République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université A.MIRA-BEJAIA

Faculté Sciences exactes
Département Informatique

THÈSE

Présentée par

HADJI Atmane

Pour l'Obtention du Grade de

DOCTEUR EN SCIENCES

Filière : Informatique

Option : Cloud Computing

Thème

Ontologies spatiales pour l'optimisation sans ambigüité de l'extraction des connaissances lors de la recherche d'Information géographique.

Soutenue le : 09/07/2025

Devant le Jury composé de :

Nom et Prénom	Grade		
Mr TARI Abdelkamel	Professeur	Univ. de Bejaia	Président
Mr KHOLLADI Mohamed-Khireddine	Professeur	Univ. El Oued	Rapporteur
Mr AMROUN Kamal	Professeur	Univ. de Bejaia	Examinateur
Mr BOUZENADA Mourad	MCA	Univ. de Constantine 2	Examinateur
Mr CHAOUI Allaoua	Professeur	Univ. de Constantine 2	Examinateur
Mr LAOUID Abdelkader	Professeur	Univ. El Oued	Examinateur

Année Universitaire : 2024/2025

Remerciements

LOUANGES à ALLAH de m'avoir appris, protégé, guidé durant toute ma vie.

Ce travail a été réalisé sous la direction du Professeur **KHOLLADI Mohamed-Khireddine**, du Département d'Informatique de l'Université Echahid Hamma Lakhdar d'El Oued et du Laboratoire MISC de l'Université Abdelhamid Mehri – Constantine 2. Je lui adresse mes plus sincères remerciements pour son encadrement scientifique, ses conseils éclairés, sa disponibilité constante et son soutien inestimable. Je lui témoigne toute ma profonde reconnaissance. Merci, Professeur **KHOLLADI**, pour tout ce que vous avez accompli pour votre équipe.

J'exprime également ma gratitude à Madame la Professeure **BORISOVA Nadezhda**, du Département d'Informatique de l'Université du Sud-Ouest « Neofit Rilski » à Blagoevgrad (2700), Bulgarie, pour avoir accepté de collaborer avec notre équipe. Je la remercie vivement pour son aide précieuse et sa disponibilité quotidienne.

Je remercie également le président du jury, Professeur **TARI Abdelkamel**, de l'Université de Bejaia et tous les membres du jury: Professeur **AMROUN Kamal**, de l'Université de Bejaia, Docteur **BOUZENADA Mourad**, Maitre de conférences à l'Université de Constantine2/MISC Lab, Professeur **CHAOUI Allaoua,** de l'Université de Constantine2/MISC Lab et Professeur **LAOUID Abdelkader**, de l'Université d'El Oued, LIAP Lab qui me font l'honneur d'examiner ce travail.

Je remercie du fond de cœur mes parents qui n'ont jamais cessé de me soutenir et de m'accompagner tout au long de mon parcours.

Je remercie mon épouse Dr BEKKA-HADJI F. pour son aide, son soutien et sa compréhension tous au long de ma préparation de ma thèse de Doctorat.

Enfin, dans le souci de n'oublier personne, que toutes celles et ceux qui m'ont aidé, de près ou de loin, trouvent ici l'expression de ma profonde gratitude.

DEDICACES

A mes très chers parents

A mon épouse et mes enfants Ahmed, Mohammed, Meriem et Fatima Zohra

A mes frères et leurs familles

A mes sœurs et leurs familles

A mon oncle Tayouche et sa famille

A ma tante et sa famille

A toute ma famille " HADJI "

Aux familles BEKKA et BENKHANOUCHE

A tous mes amis et mes collègues

A tous mes enseignants

A tous les miens

Table des matières

Chapitre 2 : Ingénierie d'ontologie et Ontologie spatiale

Chapitre 3 : Développement et création de l'ontologie ASTO

Chapitre 4 : Développement d'une approche basée règles JAPE

Chapitre 5 : Nouvelle approche hybride Ontologie-Règles

Liste des figures

Liste des tableaux

Liste des abréviations

AIS	**A**utomatic **I**dentification **S**ystem
ANNIE	a**N**early-**N**ew **I**nformation **E**xtraction **S**ystem
ARCHONTE	**ARCH**itecture for **ONT**ological **E**laborating
ASTO	**A**rabic **S**patial **T**oponymy **O**ntology
BFO	**B**asic **F**ormal **O**ntology
CSV	**C**omma-**S**eparated **V**alues
DOLCE	**D**escriptive **O**ntology for **L**inguistic and **C**ognitive **E**ngineering
EI	**E**xtraction d'**I**nformation
GATE	**G**eneral **A**rchitecture for **T**ext **E**ngineering
JAPE	**J**ava **A**nnotation **P**atterns **E**ngine
JSON	**J**ava **S**cript **O**bject **N**otation
KIM	**K**nowledge and **I**nformation **M**anagement
KML	**K**eyhole **M**arkup **L**anguage
LHS	**L**eft-**H**and-**S**ize
ML	**M**achine **L**earning
MNT	**M**odèle **N**umérique d'**A**ltitude
MUC	**M**essage **U**nderstanding **C**onference
PNG	**P**ortable **N**etwork **G**raphics
OBIE	**O**ntology **B**ased **I**nformation **E**xtraction
OBIESOF	**S**ystem for **O**rganic **F**arming
OTK	**O**N-**T**O-**K**NOWLEDGE
OWL	**W**eb **O**ntology **L**anguage
RCC-8	**R**egion **C**onnection **C**alculus
RDF	**R**esource **D**escription **F**ramework
REN	**R**econnaissance d'**E**ntités **N**ommées
RENS	**R**econnaissance d'**E**ntités **N**ommées **S**patiales
RHS	**R**ight-**H**and-**S**ize
RI	**R**echerche d'**I**nformation
SBL	**S**ervices **B**asés sur la **L**ocalisation
SIG	**S**ystèmes d'**I**nformation **G**éographique
SOBA	**S**mart **W**eb **O**ntology-**B**ased **A**nnotation
TALN	**T**raitement **A**utomatique du **L**angage **N**aturel
TIFF	**T**agged **I**mage **F**ile **F**ormat
UTM	**U**niversal **T**ransverse **M**ercator
WGS	**W**orld **G**eodetic **S**ystem
XML	e**X**tensible **M**arkup **L**anguage

Résumé

Avec l'essor rapide d'Internet, la quantité d'informations disponibles a considérablement augmenté, entraînant un problème croissant de « surcharge d'informations ». Cette situation souligne l'importance d'accéder de manière rapide et précise à des informations pertinentes, en particulier dans des domaines spécialisés tels que les systèmes d'information géographique (SIG). L'extraction d'informations spatiales à partir de textes en arabe constitue un défi majeur en raison de la complexité linguistique et de l'importance des données géospatiales dans divers secteurs.

Pour atteindre cet objectif, cette recherche s'est concentrée sur la création et l'intégration de l'Ontologie Arabe des Toponymes Spatiaux (ASTO). Cette ontologie vise à structurer les connaissances géospatiales en arabe et à améliorer la précision de l'extraction d'informations spatiales. Parallèlement, une approche basée sur les règles JAPE a été développée et évaluée.

De même, une approche hybride combinant l'ontologie avec des méthodes fondées sur des règles a été mise en œuvre. Les deux approches ont été comparées afin d'évaluer leurs performances respectives. Les résultats montrent que l'approche hybride surpasse celle basée uniquement sur les règles JAPE en termes d'efficacité, de précision et de couverture des annotations spatiales dans les SIG.

En conclusion, ce travail apporte une contribution significative à l'amélioration des technologies d'extraction d'informations spatiales en arabe, en proposant des solutions prometteuses pour la gestion des données géospatiales.

Mots Clés : Ontologie Spatiale, Extraction d'Information Spatiale, Rules JAPE, TALN arabe.

Abstract

With the rapid growth of the Internet, the amount of available information has increased significantly, leading to a growing problem of “information overload.” This situation highlights the importance of quickly and accurately accessing relevant information, especially in specialized fields such as Geographic Information Systems (GIS). Extracting spatial information from Arabic texts presents a major challenge due to the linguistic complexity and the significance of geospatial data across various sectors.

To address this issue, this research focused on the creation and integration of the Arabic Spatial Toponyms Ontology (ASTO). This ontology aims to structure geospatial knowledge in Arabic and improve the accuracy of spatial information extraction. In parallel, a rule-based approach using JAPE was developed and evaluated.

Additionally, a hybrid approach combining the ontology with rule-based methods was implemented. Both approaches were compared to evaluate their respective performances. The results show that the hybrid approach outperforms the JAPE rule-based method in terms of efficiency, accuracy, and coverage of spatial annotations in GIS.

In conclusion, this work makes a significant contribution to the advancement of spatial information extraction technologies for the Arabic language by offering promising solutions for geospatial data management.

Keywords: Spatial Ontology, Spatial Information Extraction, JAPE Rules, Arabic NLP.

الملخص

مع الانتشار السريع للإنترنت، ازدادت كمية المعلومات المتاحة بشكل كبير، مما أدى إلى ظهور مشكلة "فرط المعلومات" بشكل متزايد. وهذا يبرز أهمية البحث للوصول السريع والدقيق إلى البيانات ذات الصلة، خاصة في المجالات المتخصصة مثل نظم المعلومات الجغرافية (SIG). إن استخراج المعلومات المكانية من النصوص العربية يمثل تحديًا حاسمًا بسبب التعقيدات اللغوية والأهمية المحورية للبيانات الجغرافية في مختلف القطاعات.

وللوصول إلى الأهداف المسطر لها، تركز هذه الدراسة على إنشاء ودمج "الأنطولوجيا العربية للأماكن المكانية" (ASTO). تم تصميم هذه الأنطولوجيا لهيكلة المعرفة الجغرافية باللغة العربية وتعزيز دقة استخراج المعلومات المكانية. في الوقت نفسه،تم تطوير نهج يعتمد على قواعد JAPE.

ومن اهم المساهمات البارزة في هذا العمل كذلك، تم تطوير نهج هجين يجمع بين الأنطولوجيا والأساليب القائمة على القواعد. اثبت هذا النهج الهجين فعاليته من خلال دمج الأنطولوجيات والقواعد المتخصصة لاستخراج المعلومات المكانية. وتم مقارنته مع النموذج الذي يعتمد على قواعد JAPE لتقييم أدائه ومدى فعالته. وتشير النتائج إلى أن النهج الهجين يتفوق بشكل كبيرمن حيث الدقة وشمولية التوصيفات المكانية في نظم المعلومات الجغرافية.

و في الختام، تساهم هذه الدراسة بشكل كبير في تطوير تقنيات استخراج المعلومات المكانية باللغة العربية، وتقدم حلولًا واعدة لإدارة البيانات الجغرافية.

الكلمات المفتاحية : الأنطولوجيا المكانية، استخراج المعلومات المكانية، قواعد JAPE، معالجة اللغة الطبيعية العربية.

Introduction générale

Au cours des dernières années, l'essor massif des informations numériques, en particulier sur Internet et dans les Big Data a mis en lumière la nécessité de développer des systèmes de traitement de l'information efficaces. Cette croissance exponentielle des données, notamment géoréférencées, a accentué la problématique de la surcharge d'information, rendant d'autant plus important l'accès rapide et précis à des informations pertinentes, en particulier dans des domaines spécialisés tels que les systèmes d'information géographique (SIG). Dans ce contexte, l'extraction d'informations spatiales à partir de textes bruts est devenue un domaine de recherche essentiel, touchant à des disciplines telles que le traitement automatique du langage naturel (TALN), l'extraction d'information (EI), la recherche d'information (RI) et les SIG.

L'extraction d'informations spatiales, notamment dans la langue Arabe, présente des avantages significatifs dans divers secteurs. Elle permet d'enrichir les bases de données géospatiales, d'améliorer la précision des systèmes d'information géographique, ainsi que d'optimiser les services basés sur la localisation (SBL). Elle contribue également à la prise de décision dans des domaines capitaux tels que l'urbanisme, la gestion des ressources naturelles et la réponse aux catastrophes naturelles. Le processus d'extraction consiste à transformer des données textuelles non structurées en informations structurées, identifiant ainsi des entités géospatiales, des relations, des rôles sémantiques et des événements pertinents pour une analyse plus approfondie.

Toutefois, malgré ces bénéfices potentiels, l'extraction d'informations spatiales à partir de textes en Arabe reste un défi majeur. En raison de la richesse morphologique de cette langue et de ses ambiguïtés sémantiques, les méthodes classiques d'extraction d'information, qu'elles soient basées sur des techniques statistiques ou sur l'apprentissage automatique, se révèlent souvent insuffisantes pour relever ces défis. La langue Arabe présente des complexités linguistiques et grammaticales qui compliquent l'identification des informations géoréférencées, rendant encore plus difficile leur intégration dans les systèmes SIG. Cela souligne l'importance de développer des techniques plus avancées, capables de traiter efficacement ces spécificités linguistiques et de surmonter les limites des approches traditionnelles.

Dans ce contexte, les ontologies émergent comme une solution prometteuse pour structurer les connaissances et faciliter l'interaction homme-machine. Elles permettent de construire des bases de connaissances partagées et réutilisables, tout en soutenant

l'évolution du Web sémantique. En effet, les ontologies jouent un rôle primordial dans la représentation des concepts et des relations dans divers domaines, y compris l'extraction et la recherche d'information. Dans le domaine de l'extraction d'informations spatiales, les ontologies sont exploitées pour indexer les entités ou concepts d'un domaine spécifique ; concevoir une hiérarchie permettant de générer des règles de recherche efficaces ; modéliser les propriétés des concepts et leurs relations pour guider le processus d'extraction et mettre à jour ou enrichir l'ontologie après l'extraction.

L'exploitation des ontologies pour l'indexation, la recherche et l'extraction automatiques dans les textes Arabes reste encore limitée. Il est donc nécessaire d'explorer des approches plus intégrées et puissants pour améliorer la précision de l'extraction d'informations spatiales en Arabe. Pour cela, notre étude propose une approche hybride combinant des ontologies spatiales et des règles d'extraction basées sur la connaissance. Cette approche tire parti des forces complémentaires de ces deux techniques pour traiter les particularités linguistiques de la langue Arabe, tout en répondant aux besoins croissants des utilisateurs de SIG dans le monde Arabe.

Dans ce travail, nous avons essayé de répondre aux défis posés par l'extraction d'informations spatiales dans des textes en langue Arabe, un domaine encore peu exploré dans le contexte des SIG. Pour cela, nous avons développé des solutions innovantes qui reposent sur les ontologies et les règles JAPE.L'objectif est de couvrir les limitations des approches traditionnelles en utilisant une ontologie pour structurer les connaissances géospatiales et faciliter l'indexation et l'extraction des entités spatiales et de leurs relations.

Les principales contributions de ce travail s'articulent comme suit :

- Création de l'ontologie ASTO (Arabic Spatial Toponym Ontology) qui a été conçue pour modéliser les toponymes spatiaux Arabes. Elle structure les connaissances géospatiales en intégrant les entités spatiales (objet spatial, localisation), leurs relations géographiques (distance, direction, orientation et topologique), ainsi que des règles de désambiguïsation.
- Développement de l'approche basée sur les règles JAPE qui est une approche complémentaire basée uniquement sur les règles JAPE et qui a été développée pour démontrer l'efficacité de cette méthode dans la désambiguïsation et la classification des informations spatiales. Cette approche repose sur des règles formelles qui

exploitent les structures linguistiques spécifiques à la langue Arabe pour identifier et annoter les entités spatiales. Cependant cette méthode présente certaines limites face à des cas plus complexes, ce qui justifie le recours à l'approche hybride.

- Développement de l'approche hybride qui combine l'utilisation de l'ontologie ASTO avec des méthodes basées sur des règles JAPE (Java Annotation Patterns Engine). Cette approche a pour objectif de tirer la structuration des connaissances apportée par l'ontologie pour améliorer la désambiguïsation, l'extraction et la classification des informations spatiales.
- Évaluation et comparaison des approches qui font recours à des tests rigoureux qui ont été effectués pour évaluer les performances des deux approches (basée sur les règles JAPE et hybride) dans des contextes réels, notamment au sein de systèmes d'information géographique.

Cette thèse est subdivisée en deux parties principales. La première partie, état de l'art, regroupe le premier chapitre qui évoque les systèmes d'information géographique (SIG), en mettant l'accent sur la modélisation des données géospatiales et leur gestion dans le cadre des systèmes d'information. Une attention particulière est accordée à l'extraction des informations spatiales, notamment la reconnaissance des entités spatiales dans les textes Arabes, soulignant les défis posés par la langue. Le deuxième chapitre aborde les ontologies spatiales, en présentant leur rôle dans la structuration des connaissances et leur application aux SIG. En explorant les différentes méthodes de modélisation des ontologies, particulièrement les ontologies spatiales, les avantages et les limites de leur utilisation dans les systèmes géospatiaux sont également analysés.

Quant à la deuxième partie, consacrée à la contribution, elle se compose de trois chapitres. Le troisième chapitre est dédié à la conception de l'ontologie ASTO, en expliquant le processus de sa création et de son développement. Il décrit les étapes de modélisation, la spécification des concepts et des relations spatiales, ainsi que les choix techniques mis en œuvre.

Le quatrième chapitre présente une approche d'extraction d'informations spatiales basée uniquement sur les règles JAPE. Cette méthode repose sur des patrons linguistiques formels permettant d'identifier les entités et les relations spatiales dans les textes arabes. Elle vise à assurer une désambiguïsation et une classification précises des informations spatiales. L'efficacité de cette approche est évaluée à travers une série d'expérimentations,

mettant en évidence sa capacité à améliorer la précision des systèmes d'extraction.

Enfin, le dernier chapitre introduit une approche hybride, combinant les règles JAPE avec l'ontologie ASTO, afin d'enrichir l'analyse linguistique par une modélisation sémantique. Cette méthode vise à surmonter les limites des règles seules et à renforcer la compréhension des expressions spatiales complexes. Une étude comparative entre l'approche JAPE seule et l'approche hybride est menée pour évaluer leurs performances respectives.

Chapitre 1

Informations Géographiques et Extraction d'Informations

1. Introduction

L'extraction d'informations est une technique de traitement automatique du texte qui permet d'extraire des entités, des relations, des événements et d'autres informations spécifiques à partir de textes non structurés en langage naturel, pour les convertir en données structurées. Cette technologie vise à rendre les informations compréhensibles par les machines, constituant ainsi un élément central et fondamental du traitement du langage naturel (Yang et al., 2022).

L'extraction d'informations spatiales à partir de textes bruts est devenue un sujet de recherche crucial dans divers domaines tels que le Traitement Automatique du Langage Naturel (TALN), l'Extraction d'Informations (EI), la recherche d'informations (RI) et les Systèmes d'information géographique (SIG) (Hadji et al., 2024). L'exploitation de la technologie des SIG fournit aux utilisateurs une gamme étendue des outils et des méthodologies pour la gestion des informations géospatiales. Cette technologie facilite la collecte, le stockage, la fusion, l'interrogation, la visualisation et l'analyse des données géospatiales à différents niveaux de précision (Raihan, 2024).

Ce chapitre est consacré à la présentation de principales notions relatives aux systèmes d'information géographique (SIG).Nous présenterons en premier lieu un aperçu de l'information géographique, des différents composants et un modèle de représentation. En second lieu, des généralités sur l'extraction d'information, suivi d'une présentation sur les différentes méthodes d'extractiond'informationsbasées règles, ontologies et apprentissage automatique.

2. Information géographique

D'après Siabato et Manso-Callejo (2011),l'information géographique se compose de trois types de données, spatiale, temporelle et attributaire.

2.1. Composants principaux de l'information géographique

- **Donnée spatiale:** Cette composante représente l'emplacement géographique des objets. Elle est souvent définie par des coordonnées géographiques (latitude, longitude) et peut inclure des informations sur la forme et la taille des objets géographiques (la position d'un bâtiment, d'une route ou d'un lac sur une carte).
- **Donnée temporelle:** Cette composante indique le moment ou la période où l'information géographique est valide ou a été collectée. Elle permet de suivre les changements dans le temps et elle est importante pour les analyses temporelles, comme l'évolution d'une zone inondée au fil des jours ou des années.
- **Données attributaires:** Ce sont des informations descriptives qui qualifient les objets géographiques. Elles peuvent inclure une variété de détails, tels que le nom, le type, la hauteur, la population, ou toute autre caractéristique pertinente. Dans le cas d'une route, les données attributaires peuvent inclure le nom de la route, le type de surface, la largeur, etc. (Figure 1.1).

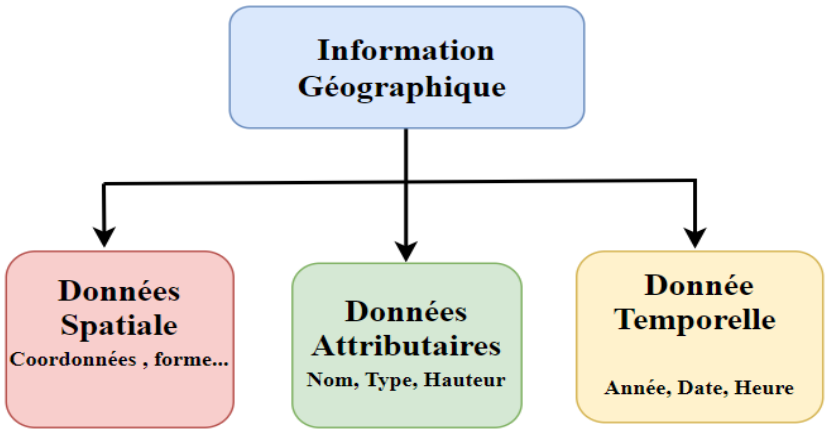

Figure 1.1 : Composants principale de l'information géographique.

Ces trois types de données permettent de caractériser un objet géographique comme un bâtiment en termes de localisation (spatiale), d'attributs descriptifs (attributaire), et d'évolution ou d'événements dans le temps (temporelle) (Figure 1.2).

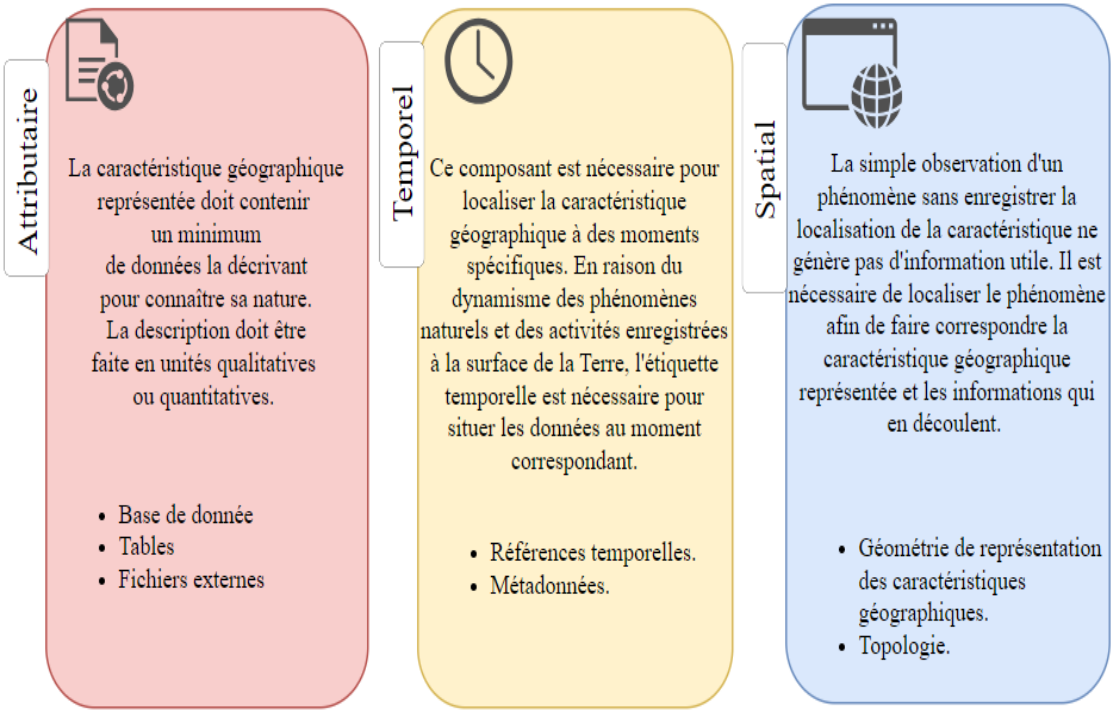

Figure 1.2 : Description des trois composantes de base de l'information géographique (Siabato et Manso-Callejo, 2011).

2.2. Emplacement géographique

L'emplacement géographique est une caractéristique fondamentale des données spatiales (Figure 1.3). Il se réfère aux coordonnées spécifiques (latitude, longitude) qui définissent la position d'une entité sur la surface terrestre. La précision des données spatiales dépend du système de référence utilisé, tel que le WGS 84 (World Geodetic System)[1]est un système de référence tridimensionnel utilisé pour définir la latitude, la longitude et l'altitude, essentiel pour la navigation, le positionnement et le ciblage.À ce jour, WGS 84 constitue le système de référence géodésique mondial le plus précis pour des applications pratiques telles que la cartographie, le géopositionnement et la navigation. D'autres systèmes de projection géographique comme UTM (Universal Transverse Mercator)[2] est un système de grille de coordonnées planes nommé d'après la projection cartographique sur laquelle il est basé (Transverse Mercator). Le système UTM se compose de 60 zones, chacune d'une largeur de 6 degrés de longitude. Les zones sont numérotées de 1 à 60, commençant à 180 degrés de longitude et augmentant vers l'est.

[1] https://earth-info.nga.mil/index.php?dir=wgs84&action=wgs84
[2] https://www.usgs.gov/faqs/what-does-term-utm-mean-utm-better-or-more-accurate-latitudelongitude

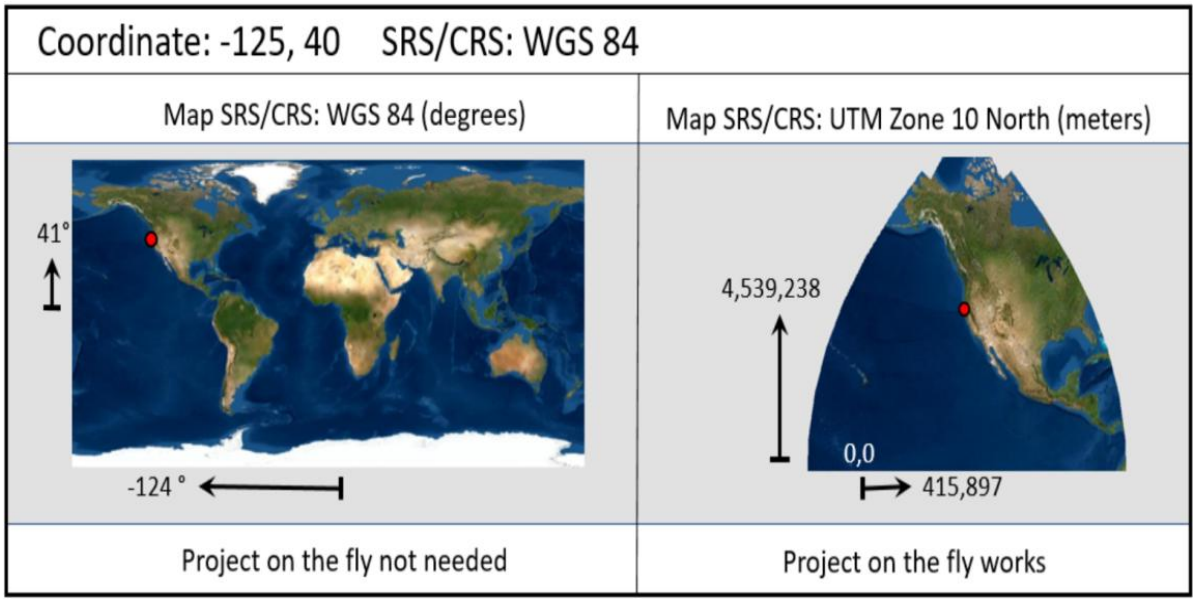

Figure 1.3 : Emplacement géographique[3]

2.3. Sources d'information géographique

Les sources d'information géographique se divisent en deux catégories principales : structurées et non structurées (Tableau 1.1). Les sources structurées incluent des bases de données géospatiales, des systèmes d'information géographique (SIG) et des modèles numériques, où les données sont organisées selon des formats définis, facilitant ainsi leur extraction et leur analyse. Cependant, les sources non structurées englobent des textes libres, des rapports, des images aériennes et des médias sociaux, où les informations spatiales sont dispersées et intégrées dans des formats moins uniformes. L'exploration de ces deux types de sources est essentiel pour obtenir une vue d'ensemble complète et précise des phénomènes géographiques, en combinant la précision des données structurées avec une masse des informations issues des sources non structurées.

Tableau 1.1 : Sources de données structurées et non structurées en géospatial(Longley et al., 2015).

Sources structurées	Sources non structurées
Données GPS : systèmes de positionnement global (GPS).	**Données des enquêtes sur le terrain** (notes, photos, relevés manuels).
Données socio-économiques : données socio-économiques.	**Données historiques et archives** (cartes anciennes, documents d'archives).

[3]https://gsp.humboldt.edu/olm/Courses/GSP_510/SpatialReferenceIssues/SRS_CRS_Define_Project.html

Données cadastrales : données cadastrales et foncières.	**Données des capteurs environnementaux** (en temps réel, données brutes).
Bases de données spatiales : bases de données spatiale.	Données issues des réseaux sociaux et des plateformes collaboratives (géotags[4]).
Systèmes d'Information : Systèmes d'information géographique (SIG).	**Données de télédétection** (images satellites, relevés lidar[5]).
Cartes et Plans : cartes et plans topographiques.	**Images** obtenues à partir camera vidéos au sol
Données Géodésiques : Données géodésiques.	**Collection de documents** rapports, article de press,….

Les quatre types fondamentaux de sources de données qui constituent la base pour la création des SIG, sont définis par Willmes (Figure 1.4)(Willmes et al., 2017).

2.3.1. Sources de données non structurées

Le terme 'non structuré' fait référence à toutes les sources de données qui ne sont pas dans un format directement exploitable par un ordinateur, comme une feuille de calcul, des données SIG, une image capturée par un capteur, ou tout autre type de données pouvant servir de base à un traitement algorithmique et informatique (Willmes et al., 2017).

2.3.2. Sources de données structurées

Pour les données numériques, leur réutilisabilité est souvent plus facile, même si elles ne sont pas au format SIG. Lorsque un contexte spatial est présent dans les données, il est possible de les convertir en formats SIG pour une utilisation ultérieure dans le contexte présenté(Willmes et al., 2017).

2.3.3. Données non SIG

Les données non SIG peuvent inclure des fichiers CSV(Comma-Separated Values), des feuilles de calcul, des fichiers XML(eXtensibleMarkupLanguage), JSON(JavaScript Object Notation) ou des bases de données relationnelles qui contiennent une référence spatiale informelle, telle qu'un champ pour les coordonnées, mais sans définition explicite d'un système de référence spatiale. Cela permet de créer un ensemble de données SIG, généralement la création d'un script pour lire le fichier CSV ou la feuille de calcul et les convertir dans un format SIG approprié(Willmes et al., 2017).

[4] https://www.pixalione.fr/glossaire/geotag/
[5] https://geoservices.ign.fr/lidarhd

2.3.4. Données SIG

Les données SIG sont des informations géospatiales qui permettent de représenter et d'analyser des phénomènes localisés sur la surface terrestre. Elles se composent de données spatiales, telles que les coordonnées géographiques et de données attributaires, qui fournissent des détails supplémentaires sur les objets représentés. Les sources de données SIG englobent toutes les données qui peuvent être directement lues et affichées dans un SIG de bureau comme QGIS (QGIS Development Team, 2016)(Willmes et al., 2017).

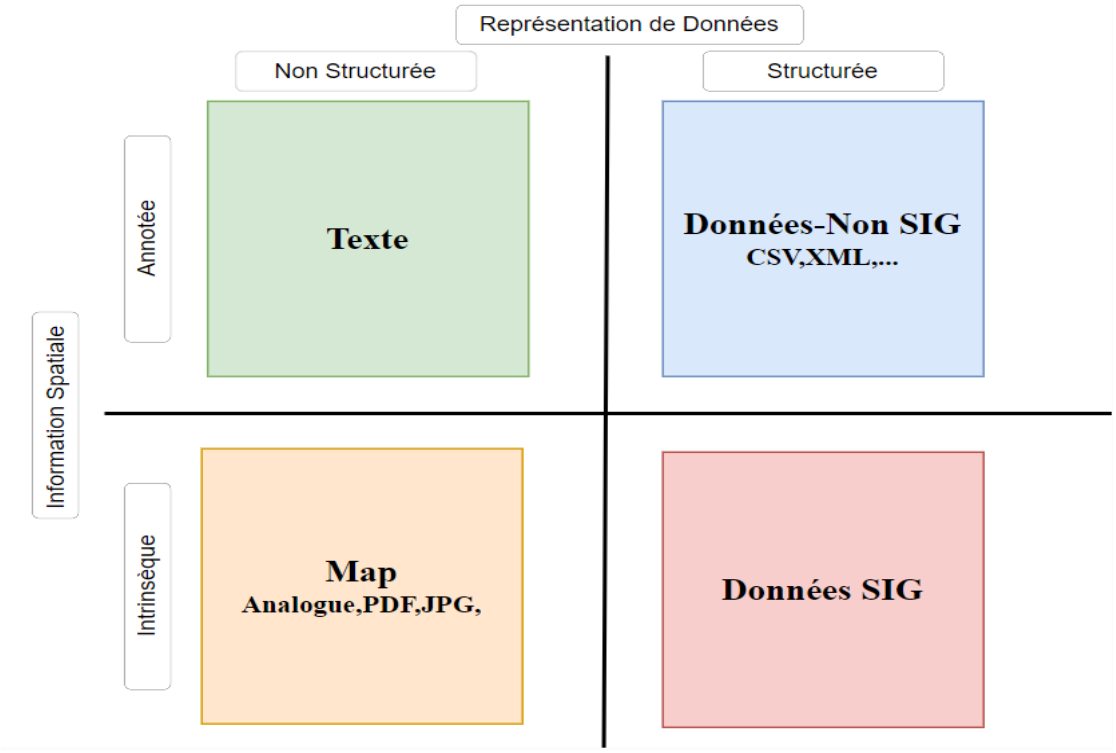

Figure 1.4 : Sources de données spatiales(Willmes et al., 2017).

3. Modélisation des données géographiques

La modélisation des données géographiques est un domaine essentiel au sein des systèmes d'information géographique(SIG), des sciences de la Terre et de l'analyse spatiale. Elle concerne la représentation, la structuration et l'analyse des informations géographiques pour comprendre, simuler et prédire des phénomènes géospatiaux. Ce processus implique l'abstraction des éléments géographiques du monde réel en des modèles de données numériques, facilitant ainsi leur manipulation informatique et leur exploitation dans divers domaines, tels que l'urbanisme, la gestion des ressources naturelles et la surveillance

environnementale(Longley et al., 2011).

3.1. Concepts fondamentaux

La modélisation des données géographiques repose sur plusieurs concepts fondamentaux, dont :

- **Représentation des entités géographiques**: Les entités géographiques telles que les rivières, les montagnes et les bâtiments sont représentées sous forme de points, de lignes et de polygones dans les SIG. Ces entités sont associées à des attributs spécifiques qui fournissent des détails supplémentaires, comme le nom d'une ville ou la longueur d'une route(Yasobant et al., 2019).
- **Relations spatiales** : Les relations spatiales entre les entités, comme la distance, la direction et la connectivité, sont importantes pour la modélisation. Commeexemple d'une route reliant deux villes ou la proximité d'une zone inondable par rapport à une rivière(Yasobant et al., 2019).

3.2. Techniques et méthodes de modélisation

- **Modèles de données vectoriels** :Les modèles vectoriels sont utilisés pour représenter des entités géographiques discrètes à l'aide de points, de lignes et de polygones. Chaque entité est stockée avec ses coordonnées géographiques, permettant une représentation précise des éléments tels que les routes, les rivières et les frontières.
- **Données vectorielles**: Elles représentent des entités géographiques à l'aide de points, de lignes et de polygones(Pandey, 2014). Chaque entité vectorielle est définie par un ensemble de coordonnées précises. Un point peut représenter un emplacement précis comme un ;, une ligne peut représenter une route et un polygone peut représenter la délimitation d'une forêt (Figure 1.5). Les formats de données vectorielles couramment utilisés incluent Shapefile,GeoJSON(JavaScript Object Notation)et KML(KeyholeMarkupLanguage)(Zhu et Tan, 2018).
- **Modèles de données raster** :Les modèles raster utilisent une grille régulière pour représenter des phénomènes continus. Chaque cellule de la grille contient une valeur représentant une caractéristique géographique, comme l'altitude ou la densité de population.
- **Données raster**: Elles représentent des phénomènes géographiques sous la forme d'une grille de cellules, où chaque cellule a une valeur qui correspond à une caractéristique géographique, comme l'altitude ou l'intensité d'une température(Pandey, 2014). Les images satellite et les photographies aériennes sont

des exemples typiques de données raster. Les formats de données raster incluent GeoTIFF(Tagged Image File Format), JPEG2000 et PNG(Portable Network Graphics)**(**Amhar et al., 2022). Un modèle d'altitude numérique (MNT) pourrait représenter la topographie d'une région sous forme de grille, chaque cellule contenant la valeur d'altitude.

- **Modélisation conceptuelle et ontologies :** Les ontologies spatiales sont utilisées pour structurer les concepts géographiques et leurs relations dans un cadre formel. Elles permettent de modéliser des connaissances géospatiales complexes et d'intégrer des informations provenant de différentes sources(Grenon et Smith,2004).Une ontologie pour un SIG pourrait définir des concepts comme "Rivière", "Montagne" et "Ville", ainsi que les relations entre eux, comme "traverse" ou "est situé près de".

La modélisation des données géographiques, en particulier dans le contexte de l'intégration de grandes quantités de données et de l'utilisation de techniques avancées comme l'intelligence artificielle, présente plusieurs défis et opportunités pour la recherche future :

- **Traitement des Big Data géospatiales** : Avec l'explosion des données géospatiales provenant de satellites, de capteurs (Internet of Things)IoT et des réseaux sociaux, il est nécessaire de développer des méthodes pour traiter et analyser efficacement ces grandes quantités de données.
- **Modélisation multi scalaire**: Les phénomènes géographiques se produisent à différentes échelles,nécessitant le développement de modèles capables de les représenter et de les intégrer de façon cohérente.
- **Intégration des techniques d'intelligence artificielle (IA)**: L'IA et en particulier l'apprentissage automatique offre de nouvelles possibilités pour améliorer la modélisation des données géographiques, notamment en automatisant l'extraction et l'analyse des données : Des réseaux de neurones sont utilisés pour classer automatiquement les types de sol à partir d'images satellites ou pour prédire les mouvements de population en fonction des données climatiques et économiques.

La modélisation des données géographiques est une discipline fondamentale qui sous-tend de nombreuses applications dans les sciences de la terre, l'urbanisme et la gestion environnementale. En intégrant des données de différentes sources et en utilisant des techniques de modélisations avancées, les chercheurs et les professionnels peuvent obtenir une compréhension plus profonde des phénomènes géospatiaux et ainsi mieux répondre aux défis du monde moderne. Les développements futurs dans ce domaine promettent d'améliorer

encore la précision, l'efficacité et l'utilité des modèles géographiques dans divers contextes(Heywood et al., 2005).

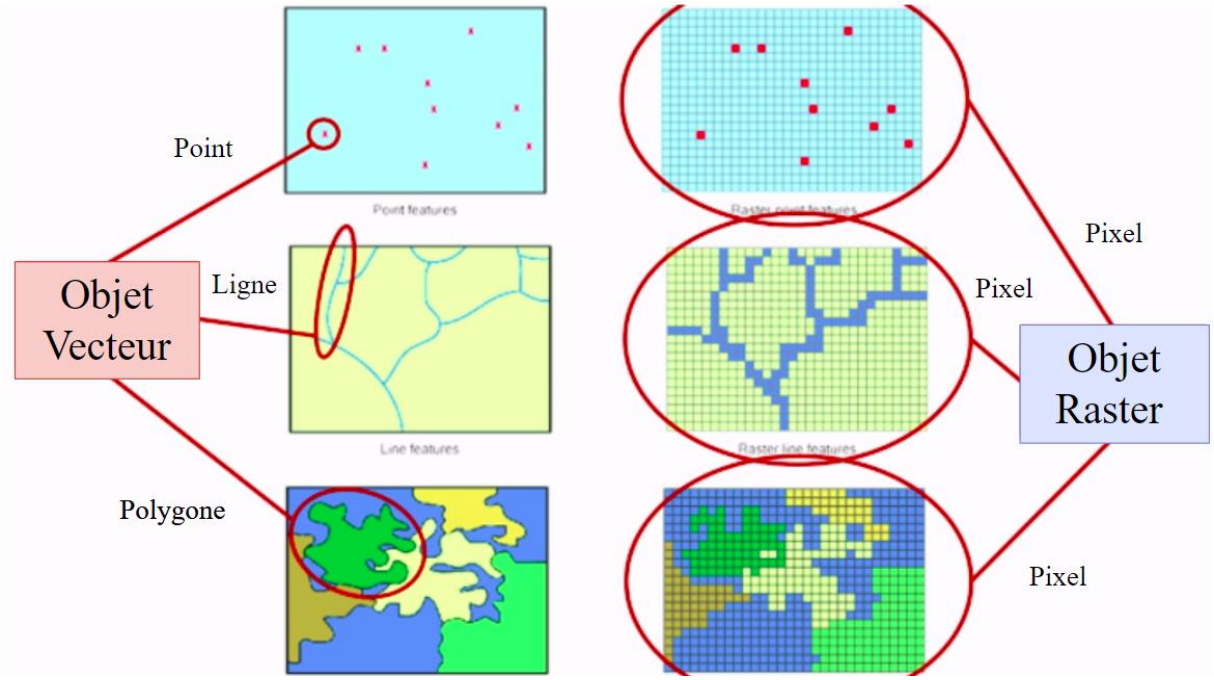

Figure 1.5 : Modèle de données vectorielles versus raster (Pandey, 2014).

4. Information spatiale

Les informations spatiales sont des informations qui sont associées à des positions géographiques sur la surface terrestre. Elles sont essentielles dans divers domaines tels que la géographie, la cartographie, l'analysespatiale,la gestion des ressources naturelles et la planification urbaine. Les données spatiales peuvent représenter des entités physiques comme des bâtiments, des routes, ou des rivières, ainsi que des phénomènes plus abstraits comme les distributions climatiques ou les réseaux de transport.Les données spatiales peuvent être classées en deux grandes catégories : les entités et les relations(Goodchild et al., 2018).

4.1. Différents types de procédés de projection Informations Spatiales

Les informations spatiales désignent des données qui comprennent des aspects géographiques ou positionnels, permettant de localiser et d'analyser des phénomènes dans l'espace géographique. L'information spatiale est caractérisée par sa capacité à intégrer des données relatives à des positions spécifiques sur la surface terrestre, ce qui permet une analyse

précise et contextuelle des phénomènes observés**(Goodchild et al., 2018).**

4.2. Forme d'informations spatiales

Les informations spatiales peuvent être représentées sous deux formes principales : sous forme d'objets discrets ou sous forme de phénomènes continus : Par exemple, un bâtiment ou une route représente un objet discret qui peut être localisé précisément sur une carte, tandis que la température ou l'altitude sur une large région constitue un phénomène continu, mesuré à travers une distribution géographique(Goodchild et al., 2018).

4.3. Entités spatiales

Les entités spatiales constituent les éléments de base de l'information géographique (Figure 1.6). Elles peuvent être des objets concrets, comme des infrastructures (bâtiments, routes), ou des zones abstraites (régions administratives, zones de recensement). Chaque entité est définie par sa localisation géographique et ses caractéristiques spécifiques(Musa et al., 2018).

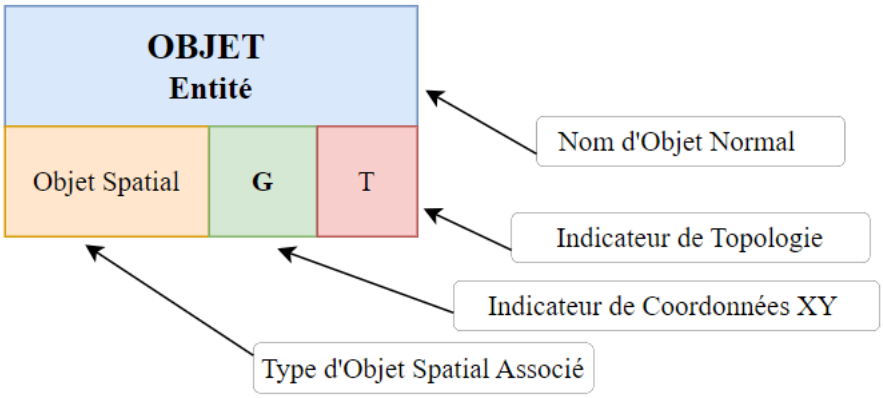

Figure 1.6 : Symbole d'entité pour objet spatial (Musa et al., 2018).

4.4. Relations spatiales

Les relations spatiales décrivent les interactions et les connexions entre ces entités dans l'espace (Figure 1.7). Ces relations peuvent inclure :

- **Proximité** : Mesure de la distance entre deux entités (la distance entre deux villes) ;
- **Connexion** : Comment deux entités sont reliées (une route reliant deux villes) ;
- **Orientation** : Direction relative entre deux entités (une ville située au nord d'une

autre).

Les relations spatiales sont essentielles pour comprendre les interactions complexes entre les différents éléments géographiques et modéliser divers phénomènes tels que la propagation de polluants, la connectivité des réseaux de transport, ou la répartition des habitats naturels(Musa et al., 2018).

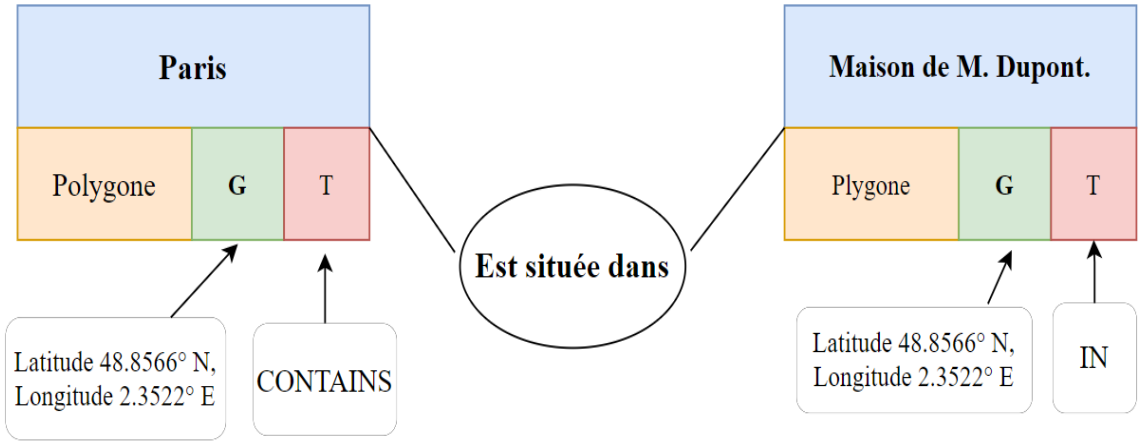

Figure 1.7 : Symbole d'une relation spatiale(Musa et al.,2018).

Explication

- **Indicateur de topologie** : La relation topologique "IN" pour la maison, indiquant qu'elle est située à l'intérieur des limites de la ville. Pour la ville, la relation est "CONTAINS", signifiant qu'elle englobe la maison dans son espace géographique.
- **Indicateur de coordonnées XY** : Les coordonnées XY précisent la localisation exacte des objets (la maison et le centre-ville).
- **Type d'objet spatial associé** : Pour la maison, un polygone représente ses limites sur un plan cadastral. Pour la ville, un polygone décrit les frontières administratives.

4.5. Reconnaissanced'Entités Nommées

La reconnaissance d'entités nommées (REN) est une tâche essentielle et bien étudiée dans le domaine du traitement automatique du langage naturel (TALN). Son objectif est d'identifier et de classifier automatiquement les mentions d'entités spécifiques telles que les noms de personnes, d'organisations, de lieux, de dates et d'autres éléments nommés dans un texte. Cette tâche est particulièrement importante pour structurer des données textuelles non structurées et permettre l'extraction d'informations précises et utiles à partir de larges volumes de données(Grishman et Sundheim,1996).

4.6. Origines et évolution de la REN

La REN est apparue comme une sous-tâche lors des premières éditions des conférences sur l'évaluation de la reconnaissance des entités dans les années 1990, notamment avec le programme MUC (Message UnderstandingConference) aux États-Unis. Les premières méthodes employées étaient principalement basées sur des règles et des lexiques, où des motifs linguistiques et des dictionnaires spécifiques étaient utilisés pour identifier des entités nommées dans les textes(Grishman et Sundheim,1996).

4.7. Conférences MUC

Le formalisme RCC-8 (RegionConnectionCalculus) est un modèle topologique largement utilisé pour décrire les relations spatiales entre deux régions dans un espace donné. Il se compose de huit relations de base : Déconnecté (Disjoint, DC), égal (Equal, EQ), externe (External, EC), chevauchement (Overlap, PO), recouvrant partiellement (Partial Overlap, TPP, TPPi) et recouvrant totalement (Total Overlap, NTPP, NTPPi). Ces relations offrent un cadre détaillé et précis pour représenter la disposition relative des régions, ce qui est essentiel pour une analyse spatiale approfondie. En permettant une modélisation fine des interactions spatiales, le RCC-8 contribue à une meilleure compréhension et interprétation des configurations géographiques ou d'image analysées(Benkirane et al., 2024).

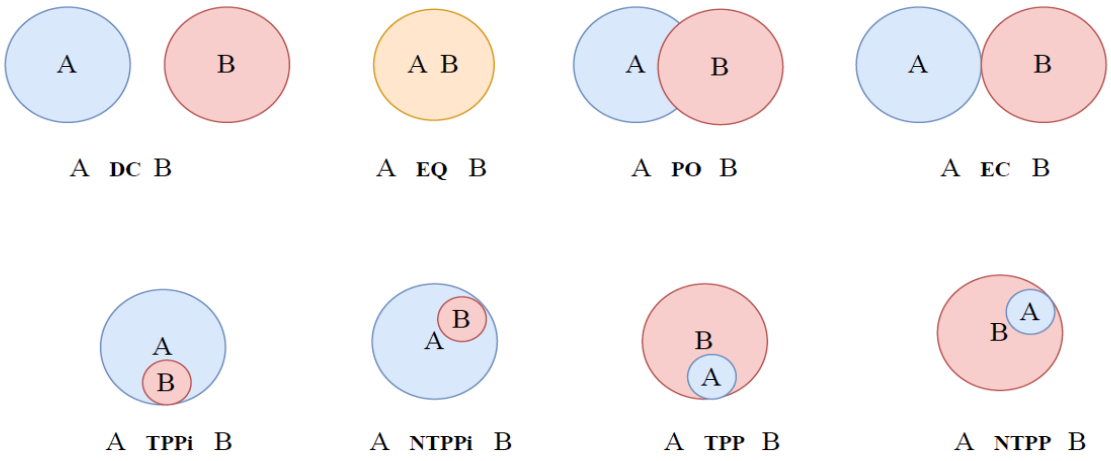

Figure 1.8 : Huit relations topologiques fondamentales(Benkirane et al., 2024).

5. Reconnaissance d'Entités Nommées Spatiales dans le contexte de la langue Arabe

La reconnaissance d'entités nommées spatiales (RENS) est une sous-catégorie spécialisée de la reconnaissance d'entités nommées qui se concentre sur l'identification et la classification des entités géographiques ou spatiales, telles que les noms de lieux, les coordonnées géographiques, les régions et les points de repère dans les textes. Dans le contexte de la langue arabe, cette tâche présente des défis uniques en raison de la complexité linguistique et des caractéristiques morphologiques spécifiques à cette langue(Abdelkaoui et Kholladi, 2015).

- **Défis de la RENS en Arabe**

L'arabe est une langue sémitique avec une structure morphologique riche, ce qui signifie que les mots peuvent avoir de nombreuses formes différentes en fonction des affixes, de la vocalisation et du contexte. Cette diversité morphologique complique l'extraction d'entités spatiales, car les noms de lieux peuvent apparaître sous différentes formes ou être fusionnés avec d'autres mots. De plus, l'arabe utilise un système d'écriture qui ne représente pas toujours les voyelles courtes, rendant l'identification précise des entités encore plus difficile.

Un autre défi est la variation dialectale, l'arabe est parlé dans de nombreuses régions avec des variantes dialectales significatives, ce qui peut entraîner des divergences dans la manière dont les entités spatiales sont nommées ou référencées dans les textes(Qu et al., 2023).

5.1. Modélisation de l'information géographique(Entité Nommée) dans le texte

La modélisation de l'information géographique dans les documents textuels est un processus essentiel en traitement automatique du langage naturel et en géomatique. Elle consiste à extraire, structurer et représenter les informations géographiques (les lieux, les entités spatiales et leurs relations) contenues dans des textes, pour permettre une meilleure compréhension et utilisation de ces informations dans diverses applications, telles que les systèmes d'information géographique(Figure 1.9)(Acheson et Purves, 2021).

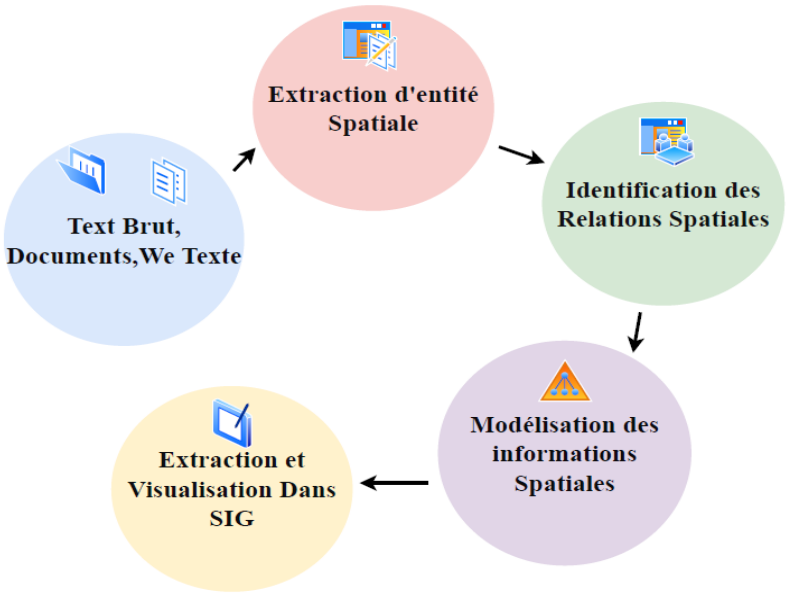

Figure 1.9 : Modélisation de l'information géographique dans les documents textuels.

5.2. Etapes de modélisation

5.2.1. Extraction d'entités géographiques

- **Identification des Lieux** : Le premier pas consiste à identifier les entités géographiques mentionnées dans le texte, comme les villes, les rivières, les montagnes, etc. Cela peut inclure l'utilisation de techniques de reconnaissance d'entités nommées pour repérer les noms de lieux spécifiques.
- **Extraction des attributs**: Une fois les lieux identifiés, il est essentiel d'extraire les attributs correspondants, tels que les coordonnées géographiques (latitude, longitude) ou des descriptions contextuelles, comme ("à l'ouest de")(Achesonet Purves, 2021).

5.2.2. Identification des relations spatiales

- **Détection des relations**: Les relations spatiales décrivent les liens entre les entités géographiques entre elles-mêmes dans l'espace. La phrase comme "La bibliothèque est à côté du parc" implique une relation de proximité entre deux entités géographiques. Ces relations peuvent inclure la distance, la direction, ou la connexion.
- **Modélisation des relations**: Après identification, les relations sont modélisées pour

permettre une analyse spatiale. Cela peut impliquer la création de graphes où les nœuds représentent les entités et les arêtes représentent les relations spatiales(AchesonetPurves, 2021).

5.2.3. Représentation des Données Spatiales

- **Visualisation Cartographique** : Les informations extraites peuvent être visualisées sur des cartes pour offrir une représentation géographique concrète. Cette visualisation aide à interpréter les relations spatiales et à identifier des motifs ou des connexions dans les données.
- **Intégration dans les SIG** : Les données structurées peuvent être intégrées dans des systèmes d'information géographique pour une analyse plus approfondie, permettant de combiner les données textuelles avec d'autres sources d'information géographique(Jones, 2014).

6. Extraction d'information

Ces dernières années, la croissance rapide de l'information numérique a mis en évidence la nécessité de systèmes de traitement de l'information efficaces, en particulier pour les langues riches en vocabulaire et complexes dans leur structure, telles que l'arabe (Alrayzah et al., 2024).Dans le contexte de l'extraction d'information, les données multimodales telles que le texte, l'image, l'audio et la vidéo offrent des perspectives variées pour extraire des informations pertinentes à partir de sources complexes et riches en contenu. Chacune de ces modalités requiert des techniques spécifiques pour capter et analyser les données et elles se complètent souvent pour fournir une compréhension plus complète d'un phénomène. La Figure 1.10 donne une explication détaillée sur les techniques ou les sources d'extraction d'information de chaque type de donnée (Adnan et Akbar, 2019).

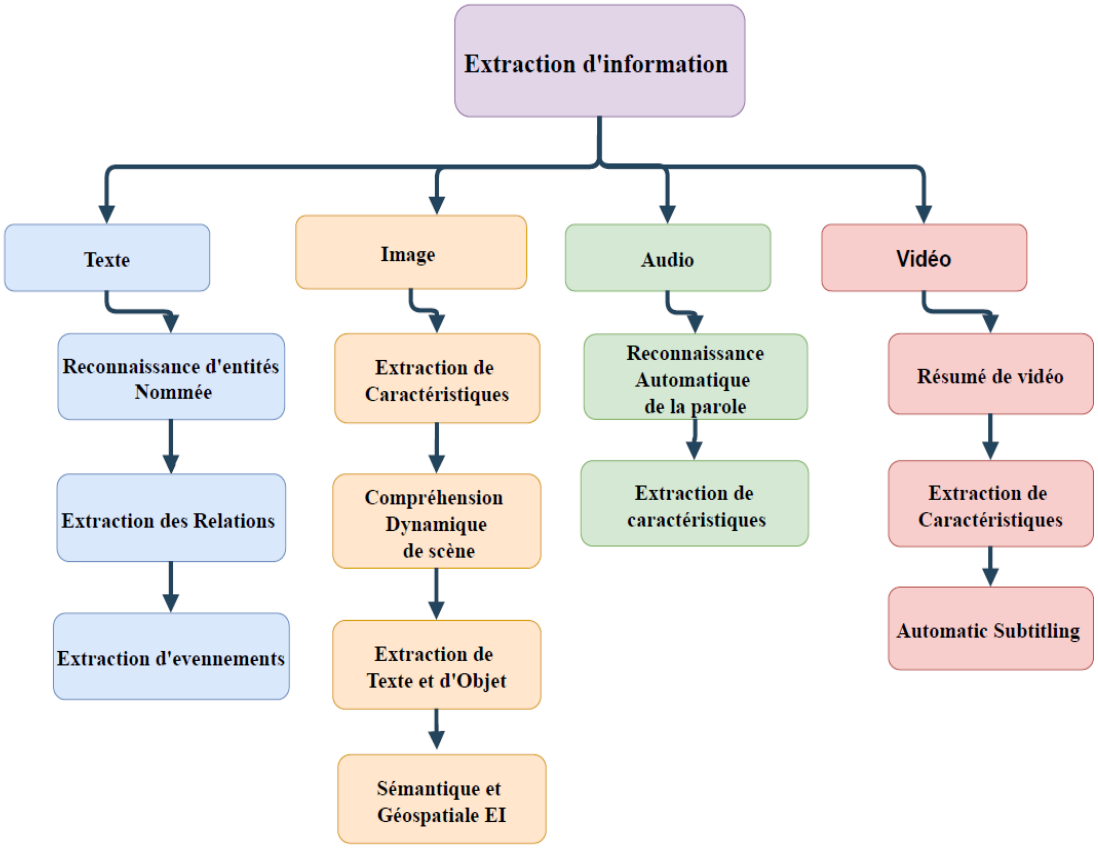

Figure 1.10 : Classification destechniques d'extraction d'information (Adnan et Akbar, 2019).

6.1. Processus d'extraction d'information

Le processus d'extraction d'information (IE) est utilisé pour extraire du contenu structuré sous forme d'entités, de relations, de faits, de termes et d'autres types d'informations, ce qui aide la chaîne de traitement des données à préparer les données pour l'analyse. La transformation efficace et précise des données non structurées améliore la performance de l'analyse des données et du processus d'IE. Différentes approches d'IE ont été proposées pour extraire des informations structurées et utiles à partir de données non structurées, ce qui contribue à la gestion, au traitement et à l'analyse de ces données. Les systèmes d'IE reposent sur le traitement du langage naturel (TALN), la modélisation linguistique et des techniques d'extraction de structures(Adnan et Akbar, 2019).

La Figure 1.11 illustre une explication détaillée des techniques d'extraction d'information pour chaque type de donnée.

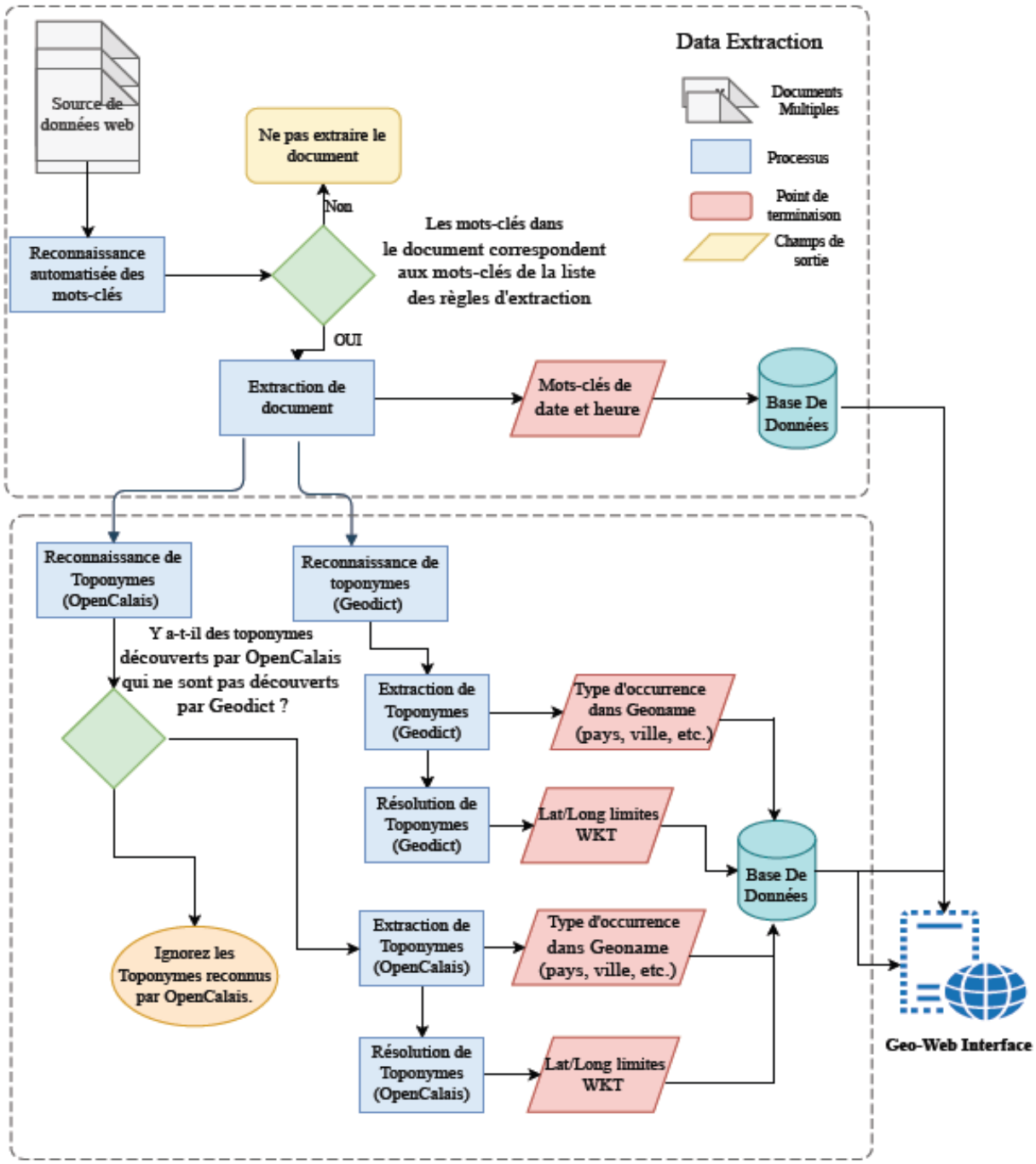

Figure 1.11 : Flux des données à travers les différentes étapes d'extraction et de traitement automatisé(Robertson et Horrocks, 2017).

6.2. Méthodes d'extraction d'information

Il existe trois méthodes principales pour l'extraction d'information et qui sont : méthodes basés sur les règles,méthodes fondées sur les ontologies et méthodes d'apprentissage automatique.

6.2.1. Méthodes basées sur les règles

Les méthodes basées sur des règles ont prouvé leur efficacité dans divers domaines de l'extraction d'information, notamment grâce à leur capacité à capturer des relations spécifiques en appliquant des règles syntaxiques et sémantiques définies. Abdelkaoui et Kholladi(2015) ont rapporté une méthode pour extraire et combiner des informations spatiales et temporelles à partir de textes arabes qui améliore les capacités de recherche et d'exploration grâce à l'architecture GATE. Eftimov et al.(2017) ont introduit "drNER", une méthode novatrice de reconnaissance d'entités nommées (NER) basée sur des règles, conçue pour extraire des concepts diététiques, et cette approche a montré des résultats importants pour l'extraction de recommandations diététiques fondées sur des preuves.

Dans le domaine bibliographique, Makhija et al.(2018) ont appliqué un processus d'extraction d'information basé sur des règles à des données bibliographiques, visant à établir une base de données de concepts pertinents, raffiner les données récupérées et automatiser le processus de récupération locale. Jusoh et al.(2020) ont développé un système combinant l'extraction d'information et la création d'ontologies pour faciliter l'extraction et la visualisation des informations cliniques.

Par ailleurs, Alamoudi(2021) a abordé le défi de l'extraction automatique de la structure de l'information à partir de livres PDF, en proposant une approche intelligente basée sur des règles pour extraire avec précision les métadonnées logiques de ces documents largement utilisés sur le web sémantique. Freitag et al.(2022),ont présenté le cadre VALET(Very Agile Language Extraction Toolkit), un système d'extraction d'information basé sur des règles , qui combine des informations lexicales, orthographiques, syntaxiques et corpus-analytiques dans une syntaxe souple.

Hassini et al.(2023) ont proposé une approche intégrant des techniques de traitement automatique du langage naturel (TALN), des règles et des répertoires géographiques pour extraire des entités spatiales et leurs relations à partir de textes, offrant ainsi une solution viable pour enrichir les SIG avec des informations spatiales précises. Zhang et al. (2024), ont démontré l'efficacité d'une approche basée sur des règles pour l'extraction des relations spatiales à partir de corpus annotés, en particulier pour les relations directionnelles simples. Liao et al.(2024) ont également proposé un système générant automatiquement des règles d'extraction à partir des caractéristiques littérales complexes du chinois. Tandis que Rios et al. (2024) ,ont démontré comment l'alignement translinguistique basé sur des règles grammaticales spécifiques peut enrichir les ensembles de données OpenIE pour des

languessous-représentées comme le portugais brésilien. Enfin, Wang et al.(2024) ont illustré comment les données AIS (Automatic Identification System) des navires de pêche peuvent être exploitées pour extraire des informations spatiales précises, visant à améliorer la gestion des ressources marines.

6.2.2. Méthodes fondées sur les ontologies

L'extraction d'information basée sur les ontologies a connu une évolution remarquable, avec des contributions importantes au fil des années. Popov et al. (2004)ont introduit le cadre KIM (Knowledge and Information Management), fournissant des services de gestion des connaissances et d'informations, facilitant l'annotation sémantique automatique, l'indexation et la récupération de documents, tout en promouvant le partage des métadonnées.Buitelaar et al.(2006)ont démontré, avec le système SOBA (SmartWeb Ontology-Based Annotation) , l'efficacité de l'extraction basée sur les ontologies OBIE (OntologyBasedInformation Extraction) pour analyser des pages web liées au football, en combinant harmonieusement l'ontologie, la base d'expertise et l'extraction de données.

Wimalasuriya et Dou, 2010 ont proposé des "extracteurs d'information" afin de favoriser la réutilisation des ontologies existantes, renforçant ainsi l'efficacité des systèmes d'extraction d'information. Nebhi, 2012 a développé un système combinant la reconnaissance d'entités nommées et un module de désambiguïsation, facilitant l'extraction d'informations à partir de messages Twitter en s'appuyant sur des ontologies telles que DBpedia.

D'après Anantharangacharet al.(2013)une approche combinant des méthodes heuristiques et des ontologies spécifiques au domaine a été employée pour structurer l'information provenant de textes non structurés. Rizvi et al. (2018)ont proposé une méthode d'extraction basée sur les ontologies pour extraire des tableaux pertinents dans des documents techniques au format PDF. Selon Jusoh et al. (2019), un système a été conçu pour faciliter l'extraction et la visualisation des informations cliniques, apportant des avancées importantes dans l'application des ontologies au domaine médical.

D'après Abayomi-Alli et al. (2021), le système OBIESOF (System for OrganicFarming)dédié à l'extraction et à la structuration des informations pour l'agriculture biologique a été développé.

Opasjumruskit et al.(2022)ont introduit l'application OntoHuman, qui permet l'extraction automatique de tuples clé-valeur-unité à partir de documents PDF, améliorant l'utilisation des ontologies dans les documents d'ingénierie.Ahaggach et al. (2023)ont modélisé les dommages automobiles à partir de rapports d'assurance et de textes en

ligne.Tandis que Etudo et Yoon (2024) ont proposé une approche pour la détection de croyances terroristes à partir de textes.

6.2.3. Méthodes d'apprentissage automatique

L'extraction d'informations basée sur l'apprentissage automatique ML (Machine Learning) a adopté aux divers domaines d'application, montrant une grande flexibilité et efficacité dans le traitement de données complexes et non structurées. Escobar et Morales-Menendez(2017) ont proposé un système de supervision intelligent capable de détecter des défauts rares liés à la qualité des processus dans l'industrie automobile. Grâce à l'utilisation de la régression logistique régularisée, cette étude a atteint une précision de 100 % dans la détection des défauts, soulignant l'importance de l'apprentissage automatique pour améliorer les normes de qualité dans un environnement hautement concurrentiel.De même, Fiebeck et al.(2018),ont mis en avant l'intégration des techniques de traitement du langage naturel (NLP) et de ML pour qualifier les fractures de Weber à partir de données radiologiques, en combinant des méthodes NLP et un modèle "bags-of-words". L'étude a montré l'efficacité de l'IA dans l'amélioration de la gestion des données cliniques, illustrant l'application pratique des techniques de ML pour le traitement des données médicales. Steinkamp et al.(2019) ont développé un système d'extraction d'informations à partir de rapports radiologiques non structurés, en mettant l'accent sur la création d'un schéma d'information pour extraire des faits contextualisés. Ce modèle a montré la faisabilité d'utiliser les techniques de ML pour structurer des informations essentielles dans des applications cliniques telles que le suivi des anomalies et la rédaction de rapports.

En revanche, Krieger et al. (2023) se sont concentrés sur l'automatisation du traitement des factures en utilisant l'apprentissage automatique pour extraire des informations à partir de formats de factures non structurés. Les résultats ont montré que le modèle LayoutLM offrait une performance prédictive supérieure par rapport à Chargrid et Random Forest, en traitant les biais de mise en page dans les ensembles de données. Fifita et al.(2024) ont introduit une approche novatrice pour la détection des ‘’fake news’’ liées au COVID-19 en combinant l'extraction d'informations biomédicales avec des modèles ML. En extrayant des caractéristiques médicales à partir de plus d'un millierd'articles de presse, l'étude a montré que l'intégration de ces informations améliore l'efficacité des modèles de détection, avec une performance notable du modèle Random Forest.

De plus, Yang et al.(2022) ont proposé des modèles d'apprentissage profond pour l'extraction des relations géospatiales en démontrant que l'approche pipeline, séparant

lesentités et les relations, surpassait les approches conjointes en termes de performance. Enfin, des études récentes montrent l'évolution des modèles Transformer pour l'extraction d'informations(Dagdelen et al., 2024 ; Luo et Yu, 2024).

Dagdelen et al. (2024) ont montré l'efficacité des modèles Transformer pour traiter des documents manuscrits numérisés.Tandis que,Luo et Yu(2024) ont présenté un modèle multimodal intégrant des graphes sémantiques d'entités pour l'extraction d'informations critiques à partir de CV en chinois. Ces recherches présentent la capacité des modèles préentraînés à simplifier l'extraction de connaissances complexes et à enrichir les bases de données spécialisées. Ces études témoignent des avancées significatives dans le domaine de l'extraction d'informations, en montrant la puissance des techniques d'apprentissage automatique pour transformer les données non structurées en informations exploitables dans divers contextes.

6.3. Mesures d'évaluation des systèmes d'extraction d'information

Les mesures d'évaluation telles que la précision, le rappel et la F-mesure sont largement utilisées pour évaluer la performance des systèmes d'extraction d'information. La précision, également connue sous le terme "precision", mesure la justesse des éléments extraits, en évaluant la proportion d'annotations correctes parmi toutes les annotations identifiées.Le rappel, ou "Recall", est une mesure d'évaluation essentielle pour les systèmes d'extraction d'information, car il évalue la capacité du système à identifier tous les éléments pertinents.Il est défini comme le rapport entre le nombre d'annotations correctes et le nombre total d'annotations pertinentes dans l'ensemble des données. Cette métrique reflète ainsi l'exhaustivité de l'extraction, c'est-à-dire la proportion d'éléments pertinents correctement identifiés par rapport à l'ensemble des éléments pertinents.

La F-mesure combine ces deux métriques en une moyenne harmonique, fournissant ainsi une évaluation globale équilibrée entre la précision et le rappel. Cela permet d'obtenir un indicateur unique qui prend en compte à la fois la justesse et l'exhaustivité des résultats.

SelonMaynard et al. (2006), la précision peut être formellement définie comme le rapport entre le nombre d'annotations correctes et le nombre total d'annotations identifiées, ce qui peut être exprimé mathématiquement comme suit :

$$\text{Précision} = \frac{(\text{Vrais Positifs})}{(\text{Vrais Positifs} + \text{Faux Positifs})} \qquad (1)$$

•**Vrais Positifs (True Positives, TP)** : Le nombre de cas où le modèle a correctement prédit une classe positive.

•**Faux Positifs (False Positives, FP)** : Le nombre de cas où le modèle a incorrectement prédit une classe positive alors que c'était en réalité une classe négative.

Le rappel peut être formellement exprimé comme suit(Gutierrezetal.,2016):

$$\text{Recall} = \frac{(\text{Vrais Positifs})}{(\text{Vrais Positifs} + \text{Faux Negatifs})} \qquad (2)$$

Avec :

Faux Négatifs (False Negatives, FN): Le nombre de cas où le modèle a manqué une instance positive et l'a classée incorrectement comme négative.

La F-mesure (ou F1-score) est une mesure combinée qui prend en compte à la fois la précision et le rappel. Elle est calculée comme la moyenne harmonique de la précision et du rappel, offrant ainsi un score global qui prend en compte à la fois la justesse et l'exhaustivité de l'extraction. Cette mesure est particulièrement utile lorsqu'il est nécessaire de trouver un compromis entre ces deux aspects(Maynard et al., 2006).

La mesure F peut être formellement exprimée comme suit :

$$\text{F} - \text{mesure} = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \qquad (3)$$

L'évaluation des systèmes d'extraction d'information géographique repose sur un ensemble de mesures qui permettent de comprendre la performance globale sous différents angles. Tandis que la précision, le rappel et la F-mesure offrent une base solide pour évaluer l'exactitude et la couverture des informations extraites, l'exactitude géométrique et l'évaluation spatiale contextuelle ajoutent une dimension essentielle, propre aux particularités des données géographiques. Ces mesures d'évaluation sont indispensables pour assurer le développement de systèmes d'EIG robustes, capables de répondre aux exigences croissantes des applications géospatiales modernes.

7. Conclusion

Ce premier chapitre a établi les fondements théoriques et conceptuels des Systèmes d'information géographique (SIG), en offrant une analyse approfondie de leurs principales composantes. Une attention particulière a été portée à la modélisation des données géospatiales et à la gestion de l'information spatiale. En outre, la reconnaissance des entités spatiales dans des contextes linguistiques complexes, comme celui de la langue arabe, a été identifiée comme un défi majeur, nécessitant le développement de techniques avancées et d'outils spécialisés. Nous avons montré, à travers l'exploration des méthodes d'extraction d'informations, l'importance croissante de l'automatisation dans domaine géographique, notamment pour les langues sous-représentées comme l'arabe.

Dans le chapitre suivant, nous approfondirons l'étude et l'ingénierie des ontologies, particulièrement les ontologies spatiales, un cadre conceptuel pour la structuration et l'organisation des connaissances géospatiales. Nous présenterons leur rôle essentiel dans l'amélioration des processus d'extraction d'informations spatiales et leur intégration efficace au sein des SIG.

Chapitre 2

Ingénierie de l'Ontologie et
Ontologie Spatiale

1. Introduction

Les ontologies sont de plus en plus étudiées et utilisées dans divers domaines, allant de l'Internet à l'intégration des systèmes d'information, en passant par l'extraction de connaissances, la gestion de la cohérence des bases de données et la recherche d'information.Elles représentent une meilleuretechnique pour décrire et partager des connaissances ainsi que des données entre différents utilisateurs, communautés, ou applications. Les ontologies visent à structurer et formaliser les connaissances d'un domaine spécifique en offrant une représentation standardisée et consensuelle(Vandecasteele et al.,2012).Cette dernières garantissant une compréhension uniforme et en facilitant l'interprétation sémantique des informations échangées. En tant qu'outil incontournable, elles assurent l'interopérabilité des systèmes, optimisent l'intégration des données hétérogènes et renforcent la cohérence conceptuelle, ce qui est essentiel pour des applications avancées en intelligence artificielle, traitement du langage naturel, et systèmes d'information géographique.

Ce chapitre se concentrera sur les concepts fondamentaux liés aux ontologies, offrant une vue d'ensemble qui constituera la base de notre recherche. Il abordera ensuite l'ingénierie des ontologies, ainsi que leur typologie et les méthodes de modélisation associées. Nous examinerons leur rôle dans l'organisation des connaissances, leur contribution à l'amélioration de la cohérence et de l'interopérabilité des systèmes, ainsi que leur utilisation spécifique dans le cadre des informations géospatiales.

.

2.Définitions

Le terme « ontologie » a été introduit dans le domaine de l'informatique, notamment en intelligence artificielle (IA), depuis les années 1990. De nombreuses définitions de ce concept existent dans la littérature. Selon Gruber (1991), « une ontologie est définie comme une représentation formelle des connaissances».En revanche, Uschold et Gruninger(1996)ont décrit une ontologie comme un « vocabulaire et des définitions des concepts d'un domaine ». Gruber (1993)a défini une ontologie comme suit : « An ontologyis an explicit specification of a conceptualization », une définition largement adoptée en ingénierie des connaissances. Cette définition est précisée parBorst(1997),comme étant «An ontologyis a formal, explicit specification of a shared conceptualisation». Selon cette définition :

- Les connaissances dans une ontologie doivent être spécifiées de manière explicite ;
- Le terme « shared » indique un consensus au sein de la communauté concernant le sens des concepts dans un domaine spécifique, permettant ainsi aux membres de la communauté de comprendre ces concepts de manière uniforme ;
- Le terme « formal » implique que l'ontologie doit être représentée dans un langage formel pour que les machines puissent la traiter efficacement.

Charlet (2002) a proposé une définition complémentaire : «Une ontologie est une spécification normalisée représentant les classes d'objets reconnues comme existantes dans un domaine». Construire une ontologie revient également à définir la manière dont les objets de ce domaine existent et interagissent.

En général, une ontologie doit satisfaire deux exigences principales :

- Exigence machine : Fournir une représentation formelle des connaissances que les ordinateurs peuvent exploiter.
- Exigence humaine : Définir une sémantique consensuelle des connaissances pour les utilisateurs humains.

2.1. Conceptualisation

Gruber(1993) et Gruber (1995) se sont référés à la définition de la conceptualisation donnée par Genesereth et Nilsson (1987), qui expliquent : "Une base de connaissances formellement représentée repose sur une conceptualisation, c'est-à-dire les objets, concepts et autres entités considérés comme existant dans un domaine spécifique, ainsi que les relations qui les lient. Une conceptualisation constitue une vision abstraite et simplifiée du monde que l'on souhaite représenter pour un objectif précis. Toute base de connaissances,

tout système de connaissance ou tout agent opérant au niveau de la connaissance est fondé sur une conceptualisation, de manière explicite ou implicite".

2.2. Spécification formelle, explicite et appropriée

Les ontologies jouent un rôle clé en tant que "spécifications explicites des conceptualisations". Elles peuvent être spécifiées de manière extensionnelle, en énumérant les relations pour tous les mondes possibles, ou de manière intentionnelle, en fixant un langage et en restreignant ses interprétations via des axiomes appropriés. Une ontologie consiste en un ensemble de tels axiomes, conçus pour capturer les modèles souhaités d'une conceptualisation et exclure ceux qui ne le sont pas**(**Guarino et al.,2009).

Une ontologie est une théorie logique qui spécifie de manière approximative une conceptualisation, en capturant les modèles souhaités et en excluant ceux qui ne sont pas désirés, tout en s'assurant que cette spécification est formelle, c'est-à-dire lisible par machine(Studer et al.,1998).

Concrètement, une ontologie comprend quatre éléments essentiels : concepts, instances, relations et axiomes.

- **Concepts** : Ce sont les éléments de base d'une ontologie, représentant des classes ou groupes d'objets partageant des propriétés communes. Les concepts sont souvent organisés hiérarchiquement.Dans certains cas, une ontologie peut être composée uniquement de concepts ;
- **Instances** : Ce sont des occurrences spécifiques de concepts.
- **Relations** : Elles expriment les liens sémantiques non taxonomiques entre deux concepts.
- **Axiomes** : Ils imposent des contraintes sur les concepts, leurs instances et leurs relations, comme la contrainte qu'une crête peut dominer plusieurs plateaux. Les axiomes sont formulés dans des langages logiques tels que la logique de description ou la logique du premier ordre et permettent de vérifier la cohérence d'une ontologie.

Une ontologie est qualifiée de légère se limite à des concepts hiérarchiques, leurs instances et les relations associées. Cependant, une ontologie lourde inclut également des axiomes (Figure 2.1).

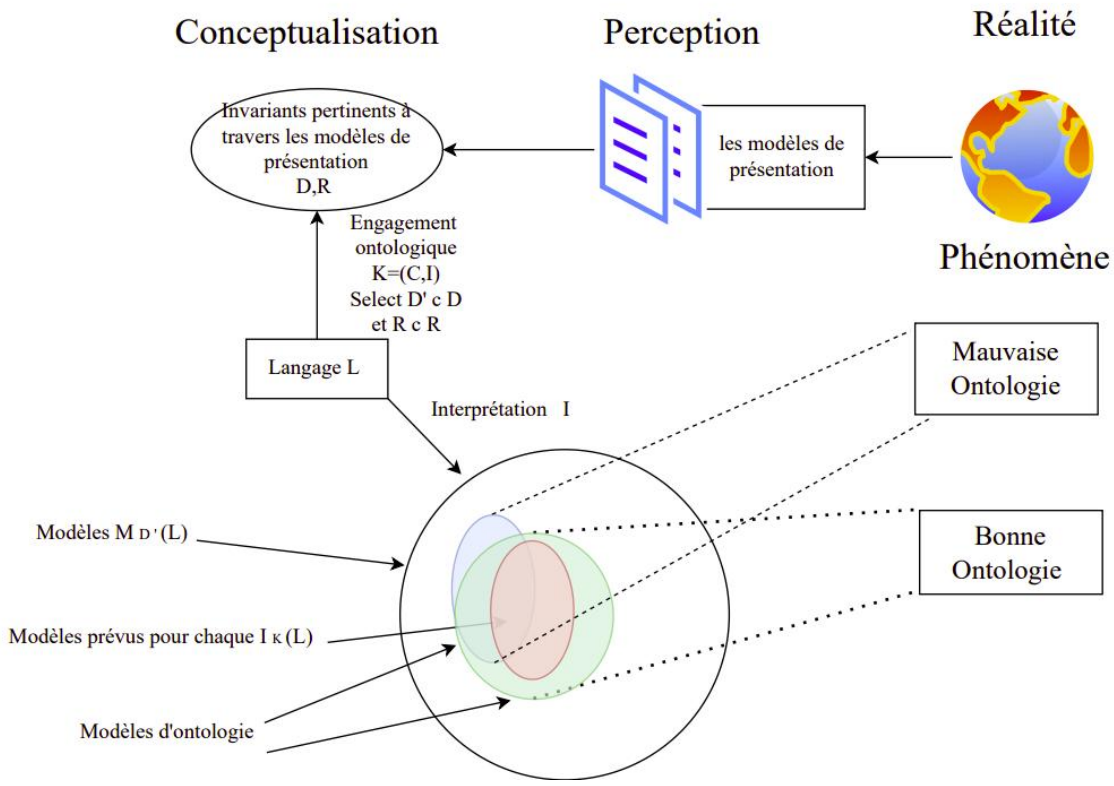

Figure 2.1 : Relations entre la réalité, sa perception, sa conceptualisation abstraite, le langage utilisé, ses modèles, et l'ontologie associée(Guarino et al.,2009).

Exemple d'ontologie

Parmi les exemples les plus connus d'ontologies figurent CYC[1], TOVE[2], WordNet[3], Gene Ontology[4]et les standards du W3C[5]. WordNet[6], développé par le Cognitive Science Laboratory de l'Université de Princeton, est un système de référence lexicale en ligne dont la conception est inspirée par les théories psycholinguistiques contemporaines sur la mémoire lexicale humaine. Il regroupe environ 100 000 mots de la langue anglaise, organisés en ensembles de synonymes représentant des concepts lexicaux sous-jacents.

Ces ensembles sont interconnectés par diverses relations sémantiques, telles que la synonymie, l'antonymie, l'hyponymie et la méronymie, qui structurent l'organisation taxonomique du système.WordNet, en tant qu'ontologie lexicale, offre ainsi une structure riche pour la compréhension des relations sémantiques au sein de la langue Anglaise, tout

[1] http://psych.utoronto.ca/users/reingold/courses/ai/cyc.html
[2] http://www.eil.utoronto.ca/theory/enterprise-modelling/tove/
[3] https://wordnet.princeton.edu/
[4] www. geneontology.org/GO.doc.shtml
[5] https://www.w3.org/TR/owl-features/
[6] https://wordnet.princeton.edu/

en facilitant la modélisation des connaissances dans divers domaines d'application (Lairini et Kazar, 2017).

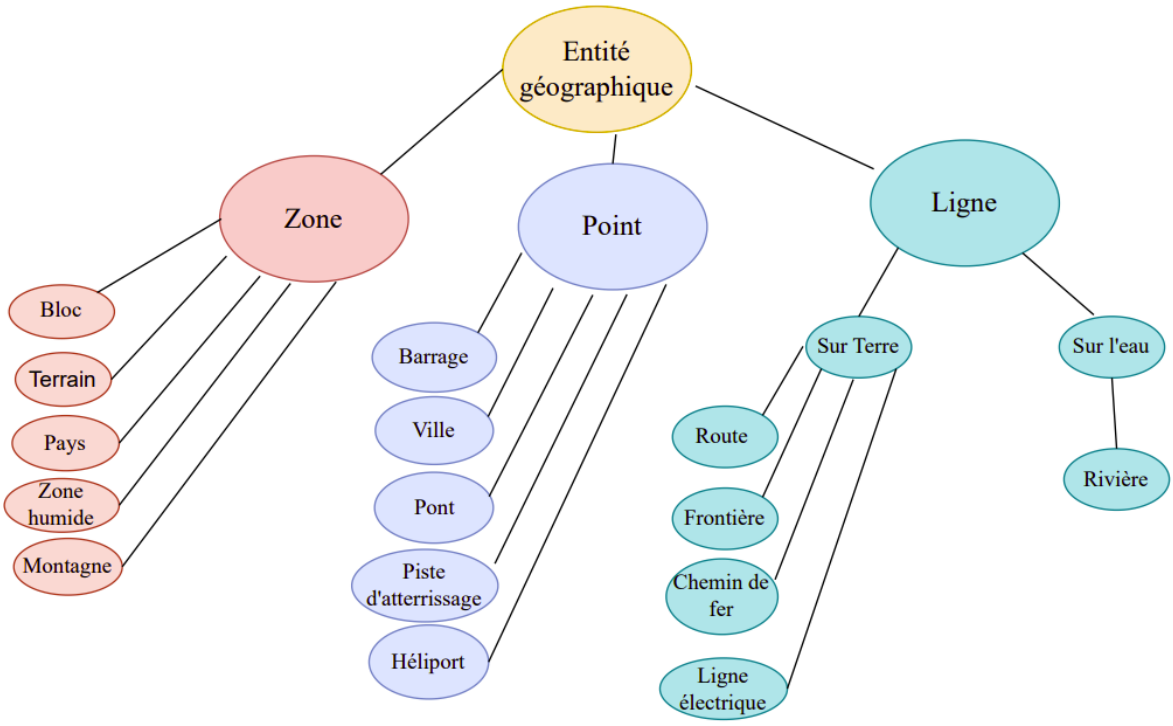

Figure 2.2: Exemple d'ontologies naïves basées sur les types géométriques de caractéristiques. (Lairini et Kazar, 2017).

3. Ingénierie de l'ontologie

L'ingénierie de l'ontologie consiste à modéliser et représenter les connaissances d'un domaine spécifique de manière structurée et compréhensible. Utilisant des langages comme OWL et SPARQL (SPARQL Protocol and RDF QueryLanguage) pour interroger et manipuler ces représentations, elle joue un rôle clé dans l'intégration des données et l'interopérabilité des systèmes. Les ontologies sont essentielles dans divers domaines tels que la recherche d'information, l'intelligence artificielle et la gestion des connaissances. Leur cycle de vie comprend la conception, le développement, l'évaluation et la maintenance, garantissant une gestion continue et efficace des connaissances.

3.1. Représentation de l'ontologie

Les ontologies jouent un rôle central dans la structuration et la formalisation des connaissances d'un domaine spécifique, facilitant ainsi une représentation standardisée et consensuelle indispensable pour la compréhension et l'interprétation sémantique des informations échangées(Gómez-Pérez et Corcho,2002). Pour accomplir cet objectif, divers langages de description d'ontologies ont été développés, notamment à partir des années

1990, dont les plus courants incluent le Resource Description Framework (RDF) et le Web OntologyLanguage (OWL), créés sous l'égide du World Wide Web Consortium (W3C) (Corcho et al., 2003). RDF, écrit en XML, est conçu pour représenter des informations sur les ressources du Web de manière lisible par les machines, en s'appuyant sur un modèle de données simple basé sur des triplets <sujet, prédicat, objet> et l'utilisation d'Identifiants Uniformes de Ressources (URI)(Champin et al., 2001). Quant à OWL, développé en 2001 par le groupe de travail WebOnt, il permet de représenter explicitement la signification des termes dans les vocabulaires et les relations entre eux, enrichissant ainsi les capacités de description des propriétés et des classes dans les documents web.

OWL se décline en trois variantes : OWL Lite, qui prend en charge les hiérarchies de classification et les contraintes simples ; OWL DL, qui garantit une expressivité optimale tout en maintenant la complétude et la décidabilité sur le plan computationnel et OWL Full, qui permet une liberté syntaxique étendue sans garanties computationnelles, s'inscrivant ainsi comme une extension d'RDF (McGuinness et al., 2004).

La Figure 2.3 illustre la hiérarchie et l'évolution de ces langages, reflétant leur importance dans le développement des ontologies modernes et leur application dans divers domaines technologiques.

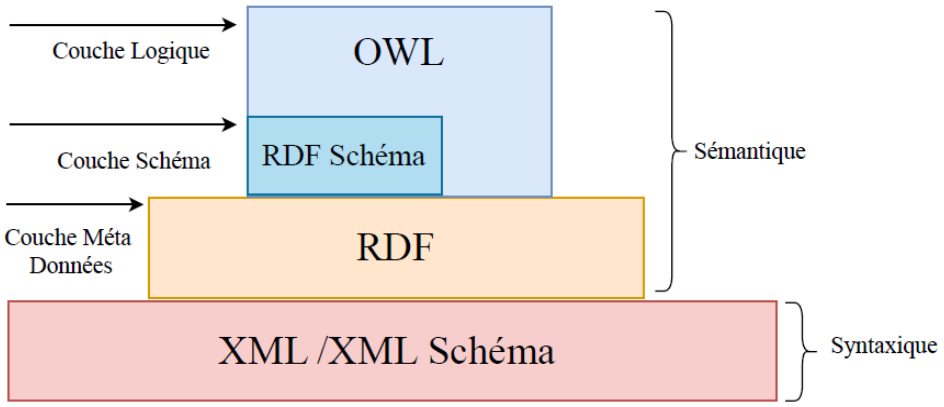

Figure 2.3 : Langage d'ontologie(Tao et al.,2013).

Les Figures 2.4, 2.5, 2.6 et 2.7 illustrentdes exemplesdistincts de la représentation d'une entité spatiale dans les langages XML, RDF, RDFS et OWL. Chaque section est isolée pour illustrer comment l'entité "City" peut être définie et utilisée dans chacun de ces

langages.

- **XML** : Utilisé pour une simple structuration de données sans sémantique formelle.
- **RDF** : Représente les données sous forme de triplets, permettant l'interopérabilité des données sur le web.
- **RDFS** : Fournit une structure hiérarchique avec des classes et des propriétés pour ajouter des relations sémantiques.
- **OWL** : Ajoute des fonctionnalités avancées comme les restrictions et la possibilité d'utiliser des langages de requête pour manipuler les données, tout en fournissant une description plus riche des entités spatiales.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<City>
    <Name>Jijel</Name>
    <Longitude>5.7667</Longitude>
    <Latitude>36.8216</Latitude>
</City>
```

Figure 2.4 : Exemple code XML

```xml
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:ex="http://example.org/ontology#">

    <rdf:Description rdf:about="http://example.org/ontology#Paris">
        <ex:Name>Jijel</ex:Name>
        <ex:Longitude>5.7667</ex:Longitude>
        <ex:Latitude>36.8216</ex:Latitude>
    </rdf:Description>
</rdf:RDF>
```

Figure 2.5 : Exemple code RDF

```rdf
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:ex="http://example.org/ontology#">

    <!-- Définition de la classe City -->
    <rdfs:Class rdf:about="http://example.org/ontology#City">
        <rdfs:label>City</rdfs:label>
        <rdfs:comment>An urban area with defined boundaries.</rdfs:comment>
    </rdfs:Class>

    <!-- Instance de la classe City -->
    <ex:City rdf:about="http://example.org/ontology#Paris">
        <ex:Name>Jijel</ex:Name>
        <ex:Longitude>5.7667</ex:Longitude>
        <ex:Latitude>36.8216</ex:Latitude>
    </ex:City>
</rdf:RDF>
```

Figure 2.6 : Exemple code RDFS

```rdf
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:owl="http://www.w3.org/2002/07/owl#"
         xmlns:geo="http://www.opengis.net/ont/geosparql#"
         xmlns:ex="http://example.org/ontology#">

    <!-- Déclaration de la classe City -->
    <owl:Class rdf:about="http://example.org/ontology#City">
        <rdfs:subClassOf rdf:resource="http://example.org/ontology#SpatialEntity"/>
        <rdfs:label>City</rdfs:label>
    </owl:Class>

    <!-- Définition de l'instance Paris -->
    <ex:City rdf:about="http://example.org/ontology#Paris">
        <rdfs:label>Jijel</rdfs:label>
        <geo:hasGeometry>
            <geo:Geometry>
                <geo:asWKT rdf:datatype="http://www.opengis.net/ont/geosparql#wktLiteral">
                    POINT(5.7667 36.8216)
                </geo:asWKT>
            </geo:Geometry>
        </geo:hasGeometry>
    </ex:City>
</rdf:RDF>
```

Figure 2.7: Exemple code OWL

3.2. Langage SPARQL pour les entités géographiques

SPARQL est un langage de requête puissant utilisé pour interroger des bases de données RDF. Il permet d'extraire et de manipuler des données structurées selon des graphes, facilitant ainsi l'exploration des relations entre différentes entités géographiques ou spatiales. Une requête SPARQL peut être utilisée pour extraire toutes les villes situées à

moins de 50 kilomètres d'un certain point géographique (figure 2.8).

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX ex: <http://example.org/ontology#>

SELECT ?city ?distance
WHERE {
  ?city a ex:City ;
        geo:hasLocation ?location .
  ?location geo:distance ?distance .
  FILTER(?distance <= 50 && ?location = "POINT(6.96 50.94)")
}
```

Figure 2.8 : Exemple code SPARQL

3.3.Importance des ontologies dans divers domaines

Les ontologies apparaissent comme une solution importante pour construire un ensemble de connaissances partagé et réutilisable, qui soutient l'interaction et la compréhension entre l'humain et la machine (Hadji et Kholladi,2012). Leur importance se manifeste particulièrement dans l'extraction de connaissances, l'intégration de systèmes d'information, ainsi que dans d'autres applications complexes nécessitant une compréhension commune des concepts et des relations.

3.3.1. Extraction de connaissances

Dans le domaine de l'extraction de connaissances, les ontologies sont utilisées pour formaliser les concepts et les relations au sein d'un domaine spécifique, ce qui facilite l'identification, l'organisation et la récupération d'informations pertinentes. Elles permettent de passer d'une simple extraction de données brutes à une compréhension plus approfondie et sémantiquement enrichie des informations. Dans le cadre de l'exploration de données ou du textmining, une ontologie peut être utilisée pour identifier des entités spécifiques, établir des relations complexes entre ces entités et interpréter les résultats de manière contextuelle(Fudholiet al.,2016).

3.3.2. Intégration de systèmes d'information

L'intégration de systèmes d'information est un autre domaine où les ontologies sont d'une importance capitale. Les systèmes d'information modernes sont souvent hétérogènes, utilisant des formats de données, des langages et des terminologies variés. Les ontologies servent de pont entre ces systèmes en fournissant un vocabulaire commun et en

standardisant les concepts utilisés. Cela permet non seulement de faciliter la communication et l'interopérabilité entre différents systèmes, mais aussi de garantir la cohérence des informations échangées. Dans des domaines tels que la santé ou l'e-commerce, les ontologies peuvent permettre l'intégration fluide de données provenant de sources diverses, améliorant ainsi la qualité et l'efficacité des services fournis(Ma et Molnár, 2020).

3.3.3. Autres applications

Les ontologies trouvent également leurs importancesdans des domaines tels que la recherche d'information, où elles améliorent la précision des moteurs de recherche en comprenant mieux l'intention de l'utilisateur et le contexte des requêtes. Dans la gestion de bases de données, elles contribuent à maintenir la cohérence des données en explicitant les règles et les contraintes qui régissent les relations entre les entités(Jain et al.,2020).

Les ontologies offrent un cadre essentiel pour la formalisation et la standardisation des connaissances, ce qui est fondamental pour l'extraction de connaissances, l'intégration de systèmes et de nombreuses autres applications qui nécessitent une gestion efficace de l'information. Leur utilisation permet d'améliorer non seulement l'efficacité des systèmes mais aussi la qualité des informations et des services qu'ils fournissent(Jain et al.,2020).

3.4. Rôle des ontologies

- *Représentation des informations et des connaissances*

Les ontologies jouent un rôle central dans la représentation des informations et des connaissances. Elles fournissent une structure formelle pour organiser et clarifier les concepts et les relations au sein d'un domaine spécifique, facilitant ainsi la compréhension et l'utilisation cohérente des données(Costa et al.,2016).

- *Communication et coordination*

Les ontologies améliorent la communication dans les projets complexes en établissant un vocabulaire commun et des définitions précises. Elles facilitent les échanges entre différentes parties prenantes, qu'il s'agisse de la communication homme-homme, homme-système, ou entre les modules d'un système. Cette normalisation aide à résoudre les problèmes de compréhension et de coordination qui peuvent surgir dans ces interactions(Costa et al.,2016).

- *Interopérabilité des systèmes*

En favorisant l'interopérabilité, les ontologies permettent à des systèmes distincts, souvent développés sur des bases technologiques différentes, de partager des informations

de manière efficace. Elles définissent les concepts et les relations de manière à ce que des applications diverses puissent échanger des données malgré des disparités de structure et de format(Kotis et al.,2020).

- *Modularité et réutilisabilité des connaissances*

Les ontologies facilitent la modularité et la réutilisabilité des connaissances en permettant la séparation des conceptualisations du formalisme spécifique utilisé pour les représenter. Elles permettent ainsi de partager et de réutiliser les données de manière indépendante du langage de programmation, de la plateforme et des protocoles de communication, tout en reconnaissant les défis techniques liés à la conception d'ontologies communes(Kotis et al.,2020).

- *Indexation et recherche d'information*

Dans le cadre du Web sémantique, les ontologies jouent un rôle clé dans l'indexation des ressources en ligne. Elles fournissent des index conceptuels qui décrivent les ressources, améliorant ainsi la recherche et la récupération d'information en offrant une compréhension contextuelle des contenus disponibles sur le Web (Asimetal., 2019).

3.5. Cycle de vie d'une ontologie

Le cycle de vie d'une ontologie comprend plusieurs étapes qui assurent son développement et son efficacité (Figure 2.9). La première étape, la capture et l'extraction, consiste à identifier et à extraire les concepts, relations et règles pertinentes à partir de diverses sources de connaissances. Ensuite, lors de l'analyse et la validation, ces éléments sont examinés pour garantir leur cohérence, pertinence et exactitude. Une fois validée, l'ontologie est structurée et stockée dans un format approprié lors de l'étape de stockage. Par la suite, elle est prête d'être utilisée dans des systèmes via la livraison et le transfert, facilitant ainsi son intégration et son exploitation. Enfin, l'étape de raffinement permet d'améliorer et d'adapter continuellement l'ontologie en fonction des nouvelles connaissances ou des évolutions du domaine (Benjamin et al.,2011).

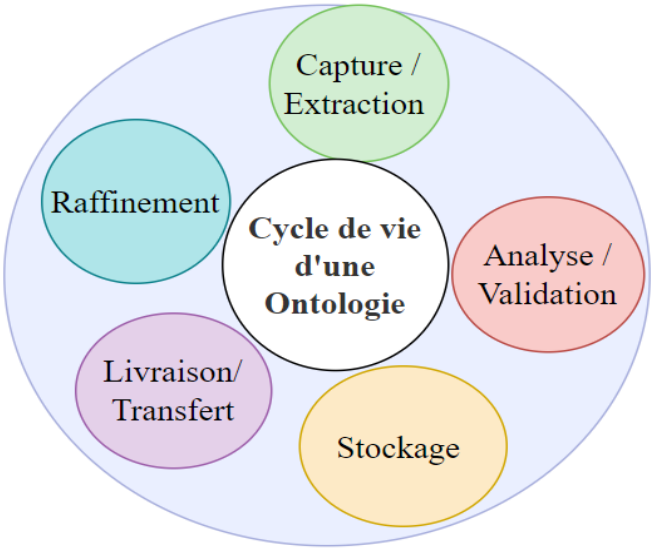

Figure 2.9 : Cycle de vie d'une ontologie (Benjamin et al.,2011).

4.Typologie des Ontologies

4.1.Ontologies de Haut Niveau

Les ontologies de haut niveaufournissent une structure générale et abstraite qui capture les concepts et relations les plus universels et fondamentaux. Elles établissent les bases sur lesquelles d'autres ontologies plus spécialisées peuvent être construites. Leur but est de définir les catégories conceptuelles les plus générales, telles que les entités, les événements, les processus et les relations qui peuvent être appliquées à de nombreux domaines différents. Par exemple, l'ontologie DOLCE[7] (Descriptive Ontology for Linguistic and Cognitive Engineering) est une ontologie fondationnelle qui offre un cadre pour la compréhension des concepts de base et de leurs relations dans divers contextes(McDaniel et Storey,2019).

4.2. Ontologies de Domaine

Les ontologies de domaine se concentrent sur des domaines spécifiques de connaissance ou de pratique. Elles fournissent une modélisation détaillée des concepts, des relations et des règles pertinents pour un domaine particulier, tel que la médecine, l'ingénierie ou l'agriculture. L'ontologie SNOMED CT[8] est une ontologie de domaine utilisée dans le domaine médical pour décrire les termes médicaux et leurs relations, facilitant ainsi l'interopérabilité et l'échange d'informations dans les systèmes de

[7] https://www.danieleporello.net/papers/BorgoEtAlAO2022.pdf
[8] https://www.snomed.org/value-proposition?lang=fr

santé(McDaniel et Storey,2019).

4.3.Ontologies de Tâche

Les ontologies de tâche sont conçues pour modéliser les activités et les processus nécessaires pour accomplir des tâches spécifiques dans un contexte donné. Elles décrivent les étapes, les procédures, les rôles et les interactions impliqués dans l'exécution d'une tâche particulière. Ces ontologies sont souvent utilisées pour structurer les connaissances nécessaires à l'automatisation des tâches ou à la gestion des processus. Une ontologie de tâche pourrait décrire les étapes impliquées dans la gestion d'un projet de recherche, incluant la planification, l'exécution et l'évaluation(McDaniel et Storey, 2019).

4.4. Ontologies d'Application

Les ontologies d'application sont adaptées à des besoins spécifiques d'application ou de systèmes particuliers. Elles intègrent des concepts et des relations qui sont directement pertinents pour une application ou un ensemble d'applications spécifiques, facilitant ainsi la communication et l'interopérabilité entre différents systèmes ou services dans un contexte donné. Une ontologie d'application pour un système de gestion de contenu pourrait inclure des concepts comme les articles, les utilisateurs, les commentaires, et leurs relations pour permettre une meilleure organisation et recherche de contenu au sein du système(Tapia-Leon et al., 2019).

Ces définitions fournissent un cadre pour comprendre comment les ontologies peuvent être utilisées pour structurer et organiser les connaissances dans divers contextes, allant des principes fondamentaux aux applications spécifiques (Figure 2.10).

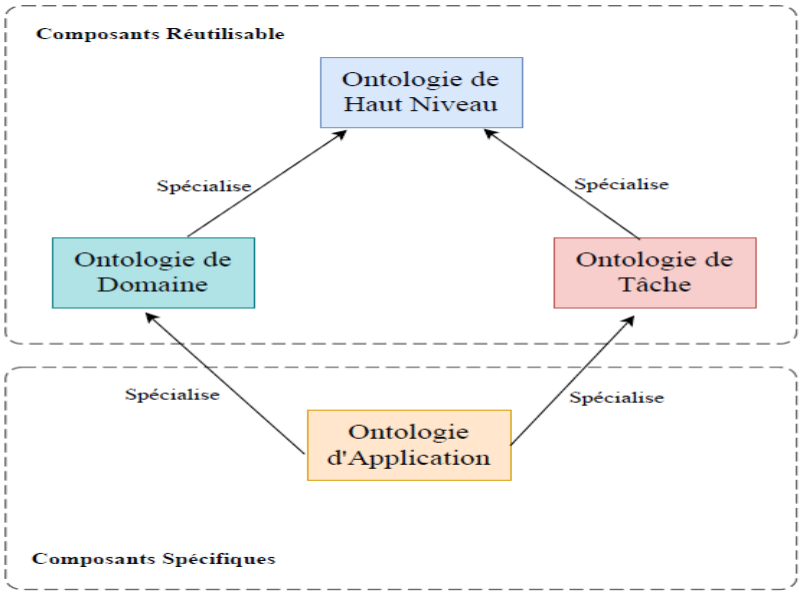

Figure2.10 : Relations entre les différentes type d'ontologies(Lando, 2006).

5. Méthodes de Modélisation d'Ontologies

5.1. Méthodologie METHONTOLOGY

La méthodologie METHONTOLOGY, élaborée par Fernández-López et al. (1997), constitue une approche complète et systématique pour le développement d'ontologies. Cette méthodologie est largement connue pour son cadre robuste, qui intègre trois types d'activités essentielles : les activités de développement, les activités de gestion, et les activités de soutien. Chacune de ces catégories joue un rôle critique dans le processus global de la création d'une ontologie de haute qualité (Figure 2.11).

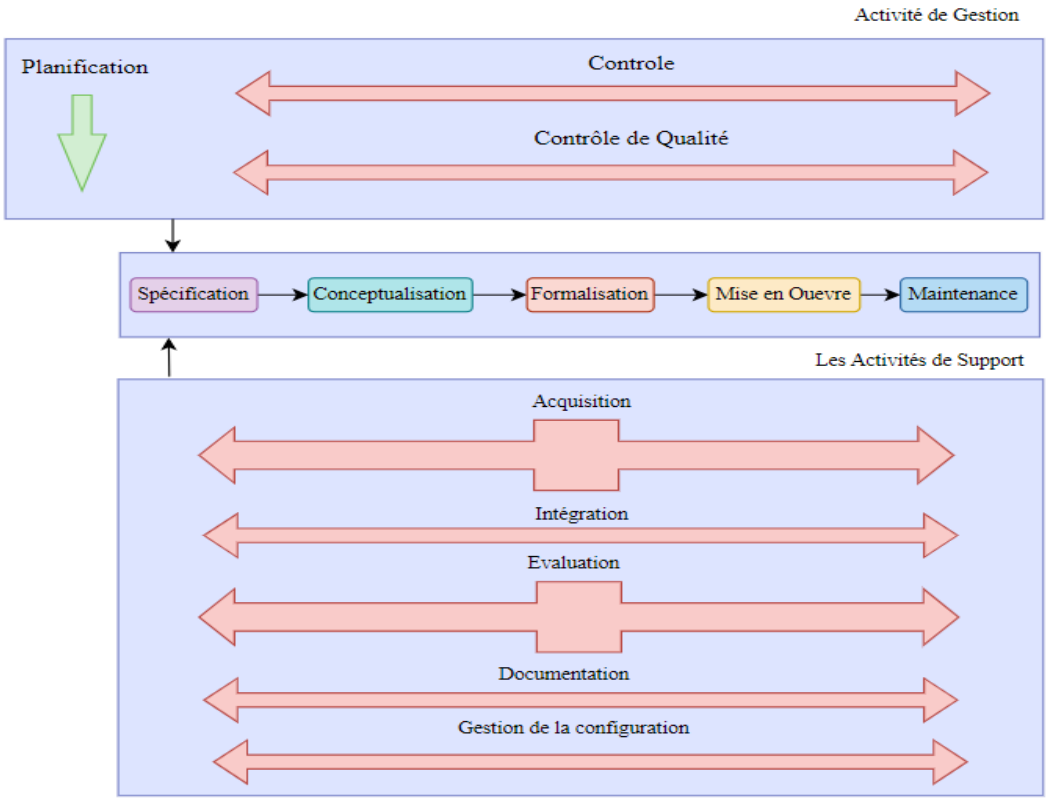

Figure 2.11 : Méthode de ModélisationMETHONTOLOGY(Fernández-López et al.,1997).

Les activités de développement constituent le noyau central de la méthodologie METHONTOLOGY, se déroulant en plusieurs phases clés :

- **Spécification** : Cette phase initiale définit les objectifs et la portée de l'ontologie, en établissant les compétences couvertes. Elle vise à identifier les besoins et les

limitations en adéquation avec les objectifs du projet ;

- **Conceptualisation** : Cette étape organise les connaissances en concepts fondamentaux, incluant :
  - La création d'un glossaire précis des termes, avec définitions, synonymes et acronymes ;
  - La classification des termes en taxonomies, structurant ainsi l'information ;
  - La définition des relations binaires entre concepts, établissant un réseau sémantique cohérent ;
  - L'élaboration d'un dictionnaire de concepts, détaillant attributs et propriétés ;
  - La précision des relations, incluant cardinalité, relations inverses et autres propriétés ;
  - La formalisation des axiomes et des règles, encadrant les logiques de l'ontologie.
- **Formalisation** : Cette étape traduit les concepts en un langage formel, facilitant leur implémentation informatique et leur manipulation automatique ;
- **Mise en œuvre** : L'ontologie est ensuite déployée dans un environnement logiciel spécifique, permettant son utilisation dans des applications pratiques ;
- **Maintenance** : Cette phase garantit l'actualisation continue de l'ontologie, assurant sa pertinence et son exactitude dans le temps.

La METHONTOLOGY se distingue par son approche exhaustive, intégrant à la fois les étapes de développement et les aspects de gestion, en faisant une méthodologie de référence dans le domaine de l'ingénierie ontologique.

5.2.Méthode ON-TO-KNOWLEDGE

La méthodologie ON-TO-KNOWLEDGE (OTK), détaillée par Sure et al. (2003),constitue une approche innovante et influente pour la modélisation d'ontologies. Elle s'appuie sur des propositions méthodologiques antérieures, notamment celles de Uschold et King (1995), ainsi que de la méthodologie METHONTOLOGY, pour offrir un cadre méthodologique robuste et adaptable. Le processus d'OTK est structuré en cinq étapes principaleset se distingue par son approche itérative et cyclique, permettant une évolution continue et un raffinement progressif de l'ontologie (Figure 2.12).

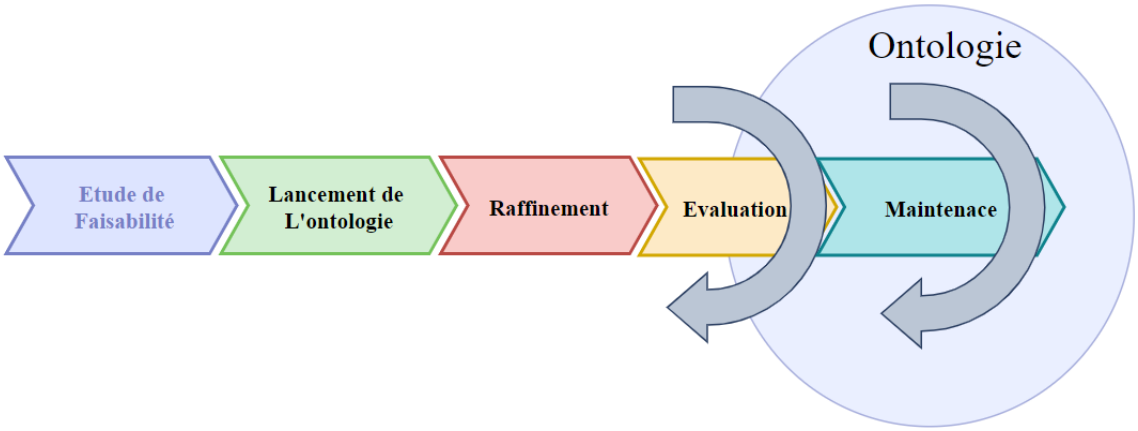

Figure 2.12 : Méthode de modélisationON-TO-KNOWLEDGE(Sure et al., 2003).

5.2.1. Étude de faisabilité

Cette phase initiale lance le développement de l'ontologie, comportant plusieurs sous-étapes:

- Définition du but, du domaine et de la portée de l'ontologie ;
- Établissement des directives de conception pour orienter les choix méthodologiques ;
- Identification et analyse des sources de connaissances, constituant un lexique préliminaire ;
- Formulation de questions de compétence pour valider la pertinence de l'ontologie ;
- Création d'une description semi-formelle, servant de premier jet pour l'évolution future de l'ontologie.

5.2.2. Raffinement

Cette étape est un processus itératif d'acquisition et de formalisation des connaissances, incluant :

- Collaboration avec des experts pour l'élicitation des connaissances, basée sur les résultats de la phase de lancement ;
- Modification ou extension de l'ontologie préliminaire pour mieux refléter les nouvelles connaissances acquises ;
- Formalisation des concepts, les transformant en structures formelles prêtes à être implémentées.

5.2.3. Évaluation

L'ontologie est soumise à une évaluation rigoureuse pour vérifier sa cohérence, sa complétude et son efficacité par rapport aux objectifs initiaux.

5.2.4. Application

L'ontologie est ensuite appliquée dans des contextes réels, permettant de tester sa robustesse et son utilité dans des situations pratiques.

La méthode ON-TO-KNOWLEDGE se distingue par son processus de raffinement itératif et son intégration continue des retours d'expérience, garantissant une ontologie à la fois rigoureuse, formellement solide et adaptée aux besoins spécifiques des utilisateurs finaux.

5.3. Méthode ARCHONTE

La méthode ARCHONTE (ARCHitecture for ONTologicalElaborating)développée par Bachimont et al. (2002) est une méthodologie innovante pour la construction d'ontologies directement à partir de textes(Figure 2.13). ARCHONTE s'inscrit dans une tradition de recherche visant à systématiser l'extraction de concepts et de relations ontologiques à partir de corpus textuels, facilitant ainsi la formalisation des connaissances implicites contenues dans les documents écrits. Cette méthode a été largement adoptée et adaptée dans plusieurs travaux de recherchequi ont démontré son efficacité et sa pertinence dans divers domaines d'application(Baneyx, 2007 ; Charlet et al.,2012).

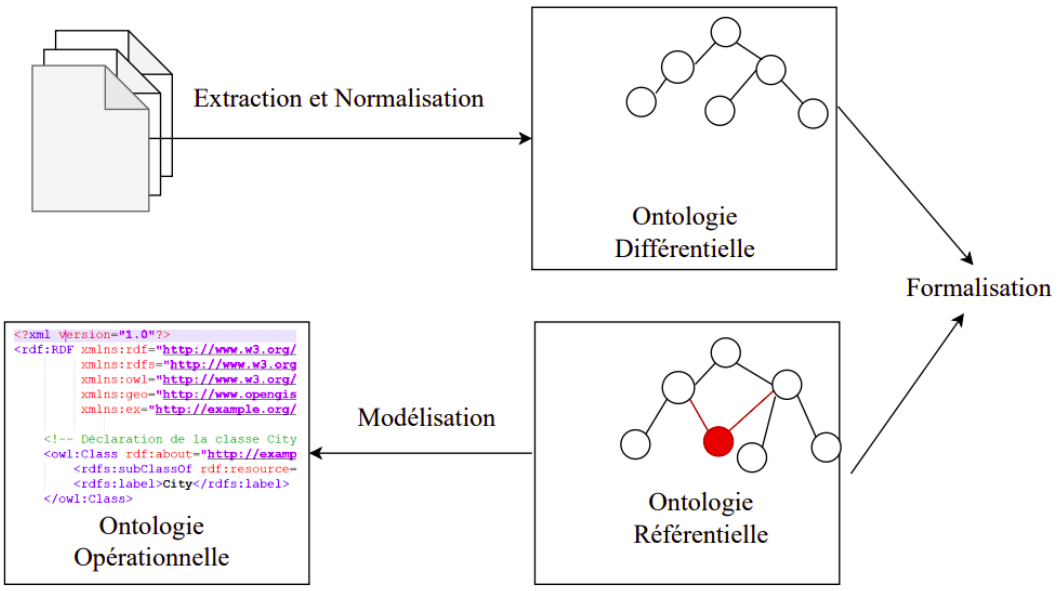

Figure 2.13 : Méthode de Modélisation ARCHONTE (Bachimont et al., 2002).

Les étapes de la méthode ARCHONTE peuvent être synthétisées comme suit :

Extraction et normalisation : Cette première étape consiste à analyser le texte pour extraire les concepts et les relations essentiels. Les concepts identifiés sont ensuite normalisés afin de garantir leur cohérence et leur adéquation aux standards terminologiques du domaine d'étude.

Formalisation : Après l'extraction et la normalisation, les concepts et les relations sont formalisés en structures ontologiques. L'objectif est de créer une représentation structurée et formelle des connaissances qui peut être facilement interprétée et exploitée par des systèmes informatiques.

Modélisation : Enfin, les éléments formalisés sont intégrés dans un modèle ontologique cohérent. Cette modélisation vise à organiser les concepts et les relations dans un cadre ontologique global qui reflète fidèlement la structure des connaissances du domaine étudié. Ce modèle ontologique sert ensuite de fondation pour diverses applications, telles que l'inférence, la recherche d'information ou d'autres processus cognitifs.

La méthode ARCHONTE offre une approche rigoureuse et méthodique pour la construction d'ontologies à partir de textes, garantissant que les ontologies résultantes sont étroitement alignées avec les informations textuelles d'origine. ARCHONTE permet de capturer, de formaliser et de modéliser les connaissances implicites contenues dans les textes, rendant ces connaissances accessibles et exploitables pour une variété d'applications.

5.4. Méthode de modélisation selon Uschold& King

Uschold et King (1995) ont proposé une méthodologie claire et structurée pour la modélisation ontologique, divisant le processus en étapes clés (Figure 2.14).

- **Définition de l'objectif** : Cette étape initiale vise à clarifier le but et le domaine d'application de l'ontologie, orientant ainsi l'ensemble du processus ;
- **Construction de l'ontologie** : Elle se décompose en plusieurs sous-étapes essentielles :
    - **Capture** : Identification et définition des concepts et relations du domaine ciblé ;
    - **Codage** : Formalisation des éléments capturés en un langage d'ontologie approprié ;
    - **Intégration** : Alignement avec des ontologies existantes pour garantir cohérence et interopérabilité ;

- **Évaluation** : Vérification rigoureuse de la précision et de la cohérence de l'ontologie par le biais de tests et de revues expertes ;
- **Documentation** : Élaboration d'une documentation détaillée pour faciliter la maintenance et l'utilisation future de l'ontologie ;

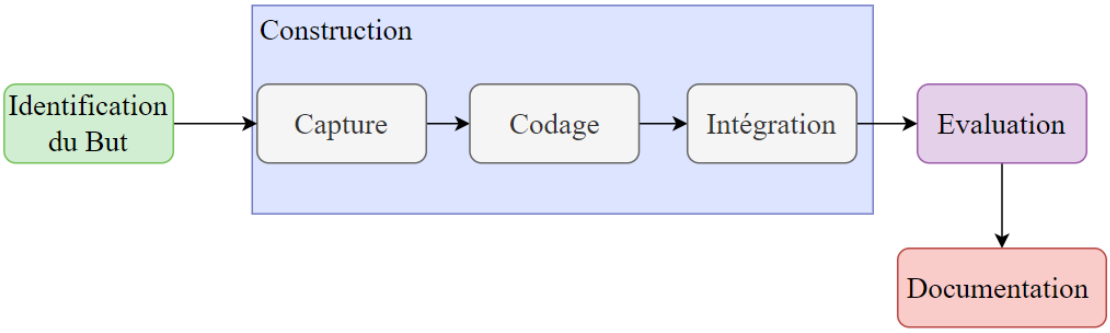

Figure 2.14 : Méthode de modélisationUschold& King(Uschold et King,1995).

6. Avantages et limites des ontologies

6.1. Avantages

6.1.1. Cohérence

Les ontologies assurent une cohérence dans la représentation des connaissances en définissant des concepts et des relations de manière rigoureuse et structurée. En établissant un cadre formel pour la modélisation des connaissances, les ontologies réduisent les ambiguïtés et les contradictions potentielles dans les données. Cette cohérence est importantepour garantir que les informations sont interprétées de manière uniforme, quel que soit le contexte d'utilisation, ce qui est particulièrement important dans des domaines comme la gestion des connaissances et les systèmes d'information(Uschold et Gruninger,1996).

6.1.2. Interopérabilité

Les ontologies favorisent **l'**interopérabilité entre différents systèmes et applications en fournissant un langage commun pour la représentation des connaissances. En utilisant des ontologies standardisées, les systèmes hétérogènes peuvent partager et comprendre les données de manière plus efficace. Dans le domaine de la santé, des ontologies telles que SNOMEDCT permettent aux systèmes cliniques, aux bases de données de recherche et aux

outils d'analyse de communiquer et d'échanger des informations de manière cohérente, facilitant ainsi la collaboration et l'intégration des données à travers les systèmes(Kotis et al.,2020).

6.1.3. Réutilisabilité

Les ontologies offrent une grande réutilisabilité des connaissances en permettant leur application dans différents contextes et domaines. Une fois qu'une ontologie est développée, elle peut être adaptée ou étendue pour répondre à de nouveaux besoins sans avoir à reconstruire les bases de connaissances à partir de début. Cette réutilisabilité permet de tirer parti des travaux antérieurs et des modèles éprouvés, réduisant ainsi le temps et les efforts nécessaires pour développer de nouvelles solutions. Une ontologie de Haut Niveau peut être utilisée comme base pour créer des ontologies spécifiques à des domaines particuliers, comme l'agriculture ou la géographie.

6.2. Limites

6.2.1. Complexité de Construction

La complexité de construction desontologies est un défi majeur, en particulier lorsqu'il s'agit de modéliser des domaines complexes ou de grande envergure. La création d'une ontologie nécessite une compréhension approfondie du domaine, une modélisation précise des concepts et des relations, et une gestion rigoureuse des exceptions et des variations. Ce processus peut être long et nécessiter l'expertise de spécialistes du domaine ainsi que de modélisateurs d'ontologies. De plus, les ontologies doivent être régulièrement mises à jour pour refléter les évolutions du domaine, ce qui ajoute une couche supplémentaire de complexité(Slimani,2015).

6.2.2. Hétérogénéité des données

L'hétérogénéité des donnéesreprésente un autre défi important pour les ontologies. Les données provenant de différentes sources peuvent varier en termes de format, de qualité et de terminologie. Les ontologies doivent être conçues pour gérer cette diversité en offrant des mécanismes pour l'intégration et l'harmonisation des données. Cependant, cette tâche peut être complexe et nécessiter des processus de normalisation et de transformation pour aligner les données hétérogènes avec l'ontologie(Yunianta et al., 2014).

6.2.3. Évolutivité

L'évolutivité des ontologies pose un problème lorsqu'il s'agit de gérer des domaines en constante évolution ou d'intégrer de nouvelles connaissances. À mesure que de nouveaux concepts et relations apparaissent, l'ontologie doit être ajustée pour les inclure

tout en maintenant la cohérence et la validité du modèle existant. Cela peut entraîner des défis liés à la mise à jour de l'ontologie et à la gestion des versions, surtout lorsque l'ontologie est utilisée par plusieurs parties prenantes ou intégrée dans des systèmes complexes(Ekaputra,2017).

7. Ontologies Spatiales

7.1. Définition et particularités des ontologies spatiales

Les ontologies spatiales sont des structures formelles qui représentent et organisent les concepts, les relations et les règles liés aux informations géospatiales. Elles capturent les connaissances spécifiques aux représentations spatiales, y compris les entités géographiques, leurs attributs et les relations spatiales entre elles. Contrairement aux ontologies générales, les ontologies spatiales intègrent des notions spécifiques telles que la localisation, la forme et la taille, ainsi que les relations spatiales comme la proximité, l'intersection et l'inclusion. Ces ontologies sont conçues pour modéliser les phénomènes géographiques de manière précise, facilitant ainsi une compréhension commune et l'échange d'informations entre différents systèmes et utilisateurs.

Le terme « ontologie géographique » regroupe deux disciplines et concepts distincts : celui des ontologies et celui de la géographie dans un sens large(Figures 2.15). Cependant, Agarwal (2005)a souligné que les systèmes d'information géographique (SIG) et les ontologies se concentrent principalement sur les forces de chaque domaine, sans véritablement créer une discipline commune. Il a été recommandé d'adopter GeoRSS[9] pour décrire les propriétés géospatiales des ressources Web (Lieberman, 2007). Ce format, inspiré de GML mais simplifié, offre une représentation spatiale simplifiée mais avec une sémantique limitée. De plus, malgré la formalisation des entités spatiales, les travaux sur les procédures de raisonnement spatial dans les ontologies restent rares (Karmacharya et al., 2010).

[9] https://georss.org/

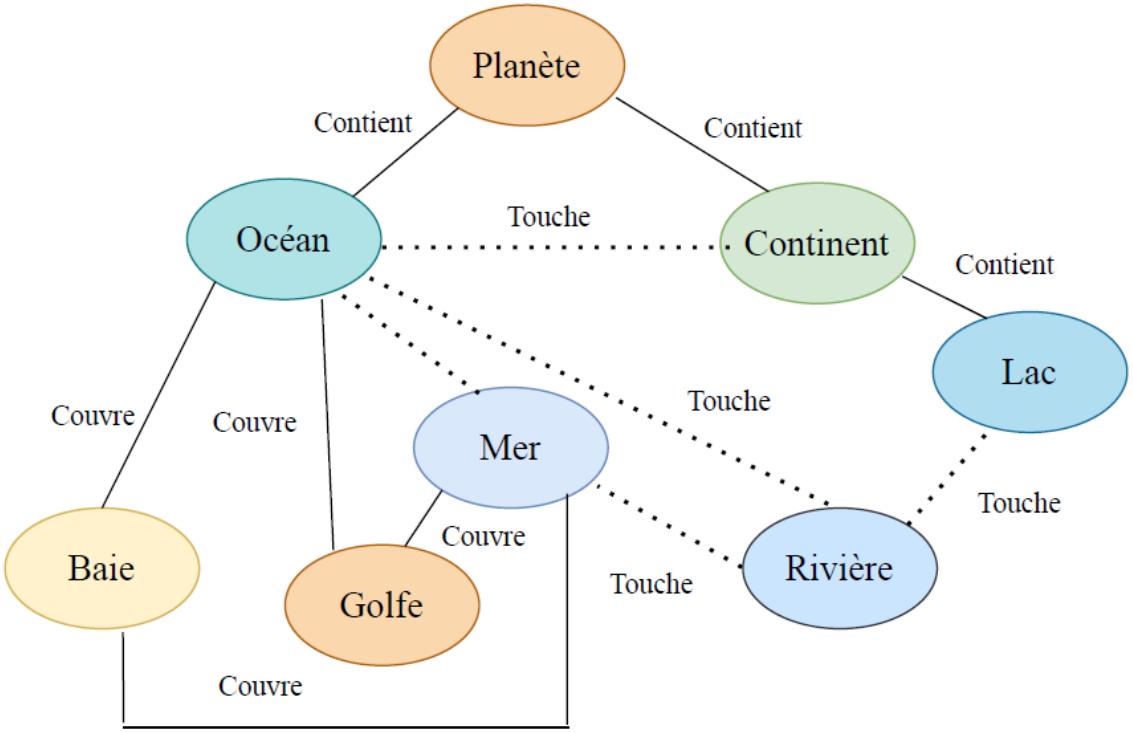

Figure2.15 : Exemple d'ontologie basée sur les relations spatiales(Laurini, 2015).

7.2. Importance du partage des données spatiales et les défis associés

Le partage des données spatiales est essentiel pour de nombreux domaines, allant de la gestion des catastrophes à la planification urbaine. Une ontologie spatiale bien conçue permet de standardiser la manière dont les données géographiques sont représentées et interprétées, facilitant ainsi leur intégration et leur utilisation dans divers contextes. Toutefois, le partage des données spatiales pose plusieurs défis.

Ces défis incluent la gestion de la diversité des formats de données, des systèmes de projection et des référentiels géodésiques. De plus, les questions liées à la confidentialité et à la qualité des données doivent être abordées pour assurer que les informations partagées soient fiables et pertinentes. Les ontologies spatiales aident à surmonter ces obstacles en fournissant un cadre commun pour la représentation des données et en facilitant la conversion entre différents formats et systèmes(KalantariOskouei et al.,2018).

7.3.Rôle des ontologies dans la gestion et l'intégration des données spatiales

Les ontologies spatiales sont largement utilisées dans la gestion et l'intégration des données spatiales en offrant une base de connaissances structurée et cohérente. Elles

permettent de décrire les entités géographiques, leurs attributs, et leurs relations de manière standardisée, facilitant ainsi l'intégration de données provenant de sources diverses. Dans un système d'information géographique, les ontologies spatiales peuvent être utilisées pour intégrer des données provenant de différentes bases de données géospatiales, en alignant les concepts et les relations définis dans chaque source de données. Cela permet de créer une vue unifiée des données spatiales, améliorant leur accessibilité et leur utilisation dans des applications telles que l'analyse spatiale, la cartographie et la prise de décision(Hasani et al.,2015).

7.4. Ambiguïté spatiale ou géographique

Le nom d'un lieu ne peut pas être utilisé comme identifiant unique dans un contexte informatique (Figure 2.16). Pour clarifier ce point, ci-dessous quelques définitions essentielles(Atoui, 1996) :

- **Toponyme** : Le nom général attribué à une entité géographique ;
- **Endonyme** : Un nom local utilisé dans la langue officielle du pays ou dans une langue bien établie dans la région où l'entité se trouve. Plusieurs toponymes peuvent coexister dans des pays avec différentes langues officielles.
- **Exonyme** : Un nom employé dans des langues autres que les langues officielles, comme Brussels en anglais ;
- **Archéonyme** : Un nom qui était utilisé dans le passé, tel que Byzance pour désigner Istanbul ;
- **Hyperonyme et hyponyme** : Des termes utilisés pour décrire des lieux dans un contexte hiérarchique. L'hyponyme est le contraire de l'hyperonyme : l'Europe est un hyperonyme de la France, tandis que la France est un hyponyme de l'Europe ;
- **Méronyme** : Un nom désignant une partie d'un lieu sans hiérarchie. Parfois, le terme "partonyme" est utilisé, par exemple, la "mer Adriatique" est un méronyme de la mer Méditerranée ;
- **Hydronyme** : Un nom désignant une étendue d'eau ;
- **Oronyme** : Un nom désignant une colline ou une montagne.

Ces définitions permettent de mieux saisir la complexité et la diversité des noms géographiques et leur usage dans différents contextes linguistiques et historiques.

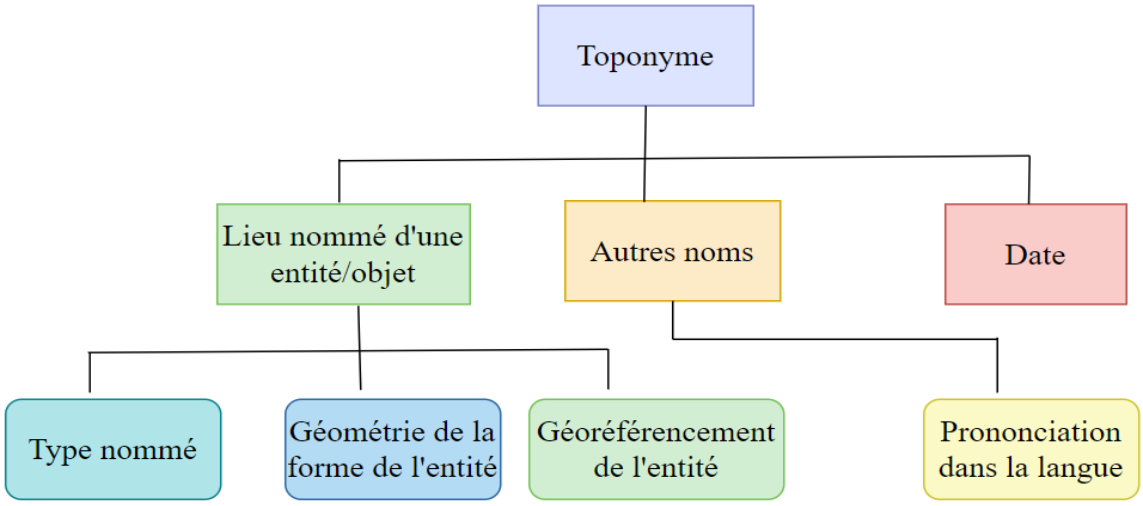

Figure 2.16: Elémentsessentiels de la toponymie(Jakir et al., 2011).

L'ambiguïté spatiale ou géographique dans la langue Arabe constitue un défi significatif dans le domaine du traitement automatique du langage naturel (TALN) et des systèmes d'information géographique (SIG). Cette ambiguïté survient lorsque des termes ou expressions utilisés pour décrire des positions, directions, distances ou entités géographiques peuvent être interprétés de différentes manières en fonction du contexte, menant ainsi à des imprécisions ou des malentendus.

- **Ambiguïté Lexicale :** Dans la langue Arabe, un mot peut avoir plusieurs significations selon son usage(Omar et Aldawsari,2020). Le mot "عين"peut signifier "source d'eau" (entité géographique) ou "œil" (partie du corps)
- **Ambiguïté contextuelle :** L'ambiguïté contextuelle en Arabe est liée à la dépendance du sens des mots au contexte environnant(Bounhas et al., 2011).L'expression "شرق"(Est) peut désigner à la fois une directioncardinale ou une région spécifique ;
- **Ambiguïté syntaxique :** La structure syntaxique de la langue Arabe peut contribuer à l'ambiguïté(Nazari, 2016).Par exemple, la phrase"على الطريق" peut se traduire par "sur la route" ou "au bord de la route". La préposition "على" peut être interprétée différemment selon la structure de la phrase, ce qui peut mener à différentes interprétations spatiales ;
- **Ambiguïté des toponymes :** Les toponymes en Arabe peuvent également prêter à confusion. Certains noms de lieux peuvent être communs à plusieurs endroits

différents(Atoui, 1996). Par exemple, **"عين" (Ain)** est un nom de lieu très courant et peut désigner plusieurs localités dans différentes régions. Sans informations supplémentaires, un système automatique pourrait ne pas être en mesure de distinguer de quel lieu il s'agit.

7.5.Résolution de l'ambiguïté spatiale

Pour traiter cette ambiguïté dans les systèmes SIG et les applications de TALN, plusieurs techniques peuvent être utilisées :

- **Désambiguïsation lexicale :** Utilisation de méthodes statistiques ou basées sur le contexte pour déterminer le sens correct d'un mot ambigu (Elayeb,2019) ;
- **Ontologies spatiales :** L'utilisation d'ontologies peut aider à structurer et à clarifier les concepts spatiaux et géographiques, réduisant ainsi l'ambiguïté (Kauppinen,2008) ;
- **Analyse contextuelle avancée** : L'intégration de l'analyse contextuelle, y compris l'utilisation des métadonnées ou des informations additionnelles pour interpréter correctement les termes spatiaux(Spinsanti et Ostermann,2013);
- **Utilisation de corpus annotés :** Les corpus annotés manuellement, où les relations spatiales et les entités géographiques sont clairement étiquetés, peuvent être utilisés pour entraîner des modèles de machine learning capables de mieux gérer l'ambiguïté(Kordjamshidi et al.,2017).

7.6.Éditeurs d'ontologie

Les éditeurs d'ontologie sont des outils offrent une interface ergonomique pour la création et la modification de classes, propriétés et relations au sein d'une ontologie, tout en assurant la cohérence et l'intégrité du modèle. Les éditeurs d'ontologie les plus couramment utilisés incluent Protégé[10], OntoStudio[11] et TopBraid Composer[12], chacun ayant des caractéristiques distinctes adaptées à différents besoins d'utilisation.

7.7.Ontologie et extraction d'information

L'utilisation d'ontologie dans l'extraction d'information présente un rôle important, particulièrement dans les contextes où la compréhension contextuelle et la précision sémantique des données sont nécessaires. En définissant clairement les concepts et les relations au sein d'un domaine spécifique, l'ontologie permet de structurer et de formaliser les connaissances de manière cohérente. Cela facilite l'extraction d'informations pertinentes à partir d'une masse de données, en garantissant des résultats précis, interprétables et exploitables.

[10] http://protege.stanford.edu/
[11] http://www.semafora-systems.com/en/products/ontostudio/
[12] https://allegrograph.com/topbraid-composer/

Dans ce processus, les ontologies agissent comme des références pour identifier, classer et relier les entités et les relations présentes dans les données textuelles ou multimodales, ce qui renforce la capacité des systèmes intelligents à extraire des informations du contexte significatives. Elles se révèlent ainsi indispensables pour améliorer l'efficacité des systèmes d'extraction d'information, en réduisant l'ambiguïté et en augmentant la pertinence des données extraites(Konys, 2018).

8. Conclusion

En conclusion, ce chapitre a fourni une vue d'ensemble approfondie des ontologies, en soulignant leur importance dans la structuration et la gestion des connaissances. À travers l'exploration des définitions, de la typologie et des méthodes de modélisation, les bases nécessaires à la compréhension de leur conception et de leur utilisation ont été établies. Bien que les avantages des ontologies, comme la cohérence, l'interopérabilité et la réutilisabilité des connaissances, soient indéniables, des défis tels que la complexité de leur construction et la gestion des données hétérogènes demeurent.

Ce chapitre pose ainsi les fondements pour les discussions ultérieures, qui se concentreront sur l'application des ontologies dans des domaines spécifiques, notamment l'extraction des informations spatiales. Ces concepts de base fourniront un cadre essentiel pour des analyses plus détaillées dans les chapitres suivants.

Dans le chapitre suivant, nous aborderons la partie dédiée aux contributions, où nous allons se focaliser sur la création de l'ontologie ASTO. Ce processus mettrait en œuvre les concepts discutés précédemment et présenterait une ontologie spécifique adaptée à la gestion des données spatiales, démontrant ainsi son efficacité dans des cas pratiques.

Chapitre 3

Développement et Création de l'Ontologie ASTO

1. Introduction

Avec l'augmentation continue de la disponibilité des informations et le coût élevé associé à l'acquisition des données spatiales, le partage de ces dernières devient un enjeu majeur. De plus, les données spatiales sont souvent complexes, imprécises et de résolutions hétérogènes, ce qui rend leur gestion et leur intégration particulièrement difficiles. Les ontologies spatiales, en fournissant un cadre structuré pour l'organisation, l'interprétation et le partage des données spatiales, sont donc essentielles pour l'intégration dans les systèmes d'information géographique (SIG), le contrôle de la cohérence des données et l'assistance à la conception (Lietal., 2012).

Dans ce chapitre, nous allons détailler les différentes étapes de la construction de l'ontologie Arabic Spatial Toponymy Ontology (ASTO). Nous allons commencer par la phase de conception, où les objectifs et les besoins seront définis. Ensuite, nous allons spécifier les concepts et les relations nécessaires avant de procéder à la création ou à la réalisation de l'ontologie proprement dite. Enfin, nous allons effectuer des tests rigoureux et une validation pour assurer la précision, la cohérence et l'efficacité de l'ontologie dans des scénarios pratiques dans le domaine de l'extraction des informations spatiales en langue Arabe.

2. Ontologie ASTO

Les ontologies sont essentielles pour développer un ensemble de connaissances partagées et réutilisables, facilitant ainsi l'interaction, l'interopérabilité, et l'intégration entre différents systèmes (Jones et al., 2001). L'extraction automatique des relations ontologiques à partir des textes est nécessaires pour représenter les documents et leur contenu de manière informatisée et lisible par des machines (Ping et Yong, 2009).D'un

point de vue philosophique, l'ontologie est une discipline qui étudie la nature et l'organisation de l'être. Avec l'évolution des technologies de l'information, l'ontologie a reçu une nouvelle définition : c'est une spécification formelle d'une conceptualisation partagée d'un domaine, représentant les concepts dans un format compréhensible par les humains et lisible par les machines, et se composant d'entités, de valeurs, de relations et d'axiomes (Guarinoetal., 2009).

L'ontologie est appliquée dans divers domaines du traitement de l'information, tels que l'ingénierie des connaissances, les bibliothèques numériques, la réutilisation des logiciels, l'extraction d'informations et le web sémantique (Buitelaar et al., 2006).

Pour la construction de l'ontologie ASTO, nous suivons la méthodologie proposée par Noy et McGuinness (2001).Ce processus est divisé en plusieurs étapes : définition du domaine ; construction de la taxonomie ; construction de l'ontologie ; description formelle ; création des classes ; processus de raisonnement ; vérification de la cohérence ; jeu de résultats (Figure 3.1).

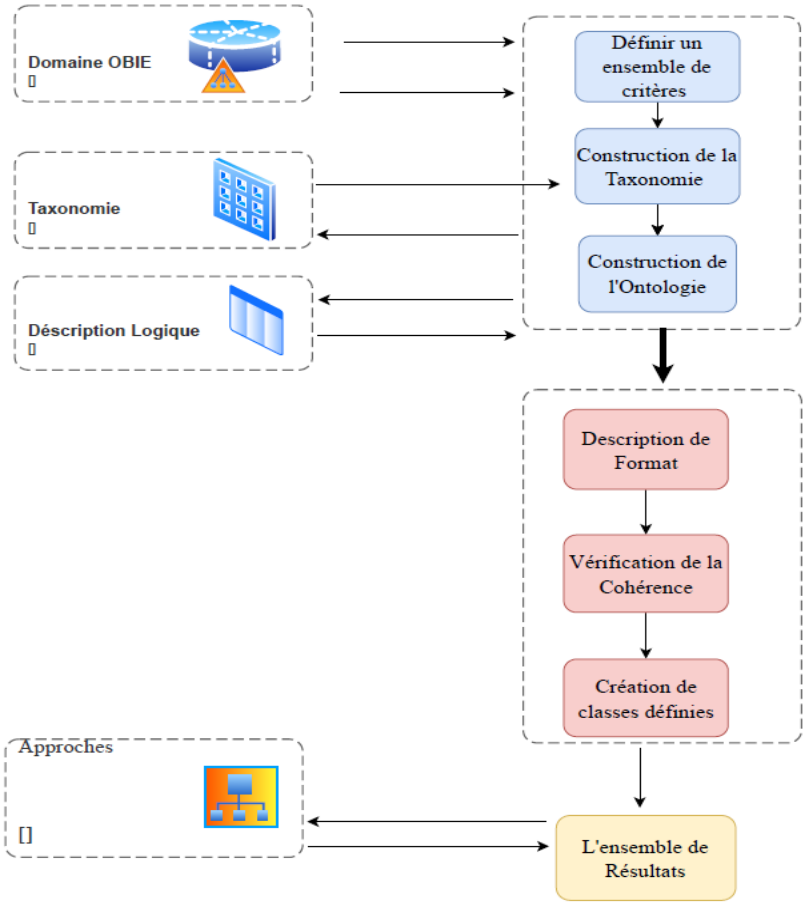

Figure 3.1 : Procédure générale de construction d'une ontologie(Noyet McGuiness 2001).

Le développement d'une ontologie implique un raffinement itératif : une version initiale est construite, évaluée à travers des applications ou des revues d'experts, puis affinée jusqu'à l'obtention d'une ontologie fonctionnelle.

Konys (2018) ont rapporté que le processus de construction d'une ontologie commence par une analyse approfondie du domaine. Cela implique la sélection d'une combinaison appropriée d'approches OBIE (Extraction d'Informations Basée sur l'Ontologie). Sur la base de cette analyse, un ensemble de caractéristiques et sous-caractéristiques est déterminé, suivi de la formation d'une hiérarchie de classes. Cette hiérarchie constitue une base générale pour la construction de taxonomies dans les systèmes et outils OBIE existants. Au cours de ce processus, divers critères et sous-critères sont établis, et des éléments tels que les concepts, les relations et les attributs sont intégrés à partir de sources différents dans le modèle de classification (Abu-Errub et al., 2014).

Par conséquent, la taxonomie est fondée sur la construction de l'ontologie. Pour implémenter l'ontologie, le langage Web OntologyLanguage (OWL) est utilisé. OWL fournit une méthode formelle et structurée pour collecter, organiser et partager des données, ainsi que de nombreuses fonctionnalités pour une gestion efficace de l'ontologie. En vérifiant la cohérence de l'ontologie créée, un ensemble de classes définies est établi. Le mécanisme de raisonnement est utilisé pour valider les classes définies ainsi que l'ensemble de l'ontologie. La cohérence et la cohésion de l'ontologie valider par le processus de raisonnement (Alaya et al., 2015).

3. Construction de l'ontologie ASTO : Méthodologie et Étapes Clés

La construction de l'ontologie ASTO (Arabic Spatial ToponymOntology) s'appuie sur la méthodologie de Noy et McGuinness (2001), une approche largement utilisée pour le développement d'ontologies. Cette méthodologie est structurée en plusieurs étapes successives, chacune étant une nécessité pour assurer la robustesse et la cohérence de l'ontologie.

3.1. Définition du domaine

La première étape consiste à définir clairement le domaine d'application de l'ontologie. Dans le cas de l'ontologie que nous allons développer, ASTO, l'objectif est de se focaliser sur les toponymes spatiaux dans les contextes géographiques Arabes, en particulier sur l'étude et la modélisation des noms de lieux (toponymes) dans les régions où la langue Arabeest utilisée.

Les toponymes spatiaux incluent les noms de villes, villages, rivières, montagnes, déserts, routes, et autres entités géographiques. Dans le contexte de l'ontologie ASTO, l'objectif est de représenter ces noms de lieux avec leurs relations spatiales, leurs caractéristiques, et leurs significations culturelles ou historiques propres à la langue Arabe.

Les contextes géographiques Arabesindiquent que l'ontologie tient compte des particularités culturelles, linguistiques et géographiques des pays arabophones. Par conséquence, certains toponymes peuvent avoir des significations ou des utilisations spécifiques en Arabequi ne se retrouvent pas dans d'autres langues ou cultures.

Exemple de toponyme en Algérie **(Hydronymes et leur distribution spatiale)**

Les toponymes qui incluent un hydronyme dans leur structure révèlent une répartition spatiale spécifique et un hydronyme est un nom propre attribué à un lieu associé à la présence d'eau, qu'elle soit permanente ou temporaire, sous forme liquide ou solide.

Dans l'étude de Atoui (1996), les hydronymes principaux ont été identifiés : Oued, Ain, Hassi, Daiet, Bir, Feid, Tala, Oglât, Hammam, Haoud, Guelta, Sebkha et Chott.

Parmi eux, le terme « Oued » (cours d'eau) est le plus fréquent, apparaissant plus de 3444 fois, suivi par « Ain » (source) avec 1825 occurrences. À eux seuls, ces deux hydronymes représentent plus de 71.6% du total des hydronymes étudiés. D'autres génériques se démarquent également, comme Hassi avec 759 occurrences, Daiet avec 431, et Bir avec 313.

L'ontologie ASTO vise donc à capturer cette richesse et cette spécificité, en fournissant une structure qui reflète étroitement la manière dont les noms de lieux sont utilisés et compris dans les contextes géographiques Arabes.Cela permet non seulement de représenter ces toponymes de manière précise dans des systèmes d'information géographique (SIG) ou d'autres applications, mais aussi de faciliter l'extraction et l'analyse de l'information géographique en Arabe, en tenant compte des nuances linguistiques et culturelles

3.2. Construction de la taxonomie

Cette étape consiste à organiser les concepts identifiés en une hiérarchie ou classification. Pour ASTO, cela implique de structurer les types géographiques, les caractéristiques naturelles et artificielles, ainsi que les relations spatiales, en classes et sous-classes (Figure 3.2). La taxonomie facilite l'organisation des connaissances et aide à établir les relations entre les différents concepts.Ci-dessous une explication détaillée des

étapes nécessaires à la création de cette taxonomie, basée sur les informations spatiale (entité spatiale et relation spatiale).

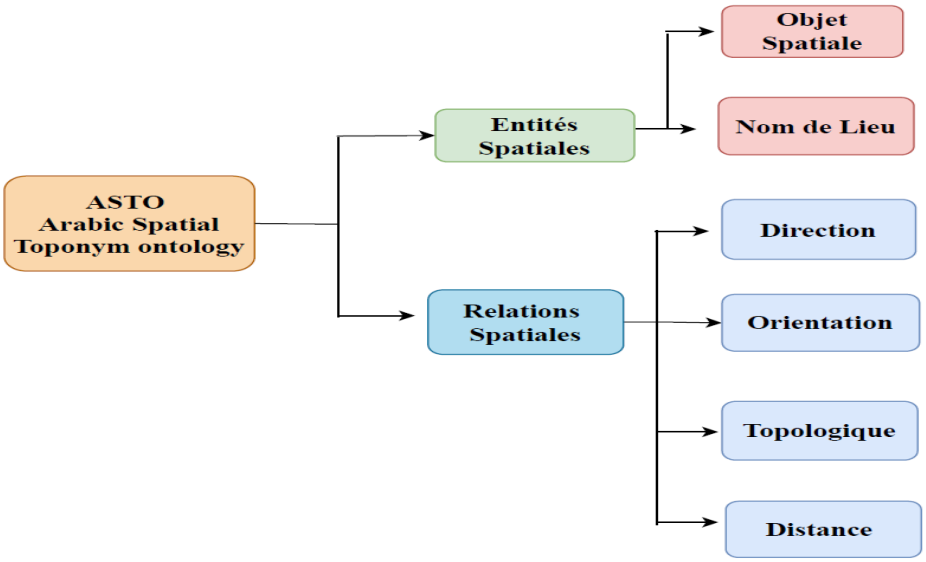

Figure 3.2 : Modèle conceptuel de l'ontologie ASTO

3.2.1. Identification des entités spatiales

- **Entité spatiale classe principale** : C'est la classe de base qui regroupe toutes les entités ayant une dimension spatiale. Elle est subdivisée en deux sous-classes, objet spatial et lieu spatial.

3.2.2. Définition des Sous-Classes

- **Objetspatial** : Cette sous-classe représente des objets physiques qui occupent un espace dans le monde réel. Elle se subdivise en :
  - **Objet naturel** : Inclut des entités géographiques naturelles telles que des montagnes, des forêts, des rivières, etc. Les instances comme "جبل" (montagne) ou "نهر" (rivière) appartiennent à cette sous-classe.
  - **Objet de construction** : Regroupe les constructions humaines telles que les écoles, hôpitaux, routes, etc. Des exemples d'instances sont "المدرسة" (école) et "الطريق" (route).
- **Lieu Spatial (localisation)**: Cette sous-classe couvre les entités qui définissent des emplacements géographiques spécifiques, comme des pays, des villes ou des régions. Des instances telles que "الجزائر" (Algérie) ou "جيجل" (Jijel) sont incluses.

3.2.3. Structuration des relations spatiales

- **Relations spatiales** : Cette classe englobe les différents types de relations qui peuvent exister entre les entités spatiales. Les sous-classes sont :
    - **Relations topologiques**: Ces relations concernent l'inclusion ou le support entre les entités, par exemple "على ضفة" (sur la rive) ou "داخل" (à l'intérieur).
    - **Relations directionnelles** : Elles décrivent les orientations cardinales (nord, sud, est, ouest) ou les chemins (comme "نحو" pour "vers").
    - **Relations de distance**: Elles mesurent les distances entre les entités, qu'elles soient quantitatives ("مسافة" pour "distance") ou qualitatives ("قرب" pour "près").
    - **Relations d'orientation :** Elles définissent la position relative des objets, soit en termes horizontaux ("أمام" pour "devant"), soit en termes verticaux ("فوق" pour "au-dessus").

3.2.4. Identification des Instances

- Pour chaque sous-classe, des instances spécifiques sont ajoutées pour représenter des exemples concrets d'entités ou de relations. Par exemple :
    - Dans **objet naturel**, des instances telles que "جبل" (montagne) ou "شاطئ" (plage) sont créées.
    - Dans r**elations direction**, des termes directionnels comme "شمال" (nord) ou "غرب" (ouest) sont inclus.

Aprèsla constructionde la taxonomie initiale, elle doit être affinée pour éliminer les redondances, clarifier les définitions, et s'assurer que toutes les relations sont logiques et cohérentes. Une validation par des experts du domaine ou via des tests empiriques est essentielle pour s'assurer que l'ontologie répond bien aux besoins du projet.

La taxonomie de l'ontologie ASTO permet de structurer les concepts géospatiaux de manière hiérarchique, facilitant ainsi l'analyse et l'interprétation des données spatiales. En organisant les entités, relations et instances de manière claire, cette ontologie devient un outil puissant pour diverses applications ou analyse de données en SIG.

3.3. Réalisation de l'ontologie

La première étape consiste à définir clairement le domaine d'application de l'ontologie. Dans notre cas, l'objectif est de se concentrer sur les toponymes spatiaux dans les contextes géographiques Arabes, en particulier sur l'étude et la modélisation des noms de lieux (toponymes) dans les régions où la langue Arabeest utilisée.

3.3.1. Objets Géographiques (entités spatiales)

Dans notre modèle, les objets géographiques sont définis comme des entités spatiales, qui peuvent être catégorisées en plusieurs types géographiques pour refléter leurs attributs physiques et naturels. Chaque caractéristique géographique peut avoir plusieurs noms. Au cours du développement de l'ontologie ASTO, plusieurs itérations ont été nécessaires : Au départ, il n'était pas évident que les termes collectés soient suffisants pour remplir l'objectif et le rôle de l'ontologie. Avec l'expérience acquise dans le domaine, de nouveaux termes ont été ajoutés en fonction des besoins, tandis que les termes inutiles ont été supprimés.

Le modèle conceptuel de l'ontologie ASTO se compose de diverses classes et sous-classes, chaque sous-classe contenant un ensemble d'instances et de propriétés (Tableau 3.1).

Table 3.1: Exemple de classeset instancesdans ASTO

Class	Subclass	Subclass	Instance
Spatial Entity	Spatial Object	Natural Object	جبل، هضبة،شاطئ، غابة،واد،بحر، صحراء، نهر. Mountain, plateau, beach, forest, valley, sea, desert, river.
	Building Object		المدرسة، المستشفى، البناية، العمارة، المسجد، المنزل، الولاية، البلدية، الطريق. School, hospital, building, architecture, mosque, home, state, municipality, Road.
	Spatial Location	Country	الجزائر،جيجل،ميلة،بجاية،الطاهير،الميلية،فرجيوة. Algeria, Jijel, Mila, Bejaia, Taher, Milia, Ferdjioua.
	Topological Relations	Inclusion	على ضفة،بعض،جزء،بضع،بين،وسط،داخل،في. On the bank, some, part, a few, between, middle, inside, in.
		Support	على،علىمستوى،على محور. On, on a level, on an axis.
Spatial Relations	Direction Relations	Cardinal	شمال،جنوب،شرق،غرب،شمال شرق ... North, south, east, west, northeast.
		Path	نحو،باتجاه، صوب، قصد، عبر، من خلال، حتى. Towards, In the direction of, Towards, Intending to, Across, Through, Until.
	Distance Relations	Quantitative	مسافة،علىبعد،تبعد. Distance, at a distance , move away.
		Qualitative	قرب،دنو،علىقرب،قصيا. Near,approach, close up.
	Orientation Relations	Horizontal	أمام، خلف، قبل، وراء، بعد، يمين،يسار،مقابل. Before, behind, before, behind, after, right, left, vs.
		Vertical	فوق، تحت،أعلى،أسفل. Above, under, up,down.

3.3.2. Relations spatiales

Dans le développement de l'ASTO, nous nous concentrons sur les relations spatiales. Cettedernières est importantesdans diverses tâches, telles que la réduction de l'ambiguïté, l'amélioration de l'extraction d'informations géographiques, l'optimisation des systèmes de navigation, la gestion du trafic, le raisonnement spatial et la réponse aux requêtes.

L'identification des relations spatiales est un élément clé dans la création d'ontologie, nous définissons les relations entre les entités géographiques en prenant en compte les relations topologiques, d'orientation, de distance et directionnelles (Tableau 3.1.). Nous avons développé une ontologie spatiale adaptée à la langue Arabe pour tenir compte de la vaste gamme de termes et de significations, ce qui améliore l'efficacité de l'extraction des informations spatiales et ajoute de l'originalité au travail présenté dans cette recherche.

3.3.3. Description formelle

La description formelle est une étape aussi nécessaire où l'ontologie est traduite en un langage formel, comme OWL[1] (Web OntologyLanguage), qui permet une interprétation automatique par les machines. Cette formalisation est essentielle pour garantir que l'ontologie peut être utilisée dans des systèmes d'information géographique (SIG) et d'autres applications logicielles. Pour ASTO, cette formalisation inclut la définition des règles et des axiomes qui régissent les interactions entre les entités spatiales.

4. Création de l'Ontologie

4.1. Création des classes

Les concepts, également appelés classes ou types, sont un élément central de la plupart des ontologies. Un concept représente un groupe d'individus partageant des caractéristiques communes, qui peuvent être plus ou moins spécifiques.

Dans cette étape, les classes identifiées lors de la construction de la taxonomie sont détaillées et enrichies. Chaque classe est définie avec ses propriétés spécifiques, ses instances, et ses relations avec d'autres classes. Pour ASTO, cela signifie la définition des classes pour différents types de toponymes (montagnes, rivières, villes) et la spécification de leurs caractéristiques spatiales et sémantiques (Figure 3.3).

[1] https://www.w3.org/2001/sw/#owl

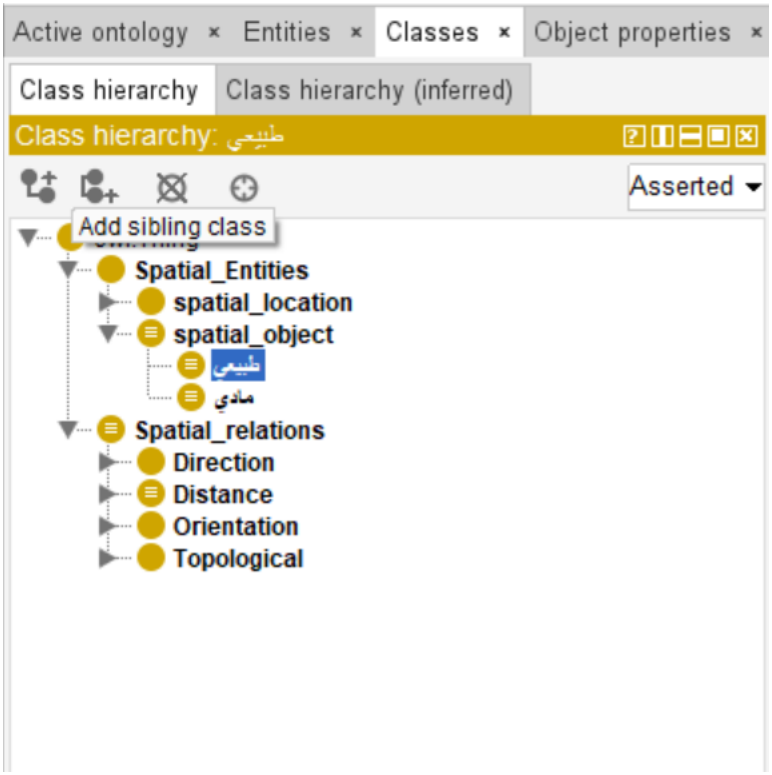

Figure 3.3 : Création des classes de l'ontologie ASTO.

4.2. Création des propriétés ou des relations

Les relations dans une ontologie décrivent la manière dont les individus sont connectés les uns aux autres. Elles peuvent généralement être exprimées directement entre individus ou bien entre concepts. Dans ce dernier cas, la relation décrit un lien entre tous les individus de ces concepts(Figure 3.4).

Figure 3.4 : Création des classes de l'ontologie ASTO.

Exemple de Relations entre les Classes

a. Entre "Spatial Object" et "Spatial Location" :

- **Relation "LocatedAt" (Situé à)** : Relation entre un objet spatial et un lieu géographique.
    - **Exemple** :
        - (جبلLocatedAtالجزائر) signifie que "la montagne est située en Algérie".
        - (مدرسةLocatedAtبجاية) signifie que "l'école est située à Bejaia".

b. Entre "Spatial Object" et "Spatial Relations" :

- **Relation "HasTopologicalRelation" (A une Relation Topologique)** : Relation entre un objet spatial et une relation topologique comme l'inclusion ou le support.
    - **Exemple** :
        - (نهرHasTopologicalRelationعلى ضفة) signifie que "la rivière est sur la berge".
        - (بنايةHasTopologicalRelationداخل) signifie que "le bâtiment est à l'intérieur".
- **Relation "HasDirectionalRelation" (A une Relation Directionnelle)** : Relation entre un objet spatial et une direction (Cardinal ou Path).
    - **Exemple** :
        - (مدينةHasDirectionalRelationشمال) signifie que "la ville est au nord".
        - (الطريقHasDirectionalRelationنحو) signifie que "la route mène vers".
- **Relation "HasDistanceRelation" (A une Relation de Distance)** : Relation de distance entre deux objets spatiaux.
    - **Exemple** :
        - (مدرسةHasDistanceRelation على بعد) signifie que "l'école est à distance".
        - (المسجدHasDistanceRelationقرب) signifie que "la mosquée est proche".

4.3. Création des instances

Les individus, également appelés instances ou particuliers, constituent l'unité de base d'une ontologie. Ils représentent les éléments que l'ontologie décrit ou est susceptible de décrire. Les individus peuvent modéliser des objets concrets tels que des personnes ou des machines, mais aussi des concepts plus abstraits comme des pays, des professions ou des

fonctions (Figure 3.5).

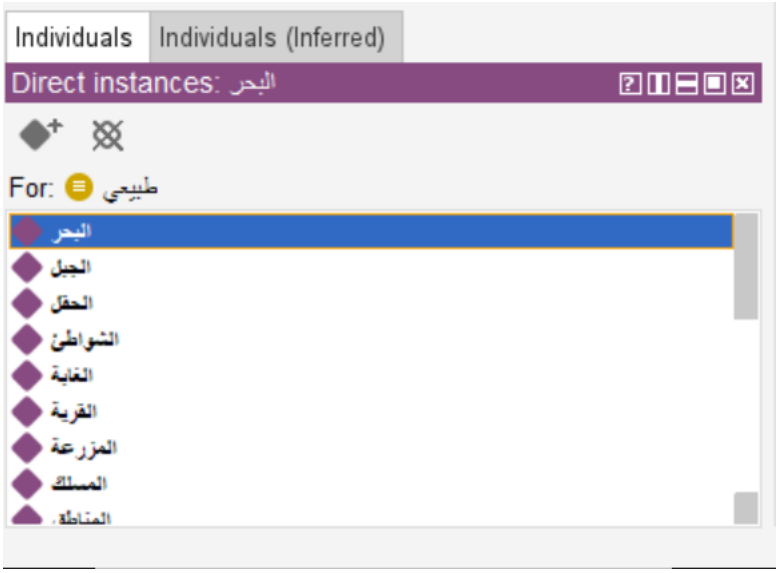

Figure 3.5 : Création des instances de l'ontologie ASTO.

4.4. Processus de raisonnement

Le processus de raisonnement consiste à utiliser les outils d'inférence pour vérifier les relations et les axiomes définis dans l'ontologie. Pour ASTO, ce processus permet de dériver de nouvelles relations spatiales implicites, de détecter des incohérences potentielles et de valider l'exactitude des relations spatiales telles que les adjacences, les inclusions, ou les distances. Ce raisonnement pour assurer que l'ontologie fonctionne correctement dans des applications pratiques (Figure 3.6).

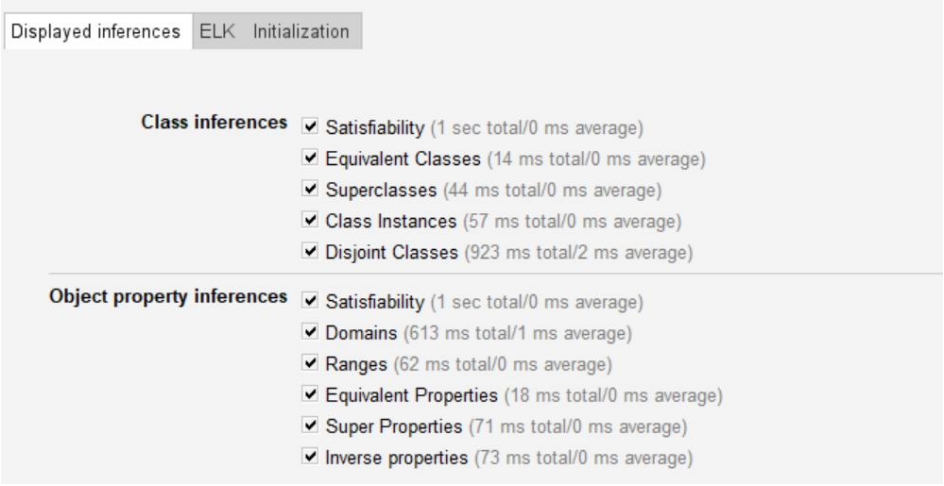

Figure 3.6 : Processus de Raisonnement ELK.0.6.0

Inférences sur les propriétés d'objet

D'après la figure 3.6, les résultats du processus de raisonnement effectué avec ELK 0.6.0, plusieurs analyses peuvent être formulées concernant les résultats obtenus :

- **Satisfiabilité** : Comme pour les classes, les propriétés d'objet sont consistantes, aucune contradiction n'a été trouvée.
- **Domaines** : Le moteur a vérifié les domaines des propriétés, assurant que les propriétés sont correctement associées aux classes appropriées.
- **Portées (Ranges)** : Les portées des propriétés ont été vérifiées, confirmant les types d'objets que les propriétés peuvent lier.
- **Propriétés équivalentes** : Le raisonnement a vérifié que les propriétés différentes ont les mêmes caractéristiques fonctionnelles.
- **Super Propriétés** : Le système a détecté les hiérarchies entre les propriétés et définissant les propriétés plus générales.
- **Propriétés inverses** : Le raisonnement a vérifié les propriétés inverses, qui relient les individus dans le sens inverse de la relation originale.

Le rapport montre que le raisonnement sur l'ontologie s'est déroulé de manière fluide, avec un bon temps de réponse pour chaque catégorie d'inférences. Aucune anomalie ou incohérence n'a été détectée, indiquant que l'ontologie est bien structurée et prête à être utilisée pour des applications comme la validation des données spatiales.

4.5. Vérification de la cohérence

La vérification de la cohérence est une étape de validation où l'ontologie est testée pour s'assurer qu'elle ne contient pas de contradictions internes. Pour ASTO, cela implique de vérifier que les relations spatiales définies sont logiquement cohérentes et que les classes et instances respectent les règles formelles établies. Cette vérification est essentielle pour garantir la fiabilité de l'ontologie dans des applications réelles (Figure 3.7).

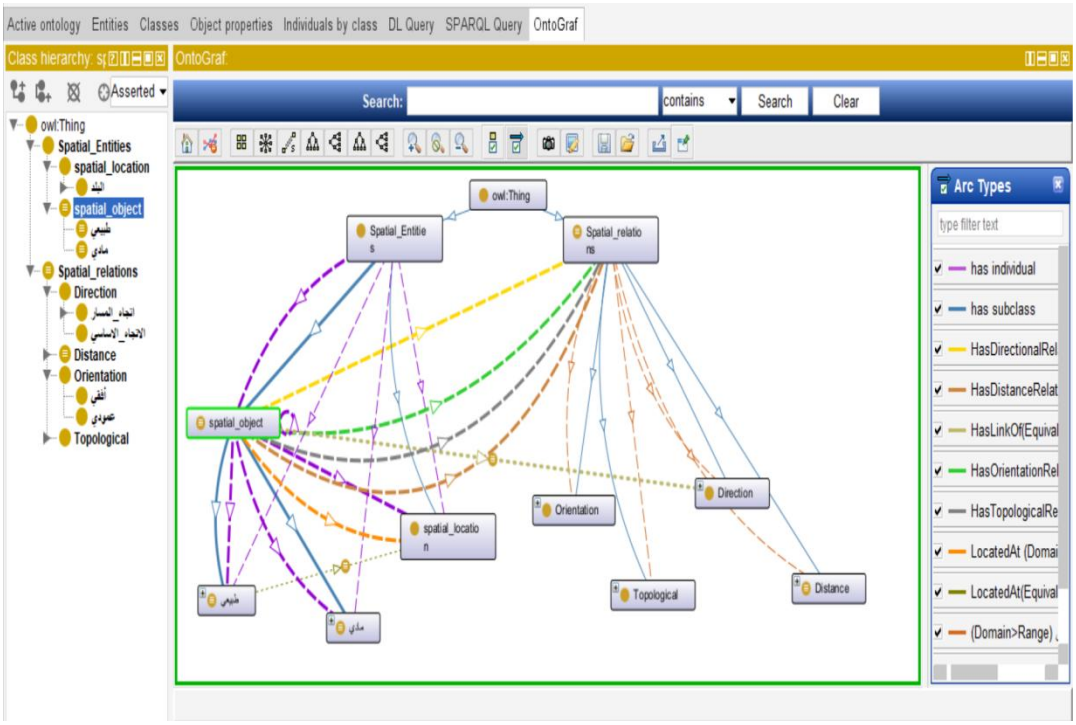

Figure 3.7 : Présentation des classes et des relations ASTO par OntoGraf.

5. Désambiguïsation et validation par ASTO

5.1. Réduction de l'ambiguïté lexicale

L'ambiguïté lexicale est fréquente dans les textes en langue Arabe, où un mot peut avoir plusieurs significations selon le contexte.Le mot "عين" peut signifier "œil" ou "source d'eau". Dans ASTO, "عين" serait catégorisé sous la classe des entités naturelles, spécifiquement comme une "source d'eau", lorsqu'il est utilisé dans un contexte géographique. Cela permet aux systèmes TALN d'éliminer l'ambiguïté en se basant sur les relations définies dans l'ontologie.

5.2. Désambiguïsation des relations spatiales

Les relations spatiales en Arabe, comme "فوق" (au-dessus) ou "تحت" (en dessous), peuvent être interprétées différemment selon le contexte. ASTO permet de définir précisément ces relations dans un contexte géographique :

- Relations Verticales : Le mot "فوق" est strictement défini comme une relation où une entité est physiquement au-dessus d'une autre.
- Relations Horizontales : De même, "تحت" est défini dans des contextes horizontaux et verticaux spécifiques, réduisant ainsi les risques de confusion.

5.3. Traitement des toponymes

Les toponymes (noms de lieux) sont souvent une source d'ambiguïté, surtout lorsque des noms identiques ou similaires sont utilisés pour désigner plusieurs localités. ASTO structure ces toponymes en les associant à des régions géographiques précises, permettant ainsi de désambiguïser les références dans les textes : Par exemple, ASTO peut associer chaque instance de "عين" à une région spécifique en utilisant des propriétés comme les coordonnées géographiques, réduisant ainsi les confusions liées aux toponymes communs.

5.4. Requête SPARQL

Dans cette section, nous mettons en œuvre des requêtes SPARQL afin de tester et valider notre ontologie ASTO. La première requête SPARQL appliquée à l'ontologie ASTO permet de récupérer des informations sur les entités spatiales, les relations spatiales, ainsi que leurs instances associées. Cette requête explore la structure de l'ontologie pour extraire les entités définies et les relations qui les lient, ce qui permet de valider la précision des connexions spatiales. Les résultats obtenus démontrent que l'ontologie est capable de gérer les informations spatiales complexes en garantissant la cohérence entre les entités et leurs relations. Les figures ci-dessous illustrent les résultats de cette première requête.

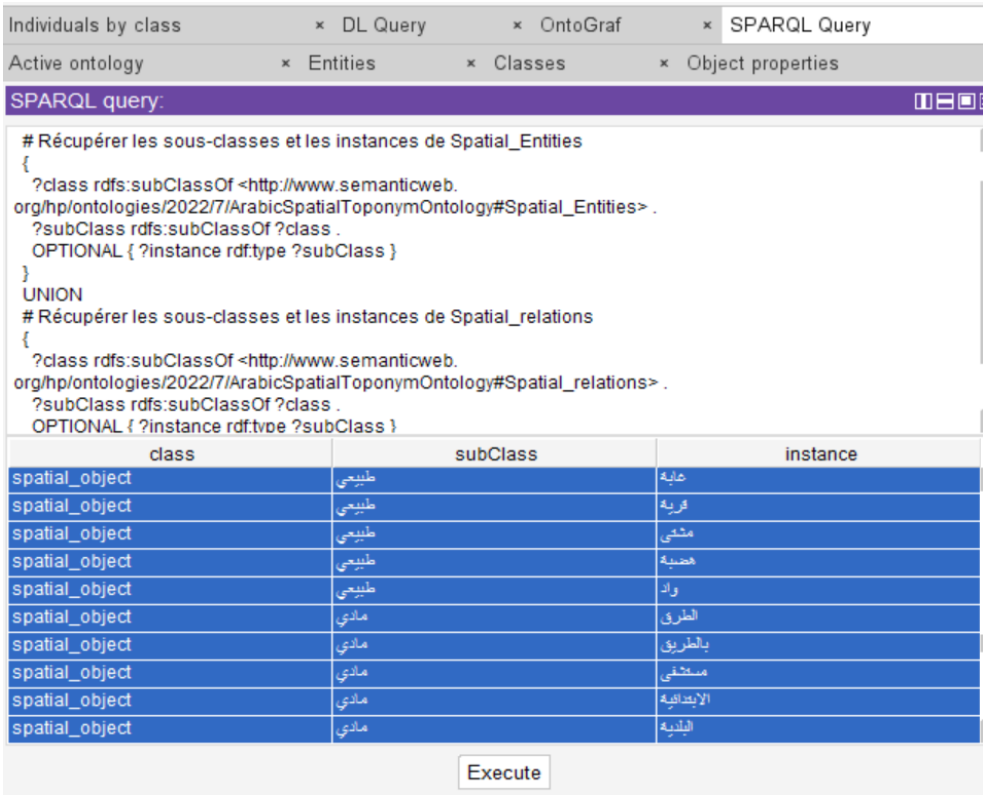

Figure 3.8 : Première requête SPARQL.

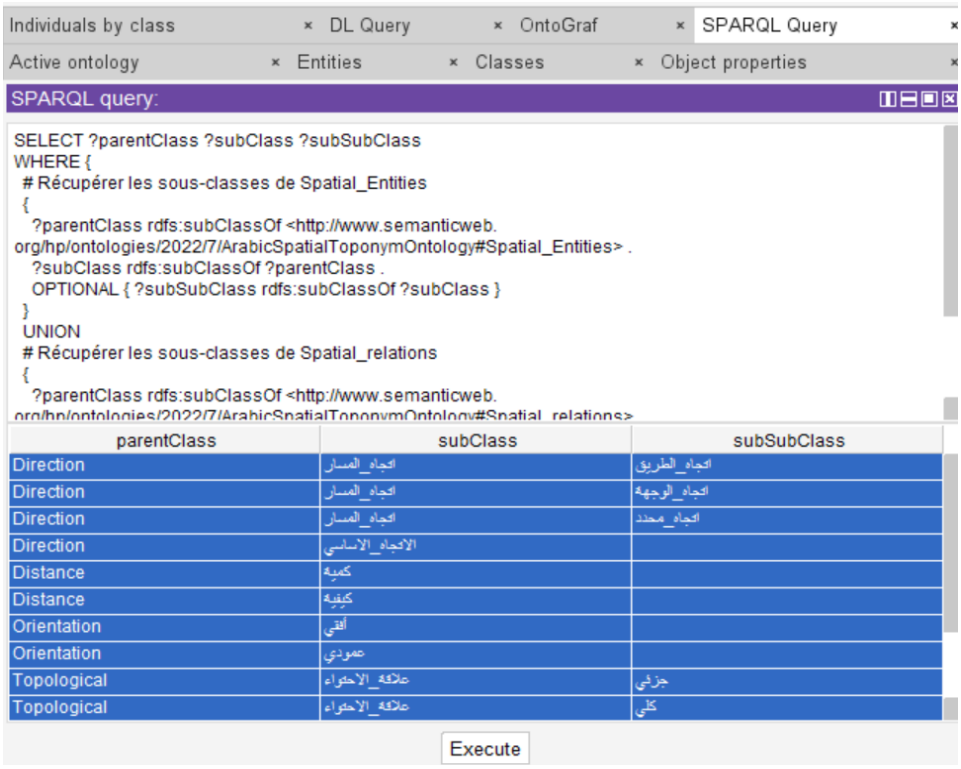

Figure 3.9 : Deuxième requête SPARQL.

La deuxième requête SPARQL interroge l'ontologie ASTO pour afficher la hiérarchie des classes, sous-classes et sous-sous-classes. Cette requête est essentielle pour vérifier la structure de l'ontologie et s'assurer que les relations de généralisation et de spécialisation entre les classes sont correctement représentées (Figure 3.9). En affichant la hiérarchie complète, la requête permet de confirmer que la modélisation suit les principes établis et que l'ontologie est correctement structurée pour représenter les concepts spatiaux.

5.5. Etude comparative ASTO et BFO

Pour effectuer une étude comparative entre l'ontologie ASTO (Arabic Spatial ToponymOntology) et BFO (Basic FormalOntology), il est important d'examiner les structures, les objectifs et les applications spécifiques de chaque ontologie, ainsi que les avantages et inconvénients qu'elles présentent. (Tableau3.2).

Tableau 3.2 : Etude comparative entre l'ontologie ASTO et BFO.

Critère	ASTO	BFO
Objectif	Modéliser les entités et relations spatiales spécifiques aux textes Arabes.	Fournir une structure ontologique générique applicable à divers domaines scientifiques.
Structure	Comprend des super-classes comme "Spatial Entity" et "Spatial Relations" avec des sous-classes spécifiques aux toponymes Arabes et aux relations spatiales.	Composée de classes abstraites telles que "Entity", "Continuant", "Occurrent", "Role", etc., pouvant être adaptées à divers domaines.
Domain Specificity	Spécifique à la langue Arabe et aux concepts géospatiaux.	Généraliste et applicable dans un large éventail de disciplines.
Relations Spatiales	Inclut des relations topologiques, directionnelles, de distance et d'orientation adaptées aux expressions linguistiques Arabes.	Fournit une structure de base pour définir des relations comme "part of", "located in", mais ne spécifie pas les relations spatiales détaillées.
Avantages	- Optimisé pour les textes en Arabe. - Précision accrue dans l'annotation des relations spatiales. - Simplicité d'utilisation pour le contexte Arabe.	- Universalité et flexibilité pour divers domaines. - Adoption large et interopérabilité avec d'autres ontologies. - Peut être adapté pour structurer des connaissances spécifiques.
Inconvénients	- Limité aux textes en Arabe et aux contextes géospatiaux. - Moins utile dans des domaines ou langues différentes sans modifications importantes.	- Complexité due à sa généralité, nécessitant des adaptations pour des applications spécifiques. - Absence de spécificité linguistique, nécessitant des couches supplémentaires pour des concepts linguistiques particuliers.

5.6. Cas d'Utilisation de l'ASTO

Cas 1 **: Extraction d'entités et relations spatiales dans les textes Arabes**(Hadji et al., 2024).

- **Contexte :** Dans une étude portant sur l'analyse de textes Arabes pour l'identification d'entités et relations géographiques, l'ASTO a été utilisée pour extrait et annoter automatiquement des entités spatiales telles que des lieux, des

objets naturel et non naturel et aussi des relations, distance, topologique, orientation et de direction.

- **Méthodologie :** En utilisant des règles JAPE associées à l'ASTO, des textes issus de journaux algériens ont été analysés. L'ontologie a permis de classifier des entités comme "جبل" (montagne), "مدرسة" (école), et "الجزائر" (Algérie) sous les catégories appropriées (Natural object, Bulding object, Location) et meme aussi pour les relations.
- **Résultats :** L'application de l'ASTO a permis d'extraire efficacement les entités spatiales des textes, facilitant ainsi une analyse plus précise des informations géospatiales. L'ASTO s'est avérée particulièrement utile pour la reconnaissance des entités toponymiques complexes, souvent négligées par les systèmes traditionnels de TALN.
- **Pertinence :** Ce cas d'utilisation démontre l'efficacité de l'ASTO pour les applications en traitement automatique du langage naturel, spécialement dans le contexte de la langue Arabe. L'ASTO améliore la précision et la cohérence des annotations spatiales, essentielles pour des analyses géospatiales plus robustes.

Cas 2:**Extraction d'informations spatiales dans les rapports de catastrophes naturelles en utilisant une partie de l'ASTO**(Hadji et Kholladi, 2024).

- Contexte : Dans le cadre d'un projet d'extraction d'informations spatiales pour la gestion des catastrophes naturelles, une partie des concepts et relations définis dans l'ASTO a été sélectionnée pour analyser et classifier les informations géospatiales relatives aux zones touchées.
- Méthodologie : Les descriptions textuelles des rapports ont été analysées en se concentrant sur les relations spatiales importantes comme "إلى الغرب من" (à l'ouest de) ou "بالقرب من" (près de). Une partie spécifique de l'ASTO, focalisée sur les relations topologiques de proximité et de distance, a été utilisée pour structurer ces informations. Les règles JAPE ont été modifiées et adaptées pour extraire ces informations de manière précise dans des documents non homogènes.
- Résultats : Grâce à l'utilisation partielle de l'ASTO, une extraction fiable des relations géospatiales critiques a été réalisée. Cela a permis de cartographier avec précision les zones d'impact et de définir des priorités d'intervention en fonction de la gravité et de la proximité des zones affectées.
- Pertinence : Ce cas d'utilisation met en avant l'avantage d'une approche modulaire

de l'ASTO dans le contexte des catastrophes naturelles. En utilisant uniquement les éléments les plus pertinents, il est possible d'améliorer l'analyse spatiale des rapports de crises, facilitant ainsi une réponse plus rapide et une meilleure allocation des ressources de secours.

Ce texte illustre l’utilisation ciblée d’une partie de l’ASTO pour optimiser l’extraction d’informations spatiales dans un contexte de gestion de catastrophes naturelles.

6. Conclusion

En conclusion, ce chapitre a fourni une vue détaillée du processus de développement de l'ontologie ASTO, en couvrant toutes les étapes critiques de sa construction. Nous avons examiné la conception initiale, la spécification des concepts et des relations, ainsi que la réalisation technique de l'ontologie. Les tests et la validation effectués ont démontré l'efficacité de l'ontologie dans la gestion des données spatiales et la représentation des relations complexes. Ce processus exhaustif assure que l'ontologie ASTO est robuste, cohérente et prête à être appliquée dans des contextes pratiques, offrant ainsi une base solide pour les applications futures dans le domaine de la gestion de l'information géographique. Cette ontologie sera exploitée pour l'indexation et l'extraction des entités et relations spatiales dans un texte en Arabe dans le chapitre suivant, où nous examinerons son application pratique et ses performances dans un contexte linguistique spécifique

Chapitre 4

Développement d'une approche basée règles JAPE

1. Introduction

L'extraction des informations spatiales, englobant à la fois les entités et les relations spatiales, représente un enjeu essentiel dans le domaine du traitement automatique du langage naturel (TALN), en particulier pour les applications géospatiales. Ces informations sont essentielles pour une variété de domaines, tels que les systèmes d'information géographique (SIG), la cartographie, la gestion des ressources naturelles et les services basés sur la localisation. Cependant, l'extraction précise et exhaustive de ces informations reste un défi majeur, en raison de la complexité des structures linguistiques et de la diversité des relations spatiales présentes dans les textes.

Dans ce chapitre, nous nous concentrons sur le développement d'une approche basée sur les règles JAPE, qui se distingue par sa pertinence dans des systèmes non compliqués d'extraction d'informations géographiques. Bien que les ontologies soient efficaces, leur mise en œuvre peut être difficile dans des contextes de ressources limitées. En revanche, les règles formelles offrent une solution flexible et performante. L'approche JAPE exploitera les structures linguistiques spécifiques à la langue Arabe pour identifier et annoter les entités spatiales, démontrant ainsi son efficacité tout en assurant une classification précise. Il est a noté que cette méthode présente des limites dans des scénarios plus complexes, où les interactions entre entités deviennent nuancées. Pour remédier à ces insuffisances, nous comparerons l'approche JAPE pure à une approche hybride intégrant les règles JAPE (Java Annotation Patterns Engine)et l'Ontologie Arabe des Toponymes Spatiaux (ASTO). Cette combinaison vise à améliorer la gestion des ambiguïtés et la précision de la classification des informations spatiales dans des

environnements complexes. Ainsi, ce chapitre se penche sur les expérimentations menées pour évaluer ces approches et leur pertinence respective dans divers contextes.

2. Contexte et motivation

L'extraction d'information constitue un sous-domaine fondamental du traitement automatique du langage naturel pour la transformation de vastes corpus de texte non structuré en données pertinentes et structurées. Dans des domaines tels que les systèmes d'information géographique, la capacité d'identifier et d'extraire des entités et des relations spatiales est essentielle pour une spécification géospatiale précise. Les méthodes traditionnelles d'extraction d'information, comme celles basées sur des règles, notamment les règles JAPE, ont été largement adoptées pour leur capacité à capturer des motifs linguistiques spécifiques. Ces règles, conçues pour offrir une couverture étendue et une application ciblée dans des contextes bien définis, exploitent des structures linguistiques propres à la langue Arabe afin d'identifier et d'annoter les entités spatiales.

Cependant, l'approche basée sur les règles a démontré son efficacité dans la désambiguïsation et la classification des informations spatiales, mais elle présente certaines limites, notamment face à des cas plus complexes où la rigidité de ces règles peut nuire à la précision de l'extraction. Cette situation souligne la nécessité d'explorer des approches alternatives et complémentaires. C'est dans ce cadre que les ontologies émergent, apportant une flexibilité et une contextualisation accrues à la modélisation des connaissances. Les ontologies améliorent la compréhension contextuelle des informations extraites. Cependant, ni les approches basées sur des règles ni celles reposant sur des ontologies ne parviennent à surmonter toutes les contraintes associées à l'extraction d'information.

L'émergence des approches hybrides, qui combinent les bienfaits des méthodes traditionnelles avec les avantages des ontologies, apparaît donc comme une solution prometteuse. En intégrant la rigueur des ontologies à la précision des règles JAPE, ces approches visent à maximiser l'efficacité de l'extraction d'information, notamment dans des contextes géospatiaux complexes.

Ce chapitre se propose d'analyser de manière comparative les deux approches basées sur des règles JAPE et la solution hybride, en mettant en lumière leurs forces, leurs faiblesses et leur applicabilité dans des situations variées. En explorant les interactions entre ces différentes méthodes, cette étude vise à identifier des pistes pour l'amélioration

des systèmes d'extraction d'information, contribuant ainsi au développement d'outils plus performants et précis pour l'analyse des données textuelles dans le domaine des informations géospatiales.

3. Méthode proposé basée sur les règles JAPE

La méthode basée sur des règles est une approche classique et largement utilisée dans le domaine de l'extraction d'information. Cette méthode repose sur un ensemble de règles prédéfinies qui sont conçues pour identifier et extraire des informations spécifiques à partir de textes ou d'autres types de données. Ces règles sont généralement exprimées sous forme de modèles ou de patrons qui correspondent à des structures linguistiques ou à des motifs spécifiques dans les données.

Architecture du Système

L'architecture générale de l'approche proposée (Figure 4.1) se compose de quatre phases distinctes.

- **Création des Règles JAPE**

Dans la première phase, les concepts liés aux entités et relations spatiales Arabes sont identifiés et collectés. Ces concepts sont ensuite utilisés pour formuler des règles JAPE spécifiques, qui permettent d'annoter et d'extraire les informations spatiales pertinentes à partir des textes en Arabe. Les règles JAPE sont des expressions régulières avancées développées en Java, permettant de détecter des modèles complexes dans le texte ;

- **Traitement du texte**

La deuxième phase consiste à appliquer des modules de traitement du langage naturel pour préparer le texte brut. Ce processus inclut des étapes telles que la normalisation, la tokenisation, et l'annotation des entités spatiales présentes dans le texte. Ces modules sont importants pour assurer que les règles JAPE puissent être appliquées de manière efficace et précise.

- **Combinaison et extraction**

La troisième phase repose sur l'application des règles JAPE créées lors de la première phase. Ces règles sont utilisées pour associer les segments du texte avec les classes, sous-classes ou instances définies. Cette phase est essentielle pour extraire automatiquement les informations spatiales structurées à partir du texte non structuré, en tenant compte des spécificités linguistiques et contextuelles de la langue Arabe.

- **Désambiguïsation et classification**

La quatrième phase se concentre sur la désambiguïsation des entités spatiales extraites et leur classification. Cette étape garantit que chaque entité et relation est correctement interprétée dans son contexte spécifique. Les règles JAPE sont également utilisées ici pour affiner les résultats, en appliquant des critères de désambiguïsation et de classification pour améliorer la précision des données extraites.

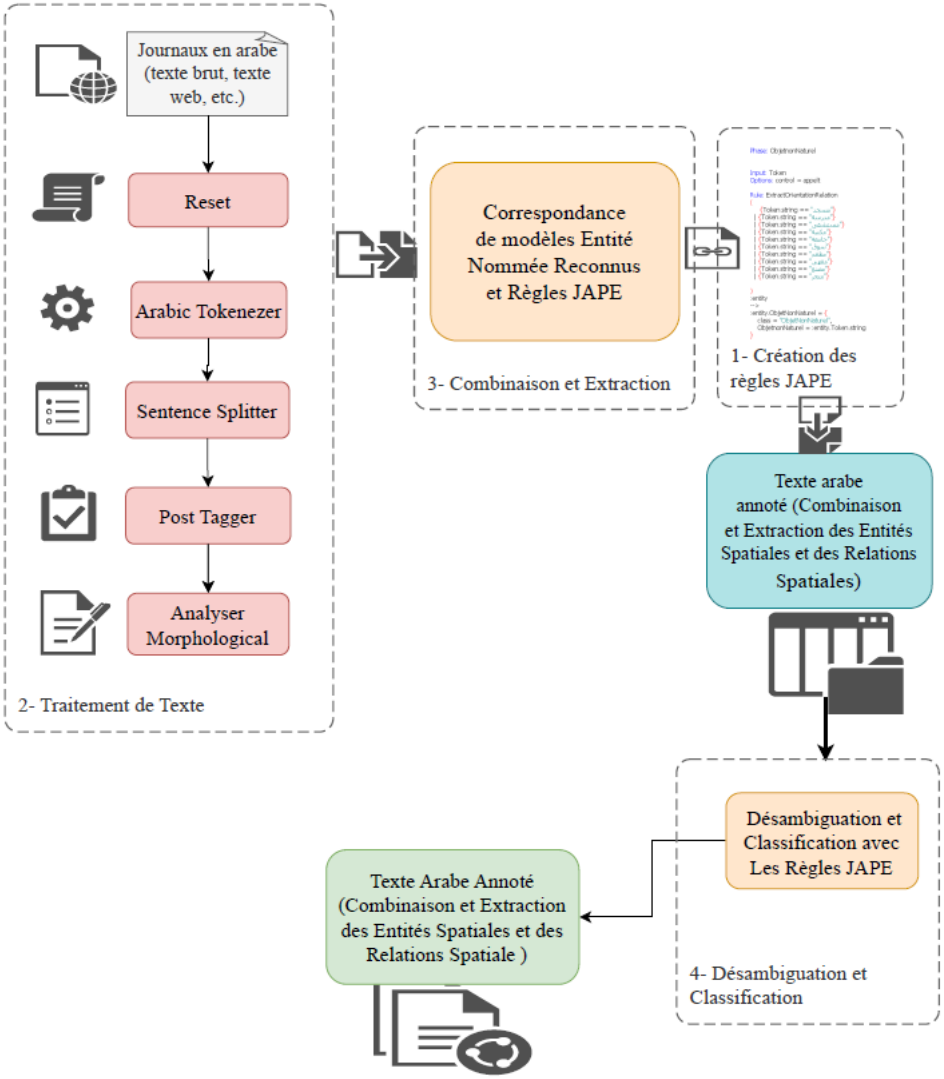

Figure 4.1 : Architecture de système proposé basé règles JAPE

4. Application et Réalisation

4.1. Phase de réalisation

Notre architecture du système basé sur les règles JAPE se compose de deux phases

principales, chacune jouant un rôle crucial dans l'extraction d'informations géographiques à partir de textes en langage naturel (Figure 4.2). La première phase utilise des techniques avancées de TALN pour préparer et normaliser les données textuelles. Cette préparation inclut le filtrage du texte, la segmentation en phraseset l'annotation initiale des éléments linguistiques, facilitant ainsi une meilleure application des règles.

La deuxième phase se concentre sur la correspondance des règles JAPE pour extraire des informations spécifiques. Cette phase implique la définition et la création des règles, l'appariement de ces règles avec le texte, la désambiguation et l'extraction des informations pertinentes. Enfin, un post-traitement est effectué pour filtrer et structurer les données extraites, les rendant prêtes pour des analyses ultérieures ou l'intégration dans des bases de données géospatiales. L'ensemble de ces phases assurent une extraction d'informations précise et efficace, adaptée aux besoins de l'analyse géographique.

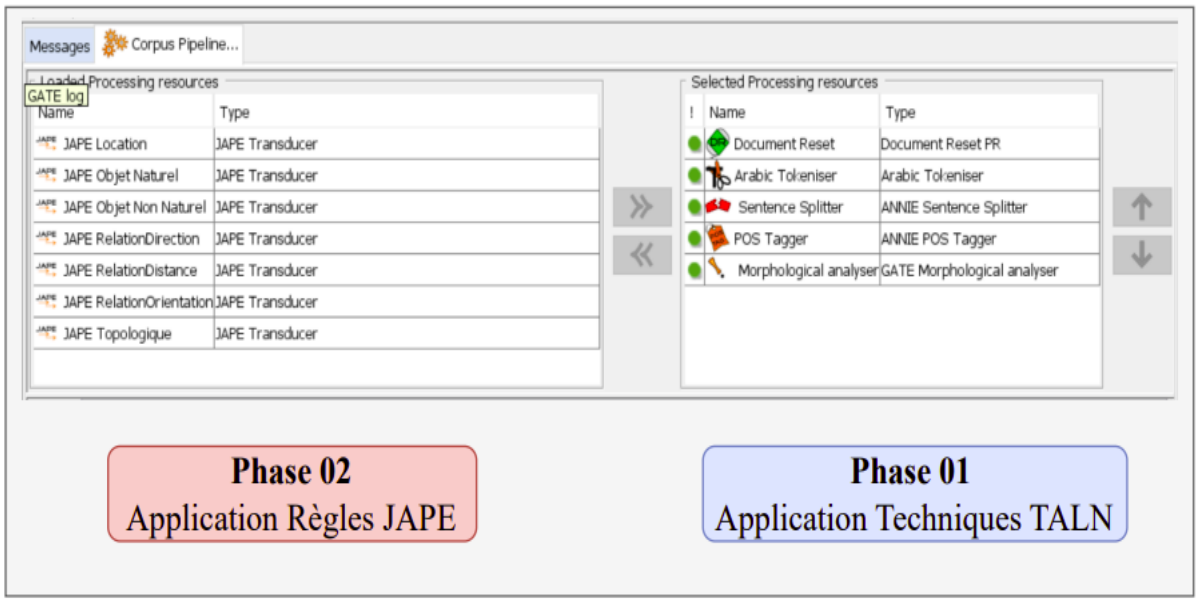

Figure 4.2 : Phases d'application de notre système

4.2.Environnement d'Application

Nous avons choisi d'utiliser l'environnement GATE (General Architecture for Text Engineering), un cadre d'ingénierie linguistique développé par l'Université de Sheffield et largement adopté depuis sa première version en 1996 pour l'enseignement et la recherche. GATE propose une suite de ressources de traitement réutilisables en JAVA, intégrées dans un système d'extraction d'informations appelé ANNIE (aNearly-New Information Extraction System).

ANNIE est par défaut configuré pour des langues autres que l'Arabe. Pour adapter

cet outil à notre langue cible, nous allons utiliser des composants spécialisés tels que le "tokenizer" Arabe, le "sentence splitter", le "POS tagger", et un analyseur morphologique Arabe. Afin d'éviter les interférences avec les exécutions précédentes, nous appliquerons l'option "reset" pour supprimer toutes les traces des processus antérieurs. Les annotations dans GATE seront effectuées en sélectionnant des mots dans le texte et en créant de nouvelles catégories d'annotation, permettant ainsi une extraction précise des entités géographiques et d'autres informations pertinentes.

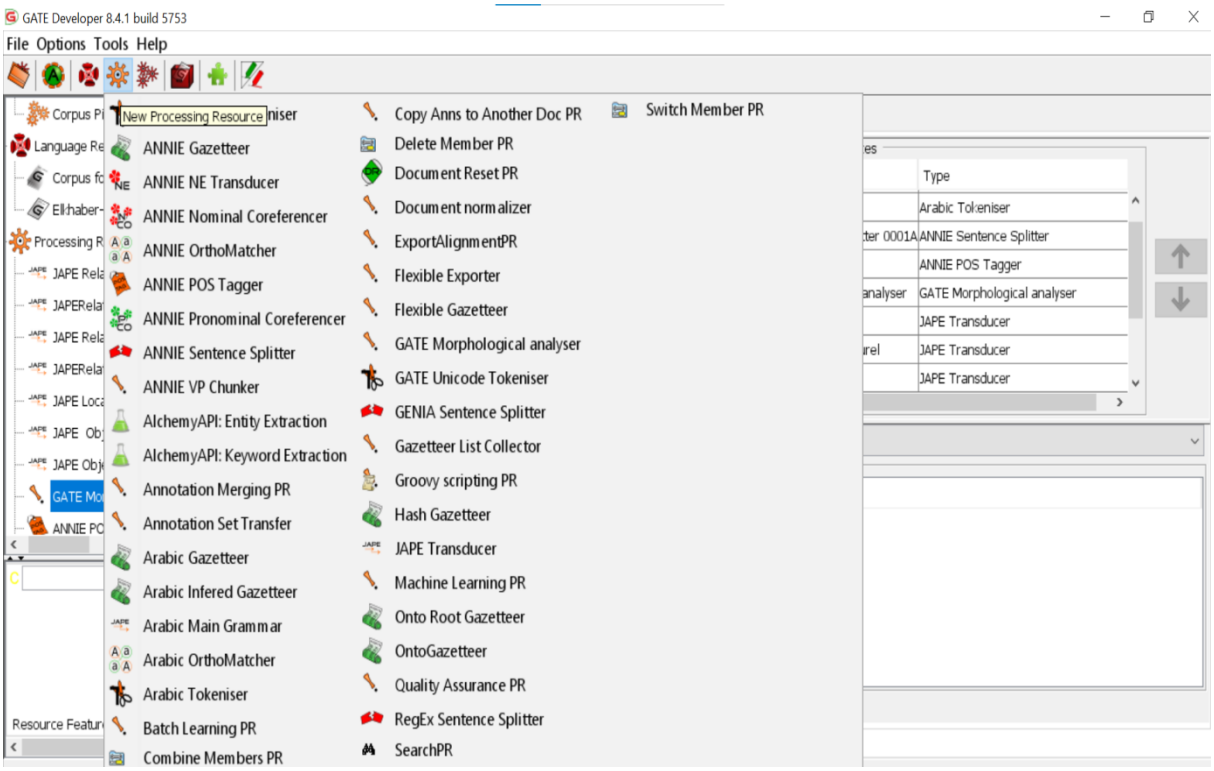

Figure 4.3 : Plateforme GATE

4.3. Première Phase : Technique de TALN

La première phase de notre architecture met en œuvre des techniques de TALN qui sont essentielles pour traiter et comprendre le texte en langage naturel. L'objectif de cette phase est de préparer les données textuelles de manière à faciliter l'extraction d'informations pertinentes sachant que nous avons utilisé les même dataset ou le corpus évoqués dans le chapitre quatre (Section 3).

4.3.1. Prétraitement linguistique

Le nettoyage du texte en langue Arabe est une étape essentielle avant d'appliquer des techniques de TALN. Ci-dessous les principales étapes spécifiques au nettoyage des textes en Arabe :

- **Suppression des caractères diacritiques (Tashkeel) :** Les textes Arabes peuvent contenir des diacritiques (harakats) comme les Fatha, Damma, Kasra, etc. Ces diacritiques peuvent être supprimés car ils ne sont souvent pas nécessaires pour l'analyse ;
- **Suppression des caractères spéciaux et des ponctuations :** Comme dans d'autres langues, les caractères spéciaux (comme !, @, #, etc.) et certaines ponctuations peuvent être supprimés pour simplifier le texte ;
- **Normalisation des caractères :** En Arabe, certains caractères peuvent être écrits de plusieurs façons. Par exemple, "آ" ،"إ" ،"أ" sont souvent normalisés en "ا". De même, le "ى" peut être transformé en "ي" ;
- **Suppression des espaces superflus :** Les textes en Arabe peuvent contenir des espaces multiples ou des espaces avant ou après les ponctuations. Ces espaces doivent être normalisés pour garantir une analyse correcte.

Ces étapes de filtrage sont nécessaires pour obtenir des résultats précis et pertinents lors de l'analyse du texte Arabe à l'aide de techniques de TALN. Un texte bien nettoyé réduit le bruit et les erreurs, ce qui permet aux algorithmes d'analyse de mieux comprendre les structures linguistiques complexes de l'Arabe, d'améliorer la précision des résultats et de garantir une meilleure interprétation des données textuelles. Ainsi, ces etapes de filtrages sont essentielles pour toute application de TALN, qu'il s'agisse de la reconnaissance d'entités nommées, de la classification de texte, ou de la traduction automatique.

4.3.2. Application des techniques de TALN

Les techniques de TALN telles que la réinitialisation de document (Document Reset), le segmentateur de phrases (SentenceSplitter), le tokeniseur Arabe (ArabicTokeniser), l'étiquetage morphosyntaxique (PostTagging) et l'analyse morphologique (Morphological Analyser) ont été expliquées en détail précédemment(chapitre 04). Dans ce chapitre, nous allons nous focaliser sur l'application pratique de ces techniques en utilisant la plateforme GATE. Cette exploration approfondie vise à fournir une meilleure compréhension de GATE et à servir de guide pratique pour son utilisation, en particulier dans le contexte de l'analyse de textes en Arabe.

La Figure 4.4 illustre le pipeline de traitement du corpus dans GATE, mettant en avant les étapes de réinitialisation (Reset) et le segmentateur de phrases (SentenceSplitter). Ces étapes permettent de segmenter le texte en phrases afin de préparer les données pour

l'extraction d'informations spatiales.

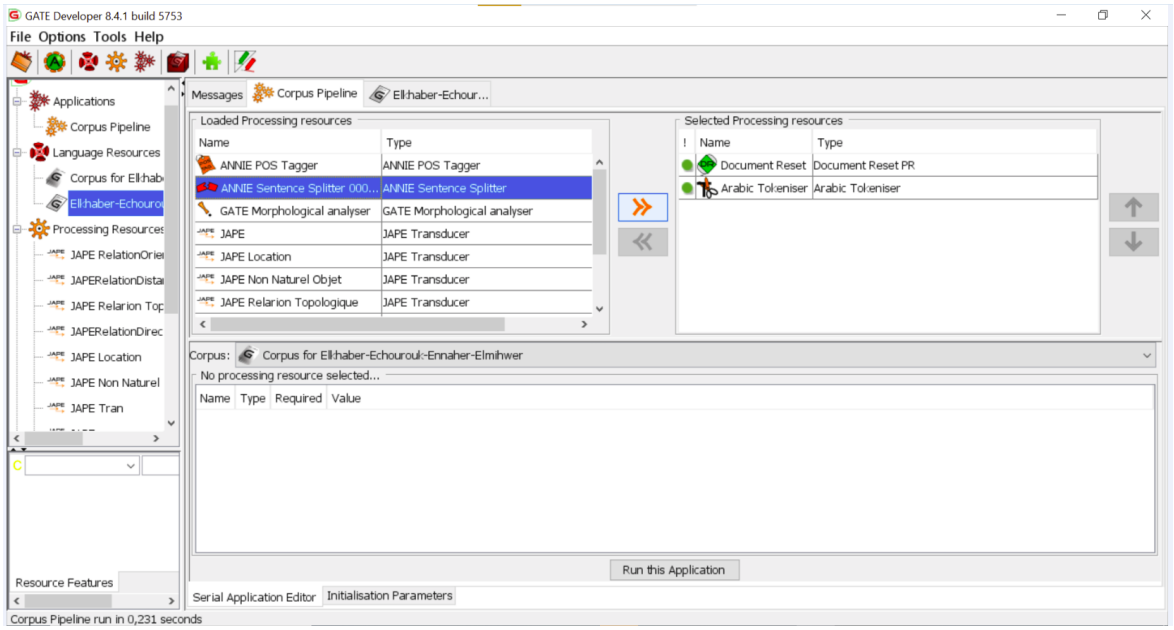

Figure 4.4 : Corpus Pipe Line TALN de GATE

4.3.3. Sentence Spliter

La figure 4.5 illustre une visualisation de GATE lors de l'étape de Sentence Splitter**.** Cette étape divise le texte en phrases distinctes, ce qui permet d'améliorer la précision des analyses syntaxiques et grammaticales.

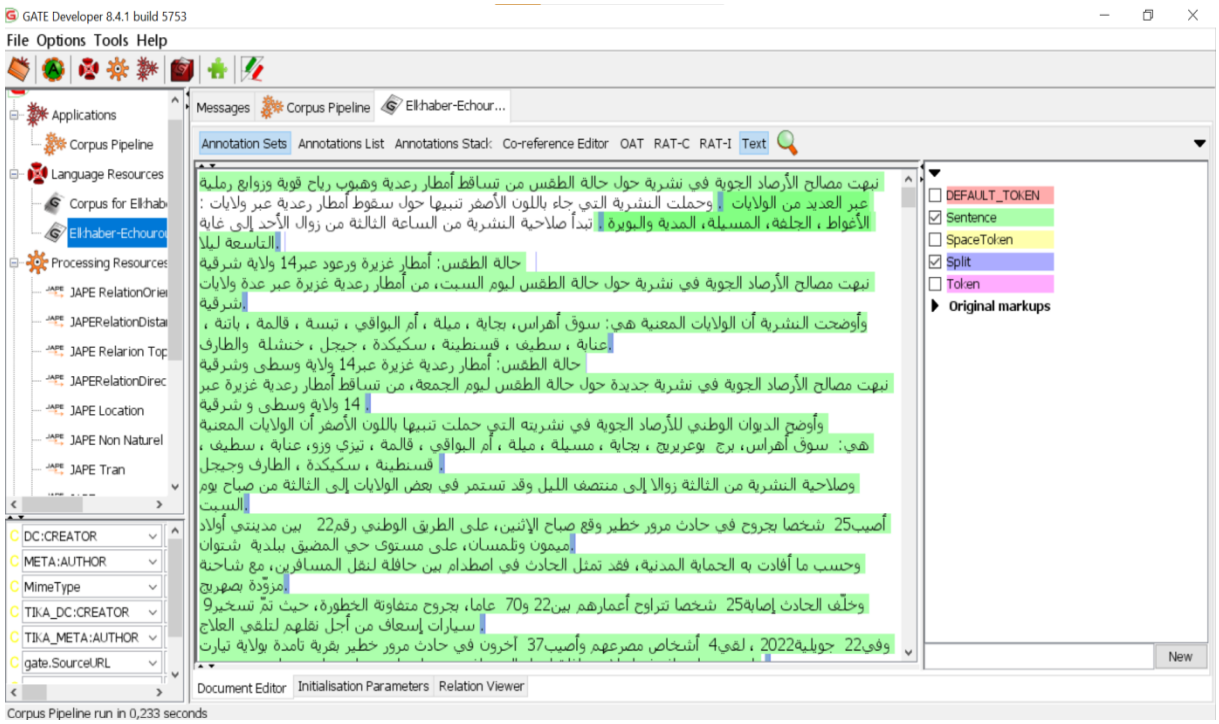

Figure 5.5 : Résultats d'Exécution Corpus Pipe Line (Reset et Sentence Splitter) GATE

4.3.4.Tokenisation

La Figure 4.6 représente une capture d'écran de la plateforme GATE, illustrant le processus de tokenisation. Elle met en évidence comment GATE segmente le texte en unités de base (tokens) pour une analyse linguistique plus approfondie.

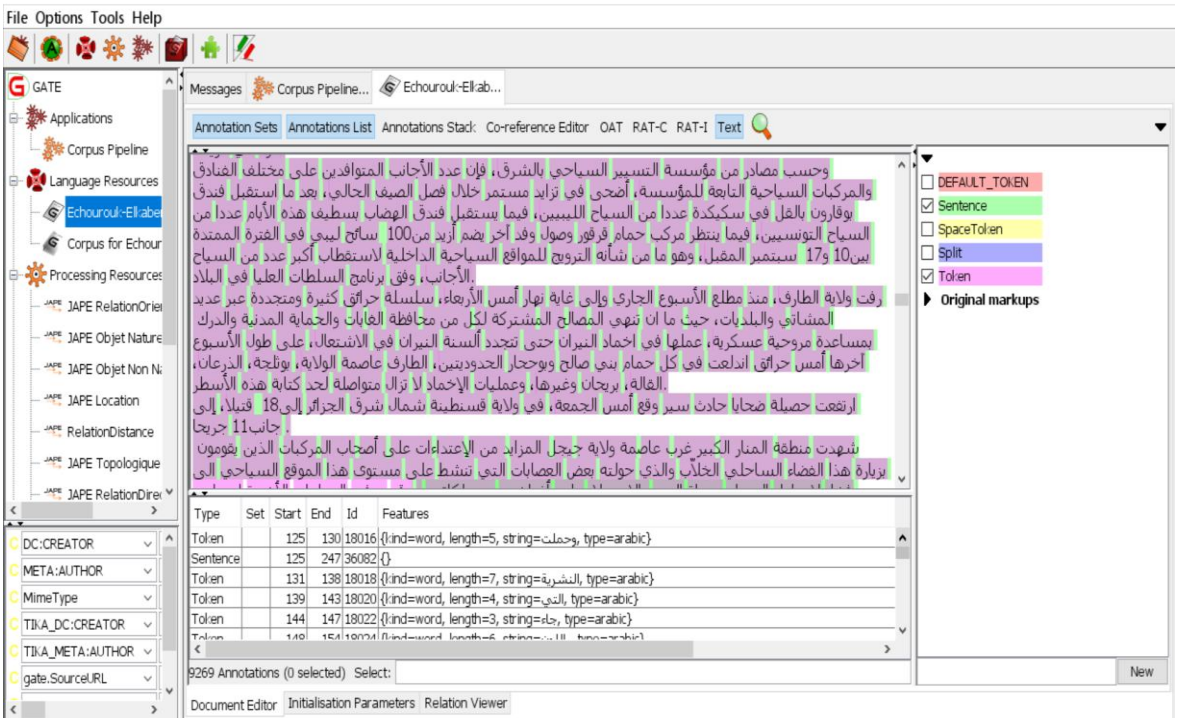

Figure 4.6 : Résultats d'exécution corpus pipe line (Tokineser) GATE

L'étiquetage morphosyntaxique (PostTagging) et l'analyse morphologique (Morphological Analyser) sont des processus essentiels dans le traitement automatique du langage naturel, particulièrement lorsqu'ils sont intégrés dans des systèmes avancés comme GATE. Dans ce contexte, ces étapes sont exploitées par les règles JAPE , qui permettent de définir des patrons d'annotations sophistiqués pour détecter des structures linguistiques spécifiques au sein d'un corpus.

Lors de l'exécution des règles JAPE, les annotations générées par le PostTagging et l'analyse morphologique enrichissent le corpus en ajoutant des métadonnées détaillées sur les catégories grammaticales et la structure morphologique des mots. Bien que ces annotations n'affectent pas l'affichage visible du texte, elles fournissent des informations invisibles mais fondamentales pour les analyses linguistiques ultérieures et l'extraction d'information précise.

4.4. Deuxième phase : Application des Règles JAPE

La deuxième phase est centrée sur l'application des règles JAPE pour l'extraction d'informations spécifiques. Cette phase suit les étapes bien définies des règles et les informations spatialessoit classés en entité spatiale et relation spatiale selon le Tableau 4.1.

Le Tableau ci-dessous présente une classification des entités spatiales, incluant les entités naturelles, les entités non naturelles, ainsi que les entités correspondant aux noms de lieux ou emplacements.

Tableau 4.1 : Classification des entités spatiales

Entité Spatiale	Objet Naturel	"جبل، واد ، هضبة ، شاطئ ، غابة ، بحر ، نهر ، صحراء ، تلة ، بحيرة ، خليج، مضيق ، جزيرة ، شبه جزيرة ، رافد ، مستنقع ، حوض ، كهف ، منخفض، تل، ينبوع، مرج، حقل، سهوب، مستنقع ملحي، رمل، شعاب مرجانية، جرف، رمال متحركة، غدير"
	Objet Non Naturel	مسجد، مدرسة، مستشفى، مكتبة، جامعة، سوق، مطعم، مقهى، مصنع، متجر، محطة ، مطار ، فندق ، قصر ، برج ، جسر ، سد ، كنيسة ، معبد ، ملعب، دار ، مسرح، سينما، متحف، قاعة مؤتمرات، مبنى حكومي، مركز شرطة، محطة إطفاء، محطة بنزين، محطة قطار، محطة مترو، مكتب بريد، حديقة عامة، حديقة حيوان، ناطحة سحاب، مرفأ، قاعة رياضية، مدرسة لتعليم القيادة، ملعب كرة قدم، محطة توليد الكهرباء، مخبز، مزرعة، مركز تجاري، سفارة، قنصلية، بنك، ملعب تنس، حمام سباحة، منتزه، حديقة مائية، سجن، دار نشر، وكالة سفر، وكالة إعلانات، وكالة توظيف، مصنع سيارات، مصنع نسيج، مصنع إلكترونيات، حديقة نباتية
	Nom de Lieu (Location)	جيجل، الطاهير، الأمير عبد القادر، الشقفة،القنار نشفي،تاكسنة، العوانة، الجمعة بني حبيبي،زيامة منصورية، بني ياجيس، سيدي عبد العزيز،جيملة، سيدي معروف،إيراقن سويسي، بني بلعيد، سلمى بن زياد، برج الطهر،وجانة، الكورنيش، كتامة، تيمزريت، بني فتح،الميلية،السطارة، بني خطاب، أولاد عسكر، بوسيف أولاد عسكر، برج بوفرة،بودريعة بن ياجيس، أولاد رابح، برج الغدير،أراقن، بني معوش، بئر الغزالة، بئر الولجة، بوطالب،بوعفرون، بني صبيح

Le Tableau 4.2 présente une classification des relations spatiales, détaillant les catégories suivantes : relations topologiques, relations directionnelles, relations de distance et relations d'orientation. Ces catégories permettent de comprendre comment les entités spatiales se positionnent, s'orientent et se relient les unes aux autres dans un espace donné. Les relations topologiques décrivent les relations de contiguïté ou d'inclusion, les relations directionnelles indiquent les orientations relatives, les relations de distance mesurent les écarts entre les entités et les relations d'orientation spécifient les alignements ou les angles

entre elles.

Tableau 4.2 : Classification des relations spatiales

Relation spatiale	Relations topologiques	على ضفة، بعض، جزء، بضع، بين، وسط، داخل، في، على، على مستوى، على محور، على حافة، بجانب، حول، قرب، خلف، أمام،على طول، بين نقطتين، ضمن...
	Relations de direction	شمال، جنوب، شرق، غرب، شمال شرق، شمال غرب، جنوب شرق، جنوب غرب، نحو، باتجاه، صوب، قصد، عبر، من خلال، حتى، نحو الأعلى، نحو الأسفل، باتجاه الشمال، باتجاه الجنوب، باتجاه الشرق.....
	Relations de distance	مسافة، على بعد، تبعد، قرب، دنو، على قرب، قصيا، قريب من، بعيد عن، على مسافة، بعيد نسبياً، قريب نسبياً، على مسافة قصيرة، على مسافة طويلة...
	Relations d'orientation	أمام، خلف، قبل، وراء، بعد، يمين، يسار، مقابل، فوق، تحت، أعلى، أسفل، على مستوى، في الأسفل، في الأعلى، فوق سطح، تحت سطح، بجانب، بعيد عن، قريب من...

5. Processus de réalisation de la méthode basée sur des règles

Le processus de la méthode basée sur des règles se décompose en quatre étapes : la définition des règles, l'appariement des règles, l'extraction des informations et le post-traitement.

5.1. Définition des règles

Les experts du domaine définissent des règles basées sur leur connaissance du langage et des structures spécifiques de l'information à extraire. Ces règles sont souvent exprimées en utilisant des expressions régulières, des grammaires contextuelles, ou des patrons de texte spécifiques. Dans notre approche, les règles utilisées sont implémentées via JAPE (Java Annotation Patterns Engine) dans la plateforme GATE.

5.2. Création des règles JAPE

Une grammaire JAPE se compose d'un ensemble de phases, chacune d'entre elles étant constituée d'un ensemble de règles de motif/action (Thakker et al., 2009). Les phases s'exécutent de manière séquentielle et forment une cascade de transducteurs à états finis sur les annotations. Le côté gauche (LHS) des règles est constitué d'une description du motif d'annotation. Le côté droit (RHS) contient des instructions de manipulation des annotations. Les annotations correspondantes sur le LHS d'une règle peuvent être référencées sur le RHS au moyen d'étiquettes attachées aux éléments du motif. Ci-dessous, un exemple de règle JAPE (Figure 4.7) pour extraire des entités nommées de la classe

"ObjetNonNaturel" contenant les instances spécifiées : جبل، هضبة ، شاطئ ، غابة ، واد ، بحر ، صحراء، نهر.

```
Phase: ObjetnonNaturel


Input: Token
Options: control = appelt

Rule: ExtractOrientationRelation

  (
   {Token.string == "جبل"}
  |{Token.string == "واد"}
  |   {Token.string == "هضبة"}
  |   {Token.string == "شاطئ"}
  |   {Token.string == "غابة"}
  |   {Token.string == "بحر"}
  |   {Token.string == "نهر"}
  |   {Token.string == "صحراء"}
  |   {Token.string == "تلة"}
  |  {Token.string == "بحيرة"}

  )
  :entity
  -->
  :entity.ObjetNonNaturel = {
     class = "ObjetNonNaturel",
     ObjetnonNaturel = :entity.Token.string
  }
```

Figure 4.7 : Exemple des règles JAPE dans GATE

5.3. Appariement des règles

Les règles définies (Figure 4.8), elles sont appliquées au texte ou aux données pour identifier les segments qui correspondent aux modèles définis. Une règle pourrait être conçue pour identifier les entités géographiques en cherchant des phrases contenant des mots-clés comme "région", "ville", ou "pays". Dans notre méthode, l'option control = appelt est utilisée pour spécifier que les règles doivent être exécutées séquentiellement. Cela garantit que chaque règle est appliquée dans un ordre précis, maximisant ainsi la précision de l'extraction.

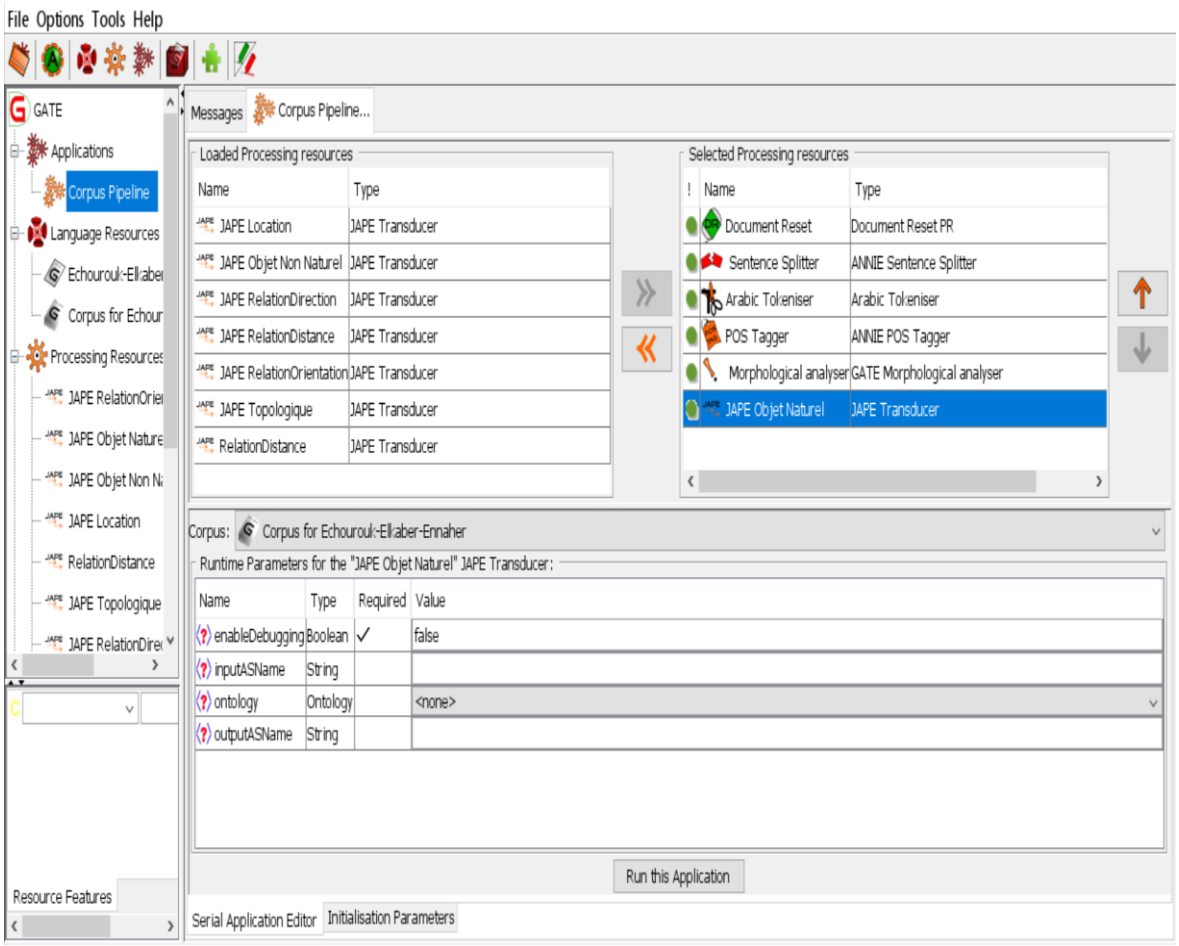

Figure 4.8 : Exemple d'exécution de règles JAPE

5.4. Extraction d'information

Lorsque les segments correspondants sont identifiés, les informations pertinentes sont extraites (Figure 4.9). Cela peut inclure la capture de mots ou de phrases spécifiques, ou l'identification de relations entre différentes entités. Par exemple, la règle "ExtractNaturelObject" dans notre script JAPE vérifie la correspondance du texte à l'une des instances spécifiées dans une liste d'objets naturels, comme "جبل" (montagne), "هضبة" (plateau), ou "بحر" (mer). Lorsqu'un token correspondant un mot dans le texte, une entité nommée "NaturelObject" est créée, et des attributs spécifiques, tels que class = "NaturelObject" et type (avec la chaîne de caractères de l'entité identifiée), sont associés à cette entité.

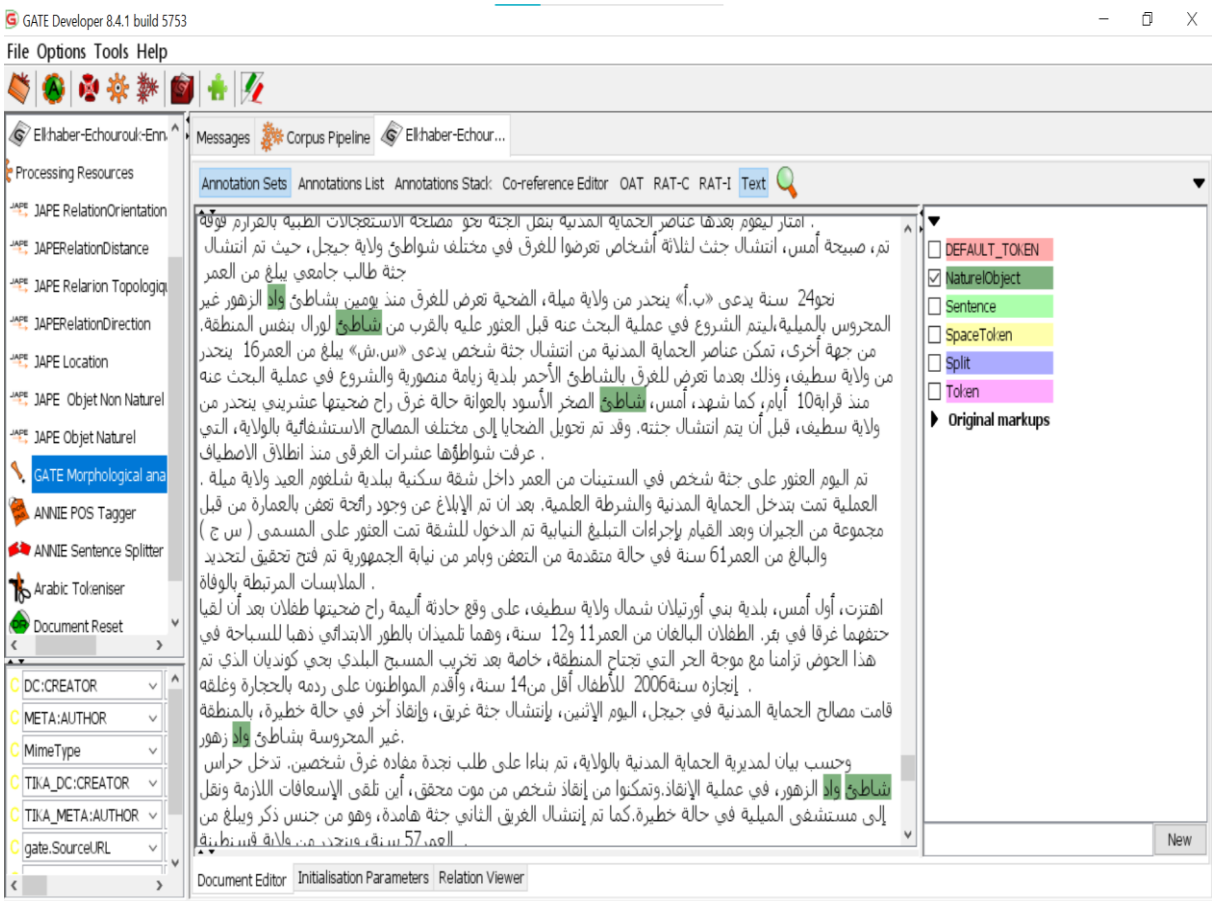

Figure 4.9 : Exemple d'extraction des entités (Objet Naturel)

5.5.Extraction et annotation des entités spatiales

La Figure 4.10 illustre le processus d'extraction et d'annotation des entités spatiales à l'aide des règles JAPE. Ce processus permet de détecter et d'annoter automatiquement des éléments tels que les noms de lieux, les objets non naturels et les objets naturels présents dans le texte. Les règles JAPE, en analysant les modèles linguistiques spécifiques, assurent une identification précise et contextuelle de ces entités spatiales.

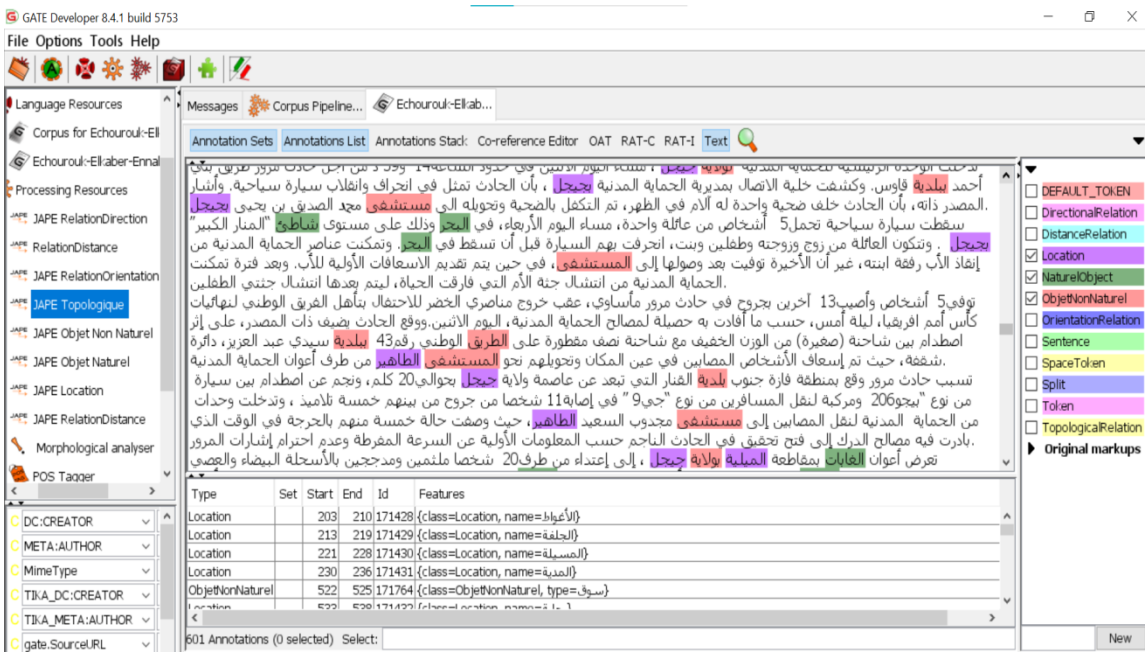

Figure 4.10 : Extraction des entités spatiales basée JAPE

5.6. Extraction et annotation des relations spatiales

La Figure 4.11 illustre le processus d'extraction et d'annotation des relations spatiales à l'aide des règles JAPE. Ce processus permet d'identifier et de classer les relations spatiales présentes dans le texte en quatre catégories distinctes : les relations de distance, de directions, topologiques et d'orientation.

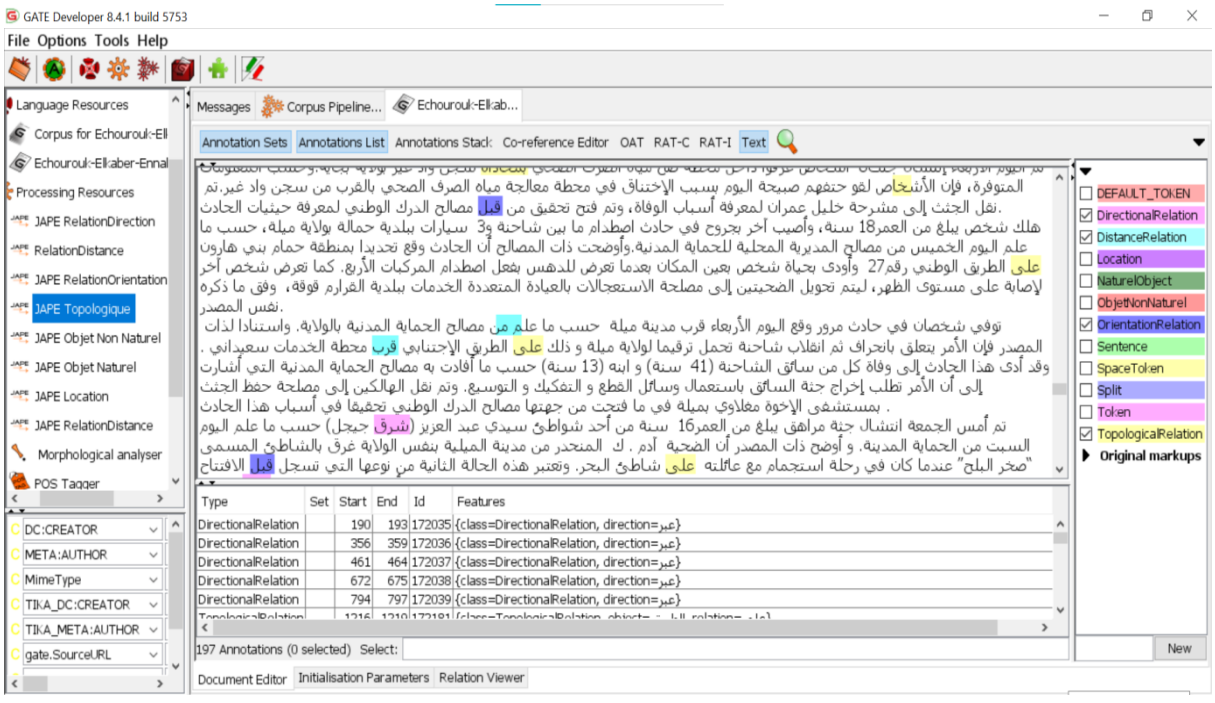

Figure 4.11 : Extraction des relations spatiales basée JAPE.

La Figure 4.12 illustre l'extraction des informations spatiales, comprenant les entités et les relations spatiales.

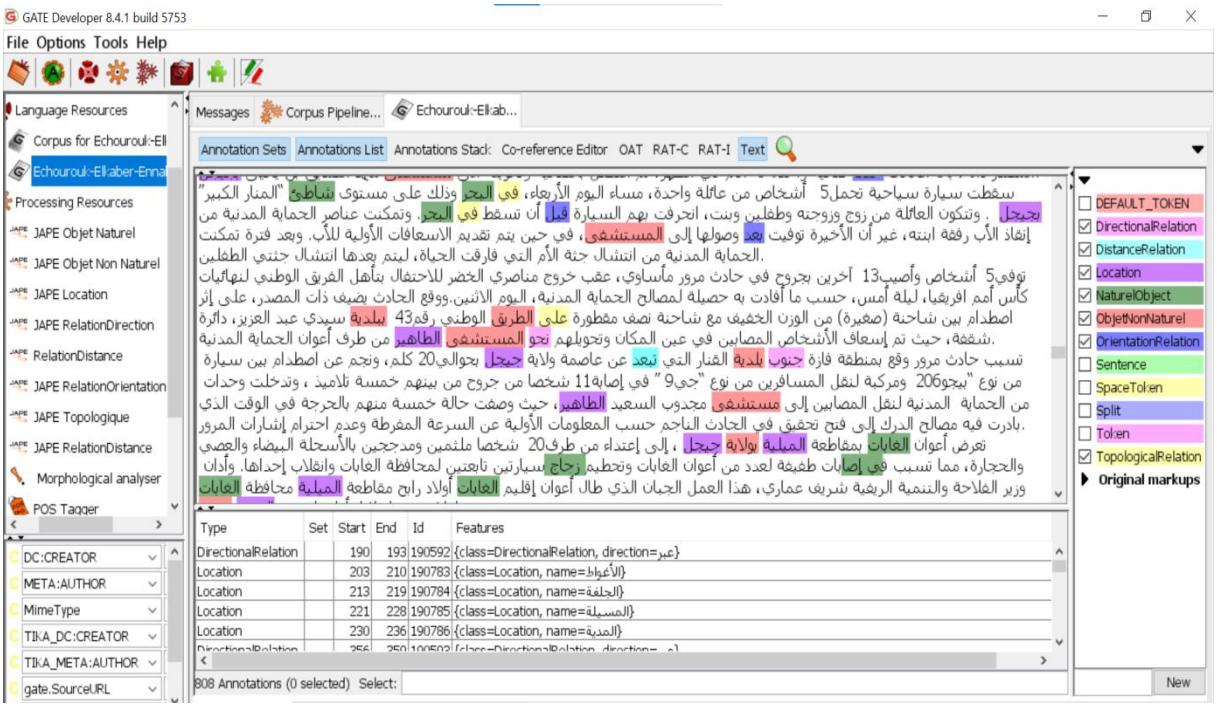

Figure 4.12 : Extraction des Informations spatiales basée JAPE

La Figure 4.13 présente une annotation détaillée réalisée par GATE dans le document. Cette annotation comprend plusieurs éléments critiques pour l'analyse sémantique : le 'type', qui spécifie la catégorie ou le type de l'entité annotée ; 'start' et 'end', qui définissent les positions exactes du début et de la fin du mot dans le texte, permettant ainsi de localiser précisément l'entité dans le document ; 'id', un identifiant unique attribué à chaque entité pour faciliter le suivi et la gestion des annotations dans le système ; 'class', qui indique la classe de l'entité, fournissant un cadre pour l'interprétation des données ; et enfin, le mot annoté lui-même, ou les instances spécifiques de cette classe, offrant une base empirique pour l'analyse linguistique. Cette structuration précise des annotations est essentielle pour une extraction efficace de l'information.

Type	Set	Start	End	Id	Features
TopologicalRelation		15460	15462	229027	{class=TopologicalRelation, object=بلدية, relation=من}
ObjetNonNaturel		15463	15468	229444	{class=ObjetNonNaturel, type=بلدية}
Location		15483	15489	229184	{class=Location, name=سكيكدة}
NaturelObject		15546	15550	229637	{class=NaturelObject, type=شاطئ}
Location		15701	15707	229185	{class=Location, name=المدية}
NaturelObject		16161	16166	229638	{class=NaturelObject, type=البحر}
ObjetNonNaturel		16210	16214	229445	{class=ObjetNonNaturel, type=مكان}
Location		16268	16275	229186	{class=Location, name=الطاهير}
Location		16341	16345	229187	{class=Location, name=جيجل}
Location		16456	16463	229188	{class=Location, name=الطاهير}
NaturelObject		16517	16524	229639	{class=NaturelObject, type=الغابات}

Figure 4.13 : Exemple des Entités Nommées Annotées par GATE

La figure 4.14 présente le corpus annoté final issu de l'extraction des informations spatiales, réalisée à l'aide des règles JAPE. Ce corpus, annoté avec précision, se compose de deux principales catégories : les entités spatiales et les relations spatiales. Les entités spatiales sont classifiées en trois sous-catégories distinctes : emplacement**,** objet naturel etobjet non naturel**.** Cette classification permet de différencier les divers types d'entités présentes dans les données spatiales, facilitant ainsi leur identification et leur traitement. Les relations spatiales, quant à elles, sont annotées en quatre classes : direction**,** orientation**,** topologique etdistance**.** Ces relations décrivent les interactions et les configurations spatiales entre les entités, fournissant un cadre pour l'analyse des relations complexes entre différents objets.

L'objectif de cette annotation est double : d'une part, elle permet de structurer les données géospatiales de manière cohérente, facilitant leur exploitation dans des applications telles que les systèmes d'information géographique (SIG) et les analyses spatiales ; d'autre part, elle améliore la précision et l'efficacité des systèmes d'extraction d'information en fournissant des repères clairs pour l'identification et la catégorisation des entités et des relations. En résumé, un corpus spatial annoté avec ces spécifications contribue significativement à la qualité des analyses et des interprétations dans divers domaines, notamment la gestion des catastrophes, l'urbanisme, et la cartographie.

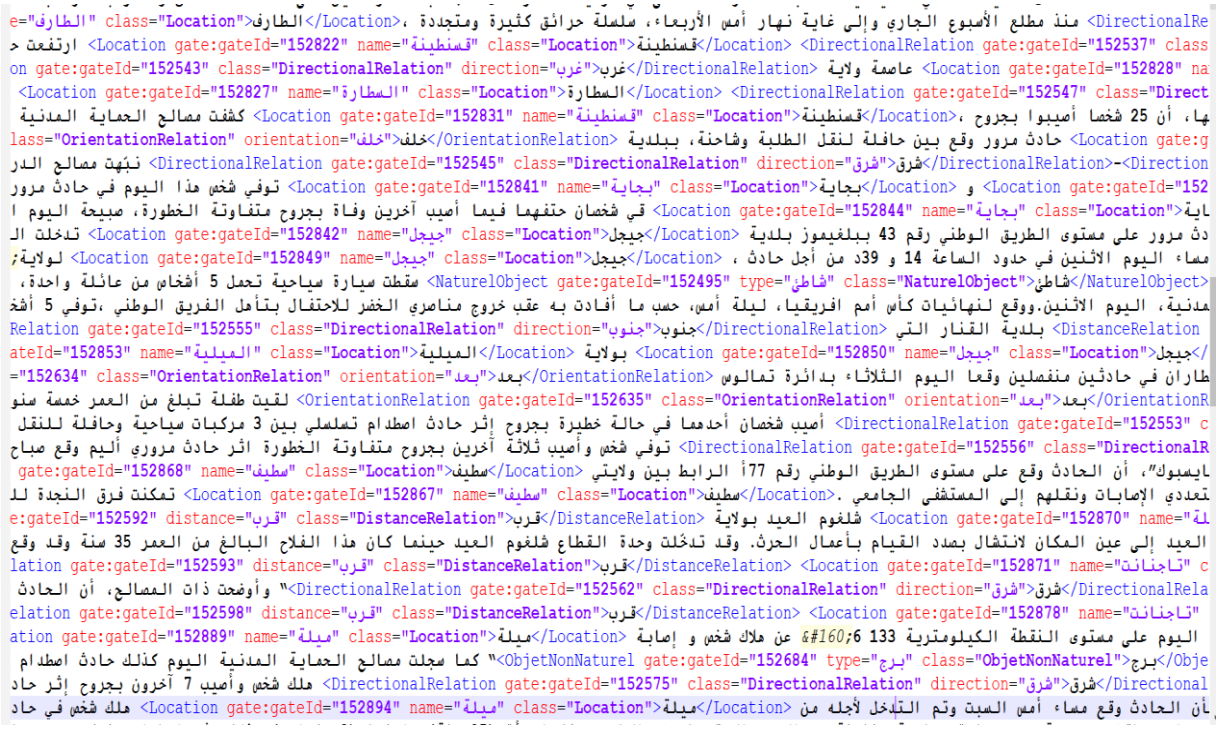

Figure 4.14 : Corpus Annoté GATE

6. Résultats et évaluation

Dans cette section, nous examinons les résultats des expérimentations visant à évaluer l'efficacité de notre système d'extraction d'informations spatiales, qui utilise des règles JAPE. Cette méthode est fondée sur l'application de ces règles pour extraire automatiquement les entités et les relations spatiales à partir d'un corpus de textes Arabes, sans l'utilisation des ontologies.

Le corpus utilisé pour cette évaluation est le même que celui décrit dans le chapitre précédent (Section 3), comprenant un total de 9 008 mots extraits de divers quatre journaux Algérien. Ce corpus a été annoté pour identifier 908 informations spatiales, dont 611 sont des entités spatiales et 197 des relations spatiales. Les résultats obtenus (Tableau 4.3) montrent la répartition des annotations parmi les différents journaux et montre que les entités spatiales constituent la majorité des annotations, représentant 68,4 % du total, tandis que les relations spatiales en représentent 31,6 %.

Cette analyse vise à évaluer la présence d'informations spatiales dans les articles de presse et leur importance dans le contenu global. Les données fournissent une perspective sur la fréquence d'apparition des entités et des relations spatiales dans ces journaux.

Tableau 4.3 : Distribution des informations spatiales dans différents journaux

Journaux	**Informationsspatiales**		**Mots Totales**
	Entitéspatiale	**Relation spatiale**	
Elkhaber	136	47	2072
Echorouk	177	42	2553
Ennaher	146	37	2280
Elmihwer	152	71	2103
Total	611	197	9008

Le Tableau 4.4 présente une analyse détaillée des entités spatiales et des relations spatiales identifiées dans quatre journaux Algériens. Les entités spatiales sont divisées en catégories telles que "LOC", "Objet Naturel" et "Objet Non Naturel", tandis que les relations spatiales sont classées par types : "DIR", "ORI", "TOP" et "DIS". Les données fournissent un aperçu de la fréquence des différentes catégories d'informations spatiales dans chaque journal, mettant en évidence la richesse et la diversité des descriptions spatiales dans ces publications.

Tableau 4.4 : Distribution détaillée des entités et relations spatiales dans les journaux Algériens

Journaux	**Entité spatiale**			**Relation spatiale**			
	LOC	Objet Naturel	Objet Non Naturel	DIR	ORI	TOP	DIS
El khaber	50	9	55	25	15	9	4
Echorouk	97	8	61	21	13	11	1
Ennaher	73	13	76	19	17	6	2
Elmihwer	81	15	73	23	13	13	4
Total	301	45	265	88	58	39	12

Les résultats obtenus montrent queles entités spatiales annotées le plus fréquemment sont les objets non naturels (265) et les lieux (301), tandis que les objets naturels sont moins représentés (Tableau 4.4). Cela pourrait refléter une focalisation sur des entités jugées plus pertinentes dans les contextes des journaux étudiés. En ce qui concerne les relations spatiales, les relations directionnelles sont les plus couramment annotées (88),

suivies par les relations d'orientation (58).

Les relations topologiques et de distance sont beaucoup moins fréquentes, ce qui pourrait indiquer qu'elles sont considérées comme moins importantes ou moins complexes dans les corpus analysés.

Le Tableau 4.5 montre le nombre d'annotations correctes, incorrectes et manquantes pour quatre journaux Algériens. Ces données permettent d'évaluer la précision des annotations spatiales effectuées sur les articles de presse, donnant un aperçu de la qualité des résultats obtenus lors du processus d'extraction

Tableau 4.5 : Résultats d'évaluation des annotations spatiales dans les journaux algériens.

Journaux	**Correcte**	**Incorrecte**	**Manquante**
EnnaherNews	157	11	19
EchoroukNews	161	21	36
ElkhaberNews	142	15	30
ElmihwerNews	175	17	24
Total	635	64	109

Le Tableau 4.6 présente les mesures de performance (précision, rappel et F-mesure) pour chaque journal. Ces indicateurs permettent d'évaluer l'efficacité globale des annotations spatiales en termes de fiabilité et d'exhaustivité des informations extraites.

Tableau 4.6 : Mesures d'évaluation des performances par journal

Journaux	**Précision**	**Rappel**	**F-mesure**
EnnaherNews	0 ,93	0,89	0,90
EchoroukNews	0,88	0,81	0,84
ElkhaberNews	0 ,90	0,82	0,91
ElmihwerNews	0,91	0,87	0,90
Total	0,90	0,85	0,87

La Figure 4.15 illustre les mesures de performance des annotations spatiales pour quatre journaux Algériens, en termes de précision, rappel et F-mesure. Ces indicateurs permettent de visualiser l'efficacité des annotations dans chaque publication, offrant une

vue comparative de la qualité des informations spatiales extraites.

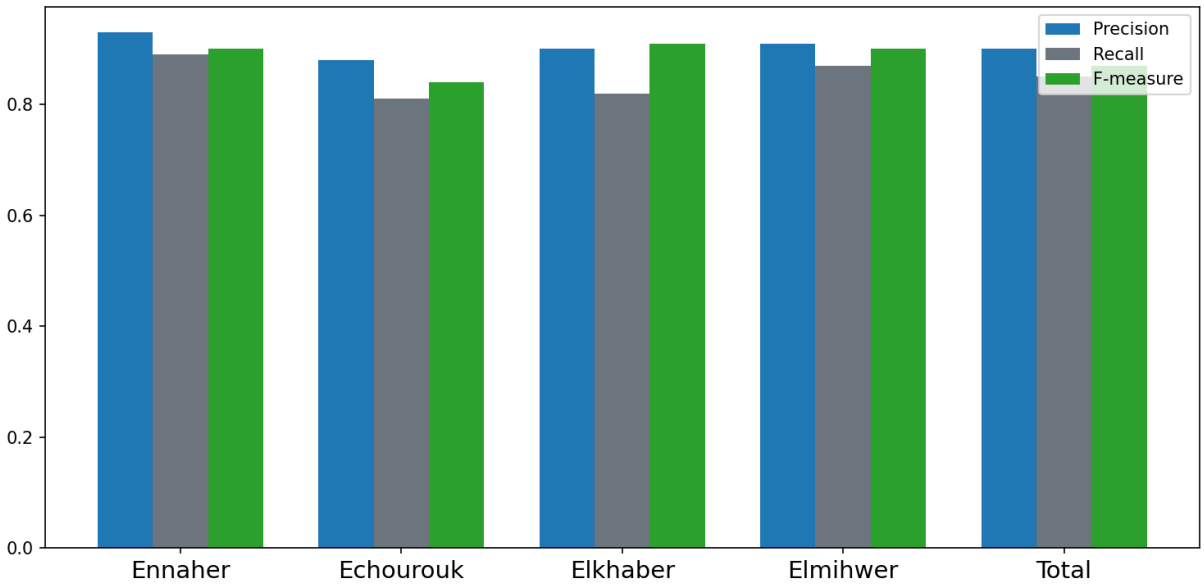

Figure 4.15 : Évaluation des performances des annotations spatiales par journal

7. Analyse et Discussion

Les résultats obtenus indiquent que le système d'annotation est généralement efficace, avec une précision moyenne de 0,90 et une F-mesure de 0,87 dans les journaux. Cependant, les variations entre les journaux montrent des différences notables dans la performance du système. Ennaher et Elkhaber se distinguent par leur meilleure performance globale, ce qui pourrait être attribué à des facteurs tels que la clarté du texte ou la cohérence dans l'utilisation des entités et des relations spatiales. En revanche, Echorouk présente des défis particuliers, notamment un rappel plus faible, ce qui pourrait nécessiter des ajustements dans les méthodes d'annotation ou une meilleure adaptation du système aux spécificités de ce journal.

Les résultats suggèrent que des améliorations ciblées pourraient être apportées pour optimiser la détection des entités et relations spatiales dans les journaux avec des performances moins élevées. Des analyses plus approfondies sont nécessaires pour identifier les sources spécifiques des erreurs et adapter le système en conséquence afin d'améliorer la couverture et la précision des annotations pour tous les types de documents.

8. Conclusion

Ce chapitre a présenté une méthode d'extraction d'informations spatiales reposant sur les règles JAPE. Cette approche utilise des patrons linguistiques pour détecter et structurer automatiquement les entités et relations spatiales dans les textes. Elle permet une annotation systématique et reproductible, particulièrement adaptée aux expressions spatiales explicites. Les résultats montrent une bonne performance dans des cas simples ou récurrents. Toutefois, des limites subsistent face à la variabilité linguistique et à l'ambiguïté contextuelle. Cette méthode repose uniquement sur des règles, sans prise en compte de la sémantique formalisée. Afin d'améliorer la couverture et la précision de l'extraction, une stratégie plus riche est nécessaire. Le prochain chapitre présentera une approche hybride combinant les règles JAPE avec une ontologie spatiale. Cette combinaison vise à enrichir l'analyse linguistique par une modélisation des connaissances spatiales.

Chapitre 5

Nouvelle Approche Hybride Ontologie-Règles

1. Introduction

L'extraction d'information est un domaine clé du traitement automatique du langage naturel (TALN), permettant de transformer des textes non structurés en données exploitables. Cette capacité est essentielle dans les systèmes d'information géographique (SIG), où l'extraction d'entités et de relations spatiales est indispensable pour une modélisation géospatiale précise. Les méthodes traditionnelles, telles que les règles, se basent sur des motifs linguistiques spécifiques, mais souffrent d'une rigidité face à la complexité et à la diversité des textes. Les ontologies apportent une structuration plus flexible, facilitant la gestion des ambiguïtés sémantiques et améliorant l'indexation et l'extraction avec une meilleure précision. Cependant, les approches hybrides, qui combinent la puissance des ontologies avec la précision des règles JAPE, émergent comme des solutions prometteuses, offrant une meilleure adaptabilité et une efficacité accrue dans les contextes complexes, tels que l'extraction d'informations géospatiales.Dans ce contexte nous proposonsdans ce chapitre une approche hybride basée Ontologie-Règle pour extraire automatiquement des informations spatiales à partir de documents en Arabe dans les systèmes d'information géographique. L'objectif principal est d'améliorer les performances des SIG en automatisant certaines tâches et en améliorant les ressources de traitement automatique du langage naturel en Arabe.

2. Approche proposée

Dans la communauté scientifique, l'information géographique est définie comme une composition de trois concepts : l'information spatiale, temporelle et thématique. L'idée principale consiste en la combinaison de ces trois types d'informations qui permet de

décrire un événement qui se produit ou s'est produit à un endroit et à un moment spécifique(De Andrade et al.,2014).

L'intégration des ontologies et des règles suscite un intérêt considérable dans la recherche liée aux ontologies et au Web sémantique (AbdelkouietKholladi,2015). L'essor des ontologies est remarquable, car elles permettent de représenter des connaissances dans divers domaines. Ces connaissances contribuent à la recherche sémantique en améliorant la précision grâce à une meilleure compréhension du but et du sens contextuel des termes tels qu'ils apparaissent dans l'espace de données. L'approche proposée se compose de deux parties principales : premièrement, la construction de l'ontologie des toponymes spatiaux Arabes (ASTO), qui représente les entités et relations spatiales ; deuxièmement, un système d'extraction d'information automatique qui exploite l'ontologie construite ainsi que les techniques de traitement du langage naturel et les règles JAPE dans GATE(Thakker et al.,2009).

L'architecture générale de l'approche proposée est présentée dans la Figure ci-dessous.

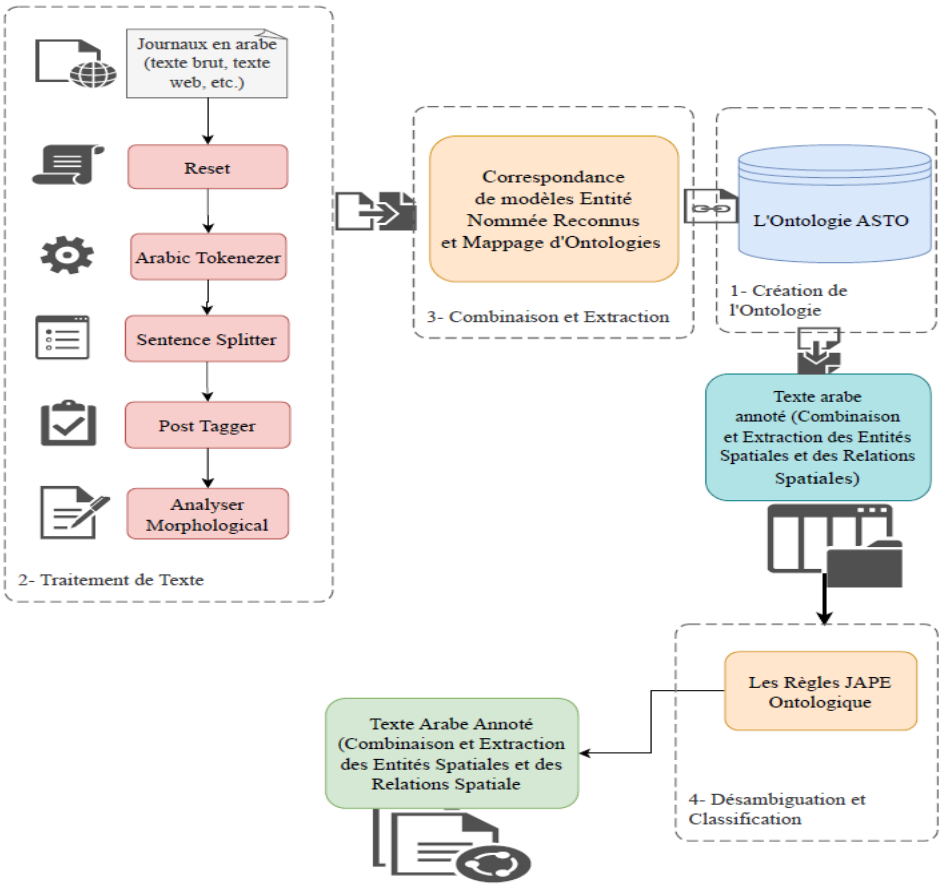

Figure 5.1 : Architecture de système proposé basé Ontologie-Règles (Hadji et al.,2024).

Cette architecture comprend quatre phases :

- Première phase (création de l'ontologie), les concepts sont collectés et des classes (représentant les entités et relations spatiales Arabes) sont créées pour être utilisées dans les étapes de correspondance de texte ;
- Deuxième phase (traitement du texte) applique divers modules pour annoter les entités nommées dans le texte ;
- Troisième phase (combinaison et extraction) utilise l'ontologie pour faire correspondre des classes, sous-classes ou instances avec le texte ;
- Quatrième phase (désambiguïsation et classification) emploie des règles JAPE, qui consiste en des algorithmes développés et implémentés en Java.

3. Source de données et techniques de prétraitement

Pour constituer notre corpus de textes en Arabe, nous avons suivi des critères de sélection rigoureux afin d'assurer diversité pertinence et qualité. Nous avons sélectionné des articles de journaux Arabes en fonction de la diversité des sources médiatiques, de la pertinence du contenu pour l'information spatiale et de la date de publication afin de garantir la pertinence temporelle des données. Les journaux ont été choisis également parmi les plus lu, les plus connus et les plus suivis.

Les articles ont été collectés manuellement à partir de quatre grandes sources d'information Arabes : Echourouk (https://www.echoroukonline.com), Elkhaber (https://www.elkhabar.com), Ennahar (https://www.ennaharonline.com) et Elmihwer (http://elmihwar.dz/ar). Le corpus final couvre divers domaines tels que les catastrophes humaines (les accidents, les incendies,…), les catastrophes naturelles (par exemple, les inondations) et les événements technologiques. Au total, le corpus comprend environ 9 000 tokens, collectés manuellement à partir des sites mentionnés (Aljabari et al.,2024).

Les techniques de prétraitement appliquées à notre corpus de textes Arabes comprennent plusieurs étapes essentielles. Nous avons commencé par le filtrage des données, en supprimant les espaces supplémentaires et les éléments de formatage pour obtenir un texte brut. Ensuite, les erreurs typographiques ont été corrigées à l'aide d'outils automatisés et manuels, suivies par la normalisation du texte, notamment la suppression des diacritiques, la normalisation des espaces et de la ponctuation. Enfin, nous avons effectué une tokenisation manuelle, où les textes ont été divisés manuellement en unités significatives (phrases ou segments) afin d'assurer la précision et l’adéquation de l'analyse

ultérieure. Ces opérations de filtrage et de normalisation sont nécessaires pour garantir que les textes soient correctement préparés pour une analyse approfondie et efficace.

4. Traitement du texte

La plateforme GATE (General Architecture for Text Engineering) est reconnue pour sa popularité dans l'analyse de texte et le traitement automatique du langage naturel. Elle a été largement utilisée dans divers projets, y compris l'extraction d'information (IE) dans différentes langues. L'avantage principal de GATE réside dans son extensibilité, car elle permet l'ajout de nouveaux composants ou applications à ses composants standards existants, élargissant ainsi les capacités d'analyse et de traitement de texte. L'objectif principal de GATE est l'annotation de documents, qui peut être réalisée par trois méthodes : entièrement automatique à l'aide d'applications, annotation manuelle par les utilisateurs, et annotation semi-automatique via une combinaison de traitement de corpus et de correction ou ajout manuel de nouvelles annotations (Al-Laith, 2021).

GATE propose une gamme de modules spécialisés pour l'analyse textuelle, accessibles via un système de plugins. Certains des modules couramment utilisés incluent les tokenizeurs (segmenteurs) pour décomposer le texte en unités plus petites, les POS Taggers (étiqueteurs morpho-syntaxiques) pour catégoriser les mots, les Gazetiers (lexiques) pour identifier des termes spécifiques et les transducteurs (JAPE) pour la correspondance basée sur des motifs. Les entités nommées extraites par GATE englobent divers types tels que les lieux, organisations, personnes et dates, entre autres. Avec GATE, les utilisateurs peuvent créer de nouvelles annotations en chargeant des ressources supplémentaires telles que des corpus, des documents et des textes, ainsi qu'en intégrant de nouveaux plugins. Cela permet de combiner et de paramétrer ces ressources au sein de la même chaîne de traitement (Attia, 2007). En tant que plateforme open-source dédiée à l'ingénierie textuelle, GATE est bien adaptée pour le développement de systèmes nécessitant des capacités de traitement du langage naturel et d'extraction d'information.

4.1. Segmentation des phrases

Dans le traitement automatique du langage naturel (TALN), diviser un texte en phrases est une étape pour extraire des informations plus significatives. Chaque phrase est délimitée par des identifiants, tels que des signes de ponctuation comme les points, points d'exclamation ou points d'interrogation. Initialement, une phrase est considérée comme une unité de discours acceptable pour l'application ou le système TALN utilisé. Segmenter un

texte en phrases permet aux systèmes TALN de traiter et d'analyser des portions plus petites du texte à la fois, facilitant ainsi la compréhension du contexte et l'extraction des informations pertinentes (Figure 5.2) (Hadji et Kholladi, 2023).

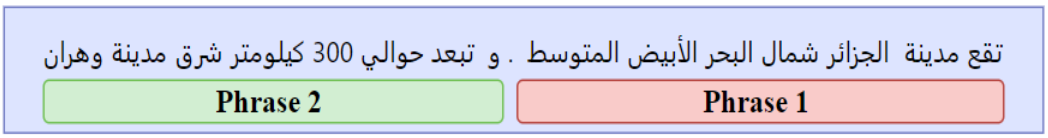

Figure 5.2 : Phase deSegmentation dans le TALN

En fragmentant le texte en phrases, les modèles TALN peuvent se focaliser sur les relations entre les mots au sein de chaque phrase, ce qui aide dans des tâches telles que la reconnaissance d'entités nommées, l'étiquetage des parties du discours, l'analyse de sentiment, et d'autres tâches de compréhension du langage. La segmentation des phrases est une étape de prétraitement essentielle dans de nombreuses applications TALN, car elle constitue la base pour une analyse plus approfondie et permet d'extraire des informations d'un texte donné.

4.2. TokeniseurArabe

Le tokeniseur est un composant essentiel utilisé pour diviser le texte en mots, nombres, symboles, espaces et signes de ponctuation. Chacune de ces unités est appelée un token, représentant un élément syntaxique distinct, qui peut être un mot complet, une partie de mot, une expression multi-mots ou un signe de ponctuation (Figure 5.3).

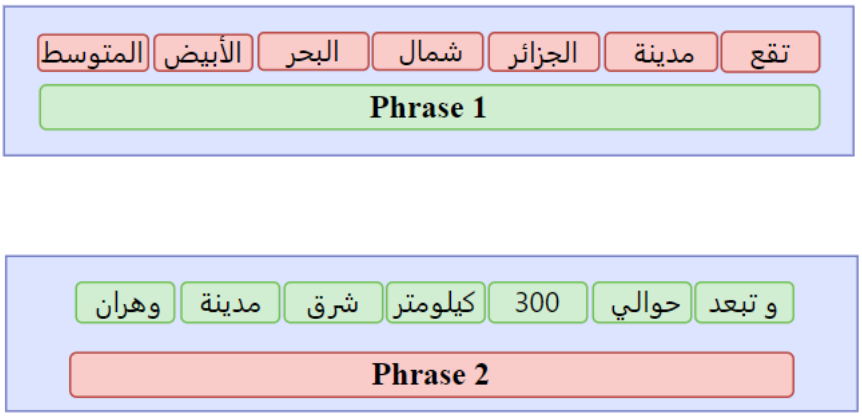

Figure 5.3 : Phase de Tokenisation dans le TALN

Le tokeniseur se base sur une liste prédéfinie de délimiteurs de mots, tels que les espaces et les signes de ponctuation, et prend en compte les nuances des tokens à l'intérieur

des mots, en particulier lorsqu'il s'agit de racines de mots et de clitiques. Dans cette recherche, les mots complets, y compris les racines avec ou sans clitiques, ainsi que les nombres, sont appelés des tokens principaux. Ces tokens principaux sont délimités par des espaces ou des signes de ponctuation. De plus, les mots en forme complète peuvent être divisés en sous-tokens en séparant les clitiques des racines (Attia, 2007).

4.3. Étiquetage morpho-syntaxique (POS tagging)

L'étiquetage morpho-syntaxique (POS tagging) est une étape importante dans le processus d'extraction et de reconnaissance des entités nommées. Il identifie et ajoute des étiquettes au modèle de texte tokenisé, c'est-à-dire qu'il identifie les noms, verbes, adjectifs et autres parties du discours pour chaque token (Figure 5.4). L'étiqueteur POS utilisé dans cette étude avec GATE est le tagueur de Hepple(Borisova,2014).

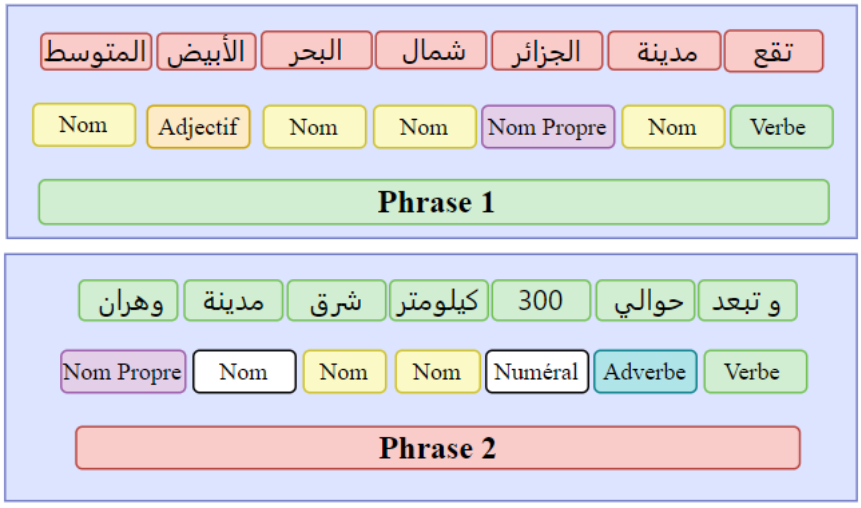

Figure 5.4 : Phase de l'étiquetage morpho-syntaxique.

4.4. Analyse morphologique

La morphologieen linguistique, est l'étude de la structure interne et de la composition des mots et de leur formation. Un morphème est défini comme la plus petite unité de sens dans une langue, qui ne peut pas être divisée en parties plus petites. L'analyseur morphologique aide à regrouper les mots qui expriment des concepts similaires(Borisova,2014). Le rôle de l'analyseur morphologique en Arabe est d'identifier les morphèmes d'un mot (racine) : les affixes (préfixe, infixe, et suffixe) et la racine.L'analyse morphologique se concentre sur la structure interne des mots, notamment en identifiant les racines et les affixes (préfixes, infixes, suffixes). Cette étape est

particulièrement importante pour la langue Arabe, où la morphologie est riche et complexe.

Exemples d'analyse

- **مدينة**: Racine = مدن, Suffixe = ة
- **تقع**: Racine = وقع
- **شمال**: Racine = شمل

L'analyse morphologique aide à normaliser les formes des mots, facilitant ainsi leur reconnaissance lors des étapes ultérieures d'extraction d'information.

4.5. Combinaison et extraction

Notre approche hybride d'extraction d'information spatiale basée sur l'OBIE est une contribution au domaine de l'extraction d'information basée sur l'ontologie. De plus, les ontologies permettent des descriptions formelles et déclaratives de termes communs et soutiennent le raisonnement automatique ou semi-automatique sur les données partagées dans un domaine.

Une fois que les phrases ont été segmentées, tokenisées, étiquetées et analysées morphologiquement, le système peut procéder à l'extraction des entités spatiales (comme "مدينة الجزائر" et "مدينة وهران") et des relations spatiales (comme "شمال" et "شرق").

La combinaison et l'extraction se fait en utilisant l'ontologie spatiale ASTO par l'identification,l'indexation et des techniques de reconnaissance d'entités nommées. Les entités et relations extraites sont ensuite annotées dans le texte, permettant une meilleure compréhension des informations géospatiales qu'il contient.

Le processus commence par le chargement de l'ontologie dans GATE en tant que ressource ontologique OWL (Figures 5.5) et fournit des moyens de charger des corpus à partir d'une URL ou d'un fichier dans l'application Pipeline. Le système, à cette étape, fait correspondre les concepts, instances et relations de l'ontologie avec un texte Arabe en entrée pour l'extraction et l'annotation des informations spatiales. Par conséquent, lorsque le système détecte un mot identique dans le texte d'entrée qui est une classe ou une instance dans l'ontologie, il l'annote. Le résultat obtenu à ce stade est un corpus annoté contenant l'ensemble des mots annotés (informations spatiales) avec une ambiguïté (Figure 5.6), car il n'est pas possible de déterminer uniquement les entités spatiales, les relations spatiales, ainsi que les catégories associées à chaque entité ou relation.

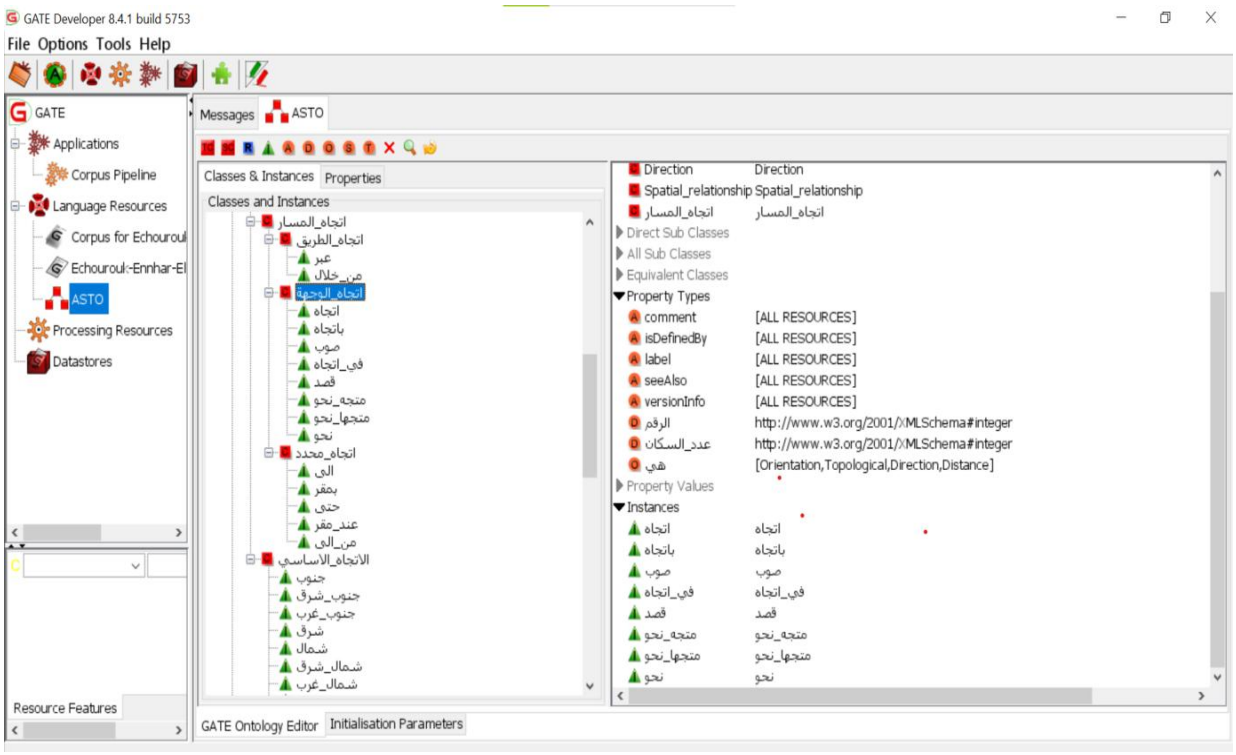

Figure 5.5 : Chargement de l'ontologie ASTO dans GATE

La Figure 5.6 illustre l'exécution d'un pipeline basé exclusivement sur l'ontologie, mettant en évidence des informations spatiales ambiguës sans distinction précise des entités ou des relations spatiales

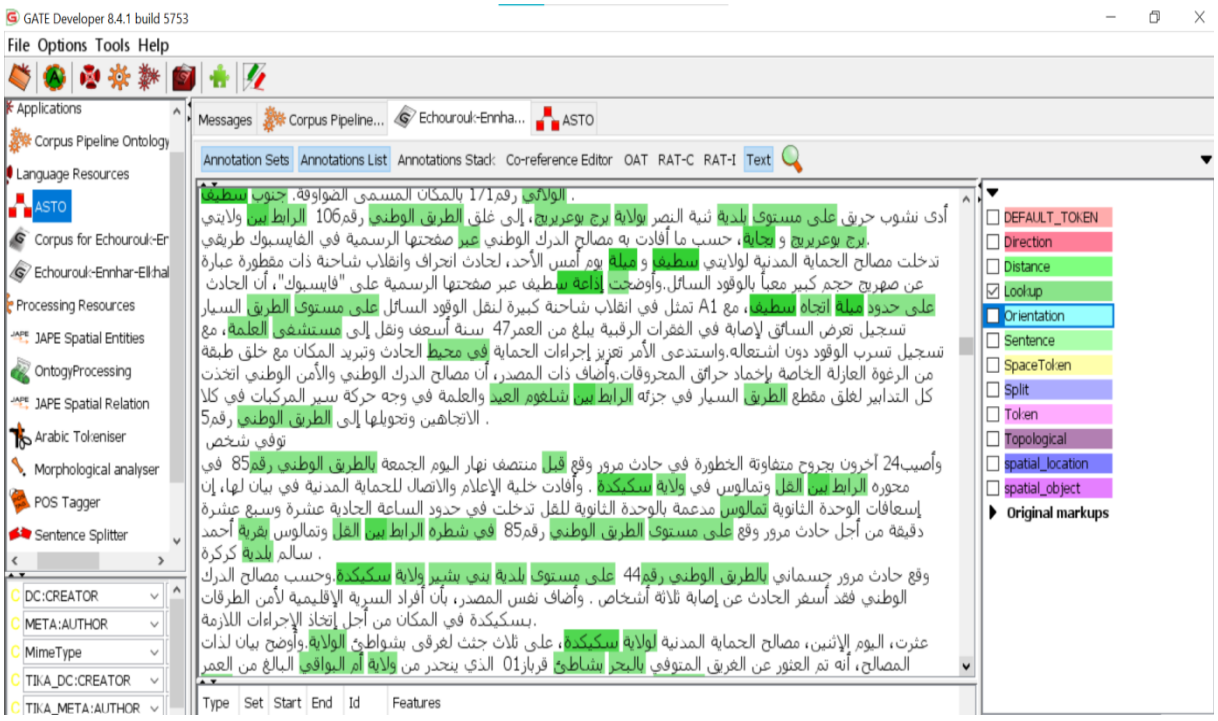

Figure 5.6 : Extraction basée sur l'ontologie sans l'utilisation des règles JAPE

4.6. Désambiguïsation et classification

Les règles JAPE[1]consistent en une séquence de phases, chacune composée d'une séquence de règles modèle/action. Chaque phase s'exécute séquentiellement et est une série de transducteurs d'état qui se termine par des annotations. Les règles se composent de deux sections : la section de gauche (LHS) décrit un modèle d'annotation et la section de droite (RHS) consiste en des instructions de manipulation d'annotations. Les annotations correspondantes sur le côté gauche d'une règle peuvent être identifiées sur le côté droit par le biais d'étiquettes liées aux éléments du modèle(Thakker et al.,2009).Les fonctions principales du transducteur JAPEdans cette étape sont de combiner et d'extraire les entités spatiales, les relations et leurs classes de l'ontologie ASTO avec le corpus d'entrée (Figures 5.7). Le système vérifie ensuite dans les règles de l'ontologie que le mot annoté (sous classe ou instance) appartient (en tant que sous-classe ou instance) à la classe correspondant à la partie gauche (LHS) de la règle JAPE. De ce fait, le système classe le mot annoté dans cette catégorie.

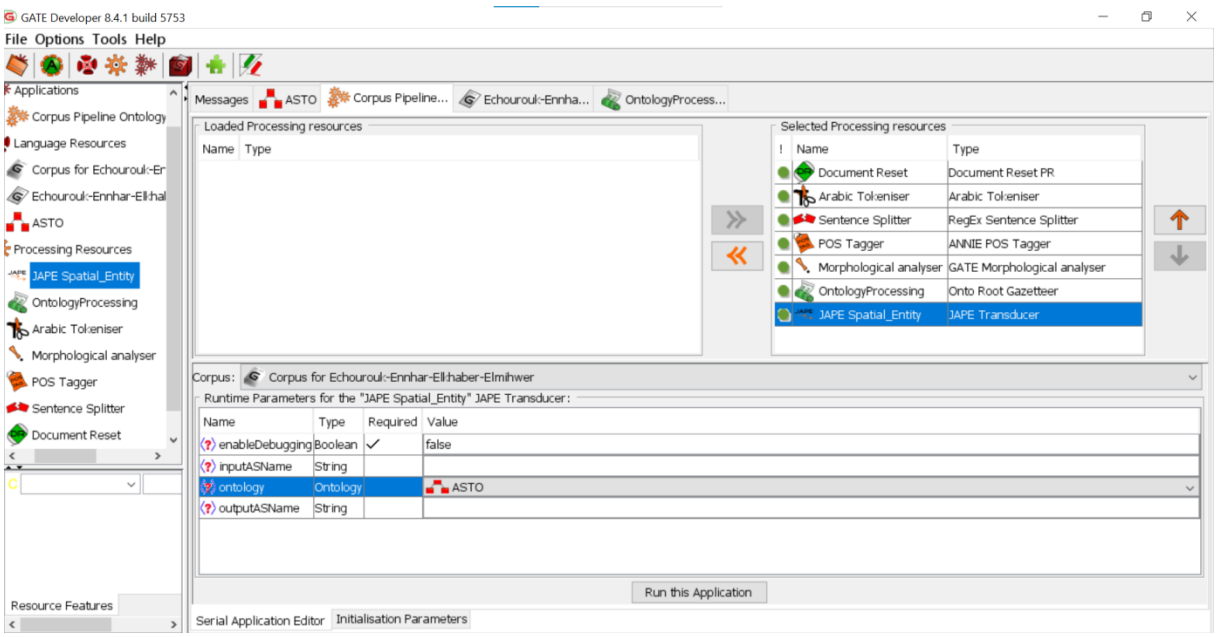

Figure 5.7 : Manipulation de règles JAPE et indexation parASTO

L'étape finale consiste à désambiguïser et classer les entités spatiales et leurs relations en utilisant des règles spécifiques (les règles JAPE dans GATE). Cela permet de déterminer avec précision la nature des entités et des relations identifiées, en distinguant les différents types de relations spatiales (comme la direction, la distance, etc.) et en les

[1]https://gate.ac.uk/sale/tao/splitch8.html

classifiant correctement dans l'ontologie (Figure 5.8).

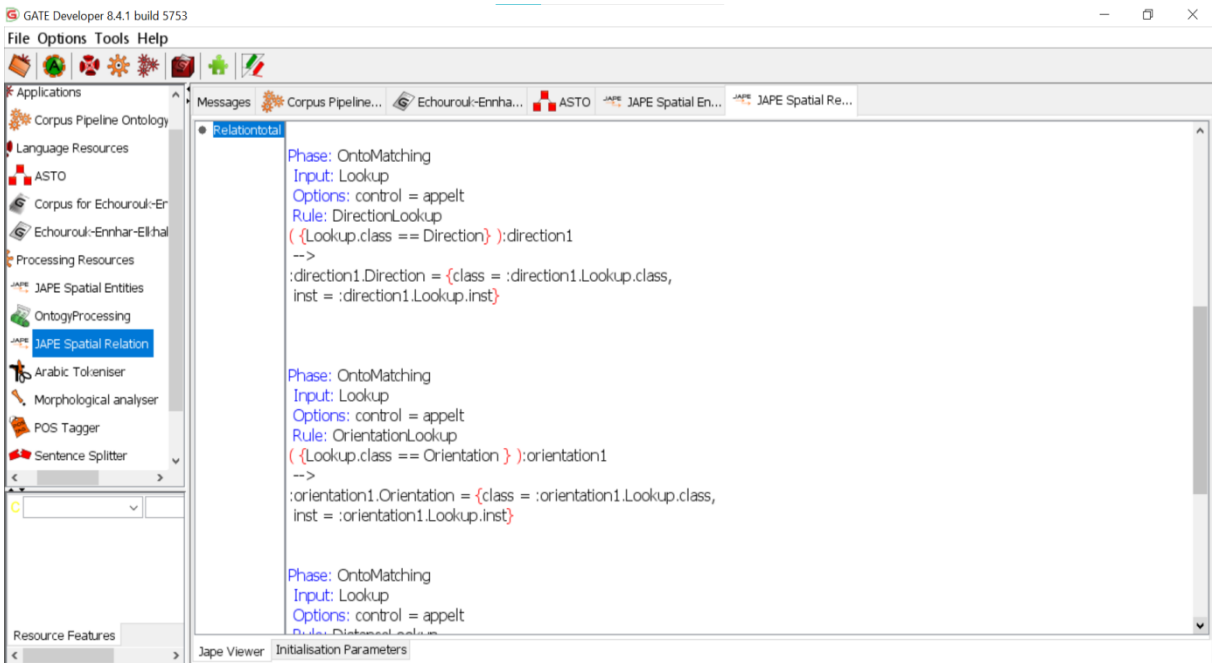

Figure 5.8 : Règles JAPE pour manipuler sur ASTO

4.7. Algorithme pour l'extraction des entités spatiales

La figure ci-dessous montre l'algorithme d'extraction des entités spatiales :

```
1.Input: TextRaw T, Ontology  Ont, Sentence S,Word W;
2.Output: Annotated-and-Extracted-Text  CorpusT;
3.Begin
4.Parse T, Read words W from the text;
5. For each sentence S in T do
6.   For each word  W in S do
7.     For each Class in Onto do
8.       If  Ont.Class == Spatial-Object  Then
9.        While (Ont.Spatial-Object.SubClass <> vide) do
10.         If (W==Ont.Spatial-Object.SubClass.inst) Then
11.           Annotate (W,Spatial-Object ,CorpusT);
12.           Extract-in (W,CorpusT , Class-Spatial-Object);
13.           Onto.Subclass=Onto.Subclass+1;
14.         Endif
15.       End while
16.     Endif
17.     If (Ont.Class == Spatial-Location ) Then
18.       While (Ont.Spatial-location.SubClass <> vide) do
19.         If (W == Onto. spatial-Location.SubClass.inst) Then
20.           Annotate (W,Spatial-Locaion ,CorpusT) ;
21.           Extract-in (W,CorpusT, Class-spatial-Location);
22.           Onto.Subclass = Onto.Subclass+1 ;
23.         Endif
24.       End while
25.     Endif
26.   EndFor
27. EndFor
28.EndFor
29.End
```

Figure 5.9 : L'algorithme d'extraction des entités spatiales

Les Entrées :T : Texte brut à analyser ; **Ont** : Ontologie contenant des classes et sous-classes ; **S** : Phrase extraite du texte ; **W** : Mot extrait de la phrase ;

Les Sorties :Corpus T : Corpus de texte annoté et classifié.

Les étapes de l'algorithme sont les suivantes :

- **Initialisation**

Le texte T est analysé pour extraire les mots individuels W. Il s'agit d'une étape de prétraitement qui prépare le texte pour une analyse plus approfondie.

- **Traitement des phrases**

Chaque phrase S du texte est traitée individuellement. Traitement des mots : Chaque mot W de la phrase est examiné pour déterminer s'il correspond à une classe de l'ontologie Ont.

- **Classes d'objets spatiaux**
  - Si la classe actuelle dans l'ontologie est un objet (spatial-Objet), l'algorithme vérifie chaque sous-classe pour voir si le mot W est une instance de cette sous-classe.
  - Si une correspondance est trouvée, le mot W est annoté et extrait comme un objet spatial dans le corpus de sortie CorpusT. L'index de la sous-classe est ensuite incrémenté pour continuer à vérifier d'autres sous-classes.
- **Classes de lieux spatiaux**
  - De la même manière, si la classe actuelle est un lieu (spatial-Lieu), l'algorithme vérifie chaque classe pour voir si le mot W est une instance de cette classe.
  - Si une correspondance est trouvée, le mot W est annoté et extrait comme un lieu spatial dans le corpus de sortie CorpusT. L'index de la classe est ensuite incrémenté pour continuer à vérifier d'autres classes.
- **Fin de l'algorithme**
  - L'algorithme se poursuit jusqu'à ce que tous les mots dans toutes les phrases aient été traités et extraits selon les classes de l'ontologie Ont. Nous avons utilisé le même algorithme pour annoter et extraire chaque relation spatiale (topologique, distance, orientation, direction).

Par exemple, le transducteur d'entités spatiales utilise une règle JAPE explicitement définie, capable de correspondre aux modèles d'entités objet (Naturel/Bâtiment) et lieu qui sont listés ci-dessous (Figure 5.10).

```
// Classification_Spatial_object

Phase: OntoMatching_Spatial_Object
Input: Lookup
Options: control = appelt
Rule: Lookup
(
 {Lookup.class == spatial_object}): Object1
 -->
: Object1.Spatial_Object = {class =: Objct1.Lookup.class,
 inst =: Objct1.Lookup.inst}


// Classification_Spatial_Location

Phase: OntoMatching_Spatial_Location
Input: Lookup
Options: control = appelt
Rule: Lookup
(
 {Lookup.class == spatial_location}): Location1
 -->
: Location1.Spatial_Location = {class =: Location1.Lookup.class,
 inst =: Location1.Lookup.inst}
```

Figure 5.10 : Exemple de règles JAPE pour l'extraction de la localisation spatiale et de l'objet spatial

Nous utilisons des règles JAPE pour la désambiguïsation et la classification des entités spatiales et des relations spatiales dans notre corpus de texte. La Figure 5.10 illustre deux phases distinctes, chacune visant à optimiser l'annotation et la classification des données spatiales.

·Phase OntoMatching_Spatial_Object

Cette première phase se concentre sur l'annotation des objets spatiaux. Les règles JAPE appliquées vérifient si une annotation correspond à la classe spatial_object définie dans notre ontologie. Lorsqu'une correspondance est trouvée, l'annotation est enrichie d'informations détaillées, y compris la classe et l'instance de l'objet. Ce processus garantit non seulement une identification précise des objets spatiaux, mais aussi une désambiguïsation efficace en associant chaque entité à une définition ontologique claire.

· Phase Onto Matching_Spatial_Location

La deuxième phase suit un processus similaire pour les lieux spatiaux. Les règles

JAPE vérifient la correspondance avec la classe spatial_location et enrichissent les annotations correspondantes en conséquence. Ce traitement améliore la précision des lieux spatiaux et assure une annotation cohérente et informée.

Ces deux phases permettent l'identification automatisée et la classification enrichie des entités spatiales, basées sur une ontologie bien définie. Nous avons utilisé les mêmes règles JAPE pour la classification et la désambiguïsation de chaque relation spatiale (topologique, distance, orientation, direction).

La Figure 5.11 illustre l'extraction des entités spatiales, telles que les noms de lieux et les objets spatiaux.

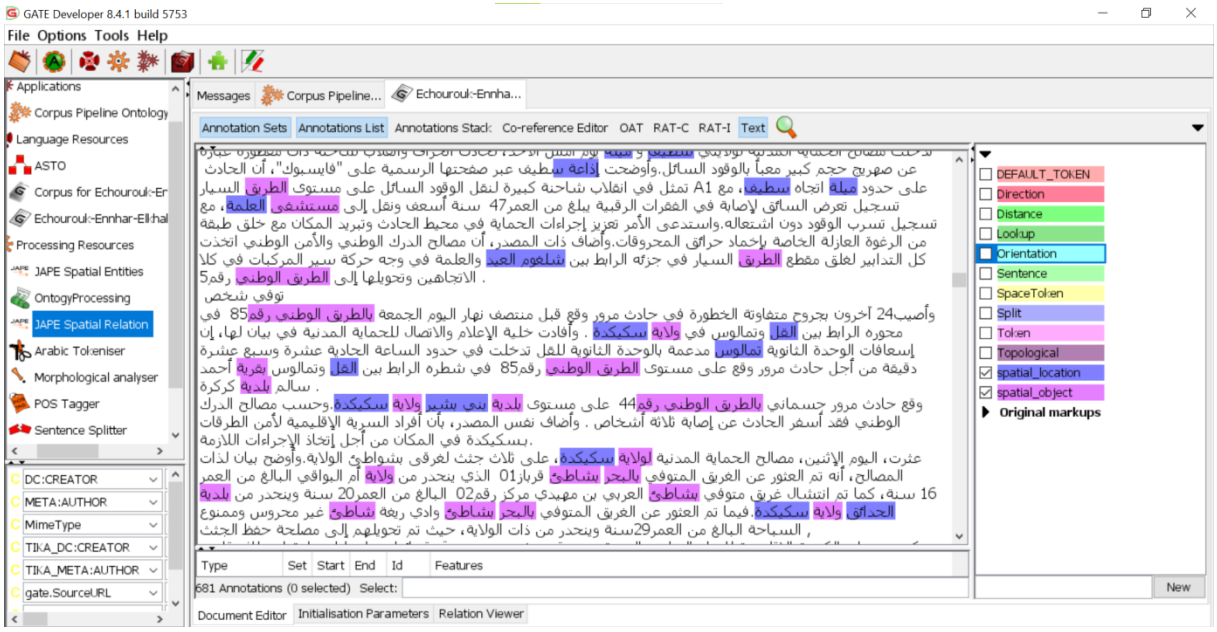

Figure 5.11 : Extraction des entités spatiales basée sur ASTO et JAPE

En complément, la Figure 5.12 montre l'extraction des relations spatiales, incluant la direction, la distance, l'orientation, et les relations topologiques**.**

Figure 5.12 : Extraction des Relations spatiales basée sur ASTO et JAPE

L'annotation des entités nommées par l'ontologie ASTO et les règles ontologiques JAPE (Figure 5.13) est un processus clé dans l'extraction des informations spatiales à partir de textes. Ce processus repose sur deux éléments essentiels : l'ontologie ASTO et les règles JAPE, qui travaillent en parallèle pour identifier et classer les entités et les relations spatiales.

L'ontologie ASTO fournit un cadre de référence structuré pour la catégorisation des entités spatiales. Elle définit des concepts et des relations spécifiques qui sont utilisés pour annoter les entités spatiales dans le texte.

Les règles JAPE (Java Annotation Patterns Engine) sont ensuite appliquées pour extraire ces entités et relations à partir des textes. Ces règles exploitent les définitions et les relations spécifiées dans l'ontologie ASTO pour identifier avec précision les occurrences d'entités spatiales dans les textes, ainsi que les relations qui existent entre elles.

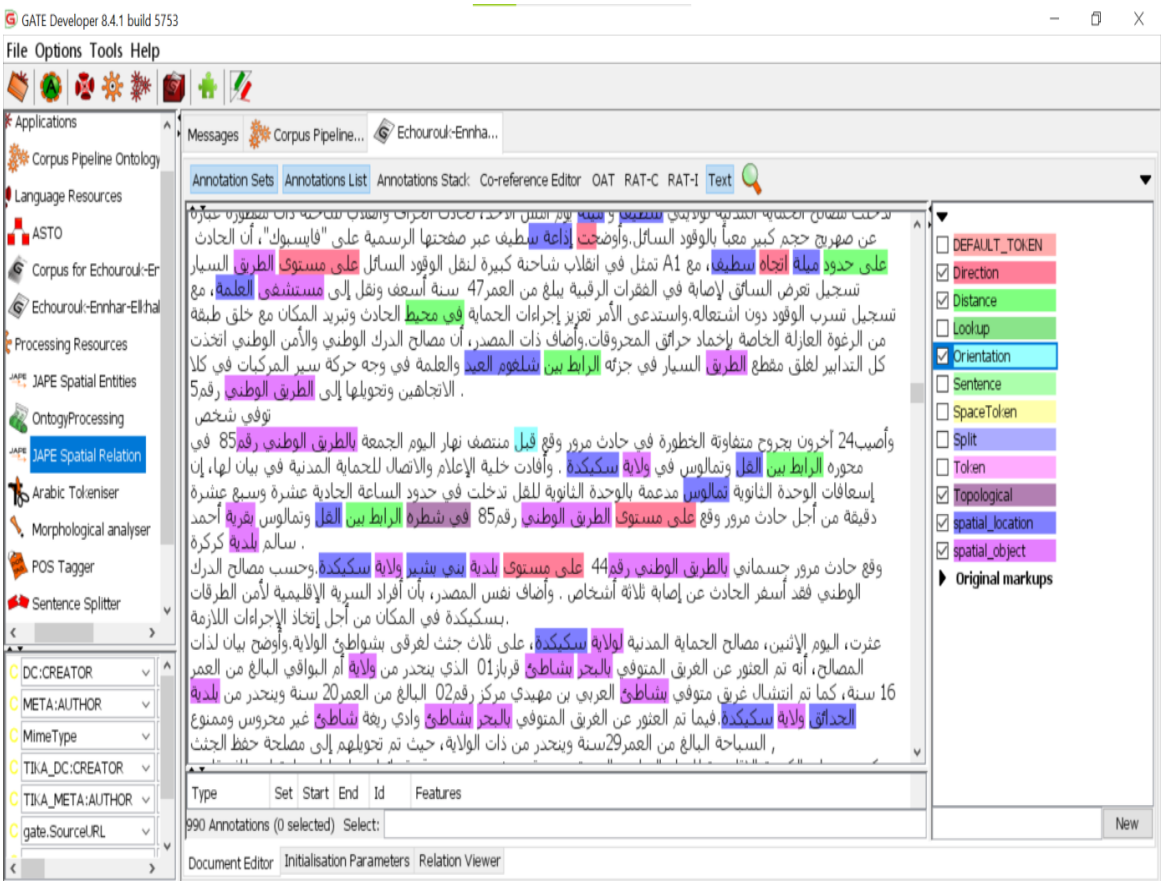

Figure 5.13 : Extraction des informations spatiales basée sur ASTO et JAPE

La Figure 5.14 illustre la structure d'annotation des entités nommées dans GATE.

Type	Set	Start	End	Id	Features
spatial_location		2122	2126	238487	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#دائرة_سطيف, inst=http://www.se
spatial_object		2297	2302	238800	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#مادي, inst=http://www.semanticw
spatial_object		2491	2497	238801	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#مادي, inst=http://www.semanticw
spatial_location		2498	2504	238488	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#دائرة_العلمة, inst=http://www.se
Direction		2507	2510	239155	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#الاتجاه_الاساسي, inst=http://www
spatial_location		2511	2515	238489	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#دائرة_سطيف, inst=http://www.se
spatial_object		2569	2576	238802	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#مادي, inst=http://www.semanticw
spatial_object		2777	2783	238803	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#مادي, inst=http://www.semanticw
spatial_location		2784	2788	238490	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#دائرة_جيجل, inst=http://www.sem
spatial_object		2827	2833	238804	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#مادي, inst=http://www.semanticw
Topological		2894	2897	239372	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#جزئي, inst=http://www.semanticw
Topological		2918	2922	239373	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#كلي, inst=http://www.semanticwe
spatial_object		2923	2929	238805	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#مادي, inst=http://www.semanticw
Orientation		3134	3137	239286	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#أفقي, inst=http://www.semanticw
Topological		3158	3161	239374	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#كلي, inst=http://www.semanticwe
spatial_object		3360	3365	238806	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#طبيعي, inst=http://www.semantic
spatial_object		3374	3380	238807	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#مادي, inst=http://www.semanticw
spatial_object		3391	3397	238808	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#مادي, inst=http://www.semanticw
spatial_location		3398	3404	238491	{class=http://www.semanticweb.org/hp/ontologies/2022/7/ArabicSpatialToponymOntology#دائرة_سكيكدة, inst=http://www.se

Figure 5.14 : Entités nommées annotées par l'ontologie ASTO et JAPE dans GATE

En combinant l'ontologie ASTO avec les règles JAPE, ce processus d'annotation permet de structurer les données géospatiales de manière cohérente, aboutissant à un corpus annoté conforme aux principes du Web sémantique (Figure 5.15).Cette structuration facilite l'exploitation des données dans diverses applications, notamment les SIG. De plus, ce cadre améliore la précision des systèmes d'extraction d'informations en fournissant des repères clairs pour l'identification et la catégorisation des entités et des relations spatiales, ce qui est essentiel dans des domaines tels que la gestion des catastrophes, l'urbanisme et la cartographie.

Figure 5.15 : Corpus annoté par l'ontologie et les règles JAPE dans GATE

5. Résultats et évaluation

Dans cette section, nous décrivons les expériences menées pour valider l'efficacité du système. Nous introduisons une méthode hybride de recherche automatique basée sur l'ontologie, impliquant la création d'une nouvelle ontologie spatiale contenant divers éléments de connaissances spatiales en Arabe, tels que des concepts, des instances et des relations. De plus, des règles ontologiques dérivées de la grammaire JAPE sont utilisées pour extraire les entités spatiales et les relations spatiales en faisant correspondre les concepts de l'ontologie avec les corpus de documents.

Le Tableau 5.1 présente le nombre d'extractions d'informations spatiales issues de divers journaux Arabes, ainsi que le total de mots analysés dans chaque source. Cette analyse permet d'évaluer la richesse des entités spatiales détectées dans chaque corpus.

Tableau 5.1: Distribution du nombre d'informations spatiales dans les journaux

Journaux	**Information Spatiale**	**Mots Totals**
Elkhaber	216	2072
Echorouk	261	2553
Ennaher	244	2280
Elmihwer	240	2103
Total	981	9008

Pour le test initial, nous avons sélectionné des textes de journaux Arabes et compilé un corpus de 9 008 tokens. Dans ce corpus, 981 mots ont été annotés avec des informations spatiales et de 671 entités spatiales ont été identifiées. De plus, 310 relations spatiales ont été spécifiquement annotées. Notamment, 68,4 % des mots annotés avec des informations spatiales sont des entités spatiales et 31,6 % de ces mots annotés représentent en réalité des relations spatiales. Ces statistiques mettent en évidence la richesse et la diversité des annotations spatiales au sein du corpus, soulignant les interconnexions complexes entre les entités et les relations spatiales. Cette analyse détaillée jette les bases de l'évaluation de l'efficacité de l'approche proposée pour capturer et comprendre les informations spatiales dans les textes Arabes (Tableaux 5.1 et 5.2.)

Le Tableau ci-dessous présente une analyse comparative des informations spatiales, des entités spatiales, et des relations spatiales extraites de plusieurs journaux Arabes. Cette comparaison met en évidence la répartition et la masse des données géospatiales dans chaque publication

Tableau 5.2 : Analyse des entités et relations spatiales extraites des journaux

Journaux	**Information spatiale**	**Entitéspatiale**	**Relation spatiale**
Elkhaber	216	134	82
Echorouk	261	190	71
Ennaher	244	182	62
Elmihwer	260	165	95
Total	981	671	310

Le Tableau 5.3 présente une analyse détaillée des entités spatiales et des relations spatiales extraites de divers journaux Arabes. Chaque catégorie d'entité et de relation spatiale est répartie en sous-catégories telles que la localisation, les objets, les directions, les orientations, les relations topologiques et les distances, permettant ainsi de mieux comprendre la diversité des informations géospatiales dans ces sources.

Tableau 5.3 :Répartition des entités et relations spatiales par catégorie

Journaux	**EntitéSpatiale**		**Relations Spatiale**			
	Location	**Objet**	**Direction**	**Orientation**	**Topologique**	**Distance**
El khaber	56	78	43	15	16	08
Echorouk	104	86	34	10	25	02
Ennaher	78	104	28	15	14	05
Elmihwer	83	82	40	13	33	09
Total	321	350	145	53	88	25

La Figure 5.16 montre la répartition des entités et relations spatiales extraites des articles de journaux en Arabe, classées selon la localisation, les objets et d'autres types de relations spatiales.

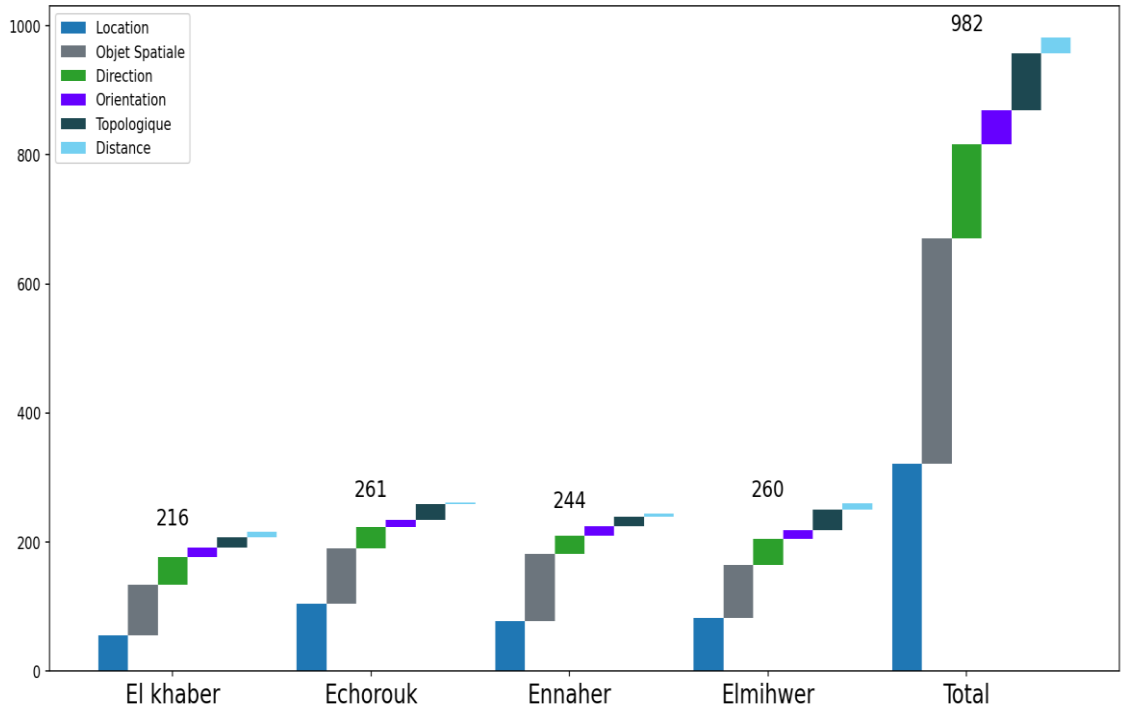

Figure 5.16 : La répartition des entités et relations spatiales extraites.

Pour évaluer et comparer notre approche hybride proposée, nous utiliserons les mesures de précision, rappel et F-mesure. La précision indique la justesse de l'extraction, tandis que le rappel indique sa complétude. La F-mesure fournit la moyenne harmonique entre la précision et le rappel.

Le Tableau 5.4 présente les résultats de l'extraction d'informations spatiales dans divers journaux arabophones. Il met en évidence le nombre de relations correctement identifiées, celles incorrectes, ainsi que les informations manquantes, permettant ainsi d'évaluer la performance des méthodes d'extraction utilisées.

Tableau 5.4 : Évaluation des performances d'extraction d'informations spatiales.

Journaux	**Correcte**	**Incorrecte**	**Manquante**
EnnaherNews	220	15	9
EchoroukNews	229	19	13
ElkhaberNews	191	14	11
ElmihwerNews	235	15	10
Total	875	63	43

Le tableau 5.5 présente les résultats des métriques de performance pour l'extraction d'informations spatiales dans divers journaux. Les mesures incluent la précision, le rappel et la F-mesure, fournissant une vue d'ensemble de l'efficacité des méthodes d'extraction utilisées.

Tableau 5.5 : Évaluation des métriques de Performance pour l'extraction d'informations

Journaux	**Précision**	**Rappel**	**F-mesure**
EnnaherNews	0,936	0,960	0,947
EchoroukNews	0,923	0,946	0,934
ElkhaberNews	0,931	0,945	0,937
ElmihwerNews	0,940	0,959	0,949
Total	0,938	0,952	0,941

Les résultats obtenus dans cette étude sont très satisfaisants, ce qui est montré par les taux élevés de précision et de rappel (Tableau 5.4 et 5.5). Cependant, il convient de

mentionner que certaines erreurs de précision sont principalement dues à des annotations incorrectes associées à des mots spécifiques tels que "بين، تحت، على، في، حول" en Anglais "between, under, on, in, around" (Tableau 5.6). Ces mots ont plusieurs significations, indiquant parfois des relations spatiales et d'autres fois dépendant du contexte de la phrase. Cette ambiguïté pose des défis pour le système d'extraction. D'autre part, l'extraction des entités spatiales n'a pas rencontré de telles ambiguïtés, et les résultats étaient précis. Il est important de considérer que la qualité de l'extraction des relations spatiales peut être influencée par des erreurs d'analyse sémantique et de parsing syntaxique.

Tableau 5.6 : Erreurs syntaxiques et sémantiques

Erreur sémantique	Liée au temps	تتراوح أعمارهم بين 20 و 35 سنة. Leur âge varie entre 20 et 35 ans.
	Relation spatiale	وقع الحادث بين ولايتي بجاية وجيجل. L'accident s'est produit entre les wilayas de Bejaia et Jijel.
Erreur Syntaxique	Verbe	خلف الحادث أربع ضحايا. L'accident a laissé quatre victimes
	Adverbe (Relation spatiale)	يقع المستشفى خلف مسجد المدينة. L'hôpital est situé derrière la mosquée de la ville.

La Figure 5.17 présente les résultats et l'évaluation des scores de précision, de rappel, et de F-mesure obtenus pour chaque journal, à savoir Ennahar, Echourouk, El Khabar et El Mihwar.

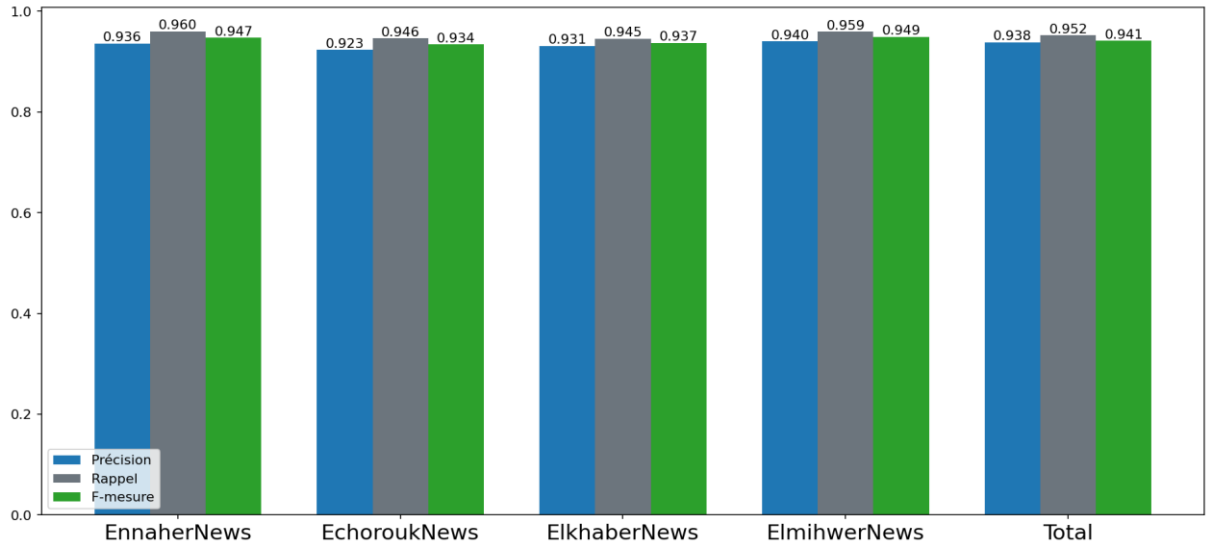

Figure 5.17 : Évaluation de l'extraction des informations spatiales en texte Arabe

6. Comparaison et Discussion

Le Tableau 5.7 présente les résultats issus de travaux précédents ainsi que nos propres résultats. La comparaison de ces derniers met en évidence les différences et les améliorations apportées par notre approche. Ensuite, nous procéderons à une analyse comparative de ces résultats afin de souligner les points forts et les limites de notre méthode par rapport aux approches existantes (Figure 5.18).

Tableau 5.7 : Evaluation de notre approche

	Precision	**Recall**	**F-mesure**
Notre Approche	0,93	0,95	0,94
(Abdelkoui et Kholladi, 2015)	0,80	0,91	0,85
(Vasilopoulos et al.,2018)	0,88	0,76	0,82
(Haris et al.,2020)	0,64	0,80	0,71
(Li et al., 2023)	0,97	0,96	0,97
(Shin et al.,2020)	0,62	0,59	0,61
(Zenasni et al.,2018)	0,83	0,86	0,84

La Figure 5.18 présente une comparaison entre notre méthode et les approches existantes. Elle met en évidence les avantages et les limites de notre méthode en la confrontant à d'autres techniques.

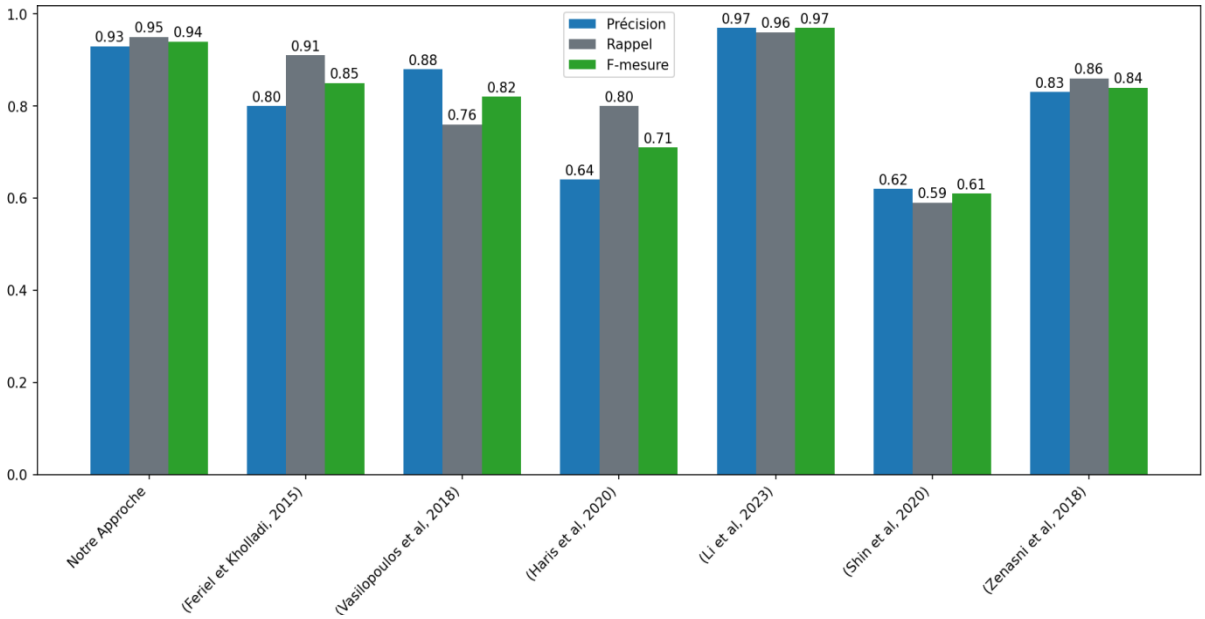

Figure 5.18 : Évaluation de notre système par rapport à d'autres systèmes

Abdelkoui et Kholladi (2015) ont présenté une méthode basée sur un corpus et des règles pour l'extraction d'informations spatio-temporelles en Arabe, qui obtient de bons résultats mais reste inférieure à notre approche, mettant ainsi en évidence l'efficacité de notre ontologie spatiale.

- Selon Vasilopoulos et al. (2018), l'extraction automatique de texte à partir de journaux Arabes" se concentre également sur l'Arabe mais montre des performances inférieures en termes de rappel et de mesure F, ce qui suggère que l'intégration des règles JAPE dans notre approche aide à mieux gérer les ambiguïtés linguistiques ;
- Comme le montre Haris et al. (2020), cette méthode en anglais, qui repose sur la co-occurrence pour l'extraction d'informations spatiales à partir de récits de voyages, montre des résultats inférieurs, soulignant la puissance de notre méthode hybride ;
- D'aprèsLi et al. (2023), bien que cette méthode est très efficace pour les textes chinois, elle n'est pas directement comparable en raison des différences linguistiques ;
- Shin et al. (2020)ont montré que les résultats obtenus sur l'extraction d'informations spatiales basée sur BERT sont acceptables.
- Zenasni et al. (2018), ont démontré que la méthode d'extraction d'informations spatiales à partir de messages courts en anglais obtient de bons résultats mais reste inférieure à notre approche.

Notre approche hybride démontre des performances puissantes et équilibrées dans l'extraction d'informations spatiales à partir de textes en Arabe. La combinaison d'une ontologie spatiale et de règles gère efficacement les complexités linguistiques et sémantiques de l'Arabe, offrant une solution prometteuse pour les applications dans les systèmes d'information géographique SIG et TALN.

7. Etude comparative entre la méthode basé sur les règles JAPE et La méthode hybride

Dans cette section, nous examinons deux approches pour l'extraction d'informations spatiales : une approche basée uniquement sur les règles JAPE et une méthode hybride combinant ontologie et règles JAPE. La méthode basée sur les règles JAPE utilise des règles fixes pour annoter les entités et relations spatiales, tandis que la méthode hybride intègre une ontologie pour enrichir ces règles, offrant ainsi une meilleure précision et

couverture. En évaluant la performance des deux approches en termes de précision, rappel, et F-mesure, cette recherche vise à déterminer l'efficacité relative de chaque méthode pour l'extraction d'informations spatiales.

Le tableau 5.8 compare deux méthodes d'extraction d'informations spatiales : la méthode hybride et la méthode basée sur des règles. Il présente la répartition des entités spatiales et des relations spatiales pour chaque méthode, en distinguant les catégories telles que "LOC", "Objet Spatiale", "DIR", "ORI", "TOP" et "DIS". Ces données permettent d'évaluer la performance et la spécificité de chaque méthode dans l'extraction des informations spatiales.

Tableau 5.8 : Répartition des entités et relations spatiales selon les méthodes d'extraction

Méthodes	**Entité Spatiale**		**Relation Spatiale**			
	LOC	Objet Spatiale	DIR	ORI	TOP	DIS
Méthode Hybride	321	350	145	53	88	25
Méthode base Règles	310	310	88	58	39	12

La Figure 5.19 compare la répartition des entités spatiales et des relations spatiales entre deux méthodes d'extraction : la méthode hybride et la méthode basée sur des règles.

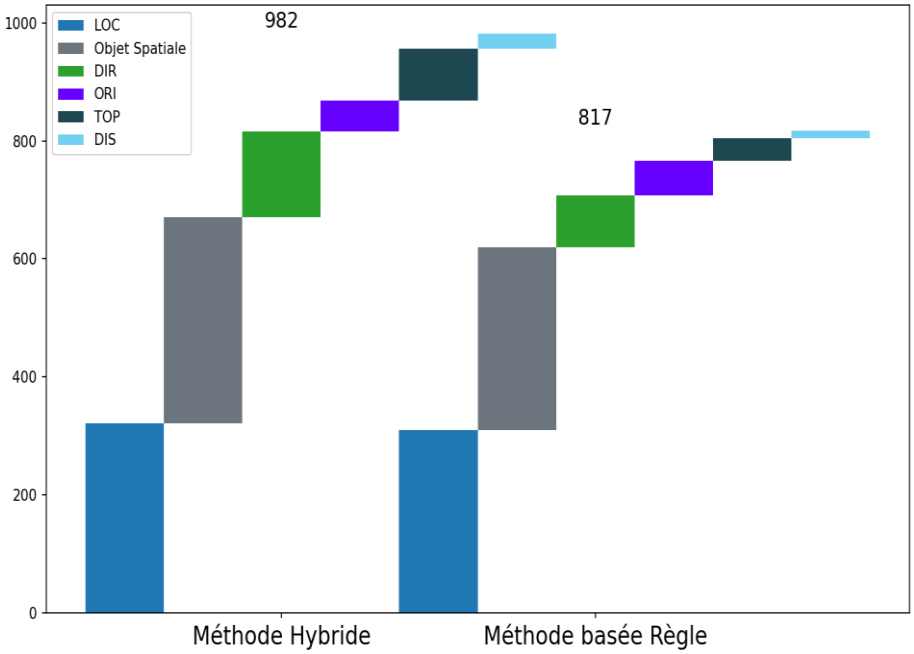

Figure 5.19 : Répartition des entités et relations spatiales selon les méthodes d'extraction.

L'analyse comparative entre les deux méthodes révèle plusieurs points clés :

- **Performance en termes de détection d'entités :** L'approche hybride montre une meilleure performance pour identifier à la fois les lieux et les objets spatiaux, ce qui peut être attribué à l'enrichissement apporté par l'ontologie. Cet enrichissement permet une meilleure compréhension des entités dans leurs contextes spécifiques.
- **Capacité à identifier des relations complexes :** La méthode hybride surpasse également la méthode basée sur les règles en termes de détection des relations spatiales, en particulier pour les relations directionnelles et topologiques. Cela démontre que l'approche hybride est plus robuste pour capturer les interactions spatiales complexes grâce à une représentation plus détaillée des relations.
- **Limites de la méthode basée sur les règles :** Bien que la méthode basée sur les règles JAPE soit efficace pour des tâches spécifiques, elle semble moins flexible pour capturer une gamme complète d'entités et de relations spatiales. Cette limitation est probablement due à la nature statique des règles définies, qui peuvent ne pas s'adapter aussi bien aux variations et aux nuances présentes dans les données.

L'approche hybride ontologie-règle JAPE offre une performance supérieure en termes de couverture des entités spatiales et de détection des relations complexes. En combinant les avantages des deux méthodes, elle fournit une solution plus complète et précise pour l'extraction d'informations spatiales. Cette approche permet de surmonter les limitations inhérentes aux règles JAPE seules, en intégrant un cadre conceptuel qui enrichit et affine l'extraction des informations. Pour des systèmes d'annotation nécessitant une précision et une profondeur accrues, l'approche hybride est clairement la méthode privilégiée.

Le Tableau 5.9 présente les mesures de performance des deux méthodes d'extraction d'informations spatiales : la méthode hybride (ontologie-règles) et la méthode basée sur des règles. Les indicateurs de performance incluent la précision, le rappel et la F-mesure, fournissant une comparaison détaillée de l'efficacité de chaque méthode.

Tableau 5.9 : Évaluation des performances des méthodes d'extraction d'informations spatiales

Méthodes	Précision	Rappel	F-mesure
Méthode hybride Ontologie-Règles	0,93	0,95	0,94
Méthode basée Règles JAPE	0,90	0,85	0,87

La figure 5.20 illustre les mesures de précision, rappel et F-mesure pour deux méthodes d'extraction : la méthode hybride (ontologie-règles) et la méthode basée sur des règles. Cette visualisation permet de comparer directement l'efficacité des méthodes en termes de qualité des annotations spatiales.

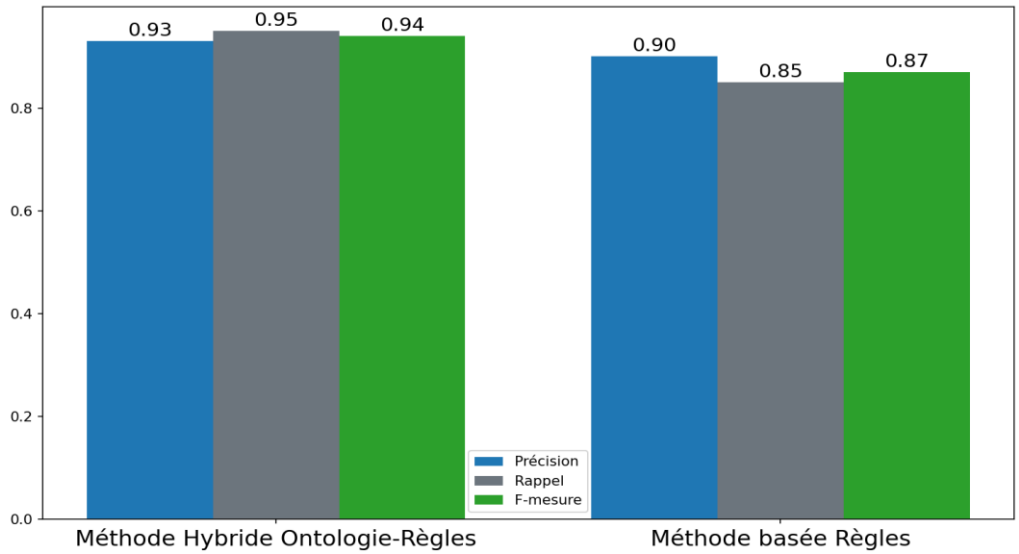

Figure 5.20 : Comparaison des performances des méthodes d'extraction d'informations spatiales

Les résultats démontrent clairement les avantages de la méthode hybride ontologie-règles par rapport à la méthode basée uniquement sur les règles JAPE. La méthode hybride atteint une meilleure précision et un rappel plus élevé, ce qui se traduit par une F-mesure plus élevée. Ces mesures indiquent que l'intégration de l'ontologie augmente la capacité du système à identifier et annoter les entités et relations spatiales de manière plus complète et précise.

- **Précision accrue**

L'intégration d'une ontologie fournit un cadre conceptuel qui permet de réduire les erreurs d'annotation, conduisant à une meilleure précision. Cela suggère que l'ontologie aide à affiner les règles JAPE, rendant l'annotation plus ciblée et pertinente.

- **Rappel amélioré**

Un rappel plus élevé avec la méthode hybride indique que cette approche est plus efficace pour identifier toutes les occurrences des entités et relations spatiales présentes dans les textes. Cela est crucial pour assurer une couverture complète des informations spatiales.

- **F-mesure équilibrée**

La F-mesure élevée de la méthode hybride montre qu'elle maintient un excellent équilibre entre précision et rappel, offrant ainsi une performance globale plus puissante.

L'approche hybride ontologie-règles se révèle être une méthode supérieure pour l'extraction d'informations spatiales, surpassant la méthode basée uniquement sur les règles JAPE en termes de précision, rappel et F-mesure. Cette supériorité est attribuable à la capacité de l'ontologie à enrichir et affiner les règles d'extraction, améliorant ainsi la qualité et la couverture des annotations spatiales. Pour des systèmes d'annotation nécessitant une performance de haut niveau, l'approche hybride est recommandée comme la solution préférée.

8. Conclusion

Cette étude introduit une approche hybride innovante pour l'extraction automatique d'informations spatiales à partir de documents textuels en Arabe, combinant l'ontologie et les règles JAPE pour améliorer les systèmes d'information géographique (SIG) et les ressources de traitement automatique de la langue Arabe (TALA). Le développement de ASTO représente une avancée significative en structurant et en formalisant les entités et relations spatiales spécifiques à l'Arabe, facilitant ainsi l'indexation, l'annotation et l'extraction des données spatiales. L'intégration des règles JAPE a permis une désambiguïsation et une classification efficaces, améliorant considérablement la gestion des ambiguïtés et des variations linguistiques. Les résultats obtenus ont confirmé l'efficacité de cette méthode hybride, démontrant son potentiel pour optimiser les systèmes SIG et améliorer la récupération d'informations en Arabe.

Les principales contributions de notre approche hybride incluent la création d'ASTO, qui structure les informations spatiales Arabes et l'utilisation des règles JAPE pour améliorer la précision de l'extraction et de la classification des données. Ces contributions optimisent non seulement les systèmes SIG, mais renforcent également les ressources du TALA et des SIG.

La comparaison entre les deux méthodes a révélé que l'intégration d'une ontologie renforce la précision et la couverture des annotations spatiales, démontrant ainsi l'avantage de l'approche hybride dans des contextes complexes. Cette analyse met en évidence les forces et limites de chaque méthode, soulignant l'importance d'une approche enrichie par l'ontologie pour des résultats plus fiables et exhaustifs dans l'extraction d'informations spatiales.

Conclusions et perspectives

Le présent travail nous a permis d'aborder de manière approfondie l'extraction d'informations spatiales à partir de textes Arabes dans le cadre des systèmes d'information géographique (SIG). Ce domaine, encore peu exploré, s'inscrit dans une intersection entre le traitement automatique du langage naturel (TALN) et la gestion des données géospatiales. Les contributions majeures de ce travail résident dans la conception de l'ontologie ASTO (Arabic Spatial Toponym Ontology) et le développement d'une approche hybride combinant des règles JAPE (Java Annotation Patterns Engine) avec des ontologies pour traiter les spécificités linguistiques de la langue Arabe.

Notre travail évoque les défis majeurs posés par la richesse morphologique et la complexité syntaxique de la langue Arabe, en particulier lorsqu'il s'agit d'extraire des entités spatiales et des relations géographiques. Ces spécificités linguistiques ont rendu inefficaces les méthodes classiques d'extraction basées sur des approches statistiques ou d'apprentissage automatique, justifiant ainsi le recours à une approche hybride et ontologique.

Les principales contributions scientifiques de ce travail peuvent être résumées dans plusieurs axes :

- L'ontologie ASTO, conçue pour modéliser les toponymes spatiaux Arabes, représente un outil novateur pour structurer les connaissances géospatiales en langue Arabe. Elle permet de capturer et de formaliser les entités spatiales (objet spatial, nom de lieu) ainsi que leurs relations (distance, orientation, topologie et direction). Ce cadre ontologique est primordial dans l'amélioration de l'indexation des informations spatiales issues des textes Arabes. Contrairement aux approches traditionnelles, l'ASTO offre une flexibilité dans la modélisation des entités et des relations géographiques complexes et permet une désambiguïsation plus précise des entités spatiales.
- L'approche basée sur les règles JAPE repose sur l'utilisation de patrons linguistiques pour identifier et extraire des entités et des relations spatiales dans les textes. Les règles JAPE sont un ensemble de règles définies manuellement pour capturer des structures linguistiques spécifiques. Dans le cadre de ce travail, des règles JAPE ont été conçues pour traiter les toponymes Arabes et les relations spatiales courantes. La simplicité et la rapidité de développement des règles JAPE en font un choix idéal pour des systèmes simples et spécifiques, nécessitant une intervention rapide. Cependant, bien que cette approche soit efficace pour des textes bien structurés et avec des expressions

géospatiales explicites, elle montre des limites face aux ambiguïtés linguistiques fréquentes dans les textes Arabes. De plus, la conception et la maintenance des règles sont coûteuses en termes de temps et nécessitent une expertise linguistique poussée. L'approche JAPE, bien qu'utile dans certains cas, reste inflexible face aux variations contextuelles et linguistiques des textes.

- Pour surmonter les limites de ces approches, une approche hybride combinant les ontologies et les règles a été développée. Cette méthode exploite la même structuration des connaissances géospatiales apportée dans l'ontologie ASTO, tout en exploitant la puissance des règles linguistiques formelles pour désambiguïser et classifier les entités spatiales. L'approche hybride s'avère plus performante dans le traitement des textes Arabes complexes, où les relations spatiales sont souvent implicites ou peu structurées.
- Une comparaison rigoureuse entre l'approche basée sur les règles JAPE et l'approche hybride a été menée pour évaluer leur efficacité dans des environnements SIG réels. Les résultats ont montré que l'approche hybride offre une précision supérieure et une meilleure couverture des entités spatiales. Contrairement aux règles JAPE seules, l'approche hybride est capable de traiter des textes plus complexes et des relations spatiales moins explicites. Cette évaluation a également mis en évidence la capacité de l'approche hybride à s'adapter à différents contextes géospatiaux, renforçant ainsi sa pertinence pour des applications variées.

Les perspectives offertes par ce travail de recherche sont nombreuses et se déclinent sur plusieurs axes d'amélioration et d'extension.

- Une première piste de travail consiste à enrichir davantage l'ontologie ASTO, tant au niveau de la couverture des toponymes que des relations spatiales complexes. Des collaborations avec des experts en géographie et en linguistique Arabe pourraient permettre d'élargir la base de connaissances intégrée dans l'ASTO, tout en améliorant sa capacité à désambiguïser des entités géospatiales dans des textes issus de différents domaines.
- Le développement d'un corpus annoté dédié à la langue Arabe dans le contexte géospatial représente une autre voie de recherche pertinente. Un tel corpus permettrait de tester et d'améliorer les performances des approches existantes, mais aussi de favoriser la création de nouvelles méthodes d'extraction basées sur l'apprentissage profond (Deep Learning) ou les modèles de langage pré-entraînés

(Machine learning).

- L'utilisation des techniques d'apprentissage profond, telles que les réseaux de neurones récurrents (RNN) ou les transformers, pourrait compléter l'approche hybride. L'intégration de modèles pré-entraînés, comme BERT adapté pour la langue Arabe (ArabeRT), pourrait permettre de capturer des informations plus subtiles dans les textes. Cette approche pourrait également automatiser la création de nouvelles règles JAPE et rendre le système plus flexible dans l'adaptation à de nouveaux types de données.
- L'approche développée dans ce travail pourrait être appliquée dans des domaines spécifiques nécessitant une extraction précise d'informations géospatiales. Par exemple, l'urbanisme, la gestion des infrastructures, ou encore la surveillance environnementale pourraient tirer parti de ces techniques pour automatiser la collecte et l'analyse de données textuelles géospatiales. Ces applications concrètes pourraient être développées en partenariat avec des institutions publiques ou privées, renforçant ainsi l'impact de ces travaux dans la réalité.
- Une perspective importante serait d'intégrer directement les techniques d'extraction d'informations développées dans ce travail au sein de plateformes SIG existantes. Cela permettrait de faciliter l'analyse des données textuelles en temps réel et d'enrichir les capacités des SIG en fournissant des outils interactifs pour la gestion des entités et des relations spatiales.

Ce travail apporte une contribution substantielle à la problématique de l'extraction d'informations spatiales en langue Arabe, en proposant des solutions innovantes et adaptées aux spécificités linguistiques de cette langue. Grâce à l'ontologie ASTO et à l'approche hybride développée, des avancées significatives ont été réalisées dans la gestion et l'intégration des données géospatiales au sein des SIG.

Références bibliographiques

Abayomi-Alli, A.A., Misra, S., Akala, M.O., Ikotun, A. M., &Ojokoh, B.A. (2021). An ontology-based information extraction system for organicfarming. *International Journal onSemantic Web and Information Systems*, *17*(2), 79-99. https://doi.org/10.4018/IJSWIS.2021040105

Abdelkoui, F., &Kholladi, M.K. (2015). Automatic extraction of spatio-temporal information from Arabic text documents. *International Journal of Computer Science & Information Technology ,7*(5), 97-107.https://doi.org/10.5121/ijcsit.2015.7507.

Abu-Errub, A., Odeh, A., Shambour, Q., &Hassan, O.A.H. (2014). Arabic roots extraction using morphological analysis. *International Journal of Computer Science Issues (IJCSI), 11*(2), 128-134.

Acheson, E., &Purves, R. S. (2021). Extracting and modeling geographic information from scientific articles. *PloS one*, *16*(1), e0244918.

Adnan, K., & Akbar, R. (2019). Limitations of information extraction methods and techniques for heterogeneousunstructuredbig data. *International Journal of Engineering Business Management, 11*, 1847979019890771.

Agarwal, P. 2005. Ontological considerations in GIScience. *International Journal of Geographical Information Science*, *19*(5), pp. 501-536.

Aguilar, A.J., Pinos-Navarrete, A., Domingo Jaramillo, C., & de la Hoz-Torres, M.L. (2024). Geographic information systems and web gis in higher education: a collaborative tool for the analysis of accessibility in the urban and built environment. In Teaching Innovation in Architecture and Building Engineering: Challenges of the 21st Century, pp. 401-415.https://doi.org/10.1007/978-3-031-59644-5_23.

Ahaggach, H., Abrouk, L.,&Lebon, E. (2023). Information extraction and ontology population using car insurance reports. In International Conference on Information Technology-New Generations, pp. 405-411. https://doi.org/10.1007/978-3-031-28332-1_46.

Alaya, N., Yahia, S. B., &Lamolle, M. (2015). What makes ontology reasoning so arduous? Unveiling the key ontological features. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics* (pp. 1-12).

Aljabari, A., Duaibes, L., Jarrar, M., &Khalilia, M. (2024). Event-arguments extraction corpus and modeling using BERT for Arabic. Computation and Language.

Al-Laith, A., Shahbaz, M., Alaskar, H.F., &Rehmat, A. (2021).Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus. *Applied Sciences*, *11*(5), 2434. https://doi.org/10.3390/app11052434.

Alrayzah, A., Alsolami, F., &Saleh, M. (2024).AraFast: Developing and evaluating a comprehensive modern standard arabic corpus for enhanced natural language processing. *Applied Sciences*, *14*(12), 5294. https://doi.org/10.3390/app14125294.

Amhar, F., Giri, E. P., Silalahi, F. E. S., Neyman, S. N., Anggrahito, Ramdani, D., ... &Murdaningsih. (2022). Ownership Protection on Digital Elevation Model (DEM) Using Transform-Based Watermarking. *ISPRS International Journal of Geo-Information*, *11*(3), 200.

Anantharangachar, R., Ramani, S., &Rajagopalan, S. (2013). Ontology guided information extraction from unstructuredtext. *International Journal of Web &SemanticTechnology (IJWesT),* *4*(1), 19-36. https://doi.org/10.5121/ijwest.2013.4102.

Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, N., &Mahmood, W. (2019). The use of ontology in retrieval: a study on textual, multilingual, and multimedia retrieval. *IEEE Access*, *7*, 21662-21686.

Atoui, B. (1996). Toponymie et espace en Algérie (Doctoral dissertation, Université de Provence-Aix-Marseille I).

Attia, M. (2007).Arabic tokenization system. *In Proceedings* of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources, Prague, Czech Republic,pp. 65-72.https://doi.org/10.3115/1654576.1654588.

B

Bachimont, B., Isaac, A.,&Troncy, R..(2002).Semantic commitment for designing ontologies: A proposal. In: Gómez-Pérez, A., Benjamins, V.R. (eds.) EKAW 2002. LNCS (LNAI), *2473*, pp. 114–121. Springer, Heidelberg.

Baneyx, A. (2007). Construire une ontologie de la Pneumologie Aspects théoriques, modèles et expérimentations (Doctoral dissertation, Université Pierre et Marie Curie-Paris VI).

Benjamin, P., Kumar, N., Fernandes, R., & Li, B. (2011). A framework for ontology life cycle management.In*Proceedings of the International Conference on Information and Knowledge Engineering (IKE)* (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Benkirane, F.E., Crombez, N., Hilaire, V., &Ruichek, Y. (2024). Spatial relationshipsknowledgeintegration in deeplearningmodeling for panoptic segmentation in urbandriving scenarios. In TechnologicalSystems, Sustainability and Safety.

Borisova, N. (2014). An approach for ontology based information extraction. *Information Technologies and Control*, *12*(1), 15-20.https://doi.org/10.1515/itc-2015-0007.

Borst, W. N. (1997). Construction of engineering ontologies for knowledge sharing and reuse(Doctoral dissertation, Institute for Telematica and Information Technology, University of Twente, Enschede, The Netherlands).

Bounhas, I., Elayeb, B., Evrard, F., &Slimani, Y. (2011). Organizing contextual knowledge for arabic text disambiguation and terminology extraction. *KO KnowledgeOrganization*, *38*(6), 473-490.

Buitelaar, P., Cimiano, P., Racioppa, S., &Siegel, M. (2006). Ontology-based information extraction with SOBA. In Proceedings of the international conference on languageresources and evaluation (LREC).

C

Champin, P. A., Prié, Y., & Mille, A. (2001). Annotating with uses: A promising way to the Semantic Web.In*Semannot@ K-CAP*.

Charlet, J. (2002). L'ingénierie des connaissances : Développements, résultats et perspectives pour la gestion des connaissances médicales (Mémoire d'habilitation à diriger des recherches, Université Pierre et Marie Curie).

Charlet, J., Declerck, G., Dhombres, F., Gayet, P., Miroux, P., &Vandenbussche, P. Y. (2012). Construire une ontologie médicale pour la recherche d'information: problématiques terminologiques et de modélisation. In *23es journées francophones d'Ingénierie des connaissances* (pp. 33-48).

Corcho, O., Fernández-López, M., & Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies: Where is their meeting point?.*Data & Knowledge Engineering*, *46*(1), 41-64.

Costa, R., Lima, C., Sarraipa, J., &Jardim-Gonçalves, R. (2016). Facilitating knowledge sharing and reuse in building and construction domain: an ontology-based approach. *Journal of Intelligent Manufacturing*, *27*, 263-282.

D

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., ... & Jain, A. (2024). Structured information extraction fromscientifictextwith large languagemodels. *Nature Communications*, *15*(1), 1418.

De Andrade, F. G., de Souza Baptista, C., & Davis, C. A. (2014). Improving geographic information retrieval in spatial data infrastructures. *GeoInformatica*, *18*, 793-818.

E

Eftimov, T., KoroušićSeljak, B., &Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, *12*(6), e0179488.

Ekaputra, F., Sabou, M., SerralAsensio, E., Kiesling, E., &Biffl, S. (2017). Ontology-based data integration in multi-disciplinary engineering environments: A review. *Open Journal of Information Systems*, *4*(1), 1-26.

Elayeb, B. (2019). Arabic word sense disambiguation: a review. *Artificial Intelligence Review*, *52*(4), 2475-2532.

Escobar, C. A., & Morales-Menendez, R. (2017). Machine learning and pattern recognition

techniques for information extraction to improve production control and design decisions. In IndustrialConference on Data Mining (pp. 286-300). *Cham: Springer International Publishing*.

Etudo, U., & Yoon, V.Y. (2024). Ontology-based information extraction for labeling radical online content using distant supervision. *Information SystemsResearch, 35*(1), 203-225. https://doi.org/10.1287/isre.2023.1223

F

Fernández-López, M., Gómez-Pérez, A., &Juristo, N. (1997). Methontology: from ontological art towards ontological engineering.

Fiebeck, J., Laser, H., Winther, H. B., &Gerbel, S. (2018). Leaving no stone unturned: using machine learningbasedapproaches for information extraction from full texts of a research data warehouse. In International Conference on Data Integration in the Life Sciences (pp. 50-58). *Cham: Springer International Publishing*.

Fifita, F., Smith, J., Hanzsek-Brill, M. B., Li, X., & Zhou, M. (2023). Machine learning-based identifications of COVID-19 fake news usingbiomedical information extraction. *Big Data and Cognitive Computing*, *7*(1), 46.

Fudholi, D. H., Rahayu, W., &Pardede, E. (2016). Ontology-based information extraction for knowledge enrichment and validation. In 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA) (pp. 1116-1123). IEEE.

G

Guarino, N., Oberle, D., &Staab, S. (2009). What is an ontology? Handbook on Ontologies, pp. 1-17. https://doi.org/10.1007/978-3-540-92673-3_0.

Genesereth, M.R., & Nilsson, N.J. (1987). Logical foundations of artificial intelligence. Morgan Kaufmann.

Gómez-Pérez, A., &Corcho, O. (2002). Ontology languages for the Semantic Web.*IEEE Intelligent Systems*, *17*(1), 54-60.

Goodchild, M. F., Longley, P. A., Maguire, D. J., &Rhind, D. W. (2005). Geographic information systems and science. *Wiley & Sons, West Sussex, UK*, *17*, 517.

Grenon, P., & Smith, B. (2004). SNAP and SPAN: Towardsdynamic spatial ontology. *Spatial cognition and computation, 4*(1), 69-104.

Grishman, R., &Sundheim, B. (1996). Message Understanding Conference-6: A Brief History. In Proceedings of the 16th International Conference on Computational Linguistics (COLING). ACL.

Gruber, T.R. (1991). The role of common ontology in achieving sharable, reusable knowledge bases. In *Proceedings of KR 1991* (pp. 601-602).

Gruber, T.R. (1993). A translation approach to portable ontology specifications.*Knowledge*

Acquisition, *5*(2), 199-220.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing.*International Journal of Human-Computer Studies*, *43*(5–6), 907-928.

Guarino, N., Oberle, D., &Staab, S. (2009). What is an ontology?. In S. Staab& R. Studer (Eds.), *Handbook on Ontologies* (pp. 1-17). *Springer*.

Gutierrez, F., Dou, D., Fickas, S., Wimalasuriya, D., &Zong, H. (2016). A hybridontology-based information extraction system. *Journal of Information Science*, *42*(6), 798-820. https://doi.org/10.1177/0165551515610989.

H

Hadji, A., &Kholladi, M.K. (2012).Traitementet fusion de donnéesdans le cadre de l'interopérabilitésémantique des systèmes d'information géographiques. *In Proceedings Conference CTIC* Adrar, Algeria, pp. 1-6.

Hadji, A., &Kholladi, M. K. (2023). Automatic Opinion Extraction from Football-Related Social Media: A Gazetteer and Rule-Based Approach. *NCAIA'2023*, 61.

Hadji, A., Kholladi, M. K., &Borisova, N. (2024). Enhancing Spatial Information Extraction from Arabic Text: A Hybrid Approach with Ontology and Rule-Based. *Ingenierie des Systemesd'Information*, *29*(4), 1261.

Hadji, A., &Kholladi, M.K. (2024).Advanced NPL Methods for Disaster Information Extraction: Analyzing JAPE Rules, Ontologies, and Machines Learning Approaches. The 3rd International Conference on Computer Science's Complex Systems and their Applications (ICCSA'2024) - *InProceedings Conference* University of Oum El Bouaghi, Algeria. *In press*.

Haris, E., Gan, K.H., & Tan, T.P. (2020). Spatial information extraction from travel narratives: Analysing the notion of co-occurrence indicating closeness of tourist places. *Journal of Information Science*, *46*(5), 581-599. https://doi.org/10.1177/0165551519837188.

Hasani, S., Sadeghi-Niaraki, A., &Jelokhani-Niaraki, M. (2015). Spatial data integration using ontology-based approach. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *40*, 293-296.

Hassini, N., Mahmoudi, K., &Faiz, S. (2023). A Hybrid Approach for Spatial Information Extraction from Natural Language Text. In *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)* (pp. 1-8). IEEE.

Heywood, I., Cornelius, S., & Carver, S. (2005). An introduction to geographical information systems. *ZarządzaniePubliczne. ZeszytyNaukoweInstytutuSprawPublicznychUniwersytetuJagiellońskiego*, (1).

Hkiri, E., Mallat, S., Zrigui, M. (2016). Events automatic extraction from Arabic texts. International Journal of Information Retrieval Research (IJIRR), 6(1):36-51. https://doi.org/10.4018/IJIRR.2016010103.

J

Jain, V., Wason, R., Chatterjee, J.M., & Le, D.N. (Eds.). (2020). Ontology-based information retrieval for healthcare systems. John Wiley & Sons.

Jakir, Ž.,Hećimović, Ž., &Štefan, Z. (2011). Place names ontologies. In A. Ruas (Ed.), *Advances in Cartography* (pp. 331-349).*Springer*.

Jones, C.B., Alani, H., &Tudhope, D. (2001). Geographical information retrieval with ontologies of place. In Spatial Information Theory: Foundations of Geographic Information Science International Conference, COSIT 2001 Morro Bay, CA, USA, pp. 322-335.https://doi.org/10.1007/3-540-45424-1_22.

Jones, C. B. (2014). *Geographical information systems and computer cartography*. Routledge.

Jusoh, S., Awajan, A., &Obeid, N. (2020). The use of ontology in clinical information extraction. In Journal of Physics: *ConferenceSeries*, *1529*(5), 052083. https://doi.org/10.1088/1742-6596/1529/5/052083

K

KalantariOskouei, A., &SaberKhoshemehr, M. (2018). Identifying and prioritizing the challenges of data sharing and spatial information. *Scientific-Research Quarterly of Geographical Data (SEPEHR)*, *27*(106), 37-55.

Karmacharya, A.,Cruz, C.,Boochs, F.,&Marzani, F.2010. Use of geospatial analyses for semantic reasoning. *In Proceedings* of the 14th International Conference on Knowledge-based and Intelligent Information and Engineering Systems: Part I, Berlin, Heidelberg, pp. 576-586.

Kauppinen, T., Henriksson, R., Sinkkilä, R., Lindroos, R., Väätäinen, J., &Hyvönen, E. (2008). Ontology-based Disambiguation of Spatiotemporal Locations. In *IRSW*.

Konys, A. (2018). Towards knowledge handling in ontology-based information extraction systems.*Procedia Computer Science*, *126*, 2208-2218.https://doi.org/10.1016/j.procs.2018.07.228.

Kordjamshidi, P., van Otterlo, M., &Moens, M. F. (2017). Spatial role labeling annotation scheme. *Handbook of linguistic annotation*, 1025-1052.

Kotis, K.I., Vouros, G.A., &Spiliotopoulos, D. (2020). Ontology engineering methodologies for the evolution of living and reused ontologies: status, trends, findings and recommendations. *The Knowledge Engineering Review*, *35*, e4.

Krieger, F., Drews, P., & Funk, B. (2023). Automatedinvoiceprocessing: Machine learning-based information extraction for long tailsuppliers. *Intelligent Systemswith Applications*, *20*, 200285.

L

Lando, P. (2006). Conception et développement d'applications informatiques utilisant des

ontologies: Application aux EIAH. *In1res Rencontres jeunes chercheurs en EIAH (RJC-EIAH).*

Laurini, R. (2015). Geographic ontologies, gazetteers and multilingualism.*Future Internet, 7*(1), 1-23.

Laurini, R., &Kazar, O. (2017). Geographic ontologies: Survey and challenges. *Meta-carto-semiotics, 9*(1), 1-13.

Li, D., Zhang, J., & Wu, H. (2012). Spatial data quality and beyond. *International Journal of Geographical Information Science*, *26*(12), 2277-2290.

Li, X., Zhang, W., Wang, Y., Tan, Y., &Xia, J. (2023).Spatio-temporal information extraction and geoparsing for public Chinese resumes. *ISPRS International Journal of Geo-Information,12*(9), 377.https://doi.org/10.3390/ijgi12090377.

Liao, Y., Hua, J., Luo, L., Ping, W., Lu, X., &Zhong, Y. (2024). APRCOIE: An open information extraction system for Chinese. *SoftwareX, 26*, 101649.

Lieberman, J.,Singh,R.,& Goad,C.2007. "W3C Geospatial Ontologies – W3C Incubator Group", W3C Incubator Group, Report.

Longley, P.A., Goodchild, M.F., Maguire, D.J., &Rhind, D.W. (2015). *Geographic information science and systems*. John Wiley & Sons.

Luo, S., &Yu, J. (2024). ESGNet: A multimodal network model incorporatingentitysemantic graphs for information extraction fromChineseresumes. *Information Processing& Management, 61*(1), 103524.

M

Ma, C., &Molnár, B. (2020). Use of ontology learning in information system integration: a literature survey. In Intelligent Information and Database Systems: 12th Asian Conference, ACIIDS 2020, Phuket, Thailand, March 23–26, 2020, *Proceedings* 12 (pp. 342-353). *Springer Singapore*.

McDaniel, M., &Storey, V.C. (2019). Evaluatingdomain ontologies: clarification, classification, and challenges. *ACM ComputingSurveys (CSUR), 52*(4), 1-44.

McGuinness, D. L. (2004).OWL Web Ontology Language overview.*W3C Member Submission*.

Maynard, D., Peters, W., & Li, Y. (2006). Metrics for evaluation of ontology-based information extraction. *In Proceedings* of the 15th International Conferenceon World Wide Web in Workshop on Evaluation of Ontologies for the Web, EON@ WWW, Edinburgh, UK.

Musa, S. H., Sauti, N. S., Ahmad, Y., Abdullah, N. A., &Nasir, F. M. (2018). Conceptual design of GIS database for heritage building in Melaka. *Politeknik&KolejKomuniti Journal of Engineering and Technology, 3*(1), 114-128.

N

Nazari, Y. (2016). Aspects of Syntactic Ambiguity in Arabic Language and Their Impacts on the Translation of the Holy Quran. *Translation Researches in the ArabicLanguage And Literature*, *6*(15), 84-59.

Nebhi, K. (2012). Ontology-based information extraction fromTwitter. *In Proceedings* of the Workshop on Information Extraction and EntityAnalytics on Social Media Data, Mumbai, pp. 17-22.

Noy, N., &McGuiness, D.L. (2001). Ontology development 101: A guide to creating your first ontology. Stanford, CA, StanfordKnowledgeSystemsLaboratory, USA.

O

Omar, A., &Aldawsari, M. (2020). Lexical Ambiguity in Arabic Information Retrieval: The Case of Six Web-Based Search Engines. *International Journal of English Linguistics*, *10*(3), 219-228.

Opasjumruskit, K., Böning, S., Schindler, S., &Peters, D. (2022). OntoHuman: Ontology-based information extraction toolswithhuman-in-the-loop interaction. *In International Conference* on Cooperative Design, Visualization and Engineering, pp. 68-74. https://doi.org/10.1007/978-3-031-16538-2_7.

P

Pandey, J. (2014). Geographic information system. The Energy and Resources Institute (TERI).

Ping, D., & Yong, L. (2009). Building place name ontology to assist in geographic information retrieval. In 2009 International Forum on Computer Science-Technology and Applications, Chongqing, China,

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., &Kirilov, A. (2004). KIM–A semanticplatform for information extraction and retrieval. Natural language engineering, 10(3-4): 375-392. https://doi.org/10.1017/S135132490400347X.

Q

Qu, X., Gu, Y., Xia, Q., Li, Z., Wang, Z., & Huai, B. (2023). A survey on arabic named entity recognition: Past, recent advances, and future trends. *IEEE Transactions on Knowledge and Data Engineering*.

R

Raihan, A. (2024). A systematicreview of Geographic Information Systems (GIS) in agriculture for evidence-baseddecisionmaking and sustainability. Global SustainabilityResearch, 3(1), 1-24.

Reddy, K.R., Sharma, V.K., Anusha, M., Jhade, S., &Dhanasekaran, B. (2024). Progressive collaborative method for protecting users privacy in location-based services. In MATEC Web of Conferences, 392,01089.http://doi:org/10.1051/matecconf/202439201089.

Ressler, J.,Dean, M.,& Kolas,D.2007.“Geospatial Ontology Trade Study”, National Geospatial Intelligence Agency, Report No. HM1582-05-C-0014,

Rios, A., Cabral, B., Claro, D., Cavalcante, R., & Souza, M. (2024). TransAlign: An Automated Corpus Generationthrough Cross-Linguistic Data Alignment for Open Information Extraction. In Proceedings of the 16th International Conference on ComputationalProcessing of Portuguese (pp. 196-206).

Rizvi, S.T.R., Mercier, D., Agne, S., Erkel, S., Dengel, A., &Ahmed, S. (2018). Ontology-based information extraction from technical documents. *In Proceedings* of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018), Funchal, Madeira, Portugal, pp. 493-500. https://doi.org/10.5220/0006596604930500.

Robertson, C., &Horrocks, K. (2017). Spatial Contextfrom Open and Online Processing (SCOOP): geographic, temporal, and thematicanalysis of online information sources. ISPRS International Journal of Geo-Information, 6(7), 193.

S

Shah, R., &Jain, S. (2014). Ontology-based information extraction: An overview and a study of different approaches. International Journal of Computer Applications, 87(4):6-8. https://doi.org/10.5120/15194-3574.

Shin, H.J., Park, J.Y., Yuk, D.B., &Lee, J.S. (2020).BERT-based spatial information extraction.*In Proceedings* of the Third International Workshop on Spatial Language Understanding, pp. 10-17.https://doi.org/10.18653/v1/2020.splu-1.2.

Siabato, W.,&Manso-Callejo, M.Á. (2011). Integration of temporal and semantic components into the Geographic Information throughmark-uplanguages. Part I: definition. In Computational Science and Its Applications-ICCSA 2011: International Conference, Santander, Spain, June 20-23, 2011. Proceedings, Part I 11 (pp. 394-409). *Springer Berlin Heidelberg*.

Slimani, T. (2015). A study investigating typical concepts and guidelines for ontology building. *arXiv preprint arXiv:1509.05434*.

Spinsanti, L., &Ostermann, F. (2013). Automated geographic context analysis for volunteered information. *AppliedGeography*, *43*, 36-44.

Steinkamp, J.M., Chambers, C., Lalevic, D., Zafar, H. M., & Cook, T.S. (2019). Towardcompletestructured information extraction fromradiology reports using machine learning. *Journal of digital imaging*, *32*, 554-564.

Studer, R., Benjamins, V.R., &Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, *25*(1-2), 161-197.

Sure, Y., Akkermans, H., Broekstra, J., Davies, J., Ding, Y., Duke, A., ...& van Harmelen, F. (2003). On-To-knowledge: semantic web-enabled knowledge management. In *Web Intelligence* (pp. 277-300). Springer Berlin Heidelberg.

T

Tao, C., Jiang, G., Oniki, T A., Freimuth, R.R., Zhu, Q., Sharma, D., ...& Chute, C.G. (2013). A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. *Journal of the American Medical Informatics Association*, *20*(3), 554-562.

Tapia-Leon, M., Rivera, A.C., Chicaiza, J., &Luján-Mora, S. (2018). Application of ontologies in higher education: A systematic mapping study. In *2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1344-1353). IEEE.

Thakker, D., Osman, T., &Lakin, P. (2009). Gate jape grammar tutorial. *Nottingham Trent University, UK, Phil Lakin, UK, Version*, *1*.

U

Uschold, M., & King, M. (1995). Towards a methodology for building ontologies (pp. 1-13).Edinburgh: Artificial Intelligence Applications Institute, University of Edinburgh.

Uschold, M., &Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The knowledge engineering review*, *11*(2), 93-136.

V

Vandecasteele, A., & Napoli, A. (2012).Spatial ontologies for detecting abnormal maritime behaviour.In 2012 Oceans-Yeosu (pp. 1-7). IEEE.

Vasilopoulos, N., Wasfi, Y., &Kavallieratou, E. (2018). Automatic text extraction from arabic newspapers. In Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, pp. 505-510.https://doi.org/10.1007/978-3-319-93000-8_57.

W

Wan, L., Cheng, T., Fan, W., Shi, Y., Zhang, H., Zhang, S., ... & Yang, S. (2023). A framework for ontology-based information extraction and integrationfromheterogeneous data sources. IEEE Access, 11, 26811-26824.

Wang, X., Zhang, S., Shen, D., Liu, J., & Wang, J. (2024). Automatic Information Extraction from Unstructured Data: A Comprehensive Survey. *ACM ComputingSurveys (CSUR)*, *56*(5), 1-37.

Willmes, C., Becker, D., Verheul, J., Yener, Y., Zickel, M., Bolten, A., ... &Bareth, G. (2017). PaleoMaps: SDI for open paleoenvironmental GIS data. *International Journal of Spatial Data Infrastructures Research*, *12*, 39-61.

Wimalasuriya, D.C., & Dou, D. (2010). Components for information extraction: Ontology-based information extractors and generic platforms. In Proceedings of the 19th ACM international conference on Information and knowledge management, Toronto ON Canada,pp. 9-18.

Wimalasuriya, D.C., &Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science, 36*(3),306-323.https://doi.org/10.1177/0165551509360123.

Y

Yang, Y., Wu, Z., Yang, Y., Lian, S., Guo, F., & Wang, Z. (2022). A survey of information extraction based on deep learning. *Applied Sciences, 12*(19), 9691.

Yasobant, S., Vora, K. S., &Upadhyay, A. (2019). Geographic information system applications in public health: Advancing health research. In Healthcare Policy and Reform: Concepts, Methodologies, Tools, and Applications (pp. 538-561). IGI Global.

Yunianta, A., Yusof, N., Aziz, A., Dengen, N., & Othman, M.S. (2014). Analysis and Identification of Data Heterogeneity on Learning Environment Using Ontology Knowledge. In *1st International Conference on Electrical Engineering, Computer Science and Informatics 2014*. Institute of Advanced Engineering and Science.

Z

Zenasni, S., Kergosien, E., Roche, M., & Teisseire, M. (2018). Spatial information extraction from short messages. *Expert Systems with Applications, 95*, 351-367.https://doi.org/10.1016/j.eswa.2017.11.025.

Zhang, X., Liu, Y., Wang, Y., & Chen, Y. (2024). Spatio-temporal Information Extraction for Urban Analytics Using Ontology and Machine Learning. *Computers, Environment and UrbanSystems, 94*, 101645.

Zhu, Z., & Tan, J. (2018). A multi-source heterogeneous vector space data integration scheme based on geojson. *In 2018 26th International Conference on Geoinformatics* (pp. 1-4). IEEE.

Résumé

Avec l'essor rapide d'Internet, la quantité d'informations disponibles a considérablement augmenté, entraînant un problème croissant de « surcharge d'informations ». Cette situation souligne l'importance d'accéder de manière rapide et précise à des informations pertinentes, en particulier dans des domaines spécialisés tels que les systèmes d'information géographique (SIG). L'extraction d'informations spatiales à partir de textes en arabe constitue un défi majeur en raison de la complexité linguistique et de l'importance des données géospatiales dans divers secteurs.

Pour atteindre cet objectif, cette recherche s'est concentrée sur la création et l'intégration de l'Ontologie Arabe des Toponymes Spatiaux (ASTO). Cette ontologie vise à structurer les connaissances géospatiales en arabe et à améliorer la précision de l'extraction d'informations spatiales. Parallèlement, une approche basée sur les règles JAPE a été développée et évaluée.

De même, une approche hybride combinant l'ontologie avec des méthodes fondées sur des règles a été mise en œuvre. Les deux approches ont été comparées afin d'évaluer leurs performances respectives. Les résultats montrent que l'approche hybride surpasse celle basée uniquement sur les règles JAPE en termes d'efficacité, de précision et de couverture des annotations spatiales dans les SIG.

En conclusion, ce travail apporte une contribution significative à l'amélioration des technologies d'extraction d'informations spatiales en arabe, en proposant des solutions prometteuses pour la gestion des données géospatiales.

Mots Clés : Ontologie Spatiale, Extraction d'Information Spatiale, Rules JAPE, TALN arabe.

Abstract

With the rapid growth of the Internet, the amount of available information has increased significantly, leading to a growing problem of "information overload." This situation highlights the importance of quickly and accurately accessing relevant information, especially in specialized fields such as Geographic Information Systems (GIS). Extracting spatial information from Arabic texts presents a major challenge due to the linguistic complexity and the significance of geospatial data across various sectors.

To address this issue, this research focused on the creation and integration of the Arabic Spatial Toponyms Ontology (ASTO). This ontology aims to structure geospatial knowledge in Arabic and improve the accuracy of spatial information extraction. In parallel, a rule-based approach using JAPE was developed and evaluated.

Additionally, a hybrid approach combining the ontology with rule-based methods was implemented. Both approaches were compared to evaluate their respective performances. The results show that the hybrid approach outperforms the JAPE rule-based method in terms of efficiency, accuracy, and coverage of spatial annotations in GIS.

In conclusion, this work makes a significant contribution to the advancement of spatial information extraction technologies for the Arabic language by offering promising solutions for geospatial data management.

Keywords: Spatial Ontology, Spatial Information Extraction, JAPE Rules, Arabic NLP.

الملخص

مع الانتشار السريع للإنترنت، ازدادت كمية المعلومات المتاحة بشكل كبير، مما أدى إلى ظهور مشكلة "فرط المعلومات" بشكل متزايد. وهذا يبرز أهمية البحث للوصول السريع والدقيق إلى البيانات ذات الصلة، خاصة في المجالات المتخصصة مثل نظم المعلومات الجغرافية (SIG). إن استخراج المعلومات المكانية من النصوص العربية يمثل تحديًا حاسمًا بسبب التعقيدات اللغوية والأهمية المحورية للبيانات الجغرافية في مختلف القطاعات.

وللوصول إلى الأهداف المسطر لها، تركز هذه الدراسة على إنشاء ودمج "الأنطولوجيا العربية للأماكن المكانية" (ASTO). تم تصميم هذه الأنطولوجيا لهيكلة المعرفة الجغرافية باللغة العربية وتعزيز دقة استخراج المعلومات المكانية. في الوقت نفسه،تم تطوير نهج يعتمد على قواعد JAPE.

ومن اهم المساهمات البارزة في هذا العمل كذلك، تم تطوير نهج هجين يجمع بين الأنطولوجيا والأساليب القائمة على القواعد. اثبت هذا النهج الهجين فعاليته من خلال دمج الأنطولوجيات والقواعد المتخصصة لاستخراج المعلومات المكانية. و تم مقارنته مع النموذج الذي يعتمد على قواعد JAPE لتقييم أدائه ومدى فعالته. وتشير النتائج إلى أن النهج الهجين يتفوق بشكل كبيرمن حيث الدقة وشمولية التوصيفات المكانية في نظم المعلومات الجغرافية.

و في الختام، تساهم هذه الدراسة بشكل كبير في تطوير تقنيات استخراج المعلومات المكانية باللغة العربية، وتقدم حلولًا واعدة لإدارة البيانات الجغرافية.

الكلمات المفتاحية : الأنطولوجيا المكانية، استخراج المعلومات المكانية، قواعد JAPE، معالجة اللغة الطبيعية العربية.