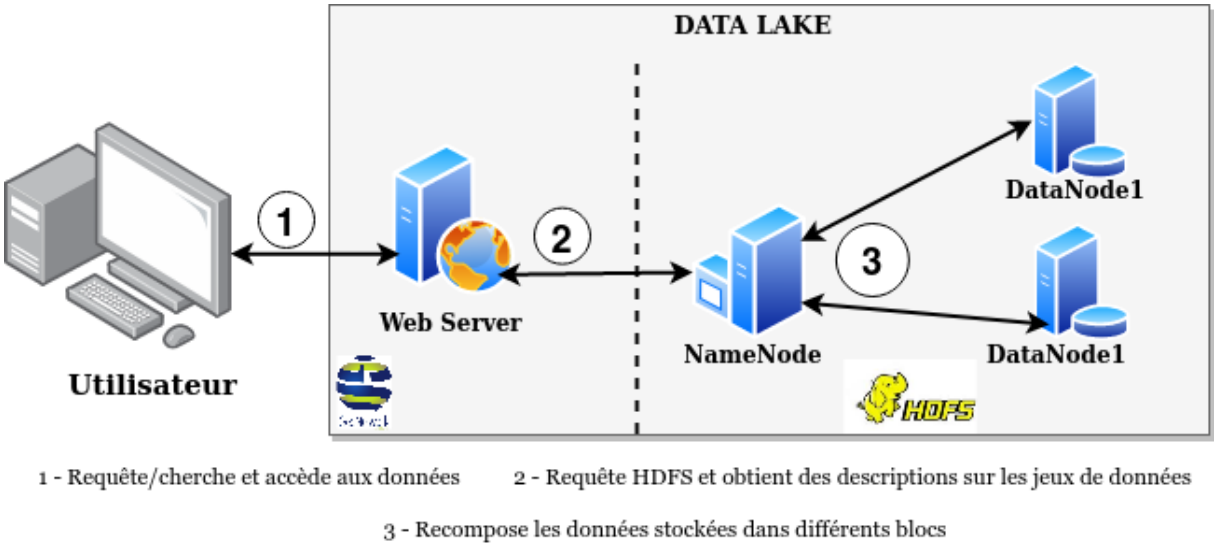


FIGURE 41 – Recherche de jeux de données

grant [6] et Ansible [7].
Cette automatisation construit et configure quatre machines virtuelles dont les trois premières appartiennent au cluster Hadoop et la dernière héberge le serveur GeoNetwork. Ceci permet d'instancier facilement notre lac de données. Nous avons aussi automatiser l'ajout de jeux de données ainsi que leur indexation. Cette automatisation prend, cette fois-ci, la forme d'un script python qui analyse les informations provenant du fichier CSV contenant les liens vers les données de 3M (voir Tableau A.1), y extrait des liens de téléchargement qui lui permettent de télécharger les jeux de données voulus. Ensuite, le script insère les fichiers de données dans le cluster HDFS sans les organiser dans une arborescence. Enfin, il crée la fiche de méta-données pour chaque donnée et les insère dans GeoNetwork afin de bénéficier de son moteur de recherche.

5.4.2.1 Déploiement automatique du cluster HDFS et du serveur GeoNetwork

L'ensemble du lac de données, c'est-à-dire le cluster HDFS et le serveur GeoNetwork, est déployé et maintenu grâce à l'utilisation de projets opensource notamment les suivants :

- Debian : Système d'exploitation utilisé par les 4 machines virtuelles. Nous avons utilisé la version 9 et non pas la dernière version à cause de problème de compatibilité avec la version java nécessaire à Hadoop et à GeoNetwork. La dernière version (version 10) de debian ne maintient plus cette version de java (version 9) ;
- VirtualBox comme hyperviseur ;
- Vagrant comme système de gestion de configuration des machines virtuelles (système d'exploitation utilisé, configuration réseau, script d'installation, ...) ;

6. Vagrant : https ://www.vagrantup.com/
7. Ansible : https ://www.ansible.com/

— Ansible comme un outil de déploiement d'application et de gestion de configuration.

Les codes sources de ce projet peuvent être retrouvés dans la Section 5.4.4. Grâce à ces dépôts logiciels, le cluster HDFS peut être déployé et configuré en quatre commandes et le serveur GeoNetwork en une commande.

De plus amples informations ou instructions techniques peuvent être retrouvées dans le fichier README.md du dépôt logiciel de notre projet. Si les variables par défaut, proposées par le dépôt logiciel, sont conservées, le cluster HDFS peut être accessible de manière graphique en se connectant à son serveur web à l'adresse http ://namenode :9870 (ou http ://10.0.0.10 :9870). D'autres informations, telles que la santé du cluster, ou l'accès aux logs ou bien encore l'accès au système de fichiers HDFS peuvent être aussi retrouvées via cette interface. Le serveur GeoNetwork est quant à lui accessible à l'adresse http ://aidmoit-geonetwork :8080/geonetwork (http ://10.0.0.9 :8080/geonetwork).

5.4.2.2 Ajout de données dans le lac de données

L'ajout de données dans le lac de données a lui aussi été automatisé. Deux scripts en python et en R ont été écrits. La complexité qu'induit le développement d'un outil basé sur deux langages de programmation différents a été motivé par les couvertures de fonctionnalités des librairies de chaque langage. En effet, python offre des libraires remarquables pour interagir avec HDFS alors que R propose des modules intéressants pour gérer des fiches de méta-données compatibles ISO 19115. Afin de faciliter l'utilisation de ces deux scripts, le programme R a été encapsulé dans le code python, permettant ainsi à l'administrateur, de ne lancer qu'un seul programme.

Comme mentionné auparavant, les fichiers sources sont indiqués dans la Section 5.4.1. L'environnement requis pour faire fonctionner ces scripts a été décrit dans le fichier "requirement.txt" présent à la racine du dépôt du logiciel. Les instructions d'installation et de lancement sont, quant à eux, présentées dans le fichier README.md.

Le script principal écrit en python opère en cinq étapes. Premièrement, il extrait les informations contenu dans le fichier datasources.csv comme le fournisseur de données, le nom du jeu de données et les mots clés associés. Le script, dans une deuxième étape, parcourt le site web du fournisseur de données afin de créer un fichier json contenant l'ensemble des liens de téléchargement des jeux de données. Ensuite, tous les fichiers constituant les jeux de données sont téléchargés, puis enregistrés dans le cluster HDFS, ce qui constitue les troisième et quatrième étapes. Enfin, le script R est lancé afin de créer des fiches de méta-données au standard ISO 19139 qui sont, ensuite, ingérées par GeoNetwork.

Les fichiers de données sont facilement récupérables à partir de GeoNetwork. En effet, le namenode du cluster HDFS offre une interface de programmation de type API REST (Application Programming Interface - REpresentational state transfer) permettant une abstraction complète des commandes HDFS.

5.4.3 Illustration d'un cas de recherche utilisateur *grand public*

L'utilisateur de notre lac de données peut créer des requêtes complexes en combinant les trois dimensions : spatiale, temporelle et sémantique. Nous exploitons l'interface web de GeoNetwork comme interface d'accès pour le grand public. La Figure 42 illustre l'interface de GeoNetwork avec les différents champs de recherche : 1) pour effectuer une recherche par mots clés prédéfinis, 2) pour effectuer des recherches pleine texte, 3) pour rechercher les données en fonction de la spatialité en sélectionnant la zone géographique. La Figure 43 illustre un exemple de recherche croisée. Les critères utilisés sont, entre autres, la mobilité comme mot clé principal, 2019 comme temporalité, et toute l'agglomération de Montpellier comme spatialité. Les données ('Resident-Mobility' et 'Car-counting') satisfaisant cette requête sont les données uniquement en relation avec la mobilité, qui sont produites en 2019, et qui concernent l'agglomération de Montpellier. Quant à la Figure 44, elle illustre le contenu du repertoire de 'Resident-Mobility' avec les liens vers le lac de données permettant à l'utilisateur de les télécharger en local dans sa machine.

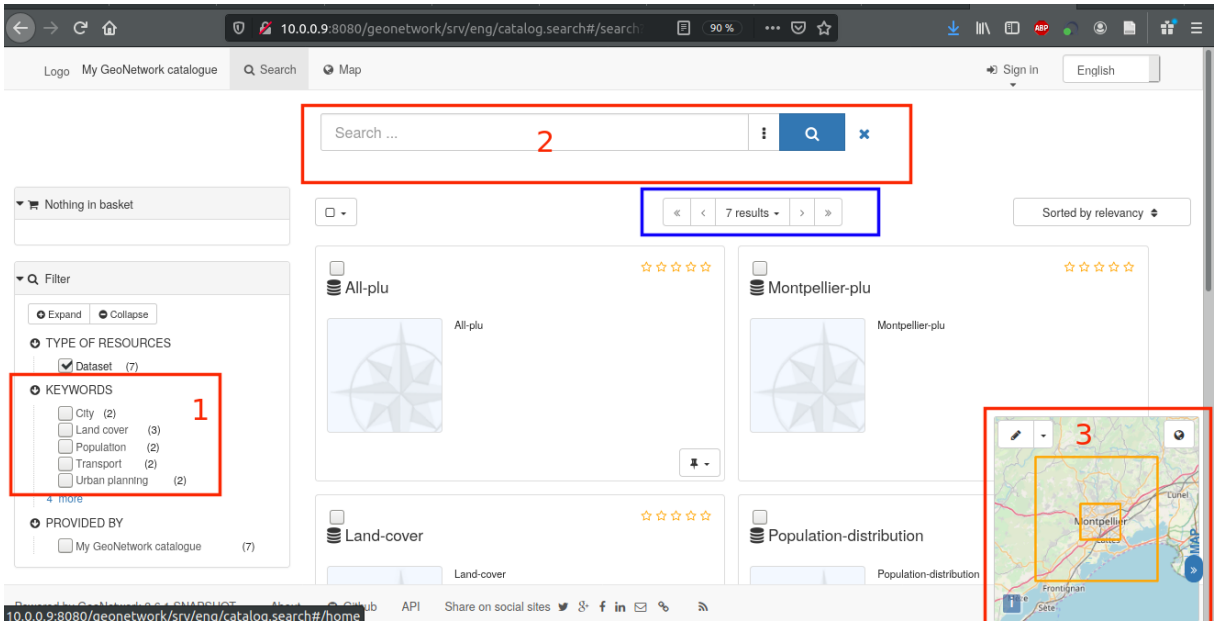


FIGURE 42 – Interface web de GeoNetwork

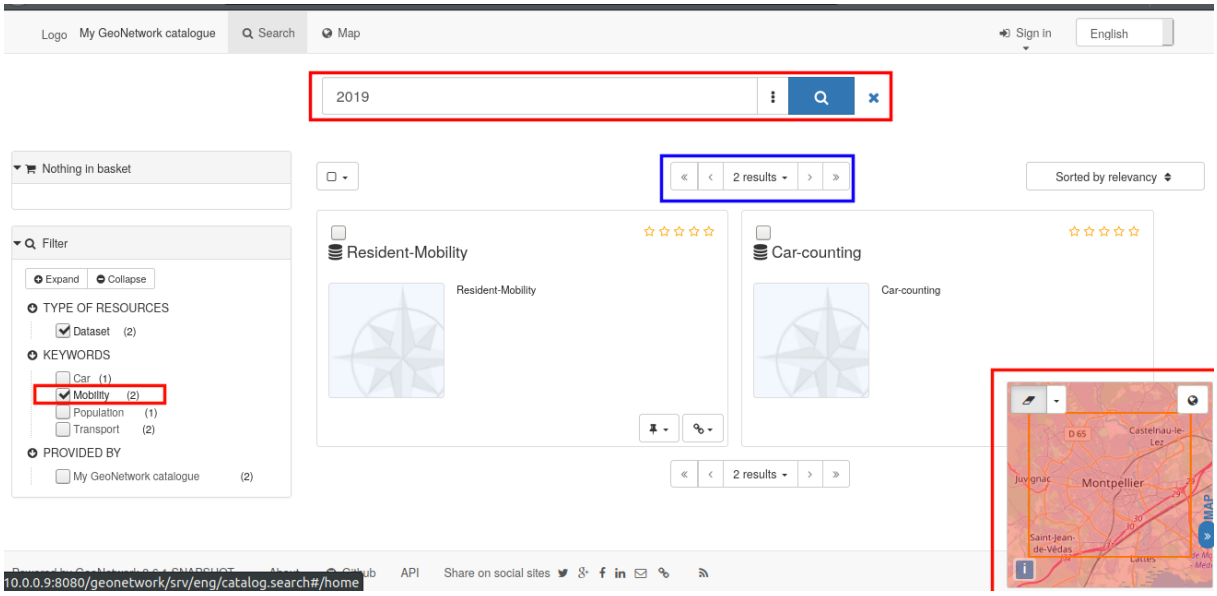


FIGURE 43 – Exemple de requêtes utilisateur

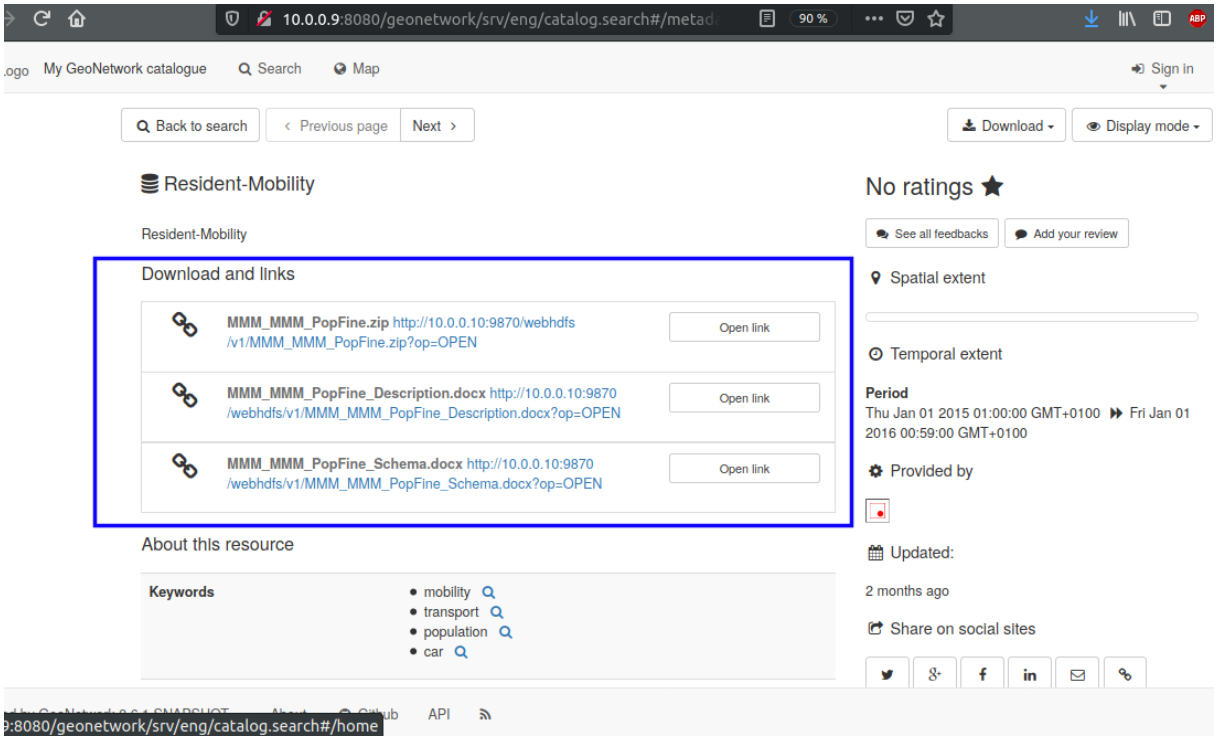


FIGURE 44 – Récupération de données du lac via GeoNetwork

5.4.4 Reproductibilité du cas d'étude

5.4.4.1 Infrastructure système

L'infrastructure du lac de données, c'est-à-dire, le cluster HDFS et le serveur GeoNetwork, est instanciable à travers l'utilisation de quatre machines virtuelles. L'installation et le lancement de ces machines ont été automatisés. Le dépôt logiciel est le suivant : https ://github.com/aidmoit/ansible-deployment. Les instructions d'utilisation sont décrites dans le fichier README.md du dépôt. Le numéro de commit utilisé pour notre implémentation est le suivant : 65de950a336ee2828cdb19db976b7946649c439c. Le dépôt est publié sous la licence GPL-3.

5.4.4.2 Logiciel et flux de traitement

L'ensemble des logiciels pour le téléchargement des données, leurs ajouts dans le cluster HDFS et leurs descriptions dans le GeoNetwork sont orchestrés par un script python. Toutes les ressources nécessaires à son exécution sont disponibles à travers ce dépôt : https ://github.com/aidmoit/collect. Le numéro de commit utilisé pour notre implémentation est le suivant : da9f63f9287a191d7e8fd24884a731bae02e1034.

Les codes sont distribués sous la licence GPL-3. Ils exploitent deux paquets R : geometa [Blondel, 2019] et geonapi [Blondel, 2018] diffusés sous licence MIT. Enfin, les scripts utilisent les données d'OpenStreetMap [8], ces données sont publiées sous licence "Open Data Commons Open Database License" (ODbL).

5.5 Conclusion

Dans cet chapitre, nous avons présenté une nouvelle approche méthodologique de conception et d'implémentation de structure de stockage des données massives et hétérogènes adaptée pour les villes intelligentes : un lac de données spatiales. Cette structure de stockage est accompagnée d'un système de gestion de méta-données géographiques permettant de prendre en compte les données ayant un fort caractère spatial.

Notre contribution principale s'inscrit dans l'introduction de la dimension spatiale dans le processus de conception de lac de données, d'où l'appellation lac de données spatiales. Nous avons également montré que les lacs de données peuvent être orientés vers les utilisateurs finaux, ce qui est possible en mettant en place une interface de requêtes, dans ce cas, un exemple d'utilisation pour la recherche et la récupération de données dans le lac via GéoNetwork en spécifiant une dimension spatiale, temporelle et thématique. Nous stockons dans le lac, l'ensemble des données précédemment collectées dans le Chapitre 3 grâce au protocole de collecte, et celles que nous avons identifiées dans l'inventaire des

8. OpenStreeMap : https ://www.openstreetmap.org

données de 3M.

Nos prochains travaux de recherche, seront consacrés à la troisième partie du processus général (voir Chapitre 2, Section 2.2.3), le Chapitre 6 pour le développement des méthodes dédiées à l'analyse et à la mise en lien des données stockées dans le lac de données spatiales mis en place dans ce chapitre.

Chapitre 6

Intégration & mise en relation de données hétérogènes

Les grandes quantités de données dont disposent les grandes villes sont d'une importance capitale car elles renferment tout l'historique de divers évènements. Une fois mises ensemble et analysées, elles peuvent contribuer à retracer ou à comprendre l'évolution de ces évènements, de même que leur comportement au fil du temps.

Vu la quantité de données et leur caractère très hétérogène (voir Chapitre 2, Section 2.1.2), il devient important de définir et de mettre en place de nouvelles méthodes et de solutions permettant de les analyser conjointement.

Relatif à la troisième partie de notre architecture générale (voir Chapitre 2, Section 2.2), nous proposons dans ce chapitre, une approche pour la mise en relation de données hétérogènes stockées dans le lac de données (voir Chapitre 5). La mise en relation est faite à partir des données textuelles obtenues grâce au protocole de collecte proposé dans le Chapitre 3, avec des images satellite disponibles sur la même zone spatiale. Cette approche que nous proposons est illustrée sur les données de la Métropole de Montpellier.

Ce chapitre est décomposé en cinq sections. La Section 6.1 correspond à l'introduction générale. Dans la Section 6.2, nous présentons les travaux relatifs à la mise en relation entre données textuelles et images de façon générale. L'approche que nous proposons est traitée dans la Section 6.3. Un cas d'application de l'approche, sur des données de 3M est proposé dans le Section 6.4. La dernière Section 6.5 conclut ce chapitre, tout en dressant les contributions et les perspectives de la solution proposée.

6.1 Introduction

Avec l'avènement de la fouille de données massives, il devient possible d'extraire des connaissances à partir de divers types de données. Suivant les besoins et les types d'analyses, des données peuvent être analysées séparément ou conjointement lorsque l'objectif visé consiste à extraire des connaissances à partir de données venant de plusieurs sources. Les données produites par les territoires ou grandes villes sont non seulement volumineuses, mais sont aussi complexes car très hétérogènes.

Les caractéristiques complexes de ces données remettent en cause la plupart des méthodes utilisées pour leur exploration et analyse dans la quête de nouvelles connaissances car elles ne permettent pas de procéder à certaines analyses sur les données hétérogènes. Comme nous l'avons déjà évoqué dans le Chapitre 1, il est nécessaire de proposer des techniques et méthodes ou d'adapter celles qui existent, afin de pouvoir analyser ces masses de données hétérogènes, après avoir traité le problème sur leur stockage détaillé dans le Chapitre 5.

Dans cette optique, nous proposons, dans ce chapitre, une approche qui vise à mettre en relation des données fortement hétérogènes : une méthode qui permet de faire un suivi spatio-temporel multimodale (images satellite et des données textuelles) d'évènement territorial. Basée sur des approches de traitement d'images et de traitement automatique de langage naturel (TALN), notre proposition vise à exploiter les informations à extraire dans chacune des données pour un enrichissement mutualisé, dans le cas de suivi d'évènements spatio-temporels. Nous partons sur la base qu'un évènement spatio-temporel peut-être représenté sous forme de graphe d'évolution grâce à des séries temporelles d'images satellite. Nous enrichissons ensuite ces graphes spatio-temporels grâce aux informations obtenues suite à l'analyse des données textuelles. Ainsi, sur une période donnée, nous avons la possibilité d'analyser la relation entre les deux types d'informations et voir s'il existe une quelconque corrélation entre elles.

6.2 État de l'art

La mise en relation entre données hétérogènes peut-être discutée sous plusieurs formes. En effet, la définition de l'hétérogénéité sur les données prend forme en fonction du cas d'étude, comme nous l'avons mentionné dans la Section 2.1.2 du Chapitre 2. En se basant sur l'hétérogénéité liée à la nature des données, nous pouvons distinguer deux approches de mise en relation : celle qui consiste à mettre en relation des données de même nature, et celle qui consiste à mettre en relation des données de nature différentes.

Pour la mise en relation entre des données hétérogènes de même nature, l'hétérogé-

néité de ces données peut se situer au niveau de leur source ou de leur contenu. Il peut s'agir d'analyser conjointement, soit des images, soit des vidéos ou des données textuelles de sources différentes. Par exemple, dans [Wang et al., 2015], les auteurs ont effectué une classification de documents à partir de données hétérogènes en s'appuyant sur des bases de connaissances. Ces bases de données étaient essentiellement 20Newsgroups [1] et RCV1 [Lewis et al., 2004] qui sont composées respectivement de 20 classes, et trois catégories de données : industrielle, économique, et des documents sur des marchés. Quant aux bases de connaissances, ils ont utilisé Freebase [2] et YAGO2 [3]. Dans ce cas, l'hétérogénéité se trouve non seulement au niveau de leur source (20Newsgroups, RCV1, Freebase, YAGO2), mais aussi dans leur contenu (différentes catégories et classes). De même, dans [Yin et al., 2006], les auteurs ont proposé LinkClus, une approche orientée graphe, permettant d'effectuer un clustering sur des données textuelles obtenues de DBLP [4] avec d'autres types de données tels que des e-mails de conférences, et des offres d'emplois. Dans [Zhao et al., 2017], des données de 19 bases de données de Stockholm EPR Corpus [Dalianis et al., 2012], ont été analysées conjointement, pour la détection des effets indésirables des médicaments sur des données cliniques en tenant compte des aspects temporels et séquentiels des données.

Dans le cadre des analyses portant sur des données de nature différente, nous distinguons les cas d'études qui se focalisent sur l'extraction d'informations textuelles pour enrichir des images dans le but d'améliorer leur classification. Nous notons les études comme [Zhu et al., 2011], où les auteurs effectuent une classification d'images à partir de données textuelles. Pour la phase d'apprentissage, les images utilisées sont de la base de données Flikr [5] et les documents textuels sont obtenus du Web, et pour la phase de test, les images de la base de données Caltech-256 [6] sont utilisées. Dans [Weiss et al., 2016], un état de l'art est réalisé sur des méthodes et cas d'études pour des données hétérogènes. Ces méthodes sont généralement basées sur du transfert learning, encore appelée apprentissage par transfert, consistant à exploiter des informations extraites d'une source pour enrichir une autre, afin de d'améliorer la classification de l'une d'elle. Dans [Liu et al., 2011, Yang et al., 2013], les auteurs ont mené une étude portant sur la classification de clips vidéos par apprentissage sur des données hétérogènes, à savoir des vidéos récupérées sur YouTube avec des images de la base de données Flikr.

D'autres analyses sur les données hétérogènes de nature différente visent à analyser des événements tels que des catastrophes naturelles. Dans [Ahmad et al., 2019], un système appelé JORD est proposé dans le but de pouvoir collecter automatiquement des données

1. http ://qwone.com/ jason/20Newsgroups/
2. https ://developers.google.com/freebase
3. https ://yago-knowledge.org/
4. https ://dblp.org/
5. https ://www.flickr.com/
6. https ://resolver.caltech.edu/CaltechAUTHORS :CNS-TR-2007-001

des médias sociaux (y compris l'analyse des textes dans les langues locales) concernant les catastrophes technologiques et environnementales, et de les relier automatiquement aux données des images satellite. Des études similaires ont été abordées dans [Said et al., 2019, Bischke et al., 2017], où les auteurs proposent la mise en lien entre des données des médias sociaux avec des images satellite, pour la détection et l'analyse des inondations. L'enrichissement d'images satellite par des données collectées sur des réseaux sociaux, est mis en évidence dans [Cervone et al., 2017]. Les auteurs ont étudié le cas où des ouragans ou des phénomènes météorologiques violents touchent de très vastes zones. Ils ont montré que la fusion ou la mise en relation entre les données recueillies des médias sociaux, tel que Twitter, et les images satellite, peuvent fournir des informations supplémentaires lorsque les données de télédétection font défaut ou ne sont pas disponibles.

Une autre forme d'analyse effectuée sur des données textuelles et des images satellite, est l'analyse d'évènements spatio-temporels. Dans [Kergosien et al., 2015, Roche et al., 2014], les auteurs réalisent leurs travaux sur l'extraction des empreintes spatiales et/ou temporelles dans des documents qu'ils mettent en relation avec des images satellite. L'empreinte spatiale ou temporelle d'un document, est désignée à partir des entités nommées spatiales et temporelles (voir Chapitre 3) qu'il contient. Ainsi, pour un corpus donné, ils ont la possibilité d'identifier et de représenter l'entité spatiale la plus abordée dans ce corpus.

La majorité des approches proposées pour la mise en relation entre données hétérogènes dans les deux cas (données de même nature, et de nature différente), abordent plus de problèmes de classification ou de mise en relation à un instant t, dans la mesure où l'étude est faite sans tenir compte de l'évolution des différents évènements. Aussi, très peu de travaux abordent le suivi des évènements territoriaux, que ce soit sur le plan spatial et/ou temporel. C'est pourquoi, nous proposons dans ce chapitre, une approche permettant d'étudier l'évolution d'un évènement territorial, disposant d'un ancrage spatial et temporel, en nous appuyant sur une analyse conjointe de séries temporelles de données textuelles et d'images satellite.

6.3 Approche méthodologique

Le principe fondamental de la mise en relation entre données hétérogènes consiste à enrichir des données de sources différentes et hétérogènes, en tirant de part et d'autre, des informations complémentaires, pour en reconstituer de plus complètes et pertinentes. Comme mentionné dans la section précédente, il existe deux types d'approche sur la mise en relation des données hétérogènes. La mise en relation entre données hétérogènes de natures différentes telles que des images avec des données textuelles, et celle qui consiste à mettre en relation des données hétérogènes de même nature, telles que des données

textuelles de diverses sources.

Dans cette partie de notre étude, nous nous focalisons sur la première approche, qui consiste à mettre en relation, des données textuelles obtenues du Web (blogs, forums, articles scientifiques, presse, etc.) avec des images satellite de très haute résolution. Elle est composée de trois étapes détaillées ci-dessous, qui sont entre autres, l'analyse des données textuelles, l'analyse des images, et l'enrichissement des images satellite à partir des données textuelles et vice versa, qui correspond à la mise en relation.

Notre approche méthodologique est basée sur le principe qu'un évènement territorial peut être décrit de façon chronologique à partir de données officielles et/ou à caractère social, à savoir les données textuelles (rapport scientifiques, forums, blogs, réseaux sociaux, etc.) et les images satellite de très haute résolution.

La mise en relation est faite à partir de série temporelle de données textuelles (STDT) avec de série temporelle d'images satellite (STIS). Les deux séries temporelles doivent être chronologiquement et spatialement liées. La relation spatiale, sous-entend que les données textuelles doivent aborder plusieurs thématiques relatives à un endroit précis. À partir des STDT, et des STIS, nous avons la possibilité, d'extraire des informations de part et d'autre que nous pouvons mettre en relation sur le plan spatio-temporel, et procéder par la suite à une analyse sémantique de ces relations.

Pour ce faire, nous représentons les informations extraites des STDT sous forme de termes et/ou d'expressions descriptifs, et celles extraites des STIS sous forme de graphes d'évolution. Les informations textuelles seront utilisées pour expliquer ou enrichir l'évolution de l'entité spatiale concernée dans le temps. Une entité spatiale ayant subi plusieurs déformations spatiales à un instant T, sera décrite par un nombre de noeuds élevé à cet instant, et vis-versa. Nous illustrons l'approche générale dans la Figure 45.

Par exemple, un projet portant sur la réalisation d'une infrastructure urbaine, peut faire l'objet de plusieurs études scientifiques (rapports d'études, recherches scientifiques, etc.) et à la fois susciter des réactions écrites auprès des habitants de la ville, en particulier ceux de la localité tout au long de sa réalisation. En plus de données textuelles, cette même réalisation peut être observée à partir d'images satellite, si toutefois la zone concernée est couverte par un satellite d'observation de la terre. Chacune de ces données, prise individuellement constitue une source importante d'informations sur l'évènement territorial. Ainsi, nous partons dans l'idée que ces données mises en correspondance pour une analyse conjointe s'enrichissent mutuellement et peuvent contribuer à retracer et à comprendre l'évolution de cet évènement sur le plan spatio-temporel.

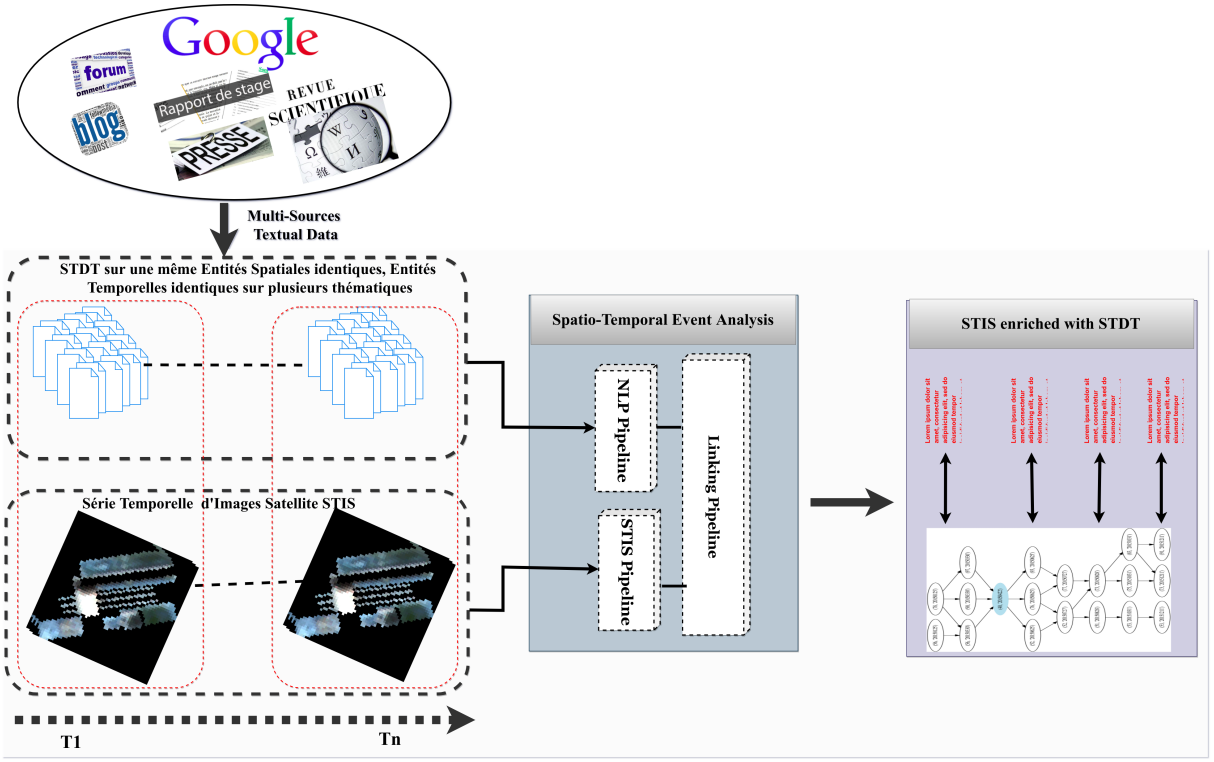


FIGURE 45 – Approche générale pour l'enrichissement d'images satellite à partir de données textuelles

6.3.1 Analyse de données textuelles - NLP Pipeline

L'analyse des données textuelles comprend trois étapes. Elle consiste à appliquer un certains nombres de traitements et d'analyse aux données textuelles, permettant de faire ressortir des informations pertinentes. Ces informations pertinentes ont pour but, de pouvoir décrire l'évolution d'un évènement ou d'une situation ayant suscité des débats ou des études scientifiques au sein du territoire.

6.3.1.1 Filtres spatio-temporels

Cette étape consiste à filtrer le corpus thématique afin de s'assurer les différents documents correspondent bien aux intervalles spatio-temporels définis. Le corpus thématique correspond à des données textuelles prenant en compte ou abordent à la fois plusieurs thématiques dans leur contenu. Premièrement, le filtre spatial sélectionne les documents de la zone d'étude ou de l'entité spatiale concernée. Lors de cette étape, pour chaque document, nous gardons uniquement le ou les paragraphes dans lesquels sont mentionnés la zone d'étude. Quant au filtre temporel, il consiste à filtrer les documents précédemment obtenus à la sortie du filtre spatial pour ne retenir que les documents dont les dates de production ou de publication sont comprises dans l'intervalle de temps, durant lequel nous souhaitons réaliser l'étude. Par exemple, pour une étude sur la réalisation d'un pont de l'année 2015 à 2017, le travail consiste tout d'abord, à filtrer les corpus thématiques

pour ne garder que les documents qui traitent de la réalisation du pont (filtre spatial). Après avoir obtenu les documents dont les contenus traitent la réalisation du pont, le filtre temporel permet à son tour, de sélectionner uniquement les documents (sous-corpus) dont les dates de publication sont dans l'intervalle 2015 - 2017. Nous avons choisi un niveau de granularité mensuelle afin d'avoir des observations assez espacées sur les images satellite. Nous considérons que le type d'évènement étudié évolue que peu sur cette durée.

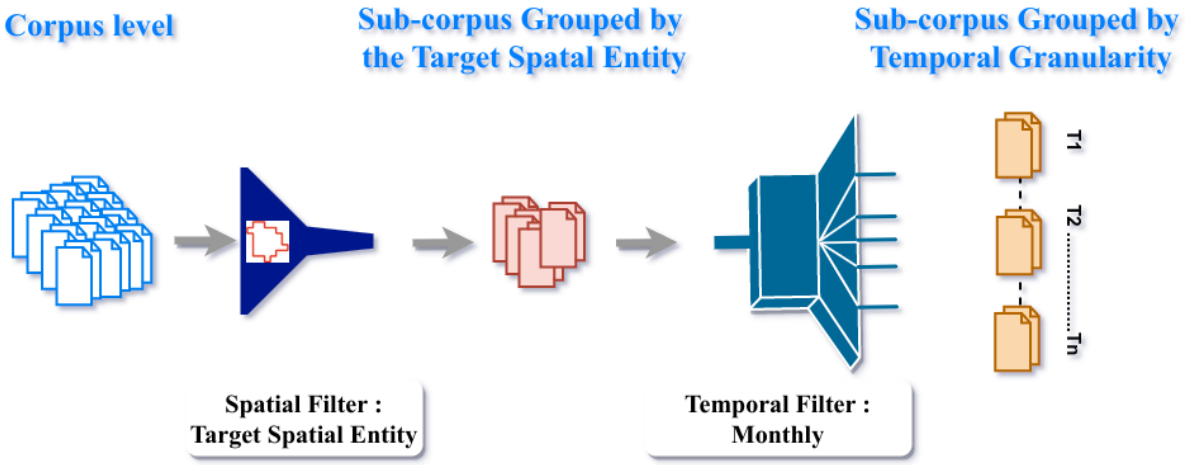


FIGURE 46 – Filtres spatio-temporels

6.3.1.2 Analyse Sémantique

L'objectif de cette étape est de faire remonter des informations à partir des sous-corpus issus des filtres spatio-temporels. Pour ce faire, les sous-corpus sont analysés afin d'y extraire des informations pertinentes.

Les analyses sont faites par deux éléments de description : le résumé et la liste des termes (noms d'organismes et de personnes). Le premier consiste à faire un résumé de chaque sous-corpus pour ne garder que les informations essentielles contenues dans chacun d'eux. Quant au second, nous nous focalisons sur l'extraction de mots ou de terminologies descriptifs, permettant de décrire ou de faire ressortir des expressions marquantes lors d'un évènement. Il s'agit, entre autres, des adjectifs, des noms de personnes ou d'organismes.

À la sortie de cette analyse, pour chaque période (en mois) de l'intervalle d'étude, une description sous forme de résumé de texte, et un ensemble de terminologies permettant de donner une susceptible description de l'entité spatiale concernée (voir Figure 47).

6.3.2 Analyse d'images satellite - STIS Pipeline

L'analyse des images satellite est un processus qui consiste à appliquer des pré-traitements et une méthode d'analyse sur les images, afin de faire ressortir des graphes

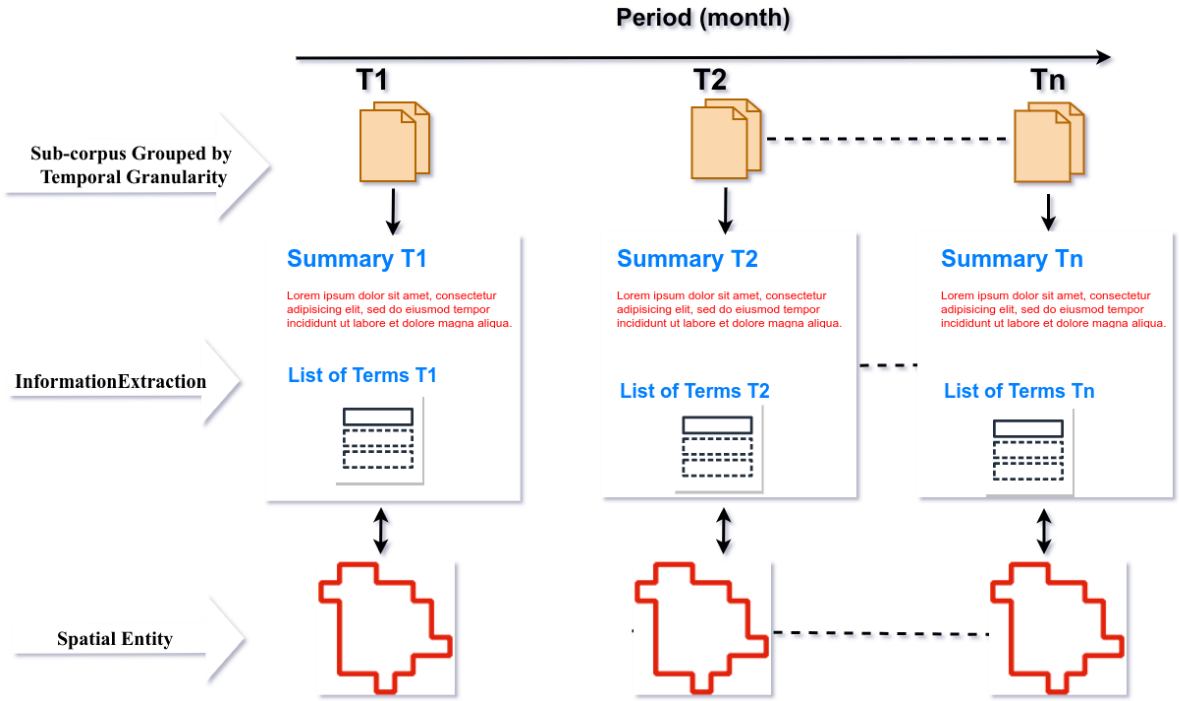


FIGURE 47 – Extraction d'information à partir des sous-corpus issus des filtres spatio-temporels

d'évolution pouvant décrire le comportement d'une entité ou d'une zone spatiale, sur le plan spatio-temporel.

Conformément à l'approche proposée par [Guttler et al., 2017] et reprise dans [Khiali et al., 2018], la méthode consiste à sélectionner tout d'abord l'ensemble des entités que l'on souhaite suivre tout au long de la période d'observation. Cet ensemble d'objets est appelé ensemble d'objets de référence noté ObjRefs. Un élément ObjRef de ObjRefs peut provenir de chacune des périodes d'observation de la série temporelle. Dans le but de pouvoir suivre l'évolution de chaque ObjRef, un graphe d'évolution est généré en tenant compte de l'ensemble des segments obtenus à partir des images de toute la série temporelle. Cette approche est motivée par le fait que l'évolution de toute entité spatiale pouvant être identifiée sur une image satellite peut-être suivie dans le temps et dans l'espace. Pour la mise en place de la méthode, nous décrivons dans les sections suivantes, les pseudo-codes que nous avons développé pour chacune des étapes.

6.3.2.1 Sélection des objets de référence ObjRefs

La sélection des objets de référence représente la première étape de la méthode. Elle se fait à partir de l'ensemble des segments obtenus lors de la segmentation des images. L'idée générale consiste à sélectionner le plus grand nombre d'objets possible parmi les segments. La sélection se fait, en commençant par les objets de plus grande taille, tout en limitant le partage d'information (redondance de pixels) entre l'objet à sélectionner avec ceux déjà

sélectionnés. La couverture donnée par les objets de référence à la fin du processus de sélection est nommée zone déjà couverte ou PAC (pixels already cover). Cette hypothèse se justifie par le fait qu'au cours d'une saison, il y a une période au cours de laquelle un objet atteint son empreinte maximale [Guttler et al., 2017], c'est-à-dire qu'il occupe un espace maximum.

Lors de la segmentation, il est fort probable que l'on se retrouve avec des objets de très petites tailles, donc difficile de suivre leur évolution. Pour ne pas considérer ces objets, nous faisons un filtrage sur les objets de départ (pour chaque pixel, nous retenons le plus grand objet qui le couvre) et le résultat de ce filtrage est appelé l'ensemble des ***objets candidats***. Le processus de sélection est décrit par l'algorithme 1. En partant de l'hypothèse que chaque pixel garde sa position sur toutes les images, il devient possible de retrouver pour chaque pixel, l'objet qui le couvre dans chaque période. Nous nous retrouvons donc avec un nombre d'objets égal au nombre de périodes pour chaque pixel. Enfin, nous gardons pour chaque pixel, l'objet ayant la plus grande taille, cet objet pouvant provenir de n'importe quelle période. Les objets candidats ainsi obtenus seront utilisés pour la sélection des objets de référence lors de la prochaine étape.

Algorithm 1 Fonction de sélection des Objets Candidats

Require: AllSegment : *# dictionnaire dont chaque clé est un tuple (segmentId,SegmentPeriod) avec pour valeurs la liste des pixels qui composent le segment*

$TempSetOfCandidateObject = \phi$
 # Tout d'abord, on recupère pour chaque pixel, les objets dans lesquels il s'y trouve
$SegmentInfo \leftarrow ((segmentId, SegmentPeriod), PixOfSegment)$
for all $(PixOfSegment \exists SegmentInfo) \in AllSegment$ **do**
  for all $pixel \in PixOfSegment$ **do**
    if $pixel \notin TempSetOfCandidateObject$ **then**
      $TempSetOfCandidateObject[pixel].APPEND(SegmentInfo[0], taille(PixOfSegment))$
    end if
  end for
end for
 # Pour finir, pour chaque pixel, on ne garde que le plus grand des segments qui le couvre
$SetOfCandidateObject = \phi$
$PixelInfo \leftarrow (SegmentInfo[0], taille(PixOfSegment))$
for all $PixelInfo \exists (pixel, PixelInfo) \in TempSetOfCandidateObject$ **do**
  $PixelInfo = DescendingOrderOf(PixelInfo)$ *# par ordre décroissant sur sur la taille des segments taille(PixOfSegment)*
  $SetOfCandidateObject[SegmentInfo[0]] = AllSegment[PixelInfo[0][0]]$ *# considérer uniquement le segment le plus grand*
end for
# SetOfCandidateObject ne contient que les objets retenus, avec les pixels qui les composent.

return $SetOfCandidateObject$

Après l'étape de la sélection des ***objets candidats***, nous procédons à la sélection des objets de référence ou ObjRefs. Étant donné que les images sont issues de la même zone d'étude, elles ont les mêmes coordonnées géographiques. Un pixel dans une image donnée

Algorithm 2 Fonction de sélection des Objets de référence

Require: CandidateObjectList, α

$PAC = \phi$
$poids(obj) = 0$
while $taille(CandidateObjectList) > 0$ **do**
  $RefObj$ = -1
  $poids\_max$ = -1
  for all $obj \in CandidateObjectList$ **do**
    if $Pix(obj) \cap Pix(PAC) = \phi$ **then**
      $poids(obj) = Pix(obj)$
    else
      $poids(obj) = \frac{taille(obj) - taille(PAC)}{taille(obj)}$ # ==> *nouveauté*
    end if
    if $poids(obj) < \alpha$ **then**
      $CandidateObjectList = CandidateObjectList - obj$
    else if $poids(obj) > poids\_max$ **then**
      $poids\_max = poids(obj)$
      $RefObj = obj$
    end if
  end for
  if $RefObj \neq$ -1 **then**
    $PAC = PAC \cap RefObj$
  end if
end while
return *PAC # ou ObjRefs*

garde la même position sur chacune des images acquises sur les autres périodes de la série temporelle.

Pour ce faire, nous construisons un *sac d'objets* qui contient tous les segments obtenus lors de la segmentation de toutes les images. La sélection se déroule en deux sous-étapes. Au cours de la première étape, les objets sélectionnés sont des objets qui n'ont aucune information (pixel) en commun. La deuxième étape commence lorsque la première s'achève. À partir de là, nous faisons intervenir un paramètre de régulation ou de seuil nommé α dans le choix des objets de référence, tout en respectant le principe de sélection sur la taille des objets. Dans cette étape, un objet est sélectionné si et seulement si sa nouveauté ou son apport d'informations est supérieur ou égal à la valeur de seuil α . La relation entre la valeur de α et les objets de référence a été mise en évidence par [Guttler et al., 2017], avec la formule 6.1 ci-dessous.

$$poids(O) = \left\{ \begin{array}{lll} \text{taille}(O) & \text{si} & \text{nouveauté (O)} = 1 \\ \text{nouveauté}(O) & \text{si} & \alpha <= \text{nouveauté(O)} < 1 \\ 0 & \text{si} & \text{nouveauté}(O) < \alpha \end{array} \right\} \quad (6.1)$$

Avec :

- taille(O) désigne la taille de l'objet ou encore le nombre de pixels qui le compose ;
- nouveauté(O) désigne la contribution de l'objet par rapport à la couverture partielle

déjà acquise (PAC)

- α est la valeur de seuil que doit satisfaire la contribution de l'objet en terme d'information ou de nouveauté pour être considéré comme objet de référence.

La nouveauté, autrement définit comme la contribution de chaque objet, est calculée comme indiqué dans la formule 6.2 ci-dessous.

$$\text{nouveauté}(O) = \frac{|taille(O) - PAC(O)|}{taille(O)} \tag{6.2}$$

Avec :

- taille(O), la taille ou encore le nombre de pixels de l'objet ;
- PAC(O), l'ensemble des pixels qui sont déjà couverts ou encore l'ensemble formé par la solution partielle (partielle lorsque le processus de sélection n'est pas encore terminé).

Cet ensemble d'objets satisfaisant les critères de sélection représente les objets de référence ObjRefs dont nous suivrons l'évolution à partir de la construction des graphes. Les différentes étapes sont décrites à partir du pseudo-code 2 [Khiali et al., 2018].

6.3.2.2 Construction des graphes d'évolution

La construction des graphes est la partie dans laquelle nous illustrons de façon explicite les liens qui existent entre un objet de référence ObjRef avec l'ensemble des objets candidats.

6.3.2.2.1 Sélection des noeuds

Pour chaque objet de référence ObjRef, nous faisons ressortir les objets candidats avec lesquels il partage des informations communes. Ce procédé permet de faire ressortir pour chaque ObjRef, son évolution au cours de la série temporelle. L'ensemble ObjRef ainsi que les objets candidats forment les noeuds du graphe qui décrit l'évolution de l'objet de référence. Pour éviter la sélection des objets non représentatifs (faible partage d'informations), deux paramètres ont été introduits lors de l'étape de sélection. Premièrement, pour qu'un objet candidat soit sélectionné comme un noeud du graphe, au moins une valeur $\sigma 1$ de l'objet candidat devrait être comprise à l'intérieur de l'empreinte de l'objet de référence ObjRef. Secondo, l'objet candidat devrait représenter au moins une valeur $\sigma 2$ de l'empreinte de l'objet de référence ObjRef. Les valeurs de $\sigma 1$ et $\sigma 2$ étant quantifiées en terme de pourcentage. La description des différentes étapes sont données par le pseudo-code 3.

6.3.2.2.2 Construction des arrêtes

La construction des arrêtes se fait en reliant les objets des différentes périodes qui ont des informations communes. Afin de pouvoir déterminer quels objets sont en relation,

Algorithm 3 Fonction de sélection des noeuds de chaque graphe

Require: $RefObjs$, $AllSegment$, $\sigma 1$, $\sigma 2$

$GraphList = \phi$ # *RefObj contient les informations : RefObjId et RefObjPeriod*
for all $RefObj \in RefObjs$ **do**
  $ListNode = \phi$
  $segment \leftarrow (segmentId, segmentPeriod)$
  for all $segment \in AllSegment$ **do**
    if $segmentPeriod \neq RefObjPeriod$ **then**
      $inters = Pix(RefObj) \cap Pix(segment)$ # *Pix(segment) constitue l'ensemble des pixels du segment*
      if $taille(inters)/taille(segment) >= \sigma 1$ **or** $taille(inters)/taille(RefObj) >= \sigma 2$ **then**
        if $segment \notin ListNode$ **then**
          $ListNode[segmentPeriod] = \phi$ # *une période donnée + l'ensemble de ces segments ne sont ajoutées qu'une et une seule fois*
        end if
        $ListNode[segmentPeriod][segment] = Pix(segment)$ #*chaque NodeListe constitue l'ensemble de noeuds d'un graphe*
      end if
    end if
  end for
  $GraphList.APPEND(ListNode)$ # *ajouter la liste de noeuds de chaque ObjRef à celle des graphes*
end for
return $GraphList$

les objets sélectionnés lors de la sélection des noeuds sont d'abord regroupés par temps d'acquisition. Une fois les objets organisés selon leur temps d'acquisition, nous calculons l'intersection (partage d'information) entre les objets de façon consécutive suivant leur période. Le partage d'information entre les objets appartenant à une période P est calculé avec ceux de la période *P+1*. Une arrête est créée entre deux objets, lorsque ces deux partagent une certaine quantité d'information. Le poids de l'arrête est égal à la taille des informations communes. Le graphe étant orienté, le sens de l'arrête est de l'objet de la période P vers l'objet de la période *P+1*. Deux objets ne peuvent être liés que lorsqu'ils ont des périodes d'acquisitions consécutives. Le nombre de couches formées par un graphe est égal au nombre de périodes considérées dans la série temporelle. Ce type de graphe est appelé graphe orienté temporel. L'une des particularités est que la couche dans laquelle se trouve l'objet de référence est composée uniquement de ce dernier comme noeud.

L'analyse du graphe est axée sur la façon dont les objets sont connectés et aussi comment leurs attributs évoluent le long de la série temporelle. Dans le dernier cas, l'accent est mis sur l'ensemble du site d'étude mais on se focalise davantage sur la répartition des entités spatiales les plus spatialement stables et les plus dynamiques. La description des différentes étapes est donnée par le pseudo-code 4.

Nous illustrons dans la Figure 48, l'ensemble des étapes qui entrent dans la construction des graphes d'évolution spatio-temporel.

Algorithm 4 Fonction pour la construction des arrêtes

Require: GraphList *#liste des graphes*

$TotalEdge = \phi$
for all $graph \in GraphList$ **do**
  $ListOfPeriod = Period(graph)$
  $ListOfPeriod = AscendingOrder(ListOfPeriod)$ *# ordonner pour chaque graphe, les periodes suivant les dates d'acquisitions par ordre croissant*
  $listeEdge = \phi$ *# contient la liste des arêtes*
  for all $time \in range(taille(ListOfPeriod) - 1)$ **do**
    $image1 = ListOfPeriod[time]$
    $image2 = ListOfPeriod[time + 1]$
    for all $node1 \in ListOfPeriod[image1]$ **do**
      for all $node2 \in ListOfPeriod[image2]$ **do**
        $inters = Pix(node1) \cap Pix(node2)$
        if $inters > 0$ **then**
          $edge = (node1, node2)$
          $listeEdge.APPEND(edge)$
        end if
      end for
    end for
  end for
  $TotalEdge.APPEND(listeEdge)$ *# contient chaque objet de reference et les noeuds qui lui sont associés*
end for
return $TotalEdge$

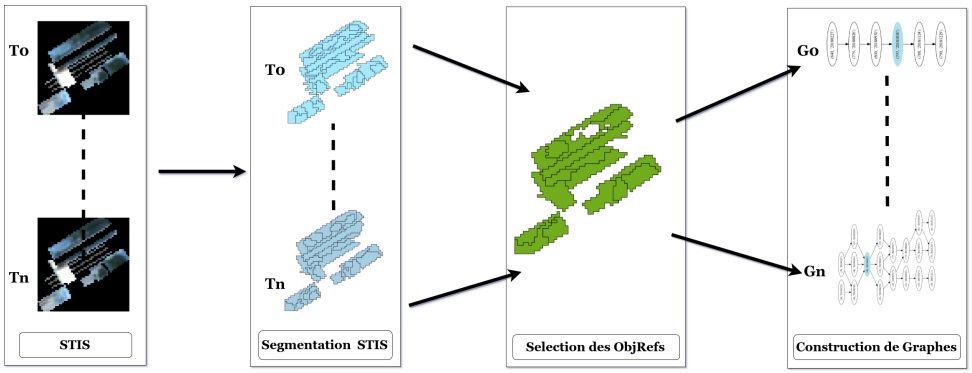


FIGURE 48 – Schéma général pour la construction de graphes d'évolution

6.3.3 Enrichissement d'images satellite par les données textuelles - Linking Pipeline

L'étape d'enrichissement consiste à mettre en relation les informations extraites d'une part à partir des données textuelles (résumé, termes), et d'autre part, celles extraites à partir des images satellite (graphes d'évolution).

La mise en relation est faite tout d'abord par une mise en correspondance entre les données d'une même période. Les informations extraites des données textuelles issues d'une période T seront mises en correspondance avec les noeuds du graphe d'évolution de cette même période, et ainsi de suite. Après la mise en correspondance, une analyse est faite sur ces couples d'informations (textes, noeuds), dans le but de faire ressortir toute information dans le texte pouvant contribuer à expliquer de façon objective, l'évolution de l'entité spatiale à une période donnée. Cette partie de l'étude est illustrée dans la Figure 49, suivie de l'approche méthodologique dans la Figure 50.

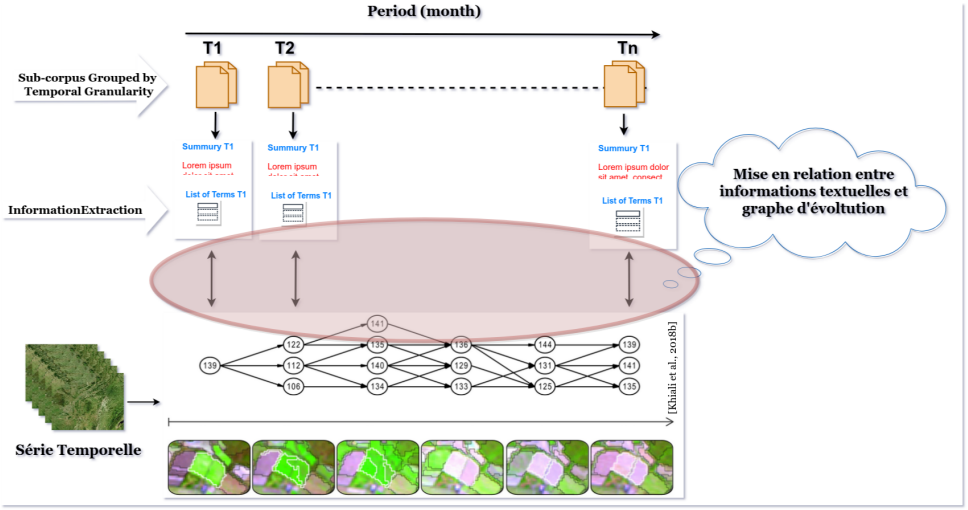


FIGURE 49 – Processus de mise en relation

6.4 Cas d'application : 3M

Afin de mettre en application l'approche précédemment présentée pour la mise en relation entre données hétérogènes, nous proposons, dans cette section, un cas d'étude pour 3M. L'objectif de l'étude consiste à mettre en relation des données relatives à un évènement territorial, à savoir la construction de la gare de train dénommée *Gare de Montpellier-Sud-de-France*[7]. Il s'agit entre autres, d'étudier l'évolution de la construction de cette gare à partir d'images satellite, et des données textuelles de diverses sources.

7. https ://fr.wikipedia.org/wiki/Gare_de_Montpellier-Sud-de-France

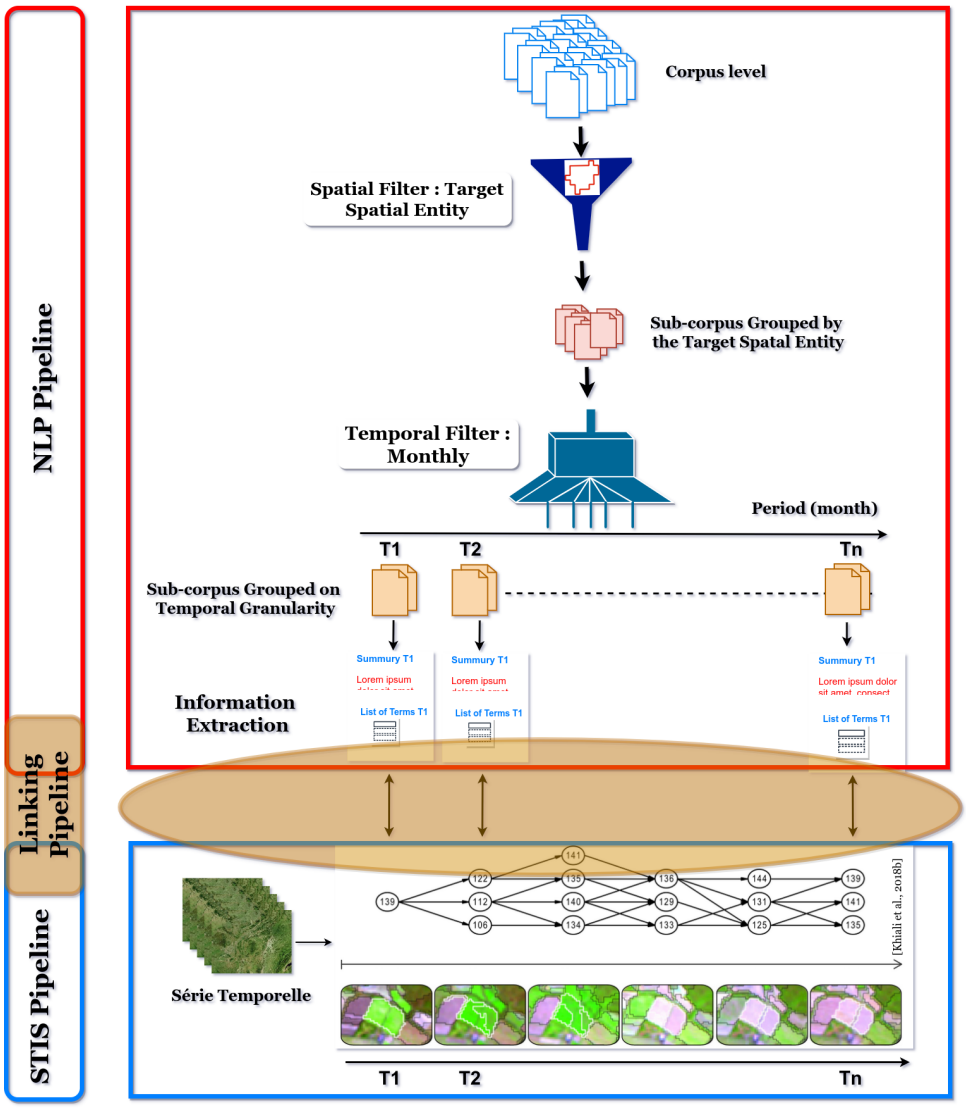


FIGURE 50 – Approche générale pour la mise en relation entre données textuelles et images satellite

La Gare de Montpellier-Sud-de-France (voir Figure 51) est constitué d'un bâtiment voyageurs qui mesure entre 10 et 15 mètres de haut, et dispose d'un hall de 3 500 m2, de 650 m2 pour des commerces, 900 m2 pour les services aux voyageurs et 800 m2 de locaux techniques [8] . Elle a été réalisée entre 2014 et 2018. Ci-dessous, nous listons quelques dates clés de la construction de cette gare.

— Juillet 2014 : lancement des travaux ;
— Avril 2016 : livraison dalle-pont ;
— Pose de la toiture en 2016 ;
— Août 2017 : livraison Bâtiment voyageur de 3500m2 ;
— Juillet 2018 : première desserte.

8. https ://fr.wikipedia.org/wiki/Gare_de_Montpellier-Sud-de-France

FIGURE 51 – Vue générale sur la gare

6.4.1 Description des données

6.4.2 Données textuelles

Les données textuelles utilisées lors de cette étude sont étroitement liées à celles abordées dans la Section 3.4.4 du Chapitre 3.

Conformément au processus de récupération et d'exploration de données, introduit dans la section citée ci-dessus, les données utilisées dans cette partie de l'étude se résument au résultat d'une requête utilisateur. Cependant, des contraintes supplémentaires s'appliquent lorsqu'il s'agit de récupérer des données qui seront utilisées en entrée de la méthode proposée. Dans le cadre de cette étude, nous avons spécifié les contraintes suivantes qui doivent être respectées lors de la récupération des données :

— une emprise spatiale obligatoire
  - lors d'une requête, l'utilisateur doit préciser l'emprise spatiale sur laquelle il souhaite obtenir des résultats.
  - de préférence à l'échelle d'un point d'intérêt (Point-Of-Interest POI). Ce qui correspond à la *Gare de Montpellier-Sud-de-France* dans notre cas.

— une emprise temporelle obligatoire
  - elle doit être sur un intervalle d'au moins un an. La réalisation de la gare, ayant pris plusieurs années, nous fixons l'intervalle minimal à un an, afin de permettre une vue générale de son évolution sur cette période.

— une expression thématique obligatoire
  - l'utilisateur doit préciser au moins un mot clé ou une expression thématique pour décrire les données qu'il souhaite récupérer.

6.4.3 Images satellite

Divers satellites offrent des données d'observation de la terre. Dans le cadre de notre étude, nous utilisons des données des satellites LANDSAT-8 [9] et Sentinel-2 [10]. Le choix de ces satellites se justifie par le fait qu'ils offrent des données dans le spectre visible de l'homme.. Soient, les bandes B2 (bleue), B3 (vert), B4 (rouge), et B5 (proche infra-rouge) pour LANDSAT-8 et B2 (bleue), B3 (vert), B4 (rouge), et B8 (proche infra-rouge) pour Sentinel-2.

Nous avons travaillé avec les données disponibles sur le portail de Theia [11]. Theia offre une base de données d'images de plusieurs satellites sur l'observation de la terre. Nous avons entre autres, des images des satellites LANDSAT-8, SPOT, Sentinel-1, Sentinel-2, etc.

Au total, nous avons récupéré 12*5 images, soit un total de 60 images pour les cinq années (voir Figure 52). Les images de chaque année sont évaluées, afin de choisir celles qui sont de bonne qualité. En effet, certaines images ne contiennent pas de données valides, dû à des problèmes de capture lors du passage du satellite. Seules les images identifiées comme valides sont utilisées lors des expérimentations. Nous décrivons ci-dessous, les pré-traitements appliqués sur ces images avant la construction des graphes.

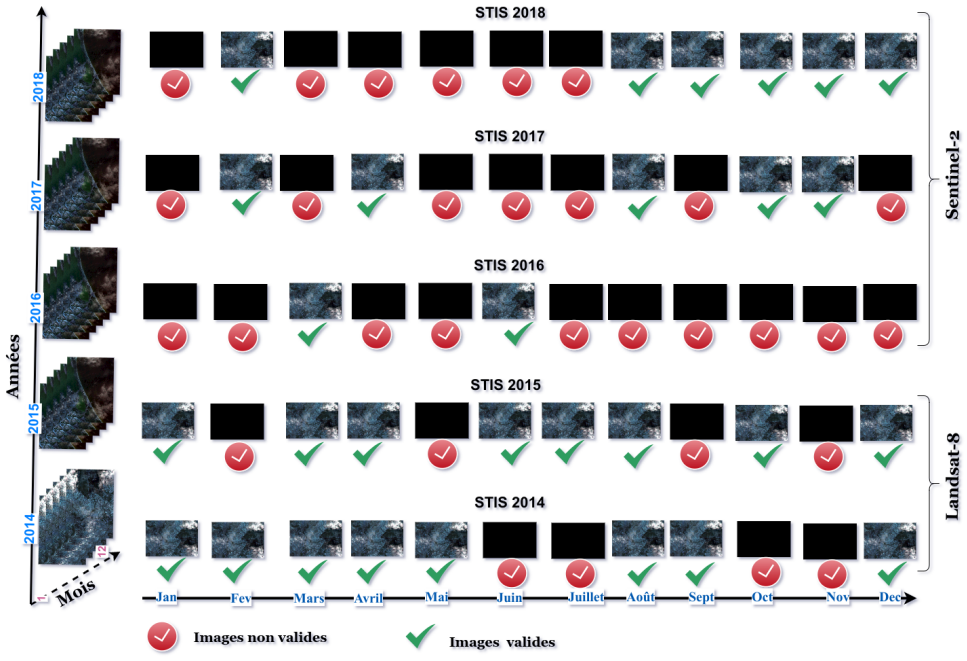


FIGURE 52 – État sur les images utilisées

Tout d'abord, nous appliquons un ré-échantillonage des images LANDSAT-8 qui ont

9. https ://earth.esa.int/eogateway/missions/landsat-8
10. https ://sentinel.esa.int/web/sentinel/missions/sentinel-2
11. https ://www.theia-land.fr/en/theia/

une résolution de 30 cm à 10 cm. Le ré-échantillonage est appliqué dans le but de pouvoir suivre les entités spatiales dont les tailles sont réduites.

En plus des quatre bandes précisées plus haut, nous introduisons une autre bande. Il s'agit de l'indice de végétation par différence normalisée (Normalized Difference Vegetation Index - NDVI) [Tucker et al., 2001]. Le NDVI est obtenu à partir de deux bandes, la bande proche infra-rouge (NIR) et la bande rouge (RED), soit respectivement B8, B4 pour Sentinel-2 et B5, B4 pour Landsat-8. L'indice NDVI est à la fois sensible à la biomasse et à l'activité chlorophyllienne. Plus le couvert végétal est développé et en pleine croissance, plus de NDVI est élevé, le NIR étant alors fort et le RED faible. Si la végétation est par contre morte ou que les données ont été enregistrées sur un sol nu, le NIR sera plus faible et le RED plus fort ce qui diminuera l'écart NIR-RED et le NDVI correspondant. Sa valeur varie entre -1 et 1. Il est calculé comme indique la formule suivante :

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)} \tag{6.3}$$

Nous reconstituons ensuite, pour chaque période, une nouvelle image composée des 5 bandes obtenues, la bande B2, B3, B4, B8 (ou B2, B3, B4, B5) et le NDVI. Enfin, nous découpons notre zone d'intérêt (la gare du sud) que nous avons extrait à partir de OpenStreetMap [12]. Les images obtenues après découpage sont d'une dimension de 52 pixels * 53 pixels (voir Figure 53).

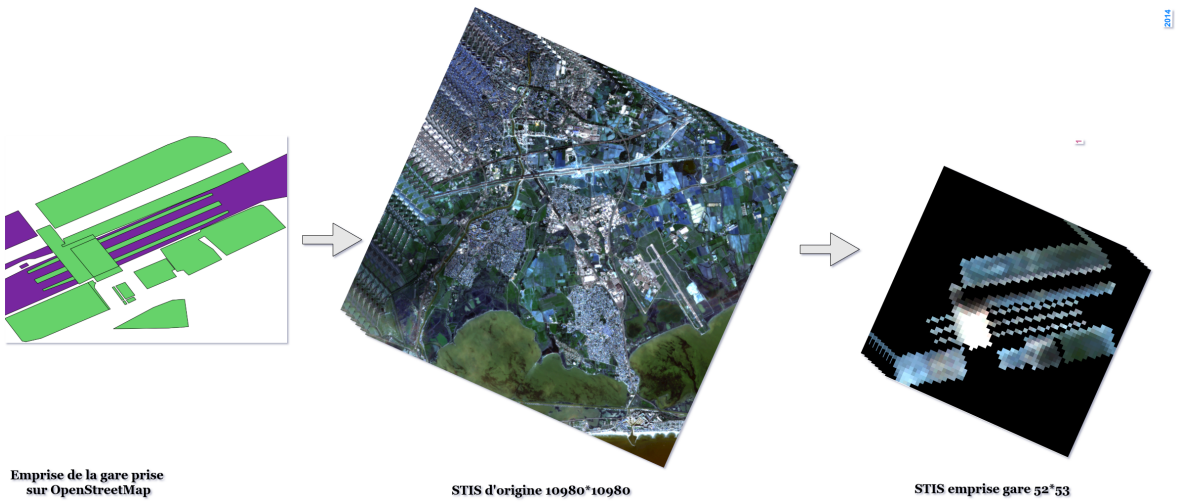


FIGURE 53 – Données de la gare : découpage de la gare à partir des STIS d'origine

À partir de la Figure 54, nous présentons les données utilisées en entrée de la méthode. Il s'agit entre autres, les séries temporelles de données textuelles, et d'images satellite en relation avec la gare du sud de Montpellier.

12. https ://www.openstreetmap.fr/

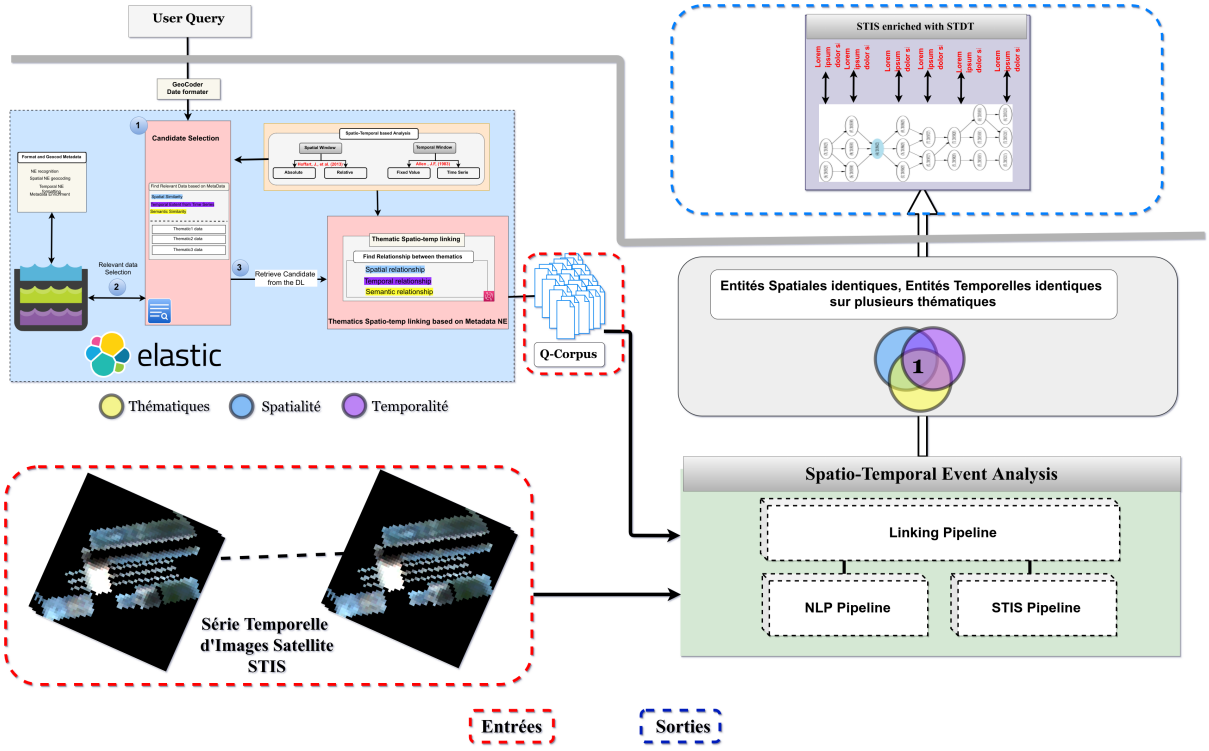


FIGURE 54 – Données utilisées en entrée de la méthode

6.4.4 Expérimentations

6.4.4.1 Analyse de corpus thématiques

Nous illustrons dans la Figure 55, le processus de filtres et d'analyses appliqués aux corpus thématiques *Q-Corpus*.

Après application du filtre spatial, nous remarquons qu'il y a environ 1,5% des documents de la base de données qui contiennent ou parlent de la *Gare de Montpellier Sud de France*, soit 105 documents sur 6854 dans la base de données. Ce filtre spatial est appliqué en tenant compte des différentes variations ou appellations communément attribuées à cette gare. Nous avons entre autres sélectionnées manuellement, *Montpellier Sud de France*, *Gare Montpellier Sud*, *Gare Sud de France à Montpellier*, *Gare Montpellier Sud de France*, *Gare Sud de France*, *Nouvelle gare Sud de France*, *Gare du Sud de Montpellier*, *Nouvelle gare de Montpellier*, *Gare Montpellier Sud de France*, *gare de Montpellier-Sud-de-France*, *Montpellier-Sud-de-France*, *nouvelle gare de Montpellier*, *Montpellier Gare de Sud de France*, *Gare de Montpellier-Sud-de-France*, *Gare du Sud de France*.

Ces données sont ensuite filtrées annuellement, puis mensuellement, en tenant compte des années de réalisation de la gare, soient 2014, 2015, 2016, 2017 et 2018. Nous obtenons par exemple pour l'année 2017, 01 document respectivement pour chacun des mois, mars, mai, juillet, octobre, et 04 documents pour le mois de novembre.

Dans le processus d'extraction des informations textuelles, nous avons utilisé respecti-

vement Xlnet [Yang et al., 2019], qui est un type de transformers, un modèle auto-régressif généralisé pour la compréhension du langage naturel, pour le résumé des corpus, et [Honnibal and Montani, 2017] pour l'extraction des entités nommées. Nous avons pris en compte uniquement deux types d'entités nommées, à savoir les noms des organismes *ORG* et des personnes *PER*, dans le but de capturer les entreprises et éventuellement des noms de personnes qui sont citées d'une manière ou d'une autre dans les documents.

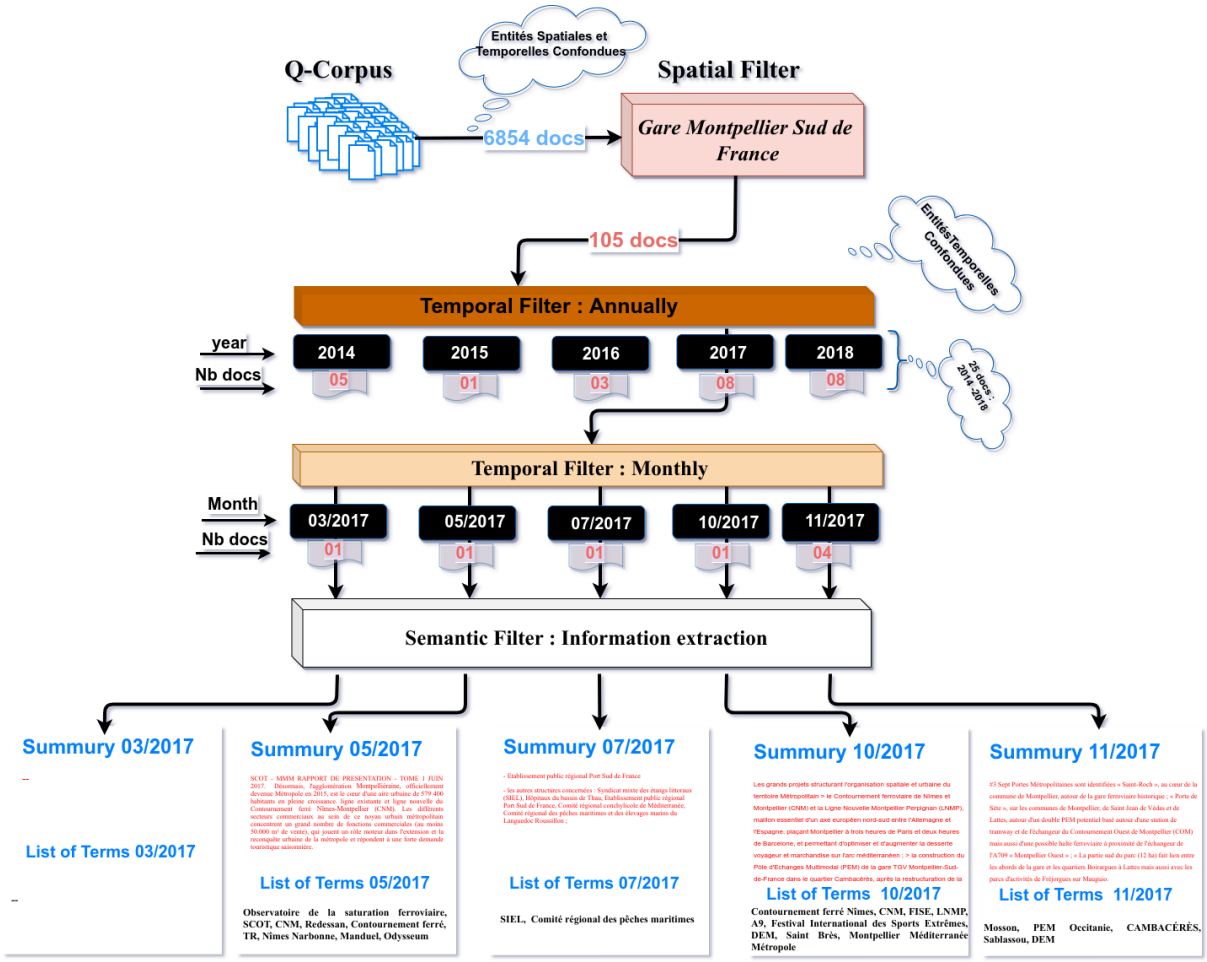


FIGURE 55 – Analyse de données textuelles de l'année 2017

6.4.4.2 Analyse d'images satellite

6.4.4.2.1 Algorithme et paramètres de segmentation

Nous avons utilisé l'algorithme LargeScaleMeanShift [Comaniciu and Meer, 2002] pour la segmentation des images. Il prend essentiellement trois paramètres qui sont :

- le spatialr : représente la valeur du rayon spatial pour le calcul de la moyenne. Plus cette valeur est grande, plus long sera le temps de calcul. Il est exprimé en nombre de pixels.
- le ranger : représente le seuil sur la distance euclidienne de la signature spectrale (exprimée en unité de radiométrie). Plus la valeur est petite, plus les pixels qui se

rapprochent en terme de valeur radio-métrique seront mis ensemble dans le rayon choisi. Ce dernier n'a pas d'impact sur le temps de calcul.

- le minsize (facultatif) : représente la taille minimale du segment à considérer au cours de la segmentation. Tout segment ayant une taille minimum à la valeur de minsize sera ajouté en extension au segment avec lequel il est le plus proche.

Afin d'obtenir des objets compacts, suite aux différentes expérimentations, nous avons fixé la valeur de la compacité à spatialr = 4 avec une tolérance de ranger = 10 sur les variations entre valeurs radio-métriques des pixels à considérer lors du calcul de la moyenne. Quant à la taille minimale, nous avons laissé le choix à l'algorithme.

6.4.4.2.2 Construction et choix des graphes

Nous procédons tout d'abord à la sélection des objets en partant de deux principaux critères pour chaque valeur de α. La première est le taux de couverture et la seconde est le taux de chevauchement ou de redondance d'informations entre objets de référence. Les bons paramètres de α sont donc ceux qui permettent d'obtenir un taux maximum de couverture avec une faible valeur de redondance entre objets de référence.

Étant donné que la construction des graphes est basée sur les objets de référence, nous nous intéressons plus à la couverture des graphes sur l'espace total et à la connexité des noeuds de ces différents graphes (chaque objet de référence constitue un graphe d'évolution). Le taux de couverture idéal est de 100% avec une connexité absolue entre les différents noeuds pour chacun des graphes. Les bonnes valeurs de $\sigma 1$ et $\sigma 2$ sont donc celles qui permettent d'obtenir un bon compromis entre ces deux critères, c'est-à-dire couverture et connexité.

En partant des différentes expérimentations que nous avons effectuées, nous fixons les valeurs de ces paramètres comme suit : $\alpha = 0.6$, $\sigma 1 = 0.5$ et $\sigma 2 = 0.3$. Avec ces paramètres, nous notons un taux de couverture de 98,5%, et un taux de chevauchement de 29,23%.

Dans les Figures 56, 57, 58, et 59, nous présentons des exemples de graphes d'évolution résultants respectivement d'objets de référence (en bleu) issus respectivement des observations de : avril 2014, avril 2015 et août 2017 et octobre 2018.

Pour chacun des objets de la gare, représenté sous forme de graphe, nous pouvons constater son évolution spatiale tout au long de la série temporelle. De prime abord, nous pouvons identifier ces objets en nous basant sur les données de la gare disponibles sur OpenStreetMap. Le graphe d'évolution de l'année 2014 correspond aujourd'hui à un espace de parking. Notons que le parking n'existait pas en 2014, et cet endroit était probablement couvert de végétation. Celui de 2015, correspond de nos jours à une partie du bâtiment voyageur de la gare et un allongement sur l'autoroute. Quant à ceux de 2017 et 2018, ils correspondent exactement au bâtiment voyageur de la gare. Les variations constatées en 2017, pourraient être justifiées par le fait que des travaux étaient toujours

en cours. Néanmoins, nous constatons la formation d'un objet unique vers fin août, ce qui pourrait justifier le fait que le bâtiment voyageur soit livré en août, selon les données que nous avons pu recueillir sur Wikipédia. Pour ce qui concerne le graphe d'évolution de 2018, son caractère linéaire se justifie par le fait qu'il n'y avait pratiquement plus de travaux à cette date. Notons que la première desserte de la gare a eu lieu en juillet 2018.

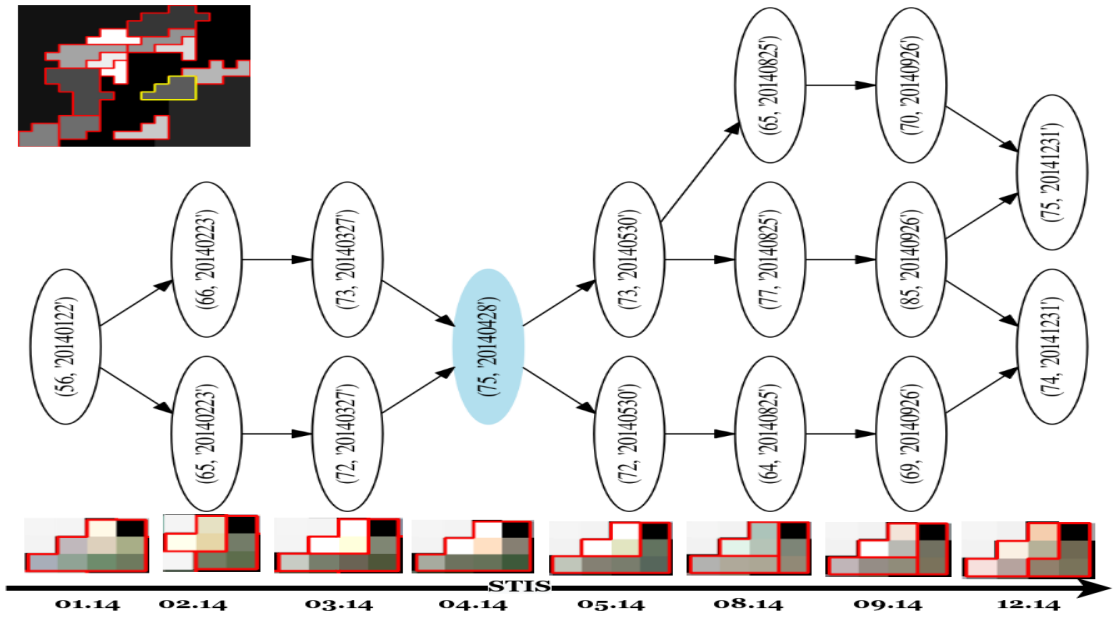


FIGURE 56 – Cas d'un graphe d'évolution de l'année 2014

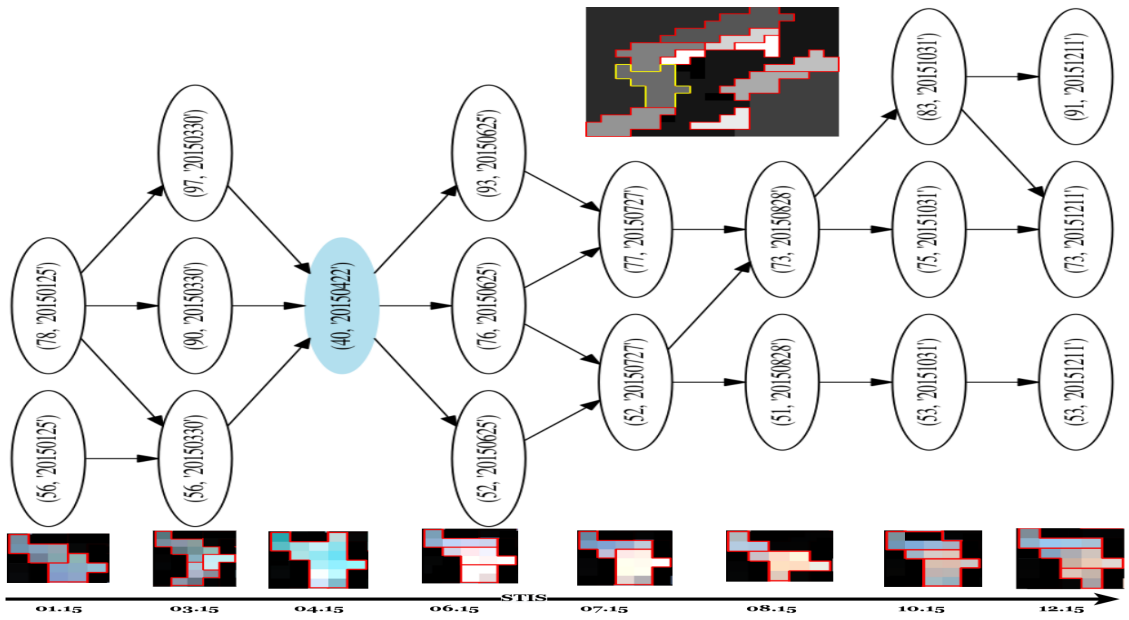


FIGURE 57 – Cas d'un graphe d'évolution de l'année 2015

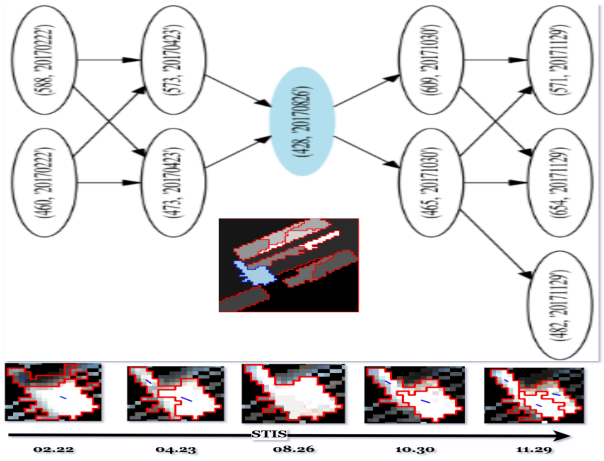


FIGURE 58 – Cas d'un graphe d'évolution de l'année 2017 : bâtiment voyageur

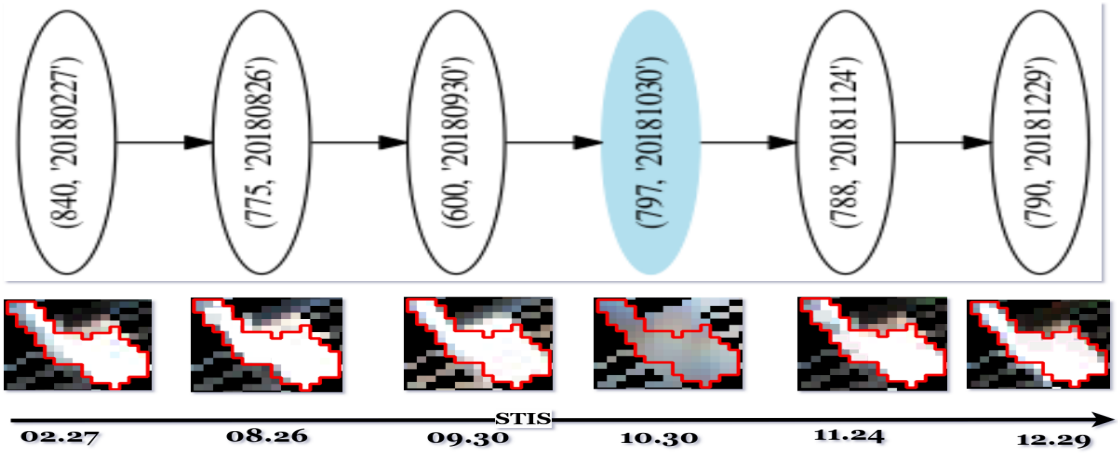


FIGURE 59 – Cas d'un graphe d'évolution de l'année 2018 : bâtiment voyageur

Les codes sources pour la construction des graphes d'évolution sont disponibles dans le dépôt *Github STIS*.

6.4.4.3 Mise en Relation et analyse

Comme décrite dans la Section 6.3, cette étape consiste à mettre en correspondance, et à analyser conjointement les informations extraites des données textuelles avec les graphes d'évolution par intervalle de temps (voir Figure 60).

Dans les Tableaux 6.1, 6.2, 6.3, 6.4, et 6.5, nous présentons les résultats obtenus de la mise en relation entre les informations textuelles et graphes des années 2014, 2015, 2016, 2017, et 2018. Pour les informations textuelles, nous renseignons une part les résumés et d'autre part les termes descriptifs (noms d'organismes et personnes), et pour les graphes d'évolution nous renseignons le nombre de noeuds à la période correspondante. Certains noeuds n'ont pas de correspondance avec des informations textuelles, et cela s'explique par le fait qu'il y a pas eu de documents collectés à cette date lors de nos expérimentations.

Période		Janvier	Février	Mars	Avril	Mai	Août	Septembre	Décembre
Infos Textuelles	Résumés	Nombre de politiques et d'actions déjà engagées constituent des acquis déterminants pour atteindre les objectifs du Plan Climat. La ville des nouvelles mobilités, visant à favoriser l'inter-modalité notamment avec la prolongation de la ligne 1 de tramway pour desservir la gare Montpellier Sud de France et le quartier ; Participation de Montpellier Agglomération = 67,2M€ (07/2011) sur 1757M€ (07/2011), soit 3,8% Sous-action 2 : Participation de Montpellier Agglomération = 6M€ sur 50M€, soit 12% Sous-action 4 : connecter une nouvelle centralité urbaine à l'ensemble de l'Agglomération : Par conséquent, grâce au prolongement de la ligne 1 de tramway, la Gare Montpellier Sud de France et le nouvel éco-quartier OZ Montpellier Nature Urbaine bénéficieront d'une très bonne accessibilité en transports en commun pour tous les habitants de Montpellier Agglomération. PROLONGER LA LIGNE 1 DU TRAMWAY JUSQU'A LA GARE MONTPELLIER SUD DE FRANCE L'extension de la ligne 1 du tramway répond à trois objectifs principaux				51% du montant des achats de fruits distribués en dehors du temps de la restauration sont ainsi remboursés par France Agri Mer, dans le cadre d'un programme européen*.	© ICADE Premières images de la Gare Montpellier Sud de France Montpellier à moins de 3 heures de Paris, c'est pour la fin de l'année 2017. Les travaux du Contournement ferroviaire de Nîmes et Montpellier, sont en cours depuis fin 2012 et ceux de la Gare de Montpellier Sud de France débuteront l'an prochain.	La desserte du quartier et de la gare Montpellier Sud de France sera assurée, dès 2017, par : Les accès viaires et notamment le cours de la gare seront le support d'une activité économique développée, en particulier, la voie principale reliant le centre-ville de Montpellier à la gare nouvelle Montpellier Sud de France 96 / 130 Juillet 2014 Gare Nouvelle Montpellier Sud France (cours de la gare)sera bordée par une activité commerciale principalement concentré au sein du quartier de gare Des itinéraires directs, exclusivement dédiés à l'usage piéton ou vélos, seront aménagés permettant ainsi de et le long de la High Street (cours de la Gare).	
	Organismes et Personnalités	TR, Montpellier Agglomération, Nîmes Manduel, Vélomagg, ZAC OZ 1, CNM, LIGNE 1, L'ACTION Direction du Foncier, ADEME, SAAM, Région Languedoc, Montpellier St Roch, PROLONGER, Transports Régionaux, OZ Montpellier Nature Urbaine				France Agri Mer	ICADE, France Montpellier, Réseau Ferré de France, Montpellier Agglomération	A9, AUTORITÉ ENVIRONNEMENTALE	
Graphes	Nb Noeuds	1	2	2	1	2	3	3	2
	Empreinte Max				Avril				
Analyse		Nous percevons le projet visant à prolonger la ligne 1 du Tramway pour desservir la gare Montpellier Sud de France, mais il n y pas d'indication précise sur l'état d'avancement de la gare elle même. L'unique noeud de graphe décrit une certaine stabilité de la zone à cette période, qui correspond aujourd'hui à un espace de parking			Période à la quelle la zone étudiée avait une empreinte maximale. Il n y a pas eu de documents collectés à cette période	Contenu en relation avec l'agriculture, donc pas de relation directe avec la gare du Sud	Dans ce résumé, nous percevons d'une part, l'annonce du début des travaux par les premières images et la fin des travaux en 2017. D'autre part, nous remarquons aussi le nom de ICADE, l'entreprise chargée de réaliser les travaux. Les trois noeuds du graphe peuvent s'expliquer par le faite que la zone à subit ou est entrain de subir des transformations (liées à la fouille, etc.) sur le plan spatial	Annonce la fin de travaux des itinéraires d'accès à la gare en 2017, soit la date de fin des travaux déjà annoncés pour la gare. Nous pouvons noter également le nom de l'A9, qui a une relation étroite avec la gare. Les 3 noeuds du graphe peuvent toujours s'expliquer par le fait que des travaux sont en cours.	

TABLE 6.1 – Tableau récapitulatif sur l'analyse des résultats : 2014

Période		Janvier	Mars	Avril	Juin	Juillet	Août	Octobre	Décembre
Infos Textuelles	Résumés	Ce plan était destiné à accompagner les mutations de la filière viti-vinicole régionale, autour de quatre axes stratégiques : la recherche – expérimentation (aide aux travaux visant l'adaptation de la filière aux exigences nouvelles du marché), l'appui aux projets d'entreprises viticoles (exploitations, coopératives et négociants), la restructuration viticole et la promotion des vins Sud de France.		Pour les demandeurs installés ou créés depuis plus d'un an à compter de la date de dépôt de dossier, ne pas présenter de fonds propres négatifs sur le dernier exercice comptable connu, ou être en procédure de sauvegarde ou de redressement judiciaire ; Ne pas avoir atteint le plafond De Minimis de 200 000 euros sur 3 ans fixé par le Règlement (UE) N1407/2013 ; Inscrire son projet dans un projet stratégique d'entreprise à 3 ans (Projet de développement agritouristique) ;					
	Organismes et Personnes	Stevia		Accueil Paysan, UE					
Graphes	Nb Noeuds	2	3	1	3	2	2	3	3
	Empreinte Max				Avril				
Analyse		pas de relation visible avec la gare dans ce résumé		pas de relation visible avec la gare dans ce résumé					

TABLE 6.2 – Tableau récapitulatif sur l'analyse des résultats : 2015

Période		Juin	Octobre
Infos Textuelles	Résumés	Montpellier (Hérault), reportage La voiture de Jacky Bessières quitte la voie rapide pour emprunter une petite route défoncée par le passage des camions. C'est ici qu'ils construisent la future gare de La Mogère. Pourtant, son financement est aujourd'hui remis en cause. Une gare à 135 millions d'euros, pour quelques TGV par jour, sans correspondance, alors que celle du centre-ville de Montpellier vient d'être refaite, « est-ce indispensable ? » s'interroge Jean-Luc Gibelin, conseiller régional communiste en charge des transports.	Ce projet présenté par la SAAM, maître d'ouvrage, consiste à la création d'un nouveau réseau de géothermie afin d'approvisionner en chaleur la nouvelle gare de Montpellier et son enveloppe urbaine ainsi que le lotissement Urban Parck situé à proximité sur le territoire de la commune de Lattes. Le projet du quartier Mogère consiste en la création de la nouvelle Gare TGV Sud de France et de son quartier environnant. le projet de gare nouvelle Montpellier Sud de Franceliée au Contournement Nîmes-Montpellier :
	Organismes et Personnes	Jacky Bessières, A9, Jean-Luc Gibelin, M. Bessières	ZAC OZ 1, CNM, A9, SAAM, Mogère, Urban Parck
Graphes	Nb Noeuds		
	Empreinte Max		
Analyse		Nous pouvons percevoir dans ce résumé, d'une part, que la gare est en cours de construction et portait une autre appellation à son temps : la gare de Mogère. D'autre part, nous notons aussi les désaccords qui animent le projet de construction en son temps. En plus de ces informations, nous avons l'A9 qui ressort dans ces textes, d'où sa relation avec la gare. Il y a aussi les noms tels que M. Bessières, un usager qui rapporte l'état de la gare, et Jean-Luc Gibelin qui se pose des questions sur la nécessité de cette gare. Nous n'avons pas de graphes d'évolution pour 2016, car nous n'avons que 2 images valides, mais nous pouvons déduire que les travaux étaient en cours à cette date	Nous percevons les travaux qui sont parallèlement en cours avec la gare, à savoir la création d'un nouveau réseau de géothermie par l'entreprise SAAM, et aussi autres relations telles que la gare et l'A9, le CNM, et la ZAC ÀZ1. Nous pouvons déduire de ce texte que les travaux étaient toujours en cours

TABLE 6.3 – Tableau récapitulatif sur l'analyse des résultats : 2016

Période		Février	Avril	Août	Octobre	Novembre
Infos Textuelles	Résumés		SCOT - MMM RAPPORT DE PRESENTATION - TOME 1 JUIN 2017 Désormais, l'agglomération Montpelliéraine, officiellement devenue Métropole en 2015, est le cœur d'une aire urbaine de 579 400 habitants en pleine croissance. ligne existante et ligne nouvelle du Contournement ferré Nîmes-Montpellier (CNM). Les différents secteurs commerciaux au sein de ce noyau urbain métropolitain concentrent un grand nombre de fonctions commerciales (au moins 50.000 m de vente), qui jouent un rôle moteur dans l'extension et la reconquête urbaine de la métropole et répondent à une forte demande touristique saisonnière.	- Etablissement public régional Port Sud de France - les autres structures concernées : Syndicat mixte des étangs littoraux (SIEL), Hôpitaux du bassin de Thau, Etablissement public régional Port Sud de France, Comité régional conchylicole de Méditerranée, Comité régional des pêches maritimes et des élevages marins du Languedoc Roussillon ;	Les grands projets structurant l'organisation spatiale et urbaine du territoire Métropolitain > le Contournement ferroviaire de Nîmes et Montpellier (CNM) et la Ligne Nouvelle Montpellier Perpignan (LNMP), maillon essentiel d'un axe européen nord-sud entre l'Allemagne et l'Espagne, plaçant Montpellier à trois heures de Paris et deux heures de Barcelone, et permettant d'optimiser et d'augmenter la desserte voyageur et marchandise sur l'arc méditerranéen ; > la construction du Pôle d'Échanges Multimodal (PEM) de la gare TGV Montpellier-Sud-de-France dans le quartier Cambacérès, après la restructuration de la gare Montpellier Saint-Roch, préfigure la montée en charge progressive de la desserte ferroviaire de Montpellier dans les dix prochaines années ; Il constitue également le site d'accueil de la future gare nouvelle Montpellier Sud de France. son aménagement relevant de la compétence métropolitaine. > premières phases de logements, bureaux & services	3 Sept Portes Métropolitaines sont identifiées « Saint-Roch », au cœur de la commune de Montpellier, autour de la gare ferroviaire historique ; « Porte de Sète », sur les communes de Montpellier, de Saint Jean de Védas et de Lattes, autour d'un double PEM potentiel basé autour d'une station de tramway et de l'échangeur du Contournement Ouest de Montpellier (COM) mais aussi d'une possible halte ferroviaire à proximité de l'échangeur de l'A709 « Montpellier Ouest » ; « La partie sud du parc (12 ha) fait lien entre les abords de la gare et les quartiers Boirargues à Lattes mais aussi avec les parcs d'activités de Fréjorgues sur Mauguio. Saint-Roch Nationale et internationale Mogère Sète Vallée de l'Hérault Pic Saint Loup Sablassou Vallée du Bérange L'amphithéâtre Lignes ferroviaires LGV/TER (principales gares de destination
	Organismes et Personnes		Observatoire de la saturation ferroviaire, SCOT, CNM, Redessan, Contournement ferré, TR, Nîmes Narbonne, Manduel, Odysseum	SIEL, Comité régional des pêches maritimes	Contournement ferré Nîmes, CNM, FISE, LNMP, A9, Festival International des Sports Extrêmes, DEM, Saint Brès, Montpellier Méditerranée Métropole	Mosson, PEM Occitanie, CAMBACÉRÈS, Sablassou, DEM, Mogère Sète Vallée, TCSP, Saint-Roch, TER, Rondelet, Sablassou
Graphes	Nb Noeuds	2	2	1	2	3
	Empreinte Max			Août		
Analyse			Nous retenons la relation entre le SCOT et la nouvelle ligne du CNM qui est en relation avec la gare. Pas d'indication explicite sur l'état d'avancement de la gare, mais nous pouvons déduire que les travaux sont toujours en cours sur le bâtiment voyageur, qui est représenté en deux noeuds.	Le contenu de ce résumé à cette période n'a pas de relation directe avec la gare, par contre, le graphe d'évolution à un noeud corrobore avec le fait que le bâtiment ait été livré dans le mois d'août.	Dans ce résumé, les grands projets structurant l'organisation spatiale et urbaine du territoire Métropolitain, dont la gare du sud y sont mentionnés. Il n y a pas de précision sur l'état d'avancement de la gare, mais nous pouvons déduire avec les deux noeuds, que des travaux continuent après sur cette dernière.	pas de relation explicite avec la gare dans le contenu de ces documents

TABLE 6.4 – Tableau récapitulatif sur l'analyse des résultats : 2017

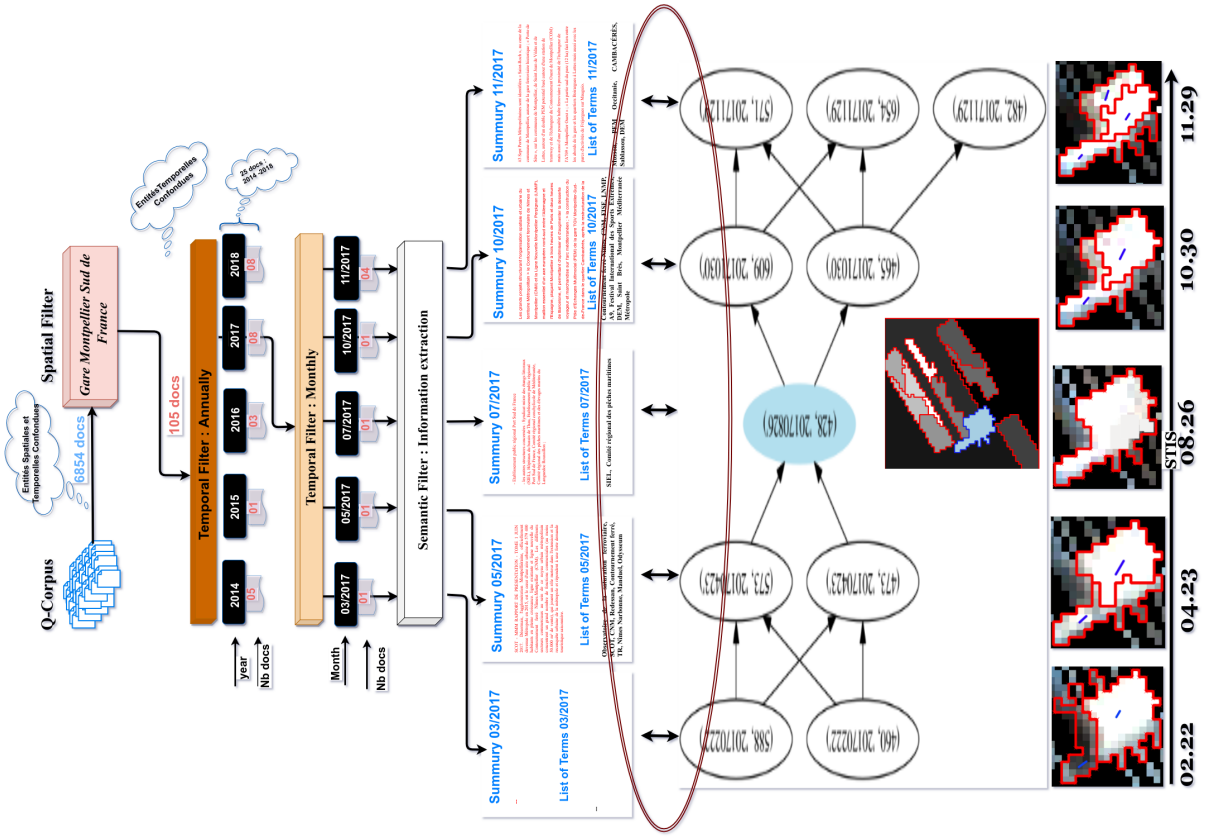


FIGURE 60 – Mise en relation : STDT & STIS de 2017

Période		Février	Août	Septembre	Octobre	Novembre	Décembre
Infos Textuelles	Résumés	Le prolongement de la ligne 1 du tramway (Mosson – Odysseum) jusqu'au pôle d'échange multimodal Montpellier Sud de France (mise en service 2022).	En matière de rayonnement, le projet prévoit la structuration de la porte d'entrée Est du territoire régional en articulant étroitement les infrastructures de grande accessibilité, aéroportuaires et ferroviaires, le débouché maritime du port de Sète-Frontignan-Sud-de-France et l'autoroute A9/A709, au rayonnement métropolitain. Il s'agit de structurer des nœuds d'échanges et d'intermodalité consacrés au travers des deux portes métropolitaines internationales et nationales de Montpellier Sud de France et Montpellier Saint Roch, de concevoir « la vitrine métropolitaine active » et la recomposition territoriale associée aux contournements routier et ferroviaire et enfin, de renforcer les liaisons ferroviaires entre le port de Sète Frontignan-Sud-de-France et la métropole. Le défi C. Il nous faut une métropole équilibrée et efficace, et, donc, accentuer l'attractivité et le rayonnement métropolitain. L'affirmation des connexions avec le port Sète Frontignan Sud de France pour renforcer les lignes ferroviaires. Nationaux, interrégionaux, métropolitains, évidemment, et locaux. De nouveaux comportements en matière de mobilité.	La question des circulations piétonnes à l'intérieur des quartiers a également été posée à Villeneuve-lès-Maguelone, Montpellier (quartier Pompignane) ou encore à Lattes (futur site d'extension urbaine entre Boirargues et la gare TGV Sud de France). De l'autogare faire partir des bus directement sur la piste (comme dans les grands aéroports)». À Montpellier, Nicolas soumet une autre proposition : « Plusieurs dizaines d'observations portent sur la gestion de l'interface entre ce futur quartier et le quartier résidentiel existant de Boirargues, tout particulièrement concernant la question de la hauteur des futurs bâtiments.	- augmentation des desertes TER, en correspondance avec les TAGV en gare de Montpellier « Saint Roch » (et également possible en gare de Montpellier « Sud de France » depuis la mise en service du CNM). Au Nord du projet, la desserte de Montpellier sera assurée à la fois par la gare actuelle de Montpellier « Saint Roch », et par la gare de Montpellier « Sud de France », implantée sur le Contournement de Nimes et Montpellier, au sein du nouveau quartier au Sud d'Odysseum. le parking Vicarello de la ville de Castelnau, a été réattribué à Ta M jusqu'au 31/12/2019. des Collectivités dans le cadre de Monsieur Pascal compte rendu des La SAEML MONTPELLIER EVENTS a en charge le développement des activités sur deux équipements appartenant à la métropole que sont le Corum-Palais des Congrès-Opéra et le Zénith Sud par délégation de service publique, et sur deux équipements appartenant à la Région que sont la Sud de France Aréna et le Parc des expositions. Cette variation n'est pas homogène et varie suivant les sites :		
	Organismes et Personnes	Mosson – Odysseum	CNM, Frontignan Sud, Montpellier Saint-Roch, TCSP	CNM	CNM, Saint Roch, TER		
Graphes	Nb Noeuds	1	1	1	1	1	1
	Empreinte Max				Octobre		
Analyse		Annonce le projet sur le prolongement de la ligne 1 (Mosson – Odysseum) du tramway jusqu'à la gare pour 2022. Nous n'avons pas de précision sur l'état d'avancement de la gare, mais le caractère linéaire du bâtiment voyageur, représenté ici avec un noeud, nous indique une stabilité spatiale à cette période.	Nous constatons la description d'un projet qui vise à une forte interconnexion ferroviaire, dont la gare du sud en fait partie. Nous pouvons faire ressortir également la relation entre la gare et le CNM et les liaisons Sète-Frontignan-Sud de France. Identique au précédent, il n'y a pas d'information explicite sur l'état d'avancement, mais le seul noeud du graphe nous indique toujours une stabilité de la partie représenté (bâtiment voyageur) à cette date	Dans ce résumé, une question sur la circulation des piétons est posée pour plusieurs quartiers, à savoir le futur site qui devrait se situer entre Boirargues et la gare. Nous ne notons pas d'information explicite sur l'état d'avancement, mais le seul noeud du graphe nous indique toujours une stabilité de la partie représenté (bâtiment voyageur) à cette date. En tenant compte des informations sur la première desserte qui a eu lieu en Juillet 2018, nous pouvons déduire qu'il n'existait plus de travaux remarquables à cette date	Dans ce résumé, nous percevons le fait qu'il ait eu une augmentation de desertes TER, avec la mise en place de la gare Sud de France, en relation avec le CNM et la gare Saint Roch. L'existence de la gare est bien perçue dans ces paragraphes et nous pouvons conclure qu'elle est en service à cette date. Le noeud du graphe reste constant à un, ce qui signifie qu'il n'y a pas eu de mouvements ou de travaux. Cette date correspond également à l'emprise maximale du bâtiment voyageur que nous avons représenté.		

TABLE 6.5 – Tableau récapitulatif sur l'analyse des résultats : 2018

6.4.5 Discussion et Conclusion

D'après les analyses effectuées précédemment dans les différents tableaux, nous notons des corrélations entre les informations textuelles et les graphes d'évolution à certaines périodes, malgré la taille réduite des corpus aux différentes périodes. Au vu des expérimentations, nous remarquons aussi que les informations issues des documents donnent des éléments de contexte au chantier de construction de la gare tels que les accès (tramway, A9, abord de la gare, etc.) mais ne décrivent pas exhaustivement le chantier en lui même.

Notons que chacun des graphes d'évolution représente une partie de la gare, et non une représentation exhaustive de cette dernière sur le plan spatial, de même, le résumé des

documents de chaque période ne donne pas une vision complète de leur contenu. En effet, les STDT sont obtenues à partir des dates de publications des documents sur le web, ce qui signifie que ces documents peuvent contenir des informations sur des évènements pouvant être datées avant (historique), pendant (description d'une situation) ou après (une projection dans le futur) la date de publication. Par conséquent, les analyses tirées des relations *information textuelles <-> graphes d'évolution* ne nous permettent pas d'avoir une vue globale sur leur évolution, mais nous permet d'avoir une idée sur la possibilité d'enrichir les images satellite avec les données textuelles en partant sur l'approche proposée.

Afin de décrire de façon exhaustive, les relations qui peuvent exister entre ces informations, il est d'abord nécessaire de disposer d'une grande quantité de données, textuelles et d'images pour chaque période. Ensuite, nous pensons qu'une étude sur l'ensemble des graphes et/ou des corpus (ex : clustering spatio-temporel, calcul de distance de similarité entre documents de périodes différentes) pourra aider à mieux catégoriser et à mieux analyser leur comportement tout au long de la série temporelle. Ainsi, nous aurons la possibilité d'associer une évolution type, à l'ensemble de la gare pour chaque période.

Pour ce qui est de l'extraction des informations textuelles, il serait nécessaire d'envisager d'autre méthodes d'extraction en lieu et place des résumés qui peuvent devenir très lourds, lorsque le nombre de documents est très élevé. Nous pouvons envisager une méthode d'extraction d'information ITEXT-TERRI (TERRI comme territoire), à l'image de ITEXT-BIO que nous avons proposé dans le Chapitre 4, en se focalisant sur le croisement de termes entre les parties spécifiques des documents obtenus.

Pour terminer, les informations extraites des différentes analyses peuvent être utilisées comme étant un support qualitatif afin d'appuyer les analyses sur les évènements spatio-temporel sur le territoire.

6.5 Conclusion

Dans ce chapitre, nous proposons une approche de mise en relation de données fortement hétérogènes, à savoir des données textuelles et des images satellites. Les analyses des résultats issus de cette approche montrent l'importance de la mise en relation entre données hétérogènes pour le suivi d'évènements territoriaux sur le plan spatial et temporel. Elle fait ressortir clairement la possibilité de tirer profit des connaissances d'une part de données textuelles, et d'autre part des images satellite pour aider à expliquer un évènement spatio-temporel. Cette étude fait ressortir également l'importance du protocole de collecte proposé dans le Chapitre 3, qui visent à facilité la constitution de corpus thématiques ancrés sur le plan spatio-temporel ou encore les séries temporelle temporelles de données textuelles, et aussi la nécessité sur des méthodes d'extraction de terminologies comme l'étude proposée dans le Chapitre 4 qui vise à faciliter l'extraction des informations les plus pertinentes à partir d'un corpus. Cette approche est généralisable pour le

suivi de tout évènement territorial, qui dispose d'un ancrage spatial et temporel.

Chapitre 7

Conclusion générale

Intitulé «intégration et mise en relation de données massives et hétérogènes pour une observation intelligente du territoire», le thème de notre sujet de recherche fait l'objet d'une étude à plusieurs niveaux que nous avons déclinée dans la Figure 4. Dans chacun des niveaux, nous sommes amenés à proposer au moins une approche méthodologique dans le but de répondre à la problématique posée. Les différentes solutions proposées ont permis de constituer une chaîne de traitement allant de la collecte de données hétérogènes jusqu'à leur mise en relation. Cette chaîne de traitement est associée à des cas d'études sur des données du territoire de la Métropole de Montpellier Méditerranée (3M).

Nous résumons dans ce chapitre, l'ensemble des travaux effectués lors de nos études. Il s'agit essentiellement des contributions sur le plan scientifique ou méthodologique et les perspectives que nous proposons pour chaque chapitre.

Ce chapitre est subdivisé en deux sections. Nous abordons tout d'abord les contributions dans la Section 7.1, et les perspectives dans la Section 7.2.

7.1 Contributions

7.1.1 Protocole de collecte

Cette première contribution répond à la problématique sur la collecte des données hétérogènes en lien avec un territoire que nous assimilons à une ville intelligente, introduite dans la Section 2.2.1 du Chapitre 2 . Comme mentionné dans la même section, la collecte des données a consisté à identifier et à récupérer les données du territoire (ex : 3M) pouvant contribuer à la mise en place de l'étude. D'une part, nous avons les données mises à disposition par 3M, et d'autre part, les données complémentaires à récupérer à partir d'autres sources comme le Web. Partant de ces deux sources de données, notre contribution est double.

La première repose sur l'inventaire des données dont dispose 3M. En effet, nous avons recensé les données de différentes thématiques, telles que l'agriculture, l'urbanisation, etc., dans le but d'avoir une idée globale sur l'état des données de 3M. Dans cet inventaire, nous avons fait ressortir des éléments tels que les caractéristiques des données, leur disponibilité, leur accessibilité, de même que leur granularité spatiale et temporelle. L'inventaire de ces données est répertorié dans le Tableau A.1.

Quant à la seconde, elle repose sur la proposition du protocole collecte proprement dit, dédié à récupérer des données textuelles, de sources diverses à partir du Web. Proposé dans le Chapitre 3 et associé à la publication [Kafando et al., 2021a], il constitue notre contribution méthodologique, et vise à offrir une plate-forme permettant de constituer des corpus thématiques relatifs à une localité, avec un ancrage spatio-temporel. Ce protocole générique permet de constituer des vocabulaires de concepts thématiques (ensemble de termes relatifs à une thématique) et de constituer des corpus thématiques relatifs à une localité. À partir d'un vocabulaire thématique, ce protocole permet à un utilisateur de constituer un corpus dédié à cette thématique avec le moins de bruit possible et fortement ancré sur le plan spatio-temporel. La solution proposée a été évaluée sur trois thématiques de la métropole de Montpellier qui sont l'agriculture, l'urbanisation et l'hydrologie.

Cet outil est important pour le suivi des territoires ou des villes intelligentes, dans la mesure où il permet de constituer des données spécifiques à un domaine, des données sur lesquelles plusieurs analyses peuvent se reposer pour faire ressortir des connaissances sur le territoire. Nous avons aussi mis en place une application Web (voir Section 3.4.4, Chapitre 3), dans le but de permettre à plusieurs types de profils (agent territorial, Data Scientist, etc), de pouvoir accéder à la base de données contenant les corpus thématiques obtenus grâce au protocole pour effectuer des requêtes et récupérer des données pour des analyses de leur choix. Les codes sources seront mis à disposition très prochainement dans le dépôt GitHub : https ://github.com/aidmoit/.

7.1.2 ITEXT-BIO

Toujours dans la dynamique de collecte de données et de suivi des territoires, nous proposons ITEXT-BIO dans le but de pouvoir suivre l'émergence d'une thématique donnée dans le temps. Cette émergence se distingue par l'apparition d'une nouvelle terminologie décrivant le plus souvent un évènement qui impacte la thématique. ITEXT-BIO se repose sur un ensemble de stratégies permettant d'extraire et d'analyser des terminologies à partir d'un corpus spécifique. Ces stratégies permettent de faire ressortir les termes les plus pertinents sémantiquement parlant, en combinant les différentes parties ou sections des documents. Cette approche a été évaluée sur des articles scientifiques du domaine biomédical.

Cette approche s'inscrit d'une part dans le long terme avec la Métropole de Montpellier Méditerranée (3M) pour le suivi des thématiques, et d'autre part sur l'extraction des terminologies pouvant servir de termes experts pour la collecte de données textuelles que nous avons décrite dans la Section 3.3.1.1 du Chapitre 3. Elle constitue notre deuxième contribution méthodologique et est traitée dans le Chapitre 4. La publication scientifique associée correspond à [Kafando et al., 2021b].

7.1.3 Lac de données spatiales

Cette contribution répond à la question de la difficulté liée au stockage des données hétérogènes avec de fortes composantes spatio-temporelles abordée dans la Section 2.1.2, Chapitre 2. Afin de répondre à la question, nous avons procédé à la définition et à l'implémentation d'un lac de données spatiales qui repose sur un système de gestion de méta-données pour le stockage de données massives et fortement hétérogènes, décrit dans le Chapitre 5. Cette solution nous permet d'intégrer et de gérer efficacement les données malgré leur hétérogénéité. Cette contribution scientifique est associée aux publications [Kafando et al., 2020b], et [Kafando et al., 2020a].

Les lac de données sont classiquement dédiés aux experts du traitement de la donnée. Pour permettre aux utilisateurs finaux de bénéficier des apports de ce système, nous avons mis en place une interface de requêtes utilisateur comme nous l'avons présenté avec GeoNetwork dans la Section 5.4.3 du Chapitre 5. Les codes sources sont ouverts et accessibles pour toute réutilisation dans le dépôt GitHub indiqué précédemment.

7.1.4 Intégration et mise en relation de données hétérogènes

Une fois les données collectées, stockées et indexées, la principale question concerne les méthodes et techniques de fouilles de données à appliquer pour l'extraction de nouvelles connaissances.

Afin de répondre à la problématique de mise en relation de données hétérogènes,

nous avons proposé, dans le Chapitre 6, une approche visant à enrichir des informations extraites à partir de séries temporelles d'images satellite par celles extraites à partir de séries temporelles de données textuelles. Ces données textuelles sont obtenues grâce au protocole de collecte proposé dans le Chapitre 3. Nous avons montré la capacité de cette approche quant à la mise en relation des données textuelles avec des images satellite pour le suivi de l'évolution d'un évènement spatio-temporel dans le territoire de Montpellier : la construction de la gare Montpellier Sud de France. Cette approche fera l'objet d'une publication afin de valoriser scientifiquement la méthode proposée. Dans [Decoupes et al., 2021], nous proposons aussi H-TFIDF qui permet de faire un suivi spatio-temporel sur les données de type réseaux sociaux. Cette méthode peut-être également utilisée dans le cadre de suivi d'évènements sur un territoire.

7.2 Perspectives

7.2.1 Protocole de collecte

Nos perspectives vis-à-vis du protocole de collecte de données textuelles sont triples. Comme nous l'avons mentionné dans la conclusion du Chapitre 3, le besoin d'avis experts pour la constitution de termes graines peut s'avérer être un grand handicap lorsque ces personnes ressources ne sont pas disponibles. Notons aussi que les propositions peuvent varier considérablement d'un expert à l'autre. Notre première perspective serait d'utiliser l'approche ITEXT-BIO du Chapitre 4 ou une autre approche à définir pour automatiser cette tâche.

La deuxième perspective vise à améliorer les fonctionnalités du protocole, en automatisant la catégorisation basée sur le genre des documents (académique, administratif, société , publicité, et hors contexte), comme indiqué dans le Tableau 3.1 et de proposer des méthodes d'extraction et de géocodages pour les empreintes spatiales et temporelles présentes dans les documents. Cette partie est importante dans la mesure où elle permet d'avoir une vision plus approfondie des corpus sans avoir à impliquer des professionnels du domaine concerné.

Quant à la dernière perspective, elle vise à estimer, de façon automatique, l'empreinte spatiale et temporelle de chaque document qui sera collecté. En effet, avec le protocole de collecte actuelle, pour tout document collecté, nous récupérons les entités nommées spatiales et temporelles, (constituant une partie des méta-données) qui y sont présentes, mais cela ne nous permet pas d'estimer directement la portée temporelle et/ou spatiale du document, car plusieurs entités spatiales peuvent se retrouver dans un même document. Cette approche permettra de réduire les ambiguïtés surtout sur les dates des documents, car nous avons pu remarquer lors de la phase de mise en relation, qu'un document relatif à un évènement peut être publié avant ou après que cet évènement ait eu lieu au sein du

territoire.

7.2.2 ITEXT-BIO

Nos perspectives se concentrent sur l'extraction et l'analyse des termes. Premièrement, il s'agira d'effectuer les analyses en prenant en compte différentes sections d'articles et en appliquant l'approche méthodologique à différents types de corpus dérivés des réseaux sociaux tels que Twitter.

Deuxièmement, effectuer les analyses en considérant des combinaisons d'outils autres que BioTex, et troisièmement, en introduisant des techniques de word embedding comme BERT [Devlin et al., 2018] ou d'autres types de transformers [Braşoveanu and Andonie, 2020] pour capturer les aspects sémantiques des termes extraits afin de réduire l'ambiguïté du contexte. Cette dernière approche permettra par exemple de clusteriser des termes sémantiquement proches afin de les regrouper dans des concepts alignables avec des ontologies du domaine.

7.2.3 Intégration et mise en relation de données hétérogènes

Nous proposons trois perspectives sur l'approche proposée pour la mise en relation des données hétérogènes. Premièrement, nous proposons d'évaluer l'approche de mise en relation sur une autre thématique (ex : Nature en ville).

Deuxièmement, nous proposons d'appliquer des méthodes basées sur l'apprentissage automatique (ex : clustering spatio-temporel), afin de pouvoir identifier des comportements similaires de l'objet étudié. En d'autre termes, cela permettra d'associer une vue globale sur l'évolution de l'entité spatiale étudiée. Dans cette même logique, nous souhaitons appliquer des méthodes avancées (word embedding) de traitement automatiquement de langage naturel. Cela permettra de créer des clusters sémantiquement proches des ceux identifiés précédemment avec les graphes d'évolution. En plus, il serait envisageable d'appliquer des méthodes d'extraction comme ITEXT-BIO pour obtenir les termes ou les expressions les plus significatifs à une période donnée. Une autre approche, serait d'appliquer différentes méthodes d'analyse en fonction du type de document pour l'extraction des informations. En effet, l'extraction d'informations dans un article scientifique peut différer de celle contenu dans un article de presse ou d'un forum, car ils ne sont pas rédigés de la même façon sur le plan structurel et aussi en terme de précision et de clarté dans le contenu.

La troisième et dernière perspective repose sur le croisement des analyses issues des résultats de l'approche avec des données de vérités terrain, pouvant être obtenues auprès de ceux qui ont suivi le déroulement du projet.

7.2.4 Visualisation

Comme nous l'avons déjà abordé dans la Section 2.2.4 du Chapitre 2, cette partie n'est pas traitée en profondeur dans cette étude. Le prototype que nous avons proposé dans la Section 3.4.4, Chapitre 3, constitue un début pour la visualisation, mais ne prend en compte que les données textuelles pour l'instant. Nos perspectives consistent : 1) à proposer des représentations visuelles accessibles à tous, sur les analyses obtenues lors de la mise en relation des données hétérogènes dans le Chapitre 6, et 2) à proposer des stratégies de restitution ou de visualisation d'analyses sur les lacs de données, afin de faciliter l'interprétation et l'accès des connaissances issues des données qui y sont stockées.

Bibliographie

Kashif Ahmad, Konstantin Pogorelov, Michael Riegler, Nicola Conci, and Pål Halvorsen. Social media and satellites. *Multimedia Tools and Applications*, 78(3) :2837–2875, 2019.

Akiko Aizawa. An information theoretic perspective of tf-idf measures. *Information Processing & Management*, 39 :45–65, 01 2003. doi : 10.1016/S0306-4573(02)00021-3.

Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, and Jameela Al-Jaroodi. Applications of big data to smart cities. *Journal of Internet Services and Applications*, 6(1), 2015. ISSN 1867-4828. doi : 10.1186/s13174-015-0041-5. URL `https://dx.doi.org/10.1186/s13174-015-0041-5`.

Vito Albino, Umberto Berardi, and Rosa Maria Dangelico. Smart cities : Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology*, 22 :3–21, 2015. ISSN 1063-0732, 1466-1853. doi : 10.1080/10630732.2014.942092.

Elena Arsevska., Mathieu Roche., Pascal Hendrikx., David Chavernac., Sylvain Falala., Renaud Lancelot., and Barbara Dufour. Identification of associations between clinical signs and hosts to monitor the web for detection of animal disease outbreaks. *International Journal of Agricultural and Environmental Information Systems*, 2016. doi : 10.4018/IJAEIS.2016070101.

Elena Arsevska, Sarah Valentin, Julien Rabatel, Jocelyn de Goër de Hervé, Sylvain Falala, Renaud Lancelot, and Mathieu Roche. Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. *PLOS ONE*, 13(8) :e0199960, August 2018. ISSN 1932-6203. doi : 10.1371/journal.pone.0199960. URL `https://dx.plos.org/10.1371/journal.pone.0199960`.

Mehdi Azarafza, Mohammad-Reza Feizi-Derakhshi, and Moosa Bagheri Shendi. Textrank-based microblogs keyword extraction method for persian language. *Conference : 3rd International Congress on Science and Engineering At : Hamburg - Germany*, 2020.

Marco Baroni and Silvia Bernardini. Bootcat : Bootstrapping corpora and terms from the web. In *Language Resources and Evaluation Conference (LREC)*, pages 1313–1316. Citeseer, 2004.

M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali. Smart cities of the future. *The European Physical*

Journal Special Topics, 214(1) :481–518, 2012. ISSN 1951-6355. doi : 10.1140/epjst/e2012-01703-3. URL https://dx.doi.org/10.1140/epjst/e2012-01703-3.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

Benjamin Bischke, Prakriti Bhardwaj, Aman Gautam, Patrick Helber, Damian Borth, and Andreas Dengel. Detection of flooding events in social multimedia and satellite imagery using deep neural networks. In *MediaEval*, 2017.

Emmanuel Blondel. geonapi : R interface to geonetwork api, August 2018. URL https://doi.org/10.5281/zenodo.1345102.

Emmanuel Blondel. geometa : Tools for Reading and WritingISO/OGC Geographic Metadata in R, October 2019. URL https://doi.org/10.5281/zenodo.3524348.

Olivier Bodenreider. The unified medical language system (umls) : integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1) :D267–D270, 2004.

David B Bracewell, Fuji Ren, and Shingo Kuriowa. Multilingual single document keyword extraction for information retrieval. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 517–522. IEEE, 2005.

Adrian M. P. Braşoveanu and Răzvan Andonie. Visualizing transformers for nlp : A brief survey. In *2020 24th International Conference Information Visualisation (IV)*, pages 270–279, 2020. doi : 10.1109/IV51561.2020.00051.

Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, page 152–155, USA, 1992. Association for Computational Linguistics. doi : 10.3115/974499.974526. URL https://doi.org/10.3115/974499.974526.

Rudi Bruchez. *Les bases de données NoSQL et le BigData : Comprendre et mettre en oeuvre*. Editions Eyrolles, April 2015. ISBN 978-2-212-30793-1.

L Campillos Llanos, A Moreno Sandoval, and JM Guirao. An automatic term extractor for biomedical terms in spanish. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM 2013)*, 2013.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509 :257–289, 2020. ISSN 0020-0255. doi : 10.1016/j.ins.2019.09.013.

Guido Cervone, Emily Schnebele, Nigel Waters, Martina Moccaldi, and Rosa Sicignano. Using social media and satellite data for damage assessment in urban areas during emergencies. In *Seeing cities through big data*, pages 443–457. Springer, 2017.

D. Chen, Y. Chen, B. N. Brownlow, P. P. Kanjamala, C. A. G. Arredondo, B. L. Radspinner, and M. A. Raveling. Real-time or near real-time persisting daily healthcare data into hdfs and elasticsearch index inside a big data platform. *IEEE Transactions on Industrial Informatics*, 13(2) :595–606, April 2017. ISSN 1941-0050. doi : 10.1109/TII.2016.2645606.

Hsinchun Chen, Roger HL Chiang, and Veda C Storey. Business intelligence and analytics : From big data to big impact. *Management Information Systems Quarterly (MIS quarterly)*, pages 1165–1188, 2012.

Dorin Comaniciu and Peter Meer. Mean shift : A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5) :603–619, 2002.

Merley Conrado, Thiago Pardo, and Solange Oliveira Rezende. A machine learning approach to automatic term extraction using a rich feature set. In *Proceedings of the 2013 NAACL HLT Student Research Workshop*, pages 16–23, 2013.

Damien Cram and Béatrice Daille. Terminology extraction with term variant detection. In *Proceedings of ACL-2016 System Demonstrations*, pages 13–18, 2016.

Hercules Dalianis, Martin Hassel, Aron Henriksson, and Maria Skeppstedt. Stockholm epr corpus : A clinical database used to improve health care. In *Swedish Language Technology Conference*, pages 17–18, 2012.

Rémy Decoupes, Rodrique Kafando, Mathieu Roche, and Maguelonne Teisseire. H-tfidf : What makes areas specific over time in the massive flow of tweets related to the covid pandemic ? *Association of Geographic Information Laboratories in Europe (AGILE) : GIScience Series*, 2 :1–8, 2021.

Barry Devlin and Lynne Doran Cote. *Data warehouse : from architecture to implementation*. Addison-Wesley Longman Publishing Co., Inc., 1996.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2018.

James Dixon. Pentaho, Hadoop, and Data Lakes, October 2010. URL `https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/`.

Swagata Duari and Vasudha Bhatnagar. Complex network based supervised keyword extractor. *Expert Systems with Applications*, 140 :112876, 2020.

Jody Foo. Term extraction using machine learning. *Linkoping University, LINKOPING*, 2009.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms :. the c-value/nc-value method. *International journal on digital libraries*, 3 (2) :115–130, 2000.

C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. HealthMap : Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association*, 15(2) : 150–157, March 2008. ISSN 1067-5027, 1527-974X. doi : 10.1197/jamia.M2544. URL `https://academic.oup.com/jamia/article-lookup/doi/10.1197/jamia.M2544`.

John Gantz and David Reinsel. The digital universe in 2020 : Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView : IDC Analyze the future*, 2007 (2012) :1–16, 2012.

Rayid Ghani, Rosie Jones, and Dunja Mladenić. Mining the web to create minority language corpora. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 279–286, 2001.

Paolo Lo Giudice, Lorenzo Musarella, Giuseppe Sofo, and Domenico Ursino. An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. *Information Sciences*, 478 : 606–626, 2019.

Fabio Guttler, Dino Ienco, Jordi Nin, Maguelonne Teisseire, and Pascal Poncelet. A graph-based approach to detect spatiotemporal dynamics in satellite image time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130 :92–107, 2017.

Maryam Habibi and Andrei Popescu-Belis. Keyword extraction and clustering for document recommendation in conversations. *IEEE/ACM Transactions on audio, speech, and language processing*, 23(4) :746–759, 2015.

Rihan Hai, Sandra Geisler, and Christoph Quix. Constance : An intelligent data lake system. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2097–2100. ACM, 2016.

Jing Han, Haihong E, Guan Le, and Jian Du. Survey on nosql database. In *2011 6th International Conference on Pervasive Computing and Applications*, pages 363–366, 2011. doi : 10.1109/ICPCA.2011.6106531.

Sung Yang Ho, Sophia Tan, Chun Chau Sze, Limsoon Wong, and Wilson Wen Bin Goh. What can venn diagrams teach us about doing data science better ? *International Journal of Data Science and Analytics*, 11(1) :1–10, 2021.

Matthew Honnibal and Ines Montani. spacy 2 : Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7, 2017.

ISO/TC 211. Norme iso 19115-1 :2014. geographic information - metadata - part 1 : Fundamentals. technical report, international organization for standardization, 2019, 2019.

Christian Jacquemin. Fastr : A unification-based front-end to automatic indexing. In *Intelligent Multimedia Information Retrieval Systems and Management - Volume 1*, RIAO '94, page 34–47, Paris, FRA, 1994. Le centre de hautes études internationales d'informatique documentaire.

Luning Ji, Mantai Sum, Qin Lu, Wenjie Li, and Yirong Chen. Chinese terminology extraction using window-based contextual information. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 62–74. Springer, 2007.

Tomcy John and Pankaj Misra. *Data Lake for Enterprises : Lambda Architecture for Building Enterprise Data Systems*. Packt Publishing, 2017. ISBN 1787281345, 9781787281349.

Junegak Joung and Kwangsoo Kim. Monitoring emerging technologies for technology planning using technical keyword based analysis from patent data. *Technological Forecasting and Social Change*, 114 :281–292, 2017.

Rodrique Kafando, Rémy Decoupes, Lucile Sautot, and Maguelonne Teisseire. Modélisation de la dynamique des territoires : méta-données et lacs de données dédiés à l'information spatiale. In *INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID2020)*, 2020a.

Rodrique Kafando, Rémy Decoupes, Lucile Sautot, and Maguelonne Teisseire. Spatial Data Lake for Smart Cities : From Design to Implementation. *AGILE : GIScience Series*, 1 :1–15, 2020b.

Rodrique Kafando, Rémy Decoupes, Maguelonne Teisseire, Lucile Sautot, and Christiane Weber. Constitution de corpus thématique : Pour un meilleur suivi du territoire de la métropole de montpellier méditerranée. In *SAGEO'21 16ème Conférence Internationale de la Géomatique, de l'Analyse Spatiale et des Sciences de l'Information Géographique.*, 2021a.

Rodrique Kafando, Rémy Decoupes, Sarah Valentin, Lucile Sautot, Maguelonne Teisseire, and Mathieu Roche. Itext-bio : Intelligent term extraction for biomedical analysis. *Health Information Science and Systems*, 9(1) :1–23, 2021b.

Kyo Kageura and Bin Umino. Methods of automatic term recognition : A review. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3(2) :259–289, 1996.

Eric Kergosien, Hugo Alatrista-Salas, Mauro Gaio, Fábio N Güttler, Mathieu Roche, and Maguelonne Teisseire. When textual information becomes spatial information compatible with satellite images. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 1, pages 301–306. IEEE, 2015.

Lynda Khiali, Dino Ienco, and Maguelonne Teisseire. Object-oriented satellite image time series analysis using a graph-based representation. *Ecological Informatics*, 43 : 52–64, 2018. ISSN 1574-9541. doi : https://doi.org/10.1016/j.ecoinf.2017.11.003. URL `https://www.sciencedirect.com/science/article/pii/S1574954117301851`.

Adam Kilgarriff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3) :333–347, 2003.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and PVS Avinesh. A corpus factory for many languages. In *Language Resources and Evaluation Conference (LREC)*. Citeseer, 2010.

Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit : The Complete Guide to Dimensional Modeling*. John Wiley & Sons, August 2011. ISBN 978-1-118-08214-0.

Ralph Kimball and Margy Ross. *The data warehouse toolkit : The definitive guide to dimensional modeling*. John Wiley & Sons, 2013.

Rob Kitchin. The real-time city ? big data and smart urbanism. *GeoJournal*, 79(1) :1–14, 2014.

Ian Lamb and Catherine Larson. Shining a light on scientific data : Building a data catalog to foster data sharing and reuse. *Code4Lib Journal*, (32), 2016.

David D Lewis, Yiming Yang, Tony Russell-Rose, and Fan Li. Rcv1 : A new benchmark collection for text categorization research. *Journal of machine learning research*, 5 (Apr) :361–397, 2004.

Shian-Hua Lin and Jan-Ming Ho. Discovering informative content blocks from web documents. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 588–593, 2002.

Xianming Liu, Hongxun Yao, Rongrong Ji, Pengfei Xu, Xiaoshuai Sun, and Qi Tian. Learning heterogeneous data for hierarchical web video classification. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 433–442, 2011.

Marilex Rea Llave. Data lakes in business intelligence : reporting from the trenches. *Procedia computer science*, 138 :516–524, 2018.

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biotex : A system for biomedical terminology extraction, ranking, and validation. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track - Volume 1272*, ISWC-PD'14, page 157–160, Aachen, DEU, 2014a. CEUR-WS.org.

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biomedical terminology extraction : A new combination of statistical and web mining approaches. In *JADT : Journées d'Analyse statistique des Données Textuelles*, pages 421–432, 2014b.

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Yet another ranking function for automatic multiword term extraction. In *International Conference on Natural Language Processing*, pages 52–64. Springer, 2014c.

Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biomedical term extraction : overview and a new methodology. *Information Retrieval Journal*, 19(1-2) :59–99, 2016.

Cedrine Madera and Anne Laurent. The next information architecture evolution : the data lake wave. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, pages 174–180. ACM, 2016.

Gjorgji Madjarov, Vedrana Vidulin, Ivica Dimitrovski, and Dragi Kocev. Web genre classification with methods for structured output prediction. *Information Sciences*, 503 :551 – 573, 2019.

Lawrence C. Madoff. ProMED-mail : an early warning system for emerging diseases. *Clinical infectious diseases*, 39(2) :227–232, 2004. URL `https://academic.oup.com/cid/article-abstract/39/2/227/327615`.

David Maier. *The Theory of Relational Databases*. Computer Science Press, 1983. ISBN 0-914894-42-0.

Yutaka Matsuo and Mitsuru Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01) :157–169, 2004.

Diana Maynard, Milena Yankova, Alexandros Kourakis, and Antonis Kokossis. Ontology-based information extraction for market monitoring and technology watch. In *ESWC Workshop" End User Apects of the Semantic Web"), Heraklion, Crete*, 2005.

Andrew McAfee, Erik Brynjolfsson, Thomas H Davenport, DJ Patil, and Dominic Barton. Big data : the management revolution. *Harvard business review*, 90(10) :60–68, 2012.

Hassan Mehmood, Ekaterina Gilman, Marta Cortes, Panos Kostakos, Andrew Byrne, Katerina Valta, Stavros Tekes, and Jukka Riekki. Implementing big data lake for heterogeneous data sources. In *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, pages 37–44. IEEE, 2019.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. Extracting information from textual documents in the electronic health record : a review of recent research. *Yearbook of medical informatics*, 17(01) :128–144, 2008.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc., 2013.

George A Miller. *WordNet : An electronic lexical database*. MIT press, 1998.

L. Minati, F. Ghielmetti, V. Ciobanu, L. D'Incerti, C. Maccagnano, A. Bizzi, and M. G. Bruzzone. Bio-image warehouse system : Concept and implementation of a diagnosis-based data warehouse for advanced imaging modalities in neuroradiology. *Journal of Digital Imaging*, 20, 2006.

A Moine. The territory as a complex system : an operational concept for land planning and geography (le territoire comme un systeme complexe : un concept opératoire pour l'aménagement et la géographie). *Esp. Géogr*, 2(35) :115, 2006.

Eric Mykhalovskiy and Lorna Weir. The Global Public Health Intelligence Network and early warning outbreak detection : a Canadian contribution to global public health. *Canadian journal of public health = Revue canadienne de sante publique*, 97(1) :42–44, 2006. doi : 10.17269/cjph.97.756.

2018 Montpellier Méditerranée Métropole. Expérimentez la Cité Intelligente, October 2017. URL `https://www.entreprendre-montpellier.com/fr/choisir-montpellier/experimentez-la-cite-intelligente`.

Taewoo Nam and Theresa A. Pardo. Conceptualizing smart city with dimensions of technology, people, and institutions. In *Proceedings of the 12th Annual International*

Digital Government Research Conference : Digital Government Innovation in Challenging Times, dg.o '11, page 282–291, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450307628. doi : 10.1145/2037556.2037602. URL https://doi.org/10.1145/2037556.2037602.

Fatemeh Nargesian, Erkang Zhu, Ken Q Pu, Renée J Miller, and Patricia C Arocena. Data lake management : Challenges and opportunities. 12(12) :4, 2019. doi : 10.14778/3352063.3352116.

Wafa Neifar, Thierry Hamon, Pierre Zweigenbaum, Mariem Ellouze Khemakhem, and Lamia Hadrich Belguith. Adaptation of a term extractor to arabic specialised texts : First experiments and limits. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 242–253. Springer, 2016.

Paolo Neirotti, Alberto De Marco, Anna Corinna Cagliano, Giulio Mangano, and Francesco Scorrano. Current trends in smart city initiatives : Some stylised facts. *Cities*, 38 :25–36, 2014.

Antoni Oliver and Mercè Vàzquez. Tbxtools : a free, fast and flexible tool for automatic terminology extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 473–479, 2015.

Lamia Oukid, Omar Boussaid, Nadija Benblidia, and Fadila Bentayeb. Tlabel : A new olap aggregation operator in text cubes. *International Journal of Data Warehousing and Mining*, 12(4) :54–74, 2016.

Vasile Pais and Radu Ion. Termeval 2020 : Racai's automatic term extraction system. In *Computational Terminology (COMPUTERM)*, 2020.

C. Paquet, D. Coulombier, R. Kaiser, and M. Ciotti. Epidemic intelligence : a new framework for strengthening disease surveillance in Europe. *Eurosurveillance*, 11(12) : 5–6, December 2006. ISSN 1560-7917. doi : 10.2807/esm.11.12.00665-en. URL https://www.eurosurveillance.org/content/\10.2807/esm.11.12.00665-en.

Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto. Terminology extraction : an analysis of linguistic and statistical approaches. In *Knowledge mining*, pages 255–279. Springer, 2005.

Cassandra Phipps and Karen C Davis. Automating data warehouse conceptual schema design and evaluation. In *Design and Management of Data Warehouses (DMDW)*, volume 2, pages 23–32. Citeseer, 2002.

Christoph Quix, Rihan Hai, and Ivan Vatov. Metadata extraction and management in data lakes with gemms. *Complex Systems Informatics and Modeling Quarterly*, (9) : 67–83, 2016.

Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142. New Jersey, USA, 2003.

Franck Ravat and Yan Zhao. Data Lakes : Trends and Perspectives. In *International Conference on Database and Expert Systems Applications*, pages 304–313. Springer, 2019.

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. Termeval 2020 : Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset. In *6th International Workshop on Computational Terminology (COMPUTERM 2020)*, pages 85–94. European Language Resources Association (ELRA), 2020.

Mathieu Roche, Maguelonne Teisseire, Bruno Crémilleux, Pierre Gancarski, Christian Sallaberry, Hugo Alatrista-Salas, Nicolas Béchet, Delphine Bernhard, Sandra Bringay, Thierry Charnois, Mauro Gaio, Fábio N. Guttler, Dino Ienco, Eric Kergosien, Pierre Maurel, Pascal Poncelet, Arnaud Sallaberry, and Christiane Weber. Animitex : Analyse d'Images fondée sur des Informations Textuelles. *Revue des Sciences et Technologies de l'Information - Série ISI : Ingénierie des Systèmes d'Information*, 19(3) :163–167, 2014. URL `https://hal-lirmm.ccsd.cnrs.fr/lirmm-01054924`.

Philip Russom. Data lakes : Purposes, practices, patterns, and platforms. *Transforming Data With Intelligence (TDWI) White Paper*, 2017.

Philip Russom et al. Big data analytics. *Transforming Data With Intelligence (TDWI) best practices report, fourth quarter*, 19(4) :1–34, 2011.

Naina Said, Kashif Ahmad, Michael Riegler, Konstantin Pogorelov, Laiq Hassan, Nasir Ahmad, and Nicola Conci. Natural disasters detection in social media and satellite imagery : a survey. *Multimedia Tools and Applications*, 78(22) :31267–31302, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert : smaller, faster, cheaper and lighter. *arXiv preprint arXiv :1910.01108*, 2019.

Marina Santini. *Cross-Testing a Genre Classification Model for the Web*, pages 87–128. Springer Netherlands, Dordrecht, 2011.

Pegdwendé N Sawadogo, Etienne Scholly, Cécile Favre, Eric Ferey, Sabine Loudcher, and Jérôme Darmont. Metadata systems for data lakes : models and features. In *European Conference on Advances in Databases and Information Systems*, pages 440–451. Springer, 2019.

Camille Schaeffer., Roberto Interdonato., and Mathieu Roche. Construction d'un corpus sur la problématique de la sécurité alimentaire guidée par un lexique et des approches de fouilles de textes, in : Toth 2020 - terminologie et ontologie : théories et applications. roche christophe (ed.). chambéry : Presses universitaires savoie mont blanc. 2020.

Himat Shah, Muhammad US Khan, and Pasi Fränti. H-rank : a keywords extraction method from web pages using pos tags. In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, volume 1, pages 264–269. IEEE, 2019.

Serge Sharoff. Creating general-purpose corpora using automated search engine queries. *WaCky*, pages 63–98, 2006.

Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, et al. The hadoop distributed file system. In *Massive Storage Systems and Technology (MSST)*, volume 10, pages 1–10, 2010.

Cristina Simone, Sergio Barile, and Mario Calabrese. Managing territory and its complexity : a decision-making model based on the viable system approach (vsa). *Land use policy*, 72 :493–502, 2018.

Compton J Tucker, Daniel A Slayback, Jorge E Pinzon, Sietse O Los, Ranga B Myneni, and Malinda G Taylor. Higher northern latitude normalized difference vegetation index and growing season trends from 1982 to 1999. *International journal of biometeorology*, 45(4) :184–190, 2001.

Sarah Valentin, Elena Arsevska, Sylvain Falala, Jocelyn de Goër, Renaud Lancelot, Alizé Mercier, Julien Rabatel, and Mathieu Roche. PADI-web : A multilingual event-based surveillance system for monitoring animal infectious diseases. *Computers and Electronics in Agriculture*, 169 :105163, February 2020. ISSN 0168-1699. doi : 10.1016/j.compag.2019.105163. URL `http://www.sciencedirect.com/science/article/pii/\S0168169919310646`.

Vedrana Vidulin, Mitja Lustrek, and Matjaz Gams. Multi-label approaches to web genre identification. *Journal for Language Technology and Computational Linguistics (JLCL)*, 24 :97–114, 01 2009.

Chenguang Wang, Yangqiu Song, Ahmed El-Kishky, Dan Roth, Ming Zhang, and Jiawei Han. Incorporating world knowledge to document clustering via heterogeneous infor-

mation networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1215–1224, 2015.

Lidong Wang. Heterogeneous data and big data analytics. *Automatic Control and Information Sciences*, 3(1) :8–15, 2017.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Michael Kinney, Ziyang Liu, William. Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. Cord-19 : The covid-19 open research dataset. *ArXiv*, abs/2004.10706, 2020a.

Rui Wang, Wei Liu, and Chris McDonald. Featureless domain-specific term extraction with minimal labelled data. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 103–112, 2016.

Xingyu Wang, Lida Zhang, and Diego Klabjan. Keyword-based topic modeling and keyword selection. *arXiv preprint arXiv :2001.07866*, 2020b.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1) :1–40, 2016.

John S Whissell and Charles LA Clarke. Improving document clustering using okapi bm25 feature weighting. *Information retrieval*, 14(5) :466–487, 2011.

Stephen T Wu, Hongfang Liu, Dingcheng Li, Cui Tao, Mark A Musen, Christopher G Chute, and Nigam H Shah. Unified Medical Language System term occurrences in clinical notes : a large-scale corpus analysis. *Journal of the American Medical Informatics Association*, 19(e1) :e149–e156, 04 2012. ISSN 1067-5027. doi : 10.1136/amiajnl-2011-000744. URL `https://doi.org/10.1136/amiajnl-2011-000744`.

Yang Yang, Yi Yang, and Heng Tao Shen. Effective transfer tagging from image to video. *ACM Trans. Multimedia Comput. Commun. Appl.*, 9(2), May 2013. ISSN 1551-6857. doi : 10.1145/2457450.2457456. URL `https://doi.org/10.1145/2457450.2457456`.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet : Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Xian-ming Yao, Jian-hou GAN, and XU Jian. Concept extraction based on hybrid approach combined with semantic analysis. *DEStech Transactions on Engineering and Technology Research*, 2017.

Xiaoxin Yin, Jiawei Han, and Philip S Yu. Linkclus : efficient clustering via heterogeneous semantic links. In *Proceedings of the 32nd international conference on Very large data bases*, pages 427–438. Citeseer, 2006.

Jing Zhao, Panagiotis Papapetrou, Lars Asker, and Henrik Boström. Learning from heterogeneous temporal data in electronic health records. *Journal of biomedical informatics*, 65 :105–119, 2017.

Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Jialin Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. Heterogeneous transfer learning for image classification. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

Annexe A

Annexes

DonnéesCaractéristiques	Description	Anal. Critiques	Intérêt	Granularité (temporelle/Spatiale)	État
		Urbanisation			
Plan local d'urbanisme PLU 3M	le PLU contient les données de planification sur l'aménagement du territoire. *accès OpenData 3M*	Utile : Oui Format : shapefile et kml, projection Lambert 93	Suivi des aménagements et l'évolution du territoire dans son ensemble	Temporalité : *numérisation de 2006* Spatialité : *communale*	Disponible : Oui Fournisseur : OpenData 3M
Plan local d'urbanisme PLU Montpellier	contient les données du projet global d'aménagement et d'urbanisme de la ville de Montpellier. *accès OpenData 3M*	Utile :Oui Format : shapefile,GDB,DWG	Suivi des aménagements et l'évolution du territoire dans son ensemble	Temporalité : *données SIG de 2006* Spatialité : *Territorial*	Disponible : Oui Fournisseur : OpenData 3M
SCot Montpellier	le Schéma de Cohérence Territoriale est contient les données de planification , les grandes orientations d'aménagement du territoire. Il fixe les limites entre, d'une part, les espaces urbains ou voués à l'urbanisation et, d'autre part, les espaces naturels et agricoles. *Inaccessible, 3M*	Utile : Oui Format : shapefile	Suivi des aménagements et l'évolution du territoire dans son ensemble	Temporalité : *données de 2006* Spatialité : *communale*	Disponible : Oui Fournisseur : 3M
Occupation du sol,	ces données renseignent l'occupation du sol sur le territoire de 3M. Elles permettent de suivre l'évolution de l'occupation du sol à différentes à de multiples échelles d'observation (1/5000 maximum). *accès OpenData 3M*	Utile : Oui Format : shapefile	Suivi de l'évolution de l'occupation du sol au sein du territoire	Temporalité : *1994, 2004, 2008, 2010, 2012, 2015 , 2016,* Spatialité : *territoriale*	Disponible : Oui Fournisseur : OpenData 3M
Distribution fine de la population	ces données renseignent la distribution de la population infra-communale 2013 à l'IRIS (INSEE) et est distribuée au prorata de la surface habitable des parcelles. *accès OpenData 3M*	Utile : Oui Format : shapefile	Suivi de l'évolution de la population au sein du territoire	Temporalité : *données de 2013,* Spatialité : *communale*	Disponible : Oui Fournisseur : OpenData 3M
Base SIRENE géolocalisée	ces données renseignent les numéros Sirene et les informations des entreprises et de leurs établissements. *accès data.gouv*	Utile : Oui Format : fichier .csv	Suivi de la mobilité des activités entrepreneuriales sur le territoire	Temporalité : *données à jour*, Spatialité : *territoriale*	Disponible : Oui Fournisseur : data.gouv 3M
Enquête Ménages Déplacements	ces données renseignent toutes les ressources liées à l'Enquête Ménages Déplacements, réalisée en 2014 sur l'Est du Département de l'Hérault. *accès OpenData 3M*	Utile : Oui Format : fichier .csv	Suivi de la mobilité des ménages sur le territoire	Temporalité : *données de 2014*, Spatialité : *territoriale*	Disponible : Oui Fournisseur : OpenData 3M
Projection de population	ces données renseignent la projection de population (2025 et 2030), calculée à partir des projets connus et des projections de l'INSEE pour le territoire. *Inaccessible, 3M*	Utile : Oui Format : inconnu	Suivi de l'évolution de la population sur le territoire	Temporalité : ; Spatialité : *territoriale*	Disponible : en cours Fournisseur : 3M
Zonage du modèle	il s'agit de diverses données agglomérées à la zone, à savoir la population, les emplois publics, les effectifs scolaires, les emplois privés et les projections de population et d'emploi. *Inaccessible,3M*	Utile : Oui Format : inconnu	Suivi de la mobilité de la population au caractère multi-échelle (emploi, scolarité,démographie) sur le territoire	Temporalité : ; Spatialité : *territoriale*	Disponible : en cours Fournisseur : 3M
Les effectifs scolaires (primaire et secondaire)	ces données renseignent les effectifs scolaires géolocalisés à partir des données de l'académie du ministère de l'éducation. accès data.gouv	Utile : Oui Format : fichier .CSV, .Json, Excel	mise en relation entre le phénomène de l'urbanisation et celui du taux de scolarisation au niveau scolaire	Temporalité : ;...; Spatialité : *Territoriale*	Disponible : Oui Fournisseur : data.gov
Les effectifs universitaires	ces données renseignent les effectifs universitaires géolocalisés à partir des données de l'académie du ministère de l'éducation. accès data.gouv	Utile : Oui Format : fichier .CSV, .Json, Excel	mise en relation entre le phénomène de l'urbanisation et celui du taux de scolarisation au niveau universitaire	Temporalité : ;...; Spatialité : *Territoriale*	Disponible : Oui Fournisseur : data.gov
Localisation de l'emploi public	ces données renseignent la localisation des emplois publics sur l'aire urbaine de Montpellier, elles ont été obtenues lors d'une enquête téléphonique réalisée en 2016 par les services de la Métropole. *Inaccessible, 3M*	Utile : Oui Format : inconnu	mise en relation entre le phénomène de l'urbanisation et celui de la disponibilité des emplois en fonction des localités	Temporalité : ;...; Spatialité : *Territoriale*	Disponible : en cours Fournisseur : 3M
Localisation de l'emploi privé	ces données renseignent à partir des informations de l'URSSAF les géolocalisation de l'emploi privé. *Inaccessible, 3M*	Utile : Oui Format : inconnu	mise en relation entre le phénomène de l'urbanisation et celui de la disponibilité des emplois en fonction des localités	Temporalité : ;...; Spatialité : *Territoriale*	Disponible : en cours Fournisseur : 3M
Projection de l'emploi	ces données renseignent la projection (2025 et 2030) d'emplois. Ces projections sont uniquement pertinentes dans le cadre des calculs de flux de déplacements du modèle multimodale. *Inaccessible, 3M*	Utile : Oui Format : inconnu	mise en relation entre le phénomène de l'urbanisation et celui de la disponibilité des emplois	Temporalité : *2025 et 2030* Spatialité : *Territoriale*	Disponible : en cours Fournisseur : 3M
		Agriculture			
Politique Agricole Commune	ensemble de données sur les prix, les subventions et les superficies des espaces cultivés pour le développement du secteur agricole *accès data.gouv*	Utile : Oui Format : fichier .CSV	Suivi de l'évolution de la production et des espaces agricoles du territoire	Temporalité : 2005 à 2009 et 2010 à 2012 Spatialité : Communale	Disponible : Oui Fournisseur : data.gouv
Production viticole	données sur la production de vignes. *indisponible*	Utile : non défini Format : inconnu	Suivi de l'évolution de la production et des espaces viticoles du territoire	Temporalité : Spatialité : Territoriale	Disponible : Fournisseur
Agro-écologie	ensemble de données agricoles en relation avec la production. *indisponible, à vérifier avec Laure*	Utile : Oui Format : inconnu	Suivi de l'évolution de la production et des espaces agricoles du territoire	Temporalité : Spatialité :	Disponible : Laure Fournisseur
Annuaires de production agricole	Annuaire regroupant les données sur les productions agricoles sur l'ensemble du territoire. *Indisponible*	Utile : Oui Format : inconnu	Suivi de l'évolution de la production et des espaces agricoles du territoire	Temporalité : *annuelle*, Spatialité : *territoriale*	Disponible : Fournisseur :
		Hydraulique			
PLU Montpellier	le PLU nous permettra d'obtenir les données relatives aux aménagement hydrauliques. *accès OpenData 3M*	Utile : Oui Format : shapefile et kml, projection Lambert 93	Suivi aménagements hydrauliques (corridors, canalisations, etc)	Temporalité : Spatialité : *communale*	Disponible : Oui Fournisseur : OpenData 3M
Relevées pluviométriques	contiennent les informations sur les données de relevés pluviométriques pour chaque saison sur l'étendu du territoire. *indisponible*	Utile : Oui Format : inconnu	Suivi de la pluviométrie, données pouvant être mis en relation avec les catastrophes naturelles liées aux inondations	Temporalité : Spatialité : *territoriale*	Disponible : Fournisseur :
Données sur la gestion intégrée de l'eau	contiennent les informations sur la gestion de l'eau sur l'ensemble du territoire. *inaccessible, voir A.Vestier*	Utile : Oui Format : inconnu	Suivi de la pluviométrie, données pouvant être mis en relation avec les catastrophes naturelles liées aux inondations	Temporalité : Spatialité : *territoriale*	Disponible :*Oui* Fournisseur : *3M (A.Vestier)*
Station MAERA	ces données renseignent le débit d'eau en entrée, en sortie, du bilan de génération et de l'énergie consommée de la station d'épuration MAERA. *Inaccessible,3M*	Utile : Oui Format : inconnu	Suivi de l'évolution des débits en entrées et sorties de la station d'épuration. Données pouvant être reliées avec les données pluviométriques	Temporalité : Spatialité : *station MAERA*	Disponible : Oui Fournisseur : 3M
		Environnement			
Météo temps réel	données météorologiques enregistrées par les (06) stations en temps réel sur le territoire. Il s'agit entre autre les données liées à l'humidité, la température, etc. *Inaccessible,3M*	Utile : non défini Format : inconnu	Etude qualitative sur les effets du changement climatique sur l'environnement	Temporalité : Spatialité : *Territoriale*	Disponible : Oui Fournisseur : 3M
Latara : données de pression, humidité	Données de mesures intérieures du Musée de Latara. *Inaccessible,voir D. Dusseaux*	Utile : non défini Format : inconnu	Suivi de l'évolution des variations de pression et d'humidité à l'intérieur du musé	Temporalité : Spatialité : *Musé de Latara*	Disponible : Oui Fournisseur : 3M (D. Dusseaux)
Capteurs d'humidité et de température	relevées de température et d'humidité permettant d'illustrer une du territoire sous forme d'une carte de chaleur. ***indisponible, horizon 2020***	Utile : non défini Format : inconnu	Suivi de variation de température et de pression à l' échelle territoire	Temporalité : *disponible en 2020* Spatialité : *territoriale*	Disponible : Oui Fournisseur : 3M
		Mobilité			
Suivi de Bus en temps réel	ces données renseignent en temps réel et l'ensemble des courses programmées pour la journée l'offre de transport Bus de la Métropole de Montpellier. Y sont inclus les la liste des arrêts, et les lignes correspondantes. *accès OpenData 3M*	Utile : non défini Format : fichier .CSV	Etude sur le comportement du trafic sur le plan de la mobilité offerte par les Bus	Temporalité : *toutes les 30s (avec historique)* Spatialité : *Territoriale*	Disponible : Oui Fournisseur : 3M
Suivi de Tram en temps réel	ces données renseignent en temps réel et l'ensemble des courses programmées pour la journée l'offre de transport Tram de la Métropole de Montpellier. Y sont inclus les la liste des arrêts, et les lignes correspondantes.	Utile : non défini Format : fichier .CSV	Etude sur le comportement du trafic sur le plan de la mobilité offerte par les Tram	Temporalité *toutes les 30s* Spatialité : *Territoriale*	Disponible : Oui Fournisseur : 3M
Disponibilité des parking	Ces données indiquent en temps réel le nombre de places disponibles dans les parkings de 3M auquel sont associées les informations telles que la date de la dernière mise à jour, nom du parking, statut d'ouverture, nombre de place libre, etc. *accès OpenData 3M*	Utile : non défini Format : fichier xml, xsd	Etude sur la disponibilité et l'influence des parking sur la mobilité dans l'ensemble du territoire	Temporalité :...; Spatialité : *Territoriale*	Disponible : Oui Fournisseur : 3M
Disponibilité et suivi des velomagg en temps réel	ces données renseignent en temps réel la position et le nombre de places disponibles dans les stations Vélomagg de Montpellier Méditerranée Métropole. *accès OpenData 3M*	Utile : non défini Format : fichier xml	Etude du comportement du trafic sur le plan de la mobilité offerte par les velomagg	Temporalité :...; Spatialité : *Territoriale*	Disponible :Oui Fournisseur : 3M
Flux de comptage de véhicules en temps réel	ces données renseignent le flux de véhicules particuliers passant dans une rue. Ces données sont issues de diverses techniques de comptage, à savoir manuelle ou automatique via un système pneumatique (piézoélectricité) ou des capteurs. *accès OpenData 3M*	Utile : non défini Format : fichier .CSV	Etude du comportement du trafic routier à partir de l'affluence des véhicules en fonction des périodes	Temporalité : *hebdomadaire* Spatialité : *Territoriale*	Disponible : Oui Fournisseur : OpenData 3M
Hiérarchie des voies	créées dans le cadre du Plan de Déplacement Urbains de 3M, ces données renseignent les tronçons de voies selon une hiérarchie calculée en partie avec la géométrie, l'usage et la fréquentation des voies. *accès OpenData 3M*	Utile : non défini Format : fichier .JSON	Etude de la mobilité en tenant compte des différents tronçons du réseau routier.	Temporalité : ;...; Spatialité : *Territoriale*	Disponible : Oui Fournisseur : OpenData 3M
Réseau routier simplifié	ces données renseignent à partir du filaire d'OpenStreetMap le réseau routier simplifié sur lequel les calculs du modèle s'opère. *Inaccessible, 3M*	Utile : non défini Format : inconnu	Etude de la mobilité en tenant compte des différents tronçons du réseau routier.	Temporalité : ;...; Spatialité : *Territoriale*	Disponible : Oui Fournisseur : 3M
Offre théorique de transport en commun	ces données renseignent l'offre de transport théorique (lignes, arrêts, horaires) de TAM, délégataire de service public des transports en commun de la Métropole de Montpellier. *accès OpenData 3M*	Utile : non défini Format : fichier GTFS	Etude de la mobilité sur l'ensemble du territoire.	Temporalité : ;...; Spatialité : *Territoriale*	Disponible : Oui Fournisseur : OpenData 3M
Enquête Origine / Destination en Transport en Commun (ODTC)	données d'enquête réalisée en 2015 auprès des usagers des Transports en Commun. Elles renseignent la fréquentation du réseau TAM en donnant pour chaque usager enquêté : l'arrêt de départ (origine), la tranche horaire de départ, le motif du trajet, le sexe, l'âge, le statut, la possession d'un titre de transport, la ligne empruntée, le mode de transport d'accès … et les mêmes informations à l'arrivée (destination). *Inaccessible, 3M*	Utile : non défini Format : inconnu	Etude de la dynamique de la mobilité sur l'ensemble du territoire.	Temporalité : *données de 2015* Spatialité : *Territoriale*	Disponible : en cours Fournisseur : 3M
Comptage vélo	ces données renseignent le nombre de vélo compté sur divers points du réseau cyclable de Montpellier au printemps 2018, 10 postes de comptage disséminés sur la commune de Montpellier ont mesuré pendant une semaine le trafic vélo. En parallèle, un relevé du stationnement vélo à différents moments de la journée a été mené dans 10 quartiers de Montpellier et Castelnau-le-Lez. *inaccessible, 3M*	Utile : non défini Format : inconnu	Etude de la dynamique de la mobilité sur l'ensemble du territoire.	Temporalité : *1 Semaine (printemps 2018)* Spatialité : *Montpellier et Castelnau-le-Lez*	Disponible : en cours Fournisseur : 3M
Comptage Flowbox	ce sont des données de comptages via le dispositif LaFlowbox, ce dernier a été installé en différents points stratégiques de la Métropole en termes d'enjeux de trafic routier. Grâce à la détection des appareils connectés, ce dispositif a permis de reconstituer la structure des flux de déplacements ainsi que les temps de parcours. *inaccessible, 3M*	Utile : non défini Format : inconnu	Etude de la dynamique de la mobilité sur l'ensemble du territoire.	Temporalité : ;...; Spatialité : *Territoriale*	Disponible : en cours Fournisseur : 3M
Comptage cordon (enquête cordon routière)	ce sont des données d'enquête menées en mars 2018. Elles ont été recueillies sur 28 postes et couvrent 3 départements et sur 3 périodes de la journée : le matin, le midi et le soir. *inaccessible, 3M*	Utile : non défini Format : inconnu	Etude de la dynamique de la mobilité sur l'ensemble du territoire.	Temporalité : ;...; Spatialité : *Territoriale*	Disponible : en cours Fournisseur : 3M
		Images Satellites			
Images satellites	Images satellites couvrant plusieurs zone y compris le territoire de Montpellier. *site Theia*	Utile : Oui Format : fichier .TIF	Mise en relation des entités nommées temporelles et spatiales	Temporalité : *série temporelle* Spatialité : *Territoriale*	Disponible : Oui Fournisseur : Theia
		Données Textuelles			
Données textuelles	Données textuelles à récolter sur des pages Web pour l'extraction et la mise en relation des entités nommées.	Utile : Oui Format : fichier .pdf, .txt, etc.	Mise en relation des entités nommées spatiales et temporelles	Temporalité : *non défini* Spatialité : *Territoriale*	Disponible : Oui Fournisseur : Web

Table A.1 – Collecte de données : synthèse des données de 3M

F-TFIDF-C_M					
Title Corpus		**Abstract Corpus**		**Content Corpus**	
terms	**rank**	**terms**	**rank**	**terms**	**rank**
respiratory syncytial virus	1.9880	public health	1.9986	additional file	1.9976
middle east respiratory syndrome coronavirus	1.9846	infectious diseases	1.9979	infectious disease	1.997
systematic review	1.9842	immune responses	1.9976	nk cells	1.997
open access	1.9819	influenza virus	1.9976	health care	1.996
zika virus	1.9819	t cells	1.9975	endothelial cells	1.9957
gene expression	1.9795	virus infection	1.9974	frequency domain	1.9957
virology journal	1.9788	respiratory tract	1.9973	ebola virus	1.9948
human coronavirus	1.976	viral infections	1.9969	influenza infection	1.9943
case report	1.9756	rna viruses	1.9967	real-time rt-pcr	1.9933
syncytial virus	1.9752	acute respiratory syndrome	1.9961	incubation period	1.99325
t cell	1.9746	95percent ci	1.996	health emergency	1.9932
infectious bronchitis	1.9726	ebola virus	1.9945	index patient	1.9932
sars coronavirus	1.9723	influenza viruses	1.9943	membrane rafts	1.9931
bmc public health	1.9701	avian influenza	1.9939	pcr products	1.9929
t cells	1.9689	respiratory tract infections	1.9938	2c atpase	1.9926
acute respiratory infection	1.9672	health care	1.9925	b cell	1.9924
mini review	1.9636	hepatitis c	1.9922	close contact	1.9924
respiratory viral infections	1.9636	type i	1.9918	final dataset	1.9922
bmc public	1.9625	cell line	1.9914	3d8 scfv	1.9921
ebola virus disease	1.9592	spike protein	1.9909	pol ii	1.992
supplementary information	1.9574	codon usage	1.9908	3c pro	1.992
community-acquired pneumonia	1.9543	pandemic influenza	1.9907	influenza pandemic	1.9919
global health	1.9543	endoplasmic reticulum	1.9904	phylogenetic tree	1.9918
peer review	1.9543	saudi arabia	1.9904	protein vi	1.9917
japanese encephalitis virus	1.9512	innate immunity	1.9903	ag nps	1.9916
innate immunity	1.9488	porcine epidemic	1.9903	influenza b	1.99125
multiple sclerosis	1.9488	global health	1.9902	ifn $\beta - 1\alpha$	1.991
human rhinovirus	1.9466	vaccine development	1.9901	ill patients	1.9908
supplementary material	1.9442	cell death	1.9898	poly tail	1.9908
cell entry	1.9417	infectious disease	1.9896	host range	1.9906
coronavirus spike	1.9417	peripheral blood	1.9895	cyclin d3	1.9903
human adenovirus	1.9417	hong kong	1.9894	sequence accession	1.9903
east respiratory syndrome coronavirus	1.9414	immune cells	1.9888	antiviral drugs	1.9897
mers coronavirus	1.9388	cell cycle	1.9886	subunit vaccines	1.9897
west africa	1.9388	clinical trials	1.9885	protein sequences	1.9895
molecular epidemiology	1.9323	infection control	1.9884	oil spill	1.9895
national natural science	1.931	mass spectrometry	1.9883	swine flu	1.9894
natural science foundation	1.931	genome sequence	1.9881	membrane proteins	1.9893
rift valley fever	1.931	clinical samples	1.9877	contact tracing	1.9891
national natural science foundation	1.9307	acute respiratory infections	1.9874	sars 3a	1.9889
influenza infection	1.9284	severe disease	1.9868	critical care	1.9888
protein response	1.9284	hepatitis b	1.9864	hk-2 cells	1.9888
science foundation	1.9284	host response	1.9864	ap2 group	1.9887
supplementary materials	1.9284	type ii	1.9864	prp sc	1.9887
natural science	1.9241	nucleic acids	1.9862	t-cell responses	1.9887
respiratory syndrome coronavirus infection	1.9241	surveillance systems	1.9859	dna vaccines	1.9886
influenza virus	1.9212	influenza virus infection	1.9852	reverse genetics	1.9886
obstructive pulmonary disease	1.92	antiviral drugs	1.9851	health system	1.9884
emerging microbes	1.9193	dna vaccine	1.9847	$b7 - h1$	1.9884
original research	1.9193	influenza infection	1.9845	hcv infection	1.9883
retrospective study	1.9193	reference genes	1.9842	lung cancer	1.9879
phylogenetic analysis	1.9153	cell types	1.984	nucleocapsid protein	1.9879
respiratory syndrome coronavirus	1.9151	b cell	1.9835	3c protease	1.9878
clinical characteristics	1.9138	vaccine candidates	1.9835	tgev infection	1.9878
mass spectrometry	1.9138	host species	1.9833	cs dna	1.9878
national natural	1.9138	respiratory viral infections	1.9832	risk perception	1.9875
rift valley	1.9138	endothelial cells	1.9829	s1 protein	1.9875
science china	1.9138	sequence data	1.9829	ring vaccination	1.9875
valley fever	1.9138	dna viruses	1.9826	syrian hamster	1.9873
respiratory virus infections	1.913	host innate	1.9826	wild mice	1.9873
syndrome coronavirus	1.9096	parainfluenza virus	1.9824	yellow fever	1.9873
classical swine fever virus	1.9087	tract infections	1.9822	climate change	1.9873
b cells	1.9074	south korea	1.9821	public health services	1.9873
host response	1.9074	acute respiratory infection	1.9817	index patients	1.9872
science foundation of china	1.9074	reproduction number	1.9816	small rna	1.9872
viral proteins	1.9074	surveillance system	1.9816	ic activity	1.9871
virus disease	1.9065	causative agent	1.9813	ebola virus disease	1.9868
clinical infectious diseases	1.9048	multiple sclerosis	1.9811	rna chaperone	1.9867
world health organization	1.9048	rsv infection	1.9809	caco-2 cells	1.9867
antiviral agents	1.9001	cellular proteins	1.9808	m2 channel	1.9865
cell culture	1.9001	west nile virus	1.9806	overlapping genes	1.9865
pulmonary disease	1.9001	respiratory diseases	1.9805	nasal mucosa	1.9865
study protocol	1.9001	tgev infection	1.9805	hepatitis e	1.9865
dengue virus	1.8946	e protein	1.9802	genetic drift	1.9865
public health	1.893	gene expression	1.9801	a7 gfp	1.9865
rna replication	1.8915	structural proteins	1.9799	tumor cells	1.9864
japanese encephalitis	1.8902	acute respiratory tract	1.9792	tanguticum nanoparticles	1.9864
syndrome coronavirus infection	1.8864	hand hygiene	1.9792	cfu ml	1.9864
human respiratory syncytial virus	1.8841	disease transmission	1.9788	ward closure	1.9861
synonymous codon usage	1.8824	human rhinovirus	1.9785	case definitions	1.9861
clinical infectious	1.8813	bacterial infections	1.9781	richards model	1.9861
health organization	1.8813	cancer cells	1.9781	epimedium koreanum	1.9861
severe pneumonia	1.8813	dna vaccines	1.9777	ms2 plp	1.9860
dengue virus infection	1.8772	type iii	1.9777	gene therapy	1.9859
clinical samples	1.8768	viral pathogenesis	1.9773	integrin b3	1.9859
classical swine fever	1.8744	zoonotic diseases	1.9773	cardiovascular diseases	1.9859
human antibody	1.869	early detection	1.9765	fourth site	1.9859
lassa virus	1.869	lung cancer	1.9756	serial interval	1.9858
pilot study	1.869	nile virus	1.9756	trm cells	1.9858
avian influenza viruses	1.8667	human disease	1.9751	electronic supplementary material	1.9857
human respiratory syncytial	1.8667	rnase l	1.9751	emergency nurses	1.9856
international health regulations	1.8667	health systems	1.9746	pet substrate	1.9856
hepatitis c virus infection	1.8661	incubation period	1.9746	fcov type	1.9856
infectious bronchitis virus strain	1.8661	rabies virus	1.9746	s1 text	1.9856
vaccine development	1.8601	adaptive immunity	1.9741	global health research	1.9854
protects hepatocytes from type i	1.8564	multiplex pcr	1.9741	ace2 activity	1.9853
type i interferon signaling disrupts	1.8564	nk cells	1.9741	β 6 ko	1.9853
adaptive immunity	1.8538	feline coronavirus	1.9735	global health	1.9852
adenovirus type	1.8538	human populations	1.9735	ham tsp	1.9851
nonhuman primates	1.8538	common cold	1.9723	blood culture	1.9849

TABLE A.2 – Termes les mieux classés extraits du corpora Paper1 en utilisant F-TFIDF-C_M

C-Value					
Title Corpus		**Abstract Corpus**		**Content Corpus**	
terms	**rank**	**terms**	**rank**	**terms**	**rank**
respiratory syndrome	386.7309	public health	1393.182	t cells	2063.1457
virus infection	366.1263	respiratory syndrome	1095.2091	public health	1644.7156
porcine epidemic diarrhea virus	329.7138	infectious diseases	952.5625	amino acid	1409.82415
porcine epidemic diarrhea	318.0	immune response	908.1835	immune response	1400.94835
epidemic diarrhea virus	306.0	immune responses	841.6151	influenza virus	1185.8689
east respiratory syndrome	284.0	influenza virus	841.6151	immune responses	1056.536
middle east	261.5188	t cells	803.576	t cell	1056.37753
epidemic diarrhea	256.7639	virus infection	760.7811	gene expression	1050.6716
diarrhea virus	245.6692	respiratory tract	727.4978	viral replication	1021.5083
infectious diseases	245.6692	viral infection	668.8542	infected cells	939.72426
respiratory syndrome coronavirus	240.0	viral replication	665.6843	cell lines	897.4057
influenza a	225.0647	viral infections	640.3249	viral infection	888.6884
public health	209.2151	east respiratory syndrome	638.0	virus infection	872.68035
syndrome coronavirus	191.7805	respiratory syndrome coronavirus	636.0	amino acids	866.816
porcine epidemic	190.1955	middle east	630.8151	mg ml	824.4975
influenza virus	182.2707	gene expression	627.6452	infectious diseases	822.27855
respiratory tract	180.6857	infectious disease	613.3805	present study	812.45177
middle east respiratory syndrome	174.1446	rna viruses	603.8707	respiratory tract	812.13477
middle east respiratory	170.0	present study	575.3414	epithelial cells	759.03855
respiratory syncytial virus	166.0	respiratory viruses	551.567	previous studies	732.41119
infectious bronchitis	160.0812	acute respiratory syndrome	516.0	room temperature	714.3426
infectious disease	156.9113	t cell	513.5279	cell culture	673.60907
infectious bronchitis virus	156.0	syndrome coronavirus	511.9429	additional file	657.75946
east respiratory	136.3068	porcine epidemic diarrhea	506.0	viral infections	635.72848
syncytial virus	134.7218	95percent ci	502.4331	immune system	617.97689
avian influenza	131.5519	viral rna	499.2632	respiratory syndrome	617.3429
respiratory viruses	131.5519	amino acid	489.7534	cell line	611.16155
east respiratory syndrome coronavirus	130.028	respiratory syncytial virus	472.0	infectious disease	607.04063
middle east respiratory syndrome coronavirus	129.2481	cell lines	443.7895	μ g ml	576.13388
influenza a virus	126.0	respiratory infections	426.3549	western blot	568.36754
bronchitis virus	125.212	epithelial cells	424.77	rnase l	565.0391
respiratory infections	125.212	virus replication	420.0151	virus replication	560.6012
systematic review	125.212	polymerase chain reaction	408.0	cell surface	543.9591
ebola virus	120.4572	epidemic diarrhea virus	406.0	xx	542.0572
acute respiratory	117.2872	epidemic diarrhea	402.5805	host cell	539.83825
viral infections	117.2872	host cell	396.2406	codon usage	523.03765
virus replication	115.7023	syncytial virus	378.806	viral proteins	520.6391
open access	109.3624	porcine epidemic diarrhea virus	376.1524	respiratory viruses	515.4298
zika virus	109.3624	antiviral activity	374.0512	nk cells	503.2256
respiratory tract infections	102.0	risk factors	374.0512	time points	497.8367
viral infection	101.4376	immune system	369.2963	influenza viruses	492.7648
immune response	99.8526	ebola virus	364.5414	important role	491.0213
hepatitis c virus	98.0	chain reaction	355.0316	allergic rhinitis	486.5835
gene expression	96.6827	influenza viruses	348.6918	antiviral activity	481.3531
pandemic influenza	96.6827	infected cells	347.1068	global health	473.9038
respiratory syndrome virus	96.0	diarrhea virus	340.7669	mg kg	470.0998
epithelial cells	95.0978	host cells	334.4271	frequency domain	469.1489
complete genome	93.5128	important role	331.2572	control group	466.13749
syndrome virus	93.5128	phylogenetic analysis	331.2572	viral load	465.34499
virology journal	93.5128	polymerase chain	331.2572	binding site	459.6391
hepatitis c	91.9278	respiratory disease	326.5023	expression levels	453.6162
immune responses	90.3429	avian influenza	324.9173	hong kong	450.7237
genome sequence	88.7579	respiratory tract infections	320.0	clinical signs	448.8613
dengue virus	87.1729	infectious bronchitis	285.2933	protein expression	448.2274
molecular sciences	84.0029	cell culture	272.6136	wild type	446.7833
type i	84.0029	hepatitis c virus	268.0	endothelial cells	441.4120
acute respiratory syndrome	84.0	health care	264.6887	table s1	438.4006
complete genome sequence	84.0	zika virus	264.6887	flow cytometry	437.4496
human coronavirus	82.4181	infectious bronchitis virus	260.0	saudi arabia	433.4872
respiratory infection	82.4181	tract infections	258.3489	viral genome	433.3992
case report	80.8331	hepatitis c	255.179	negative control	433.2230
tract infections	80.8331	innate immune response	252.0	–	431.7890
risk factors	79.2481	monoclonal antibodies	248.8391	cell types	431.1098
spike protein	77.6632	viral genome	247.2542	viral entry	427.9399
t cell	77.6632	type i	242.4993	cell death	425.24544
acute respiratory infections	76.0	central nervous system	242.0	er stress	423.185
coronavirus infection	74.4932	amino acids	239.3293	significant differences	420.6490
rna viruses	74.4932	animal models	237.7444	health care	420.4905
severe acute respiratory	72.0	real-time pcr	236.1594	tcid 50	417.3734
sars coronavirus	71.3233	dengue virus	232.9895	cathepsin l	410.5053
isothermal amplification	69.7384	viral load	232.9895	risk factors	408.9203
respiratory disease	69.7384	world health organization	232.0	positive selection	405.7504
bmc public health	66.0	cell line	231.4045	cell cycle	400.9955
disease virus	64.9835	viral proteins	229.8196	nucleotide sequences	397.8256
t cells	63.3985	nervous system	226.6496	plasma membrane	393.5990
influenza viruses	61.8135	wide range	223.4797	intensive care	392.2782
acute respiratory infection	60.0	virus infections	221.8948	host cells	384.82889
type i interferon	60.0	middle east respiratory syndrome	220.5832	hand hygiene	383.5609
journal frontiers	58.6436	immunodeficiency virus	218.7248	significant difference	382.6099
fever virus	57.0587	spike protein	218.7248	immune cells	381.02498
respiratory syncytial	57.0587	life cycle	217.1399	reference genes	380.3909
severe acute	57.0587	recent years	217.1399	hiv aids	377.2211
respiratory tract infection	56.0	codon usage	215.5549	avian influenza	376.8688
antiviral activity	55.4737	viral pathogens	215.5549	serum samples	375.8625
bmc infectious	55.4737	pandemic influenza	213.9699	body weight	375.0021
hong kong	55.4737	clinical signs	212.385	figure 1a	374.0511
viral replication	55.4737	dendritic cells	209.2151	membrane fusion	374.0511
virus infections	55.4737	acute respiratory syndrome coronavirus	208.9735	clinical trials	373.8750
bmc infectious diseases	54.0	bronchitis virus	207.6301	time point	373.3719
respiratory viral infections	54.0	endoplasmic reticulum	207.6301	protein synthesis	369.2962
case study	53.8887	rna virus	207.6301	dengue virus	367.7113
dendritic cells	53.8887	saudi arabia	207.6301	e protein	367.7113
mini review	53.8887	innate immunity	206.0451	high levels	365.3339
rna virus	53.8887	recent studies	206.0451	virus particles	364.5414
transmissible gastroenteritis	53.8887	economic losses	204.4602	target cells	362.5601
bmc public	52.3038	porcine epidemic	204.4602	viral particles	360.4204
monoclonal antibodies	52.3038	world health	204.4602	dendritic cells	357.5675
creative commons cc-by 4	51.0824	global health	202.8752	total number	356.4580
influenza pandemic	50.7188	type 1	202.8752	cancer cells	356.0883
type 1	50.7188	vaccine development	201.2902	disease control	355.2957

TABLE A.3 – Termes les mieux classés extraits du corpora Paper1 en utilisant C-Value

sous_corpus	measure	term	domain relevant	COVID-19 surveillance	syndromic surveillance	Incomplet disease name
title	C-value	respiratory syndrome coronavirus	n	n	n	y
		porcine epidemic diarrhea	y	n	n	n
		syndrome coronavirus	n	n	n	y
		epidemic diarrhea virus	n	n	n	y
		acute respiratory syndrome	n	n	n	y
		public access	n	n	n	n
		diarrhea virus	n	n	n	y
		infectious bronchitis	y	n	y	n
		acute respiratory	n	n	n	y
		bronchitis virus	y	n	y	n
	F-TFIDF-C	journal pre-proof	n	n	n	n
		virology journal	n	n	n	n
		influenza pandemic	y	n	n	n
		coronavirus spike	y	y	n	n
		bmc public health	n	n	n	n
		influenza virus infection	y	n	n	n
		emerging infectious	y	n	y	n
		porcine circovirus type	n	n	n	y
		codon usage	n	n	n	n
		respiratory syndrome	n	n	n	y
abstract	C-value	acute respiratory syndrome	n	n	n	y
		respiratory syndrome coronavirus	n	n	n	y
		east respiratory syndrome	n	n	n	y
		syndrome coronavirus	n	n	n	y
		present study	n	n	n	n
		chain reaction	n	n	n	n
		syncytial virus	n	n	n	y
		porcine epidemic diarrhea	y	n	n	n
		polymerase chain	n	n	n	n
	F-TFIDF-C	virus infections	y	n	y	n
		porcine epidemic	n	n	n	y
		clinical samples	n	n	n	n
		codon usage	n	n	n	n
		mers-cov infection	y	y	n	n
		pandemic influenza	y	n	n	n
		viral entry	y	n	y	n
		95percent confidence interval	n	n	n	n
		immune cells	n	n	n	n
		influenza pandemic	y	n	n	n
		sono stati	n	n	n	n
content	C-value	infected cells	n	n	n	n
		respiratory syndrome	n	n	n	y
		present study	n	n	n	n
		individual components	n	n	n	n
		essential medicines	n	n	n	n
		previous studies	n	n	n	n
		de los	n	n	n	n
		functional task	n	n	n	n
		der schwangerschaft	n	n	n	n
	F-TFIDF-C	health emergency	y	n	y	n
		membrane rafts	n	n	n	n
		pcr products	n	n	n	n
		afa dr	n	n	n	n
		cod trypsin	n	n	n	n
		2c atpase	n	n	n	n
		naked mole	n	n	n	n
		intracellular delivery	n	n	n	n
		close contact	y	n	y	n
		final dataset	n	n	n	n
		respiratory syndrome	n	n	n	y
title + abstract	C-value	acute respiratory syndrome	n	n	n	y
		respiratory syndrome coronavirus	n	n	n	y
		east respiratory syndrome	n	n	n	y
		syndrome coronavirus	n	n	n	y
		syncytial virus	n	n	n	y
		porcine epidemic diarrhea	y	n	n	n
		antiviral activity	y	n	n	n
		acute respiratory syndrome coronavirus	n	n	n	y
		infectious bronchitis	y	n	n	n

TABLE A.4 – Termes étendus à partir du tableau 4.6. Chaque terme a été évalué par un expert selon 4 critères : pertinence du domaine, surveillance COVID-19, surveillance syndromique, nom de maladie incomplet (y : yes, n : no)