Université Grenoble Alpes

# Construction automatique d'un graphe de connaissances géo-historiques à partir d'entrées encyclopédiques

**Bin YANG**

Septembre, 2025

**Sous la direction de**
Ludovic MONCLA, INSA Lyon, Liris UMR 5205
Fabien DUCHATEAU, Université Claude Bernard Lyon 1, Liris UMR 5205
Frédérique LAFOREST, INSA Lyon, Liris UMR 5205

Laboratoire d'Informatique en Images et Systèmes d'Information (INSA Lyon)

---

Mémoire de Master 2
UFR LLASIC, Département Sciences du langage
Parcours Industrie de la Langue, orientation professionnelle
Référent pédagogique : Claude PONTON, Université Grenoble Alpes
Année universitaire 2024-2025

Université Grenoble Alpes

# Construction automatique d'un graphe de connaissances géo-historiques à partir d'entrées encyclopédiques

Bin YANG

Septembre, 2025

**Sous la direction de**
Ludovic MONCLA, INSA Lyon, Liris UMR 5205
Fabien DUCHATEAU, Université Claude Bernard Lyon 1, Liris UMR 5205
Frédérique LAFOREST, INSA Lyon, Liris UMR 5205

Laboratoire d'Informatique en Images et Systèmes d'Information (INSA Lyon)

---

Mémoire de Master 2
UFR LLASIC, Département Sciences du langage
Parcours Industrie de la Langue, orientation professionnelle
Référent pédagogique : Claude PONTON, Université Grenoble Alpes
Année universitaire 2024-2025

# Remerciements

Je tiens à exprimer ma profonde gratitude à toutes les personnes qui m'ont accompagné et soutenu tout au long de mes études de master et de mon stage. Leur présence, leurs encouragements et leur bienveillance ont joué un rôle déterminant dans la réussite de ce parcours.

Je remercie tout d'abord l'Université Grenoble Alpes ainsi que l'ensemble de l'équipe pédagogique du master Sciences du langage – parcours Industrie de la langue pour la qualité de leur enseignement, la richesse des contenus proposés et leur encadrement attentif tout au long de ma formation. Grâce à eux, j'ai pu acquérir des compétences solides en traitement automatique des langues, qui m'ont été précieuses tant sur le plan académique que professionnel.

Je souhaite ensuite adresser mes sincères remerciements au Laboratoire d'Informatique en Images et Systèmes d'Information (LIRIS) qui m'a offert l'opportunité de réaliser mon stage de fin d'études dans un environnement de recherche stimulant, bienveillant et intellectuellement enrichissant.

Je suis particulièrement reconnaissant envers mes trois tuteurs :

**Ludovic Moncla**, maître de conférences à l'INSA Lyon,

**Fabien Duchateau**, maître de conférences à l'Université Claude Bernard Lyon 1,

et **Frédérique Laforest**, professeure à l'INSA Lyon,

pour leur supervision rigoureuse et bienveillante, leur patience, ainsi que pour leurs conseils pertinents et éclairés. Leur disponibilité, leur confiance et leur accompagnement constant ont grandement facilité mon intégration au sein de l'équipe et ont contribué de manière décisive à l'avancement et à la qualité de ce mémoire. J'ai énormément appris à leurs côtés, tant sur le plan scientifique que méthodologique, et je leur en suis profondément reconnaissant.

Je tiens également à remercier chaleureusement mes collègues doctorants du laboratoire pour leur accueil convivial, leurs échanges toujours intéressants et leur précieuse aide au quotidien. Leur esprit d'équipe et leur enthousiasme ont largement contribué à rendre cette expérience aussi agréable qu'enrichissante.

Enfin, je remercie toutes les personnes qui, de près ou de loin, ont contribué à mon parcours, par leurs encouragements, leurs conseils ou leur soutien moral. plm Je remercie ChatGPT et Deep-seek pour leur aide à la reformulation de quelques paragraphes de ce document.

Université Grenoble Alpes

# DÉCLARATION ANTI-PLAGIAT

1. Ce travail est le fruit d'un travail personnel et constitue un document original.

2. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.

3. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).

4. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

PRENOM : BIN

NOM : YANG

DATE : 01/09/2025

# Table des matières

# Chapitre 1 - Introduction

# 1 Introduction

## 1.1 Problématique

Les dictionnaires encyclopédiques anciens, comme ceux du siècle des Lumières, ont joué un rôle important dans la diffusion des savoirs. Aujourd'hui, ils constituent une source précieuse pour les linguistes et les historiens qui étudient l'évolution des connaissances et des représentations du monde à travers le temps. Cependant, la richesse et la densité de ces corpus rendent toute analyse manuelle difficile, voire irréalisable à grande échelle.

Les outils du traitement automatique des langues (TAL) offrent de nouvelles perspectives pour explorer et exploiter ces corpus. La construction de graphes de connaissances est une approche particulièrement pertinente : elle permet de structurer les informations sous une forme interopérable, exploitable, et adaptée aux analyses comparatives ou diachroniques. Spécifiquement, un graphe géographique facilite l'identification des entités (e.g. pays, régions, villes, rivières) et de leurs relations spatiales (e.g. topologiques, directionnelles), tout en fournissant un support visuel et analytique à la recherche en sciences humaines.

Dès lors, comment pouvons-nous automatiser la construction d'un graphe de connaissances géographiques à partir de textes encyclopédiques anciens, tout en tenant compte des spécificités linguistiques et sémantiques propres à ce type de corpus ?

## 1.2 Objectifs

Le projet ECoDA [1] (2025-2026), financé par la Fédération Informatique de Lyon (FIL), a pour objectif d'étudier les évolutions géographiques dans les dictionnaires anciens. Dans le cadre de ce projet, mon stage porte sur la construction automatique d'un graphe de connaissances géo-historiques à partir d'articles du domaine géographique issus de l'EDdA.

En mettant en application des méthodes de traitement automatique des langues (TAL) adaptées aux textes en français ancien, ainsi que des approches d'apprentissage automatique notamment apprentissage pronfond (deep learning), l'objectif de mon stage est de reconnaître et classifier toutes les entités géographiques dans les

1. https ://liris.cnrs.fr/projet-institutionnel/fil-2025-projet-ecoda

textes, détecter les relations spatiales entre elles, puis modéliser et structurer toutes les informations extraites pour en construire un graphe de connaissances. Ainsi, nous pourrons transformer ce corpus patrimonial en une ressource exploitable sous forme de graphe, facilitant l'analyse et la valorisation du savoir géographique tel qu'il était représenté au XVIIIe siècle.

### 1.3 Laboratoire d'accueil

Le laboratoire de recherche qui m'a accueilli est le **Laboratoire d'InfoRmatique en Images et Systèmes d'Information (LIRIS), URM CNRS 5205, sur le site de l'INSA Lyon**[2]. Ce laboratoire, créé en 2003, est spécialisé en informatique et plus généralement en gestion de données et en traitement et analyse des images. Il regroupe aujourd'hui 330 membres répartis en 12 équipes.

Le LIRIS se structure autour de six grands axes d'expertise :
1. Données, systèmes et sécurité (équipes BD, DRIM, SOC et DM2L)
2. Informatique graphique et géométrie (équipe ORIGAMI)
3. Images, vision et apprentissage (équipe IMAGINE)
4. Interaction et cognition (équipes SICAL, SyCoSMA et TWEAK)
5. Algorithmes et combinatoire (équipe GOAL)
6. Simulation et sciences du vivant (équipes SAARA et BEAGLE)

Durant ces six mois, j'ai travaillé au sein de l'équipe **DM2L** (Data Mining and Machine Learning) sous la direction de **Ludovic Moncla** (maître de conférences, équipe DM2L, INSA Lyon), **Fabien Duchateau** (maître de conférences, équipe BD, Université de Claude Bernard Lyon 1) et **Frédérique Laforest** (professeure, équipe TWEAK, INSA Lyon) avec au moins une réunion par semaine.

2. https ://liris.cnrs.fr/

# Chapitre 2 - État de l'art

# 2 État de l'art

Cette section présente les travaux liés aux graphes de connaissances géographiques (2.1) ainsi que ceux portant sur l'extraction d'informations à partir de textes (2.2).

## 2.1 Travaux sur les graphes de connaissances géographiques

Les graphes de connaissances (ou *knowledge graphs*) sont des structures formelles destinées à représenter des entités et des relations qui les unissent sous forme de triplets ⟨sujet, prédicat, objet⟩. Le standard RDF (*Resource Description Framework*) est largement utilisé pour modéliser ces graphes, notamment dans le Web sémantique.

Plusieurs ontologies [3] génériques, comme FOAF [4], DBpedia ou Wikidata, ont permis de structurer de vastes ensembles de connaissances dans des domaines variés. Cependant, lorsqu'il s'agit de domaines spécialisés, comme la géographie historique, des modèles plus adaptés sont nécessaires. Des travaux comme GeoSPARQL [11, 7] ont été proposés pour intégrer des aspects spatiaux dans des graphes RDF, notamment en permettant la représentation de géométries, la manipulation de coordonnées spatiales et l'interrogation spatiale via des opérateurs topologiques.

La modélisation des entités géographiques dans les graphes de connaissances permet de formaliser les types d'objets géographiques (pays, villes, rivières, montagnes, etc.) et leurs propriétés spatiales. Des ontologies comme GeoNames Ontology [15], SWEET [13] ou encore GEO (Geographic Ontology) ont été élaborées pour structurer l'information géographique et faciliter l'interopérabilité entre sources hétérogènes. Ces modèles permettent notamment l'instanciation explicite des entités géographiques selon leur nature, leur position, leur hiérarchie spatiale et leurs relations avec d'autres objets.

Des travaux récents, tels que ceux de Rawsthorne et al. [12], proposent une approche de référence pour l'identification d'entités spatiales imbriquées ainsi que des relations spatiales à partir de textes. Cette méthode s'appuie sur un jeu de données annoté spécifiquement conçu pour évaluer la performance des modèles, et fournit une base solide pour comparer différentes stratégies d'extraction. De telles contributions facilitent la structuration de l'information spatiale textuelle et ouvrent la voie à une

3. `https://fr.wikipedia.org/wiki/Ontologie`
4. `https://fr.wikipedia.org/wiki/FOAF`

meilleure intégration des connaissances géographiques dans des systèmes intelligents.

Enfin, certains travaux ont exploré l'enrichissement de graphes de connaissances avec des dimensions temporelles et spatiales (cf. T-GK [4], SpatioTemporal RDF [14]), mais ceux-ci restent souvent centrés sur des données contemporaines, bien structurées et issues de corpus modernes.

## 2.2 Travaux sur l'extraction d'information

L'extraction d'information (EI) vise à transformer un texte non structuré en données exploitables, typiquement sous forme de triplets ou d'annotations catégorisées. Les principales tâches incluent la reconnaissance d'entités nommées (NER), la classification d'entités, et l'extraction de relations (RE) entre ces entités [6].

Les approches classiques reposent sur des méthodes symboliques ou basées sur des règles [2, 8] (ex. expressions régulières, grammaires d'extraction). Elles ont progressivement été supplantées par des méthodes statistiques puis neuronales. Les architectures basées sur les réseaux de neurones profonds et les modèles pré-entrainés (notamment BERT [5] et ses variantes multilingues) ont permis des avancées significatives, notamment pour la RE en contexte multilingue ou faible supervision. Plus récemment, les grands modèles de langage (LLMs, comme GPT-3, GPT-4, LLaMA, Mistral, etc.) ont introduit une nouvelle façon d'aborder l'extraction d'information, via des méthodes dites zero-shot ou few-shot, fondées sur le prompt engineering [10].

Dans le domaine géographique, plusieurs travaux se sont concentrés sur l'extraction de relations spatiales à partir de textes, en particulier des relations telles que la contenance, la proximité ou la direction. Ces approches reposent souvent sur des modèles spécialisés, comme le Spatial Role Labeling (SpRL), qui attribuent des rôles sémantiques aux entités géographiques, par exemple la trajectoire, le lieu de référence ou l'objet de la relation [5]. D'autres méthodes combinent des règles linguistiques avec des techniques d'apprentissage automatique pour améliorer la précision de l'extraction. Ces approches peuvent être appliquées à divers types de corpus géographiques, tels que des textes historiques, des documents administratifs ou des articles encyclopédiques, et permettent de construire des graphes de connaissances spatiales représentant automatiquement les relations entre lieux.

Toutefois, la majorité de ces ressources sont en anglais contemporain, ce qui limite

5. https://en.wikipedia.org/wiki/BERT_(language_model)

leur réutilisabilité dans des contextes historiques ou multilingues. De plus, les outils existants s'appuient souvent sur des connaissances linguistiques et encyclopédiques modernes, difficilement transposables à des textes anciens.

Des travaux récents, tel que le projet **GEODE**[6] se sont intéressés à des corpus historiques en français avec l'objectif d'étudier des évolutions survenues dans les discours géographiques à travers le temps dans un corpus encyclopédique. Ce projet a notamment proposé une étude limitée à l'identification d'entités nommées dans les textes anciens de l'Encyclopédie de Diderot et d'Alembert [10]. Différentes approches ont été expérimentées telles que le modèle à champs aléatoires conditionnels (CRF), le *fine-tuning* du modèle pré-entraîné CamemBERT, le modèle Flair, et le *few-shot prompting* de LLMs.

6. https ://geode-project.github.io/

# Chapitre 3 - Positionnement

# 3 Positionnement

Les approches présentées dans les sous-sections précédentes permettent d'extraire et de structurer les connaissances issues de textes modernes et bien formatés, en particulier dans les contextes génériques ou encyclopédiques. Cependant, elles s'avèrent insuffisantes pour notre objectif spécifique : la construction d'un graphe de connaissances géo-historiques à partir de textes issus de l'Encyclopédie de Diderot et d'Alembert. Plusieurs limites se posent :

- **Absence d'ontologie spécialisée** : Il n'existe pas, à notre connaissance, d'ontologie dédiée à la représentation fine des connaissances géographiques. Les relations spatiales (notamment les dépendances territoriales, les parcours des cours d'eau) sont rarement modélisées de manière systématique.
- **Manque de pipeline pour la détection de relations spatiales et le relation-linking** : Les outils d'extraction de relations actuels sont peu adaptés à la détection de relations spatiales complexes, notamment lorsqu'elles sont exprimées de manière indirecte ou floue. Par ailleurs, il n'existe pas de pipeline robuste permettant d'effectuer un linking fiable entre les entités géographiques identifiées, en particulier dans les cas d'ambiguïté sémantique (e.g. ville de Vienne en France et celle en Autriche) ou de changements au cours du temps.
- **Spécificité du corpus** : Les textes de notre corpus sont rédigés en français du XVIIIe siècle, avec une structure syntaxique et un vocabulaire différents du français moderne. Cette spécificité linguistique limite l'efficacité des outils d'EI actuels, entraînés majoritairement sur des corpus contemporains.

Pour répondre à notre problématique, nous avons identifié trois grandes étapes :

- **Modélisation du graphe :** conception de la structure du graphe
- **Peuplement du graphe :**
  - Pré-traitement de données
  - Classification des vedettes des articles (pays, ville, rivière, etc.)
  - Reconnaissance des entités nommées au sein des articles
  - Classification des entités nommées de lieux (pays, ville, rivière, etc.)
  - Détection de relations spatiales entre les entités nommées de lieux
  - Construction du graphe en RDF
- **Évaluation du graphe :** Analyse quantitative et qualitative du graphe

La figure 1 illustre notre approche avec 5 exemples d'article. En entrée, nous avons les articles de l'encyclopédie sous forme de tableur. Le pré-traitement de données vise à sélectionner des articles de type "Place" puis à segmenter ceux dont les vedettes sont polysémiques (e.g. la vedette THESTIS décrit deux villes différentes portant le même nom) en sous-articles indépendants.

Pour chaque article de type "Place", nous effectuons ensuite la classification de type de lieu des vedettes. L'encart numéroté ① (à gauche) montre que la vedette Paris et 2 THESTIS ont été typées comme Ville alors que la France a été typée en Pays. Les textes sont ensuite traités pour identifier les entités nommées de lieux et classifier leurs types de lieu. L'encart numéroté ② montre 2 nouvelles entités, `E5`(la Seine) et `E6`(Londres), et le type de ces entités est spécifié dans l'encart ③. Nous cherchons ensuite à détecter les relations spatiales entre les vedettes et les entités identifiées, comme illustré dans l'encart numéroté ④ par les relations `crosses`, `inclusion` et `orientation` dans l'exemple. Nous finissons par la structuration des entités et relations spatiales en RDF.

En particulier, chaque fois que nous identifions un triplet de relation, il sera associé à son article d'origine pour tracer la provenance de cette information extraite. A cause de la limite d'espace dans le figure, ce point n'y est pas illustré et nous allons détailler dans la section 2 - Modélisation du graphe.

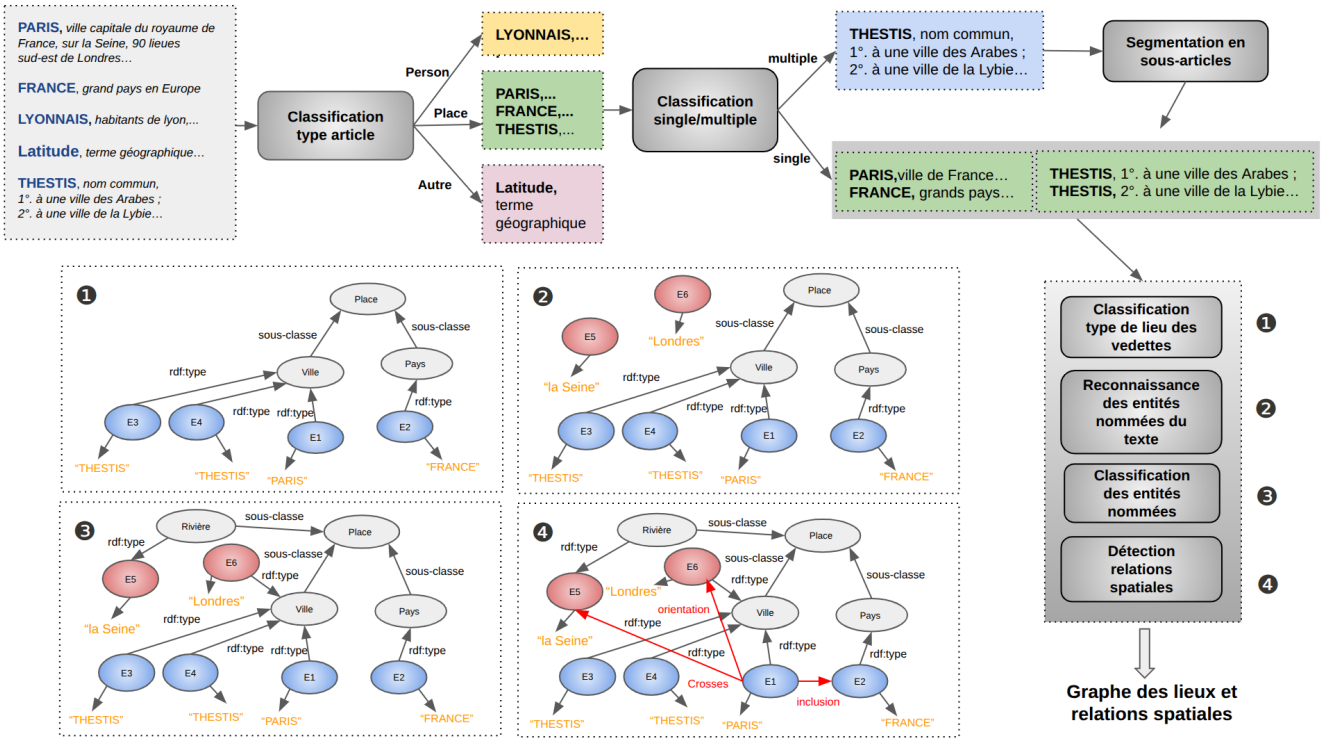


FIGURE 1 – Aperçu du workflow de notre approche

# Chapitre 4 - Modélisation du graphe

# 4 Modélisation du graphe

Dans cette section, nous nous concentrons sur la conception de la structure du graphe. La section 4.1 présente l'ontologie spatiale, un modèle formel qui permet de représenter et d'organiser les concepts et les relations spatiales dans l'espace géographique. La section 4.2 présente l'ontologie de provenance qui permet d'identifier la source de toutes les informations extraites de l'EDdA.

## 4.1 Ontologie spatiale

L'ontologie spatiale définit et structure les concepts, relations et propriétés liés à l'information géographique. Elle est donc une représentation sous forme de graphes où :

— Les **Classes** représentent des concepts géographiques de différents types (e.g. Pays, Région, Ville, Rivière).

— Les **Relations** représentent les liens spatiaux existants entre ces entités (e.g. Inclusion, Adjacence, Orientation, Distance) ou ceux qui désignent des attributs géographiques des entités (e.g. latitude, longitude, surface, longueur)

Dans l'ontologie spatiale, nous avons défini trois classes de premier niveau :

— `Place` : tout ce qui relève de la localisation géographique,

— `Person` : des collectifs humains ou communautés (ex. "les Gaulois", "les Parisiens", "les marchands", "les paysans"), pertinents pour suivre l'évolution démographique,

— `Misc` : une catégorie de secours pour des entités ou termes géographiques particulières qui ne rentrent pas clairement dans les deux autres (ex. "Latitude", "Longitude").

Comme nous nous focalisons sur la géographie, la classe “Place” contient ensuite dix sous-classes : Pays, Région, Ville, Mer, Rivière, Lac, Île, Montagne, Construction Humaine, Autrelieu. La figure 2 montre la hiérarchie de nos classes pour les deux niveaux.

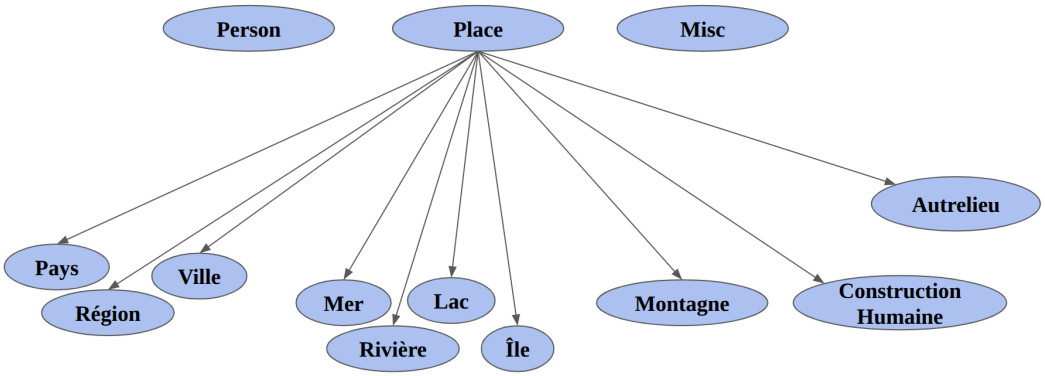


FIGURE 2 – Classes de l'ontologie spatiale

Concernant les relations dans l'ontologie spatiale, nous avons d'abord défini finement les types de relations entre les différents types d'entités :

- **Inclusion** (e.g. dans le duché, en Allemagne, au royaume de France)
- **Adjacence** (e.g. à côté de, près de, proche de, sur la côte de)
- **Orientation** (e.g. au sud de, au couchant de, au midi, au levant de, au S.O. de)
- **Distance** (e.g. à deux lieues de, à cinq milles de, à deux parasanges de)
- **Mouvement** (e.g. se jette dans, prend sa source de, coule dans la mer)
- **Crosses** (e.g. se situe sur la rivière, traverse la ville)
- **Autre-relation** (e.g. entre le Liban et l'Antiliban)

Comme les informations d'attributs (longitude, latitude, longueur) des entités sont importantes pour étudier des évolutions du savoir géographique, nous avons ensuite défini des types de relations entre une entité et ses attributs (valeurs littérales) :

- **a-longitude** (e.g. long. 21. 24.)
- **a-latitude** (e.g. lat. 51. 17.)
- **a-surface** (e.g. 2000 mètres carrés)
- **a-longueur** (e.g. 100 milles)

## 4.2 Ontologie de provenance

L'ontologie de provenance a pour objectif de tracer l'origine des informations extraites de l'encyclopédie et stockées dans l'ontologie spatiale. En effet, à terme, il

est envisagé d'intégrer d'autres encyclopédies (e.g. le dictionnaire du Trevoux[7] en plusieurs éditions). De nouvelles classes et relations permettent donc de préciser la source des informations extraites, c'est-à-dire, de quels articles, quels volumes et quelles encyclopédies les informations extraites sont issues :

— **Classes** : Encyclopédie, Volume, Article
— **Relations** : articleDe, volumeDe, numArticle, numVolume

La figure 3 illustre ces classes et relations liées à la provenance.

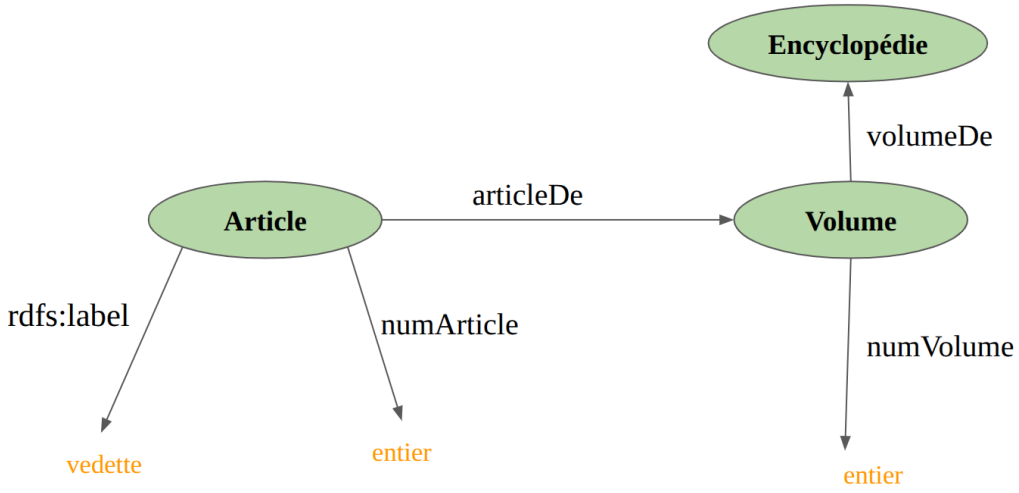


FIGURE 3 – Classes et relations de l'ontologie de provenance

En fonction de ces deux ontologies, tout triplet extrait du dictionnaire sera représenté par l'intermédiaire d'un statement (procédé de réification) et intégré dans le graphe avec l'identification de sa source. La figure 4 illustre l'intégration du triplet (Lyon, Adjacence, Grenoble) dans un graphe en joignant deux ontologies définies.

7. https://fr.wikipedia.org/wiki/Dictionnaire_de_Trevoux

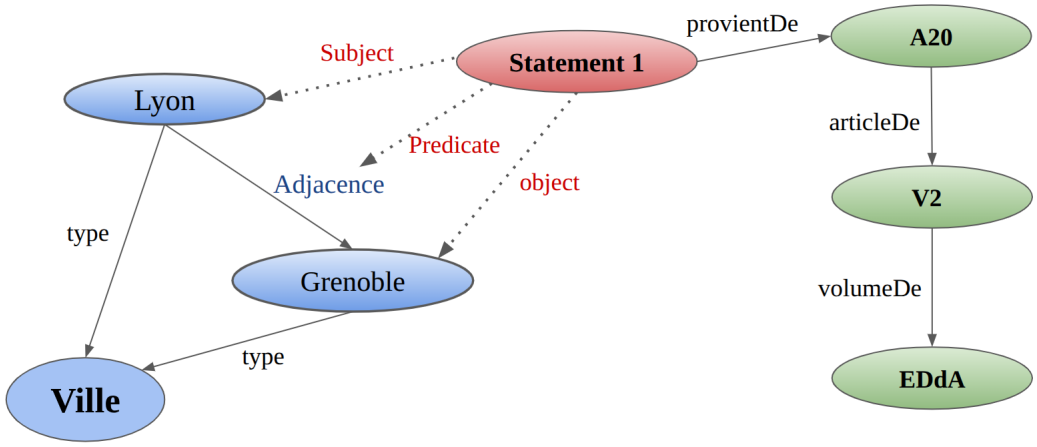


FIGURE 4 – L'exemple de l'intégration du triplet (Lyon, Adjacence, Grenoble) au sein du graphe

# Chapitre 5 - Peuplement du graphe

# 5 Peuplement du graphe

Une fois la modélisation définie, nous avons pu mettre en place le processus d'instanciation des concepts et des relations sous la forme d'un graphe de connaissances. Cette section décrit donc en détails chacune des étapes du processus de préparation, d'extraction d'information et de peuplement du graphe de connaissances.

## 5.1 Pré-traitement de données

Cette étape consiste en la sélection des articles géographiques qui décrivent des lieux, car la catégorie géographie contient d'autres types d'articles comme par exemple les noms de peuples ou de communautés (e.g. Salyens [8]) ainsi que des noms de concepts géographiques (e.g. latitude [9]). Pour cela, nous avons mené trois expérimentations successives :

— la classification des articles selon leur type — **Lieu** (e.g. Paris, France, la Seine), **Personne** (e.g. Lyonnais, Salyens) et **Autre** (e.g. Actium [10] ) (section 5.1.1)

— la classification de vedettes en single ou multiple (section 5.1.2)

— lorsqu'ils décrivent plusieurs lieux, la segmentation des articles en sous-articles (section 5.1.3)

Ces expérimentations s'appuient sur un jeu de données annoté manuellement par moi-même, appelé **GeoEDdA-TopoRel** [11], composé de 2,750 articles : 2,250 de type Place, 250 de type Person, et 250 de type Misc. Le tableau 1 décrit la répartition de ces trois classes en train, validation et test.

En particulier, pour mener des expérimentations de classification single/multiple selon le nombre de lieux décrit par la vedette (section 5.1.2) et de classification selon le type de lieu des vedettes (section 5.2), nous avons annoté les nombres de lieux et les types des vedettes dans les articles du sous-ensemble "Place" du GeoEDdA-TopoRel. Le tableau 2 décrit la distribution des vedettes "single" et "multiple". Le tableau 3 décrit la distribution des types des vedettes pour les 2000 vedettes "single".

8. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/14/3429
9. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/9/1466
10. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/1/729?byte=1775742
11. https://huggingface.co/datasets/GEODE/GeoEDdA-TopoRel

| | train | validation | test | total |
|---|---|---|---|---|
| Place | 1800 | 225 | 225 | 2250 |
| Person | 200 | 25 | 25 | 250 |
| Misc | 200 | 25 | 25 | 250 |

TABLE 1 – Description du jeu de données GeoEDdA-TopoRel

| | train | validation | test | total |
|---|---|---|---|---|
| single | 1600 | 200 | 200 | 2000 |
| multiple | 200 | 25 | 25 | 250 |

TABLE 2 – Description du sous-ensemble "Place" du GeoEDdA-TopoRel

| | train | validation | test | total |
|---|---|---|---|---|
| Ville | 921 | 51 | 40 | 1012 |
| Île | 216 | 20 | 27 | 263 |
| Région | 138 | 22 | 28 | 188 |
| Rivière | 133 | 20 | 28 | 181 |
| Montagne | 63 | 29 | 22 | 114 |
| Construction Humaine | 38 | 12 | 9 | 59 |
| Autrelieu | 27 | 12 | 12 | 51 |
| Mer | 26 | 13 | 12 | 51 |
| Lac | 22 | 9 | 9 | 40 |
| Pays | 16 | 12 | 13 | 41 |
| Total | 1600 | 200 | 200 | 2000 |

TABLE 3 – Description du sous-ensemble "single" du GeoEDdA-TopoRel

#### 5.1.1 Classification de type d'article (Lieu/Personne/Autre)

Avec ce jeu de données de 2750 articles, nous avons fine-tuné le modèle "bert-base-multilingual-cased" [12] pour entraîner un modèle qui permet de prédire automatiquement les types d'articles de l'EDdA.

Le tableau 4 montre que cet entraînement atteint une très bonne performance avec une F-mesure moyenne de 96.91%.

| | Précision | Rappel | F-mesure |
|---|---|---|---|
| Place | 94.74% | 99.56% | 98.25% |
| Person | 100% | 100% | 100% |
| Misc | 96.97% | 72.00% | 81.82% |
| average | 97.04% | 97.09% | 96.91% |

TABLE 4 – Précision, rappel et F-mesure du modèle de classification de type d'articles en Place, Person et Misc

Nous avons donc mis en application ce modèle pour classifier les 15,384 articles géographiques initiaux et nous avons obtenu 14,204 articles de type "Place". Les

12. `https://huggingface.co/google-bert/bert-base-multilingual-cased`

vedettes de ces 14,204 articles sont prêtes pour la classification selon le nombre de lieux en single ou multiple (section 5.1.2).

### 5.1.2 Classification de vedettes en single ou multiple

Cette section a pour objectif de classifier en single ou multiple chaque vedette selon le nombre de lieu qu'elle décrit. En effet, certaines vedettes présentent une polysémie, c'est-à-dire qu'elles renvoient à plusieurs lieux différents ou plusieurs types de lieux portant le même nom. Par exemple :

— **HUNOLDSTEIN** [13], (Géog.) **ville & château** d'Allemagne, dans l'électorat de Trèves.

— **MÉGARSUS, ou MAGARSUS** [14], (Géog. anc.) **1°** une ville de Cilicie, près du fleuve Pyrame ; **2°** une rivière de Scythie, selon Strabon ; **3°** un fleuve de l'Inde, selon Denys le Périégète.

— **SYCAE** [15], (Géog. anc.) nom **d'une ville** de la Cilicie, **& d'une ville** de la Thrace, selon Étienne le géographe. (D. J.)

— **BRANDEIS** [16], (Géog.) **petite ville & château** de Bohême sur l'Elbe, & située à trois lieues de Prague. **Il y a encore une autre ville** de ce nom en Bohême, située sur la rivière d'Orlitz.

Nous avons employé la même méthode de fine-tuning, toujours basée sur le modèle "bert-based-multilingual-cased", pour entraîner un modèle de classification selon le nombre de lieux - single ou multiple. Cet entraînement s'est appuyé sur le sous-ensemble décrit dans le tableau 2.

La figure 5 montre les résultats de l'évaluation sur cet entraînement. Nous pouvons ici aussi constater une F-mesure élevée de 97.76%. Nous avons mis en application ce modèle pour classifier les 14,204 articles typés "Place" obtenus dans l'étape précédente. Nous avons obtenu **13,476** single et **728** multiple.

13. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/8/1718
14. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/10/1365
15. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/15/3468
16. https ://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/2/3535

| | Précision | Rappel | F-mesure |
|---|---|---|---|
| single | 98.51% | 99.00% | 98.75% |
| multiple | 91.67% | 88.00% | 89.80% |
| average | 97.75% | 97.78% | 97.76% |

TABLE 5 – Précision, Rappel et F-mesure du modèle de classification selon le nombre de lieux single/multiple

#### 5.1.3 Segmentation des articles multiples en sous-articles

Cette étape vise à segmenter les articles de type multiple en plusieurs sous-articles de type single. Cependant, la grande diversité des formulations employées pour désigner plusieurs lieux telles que les énumérations, l'usage du symbole «&», ou encore des expressions comme «il y a encore une autre ville», rend difficile l'apprentissage d'un modèle de classification apte à effectuer une segmentation pertinente. Pour contourner cette difficulté, nous avons exploré une approche fondée sur l'instruction de grands modèles de langage (LLMs) à l'aide de prompts few-shot.

Nous avons utilisé les instructions ci-dessous :

""" *Tu es un assistant linguistique expert en langue française. Si une vedette signifie plusieurs lieux, tu dois segmenter l'article en sous-articles numérotés (1°, 2°, 3°, etc.). Chaque sous-article doit commencer une nouvelle ligne, la sortie doit garder la structure :*
*1° NOM, description...*
*2° NOM, description...*
*3° NOM, description...*

Exemple 1 :

**Input :** CAKET, (Géog.) ville & château d'Asie, dépendant du roi de Perse, près du Caucase. Long. 63. 50. lat. 43. 32.

**Output :** 1° CAKET, (Géog.) ville d'Asie, dépendant du roi de Perse, près du Caucase. Long. 63. 50. lat. 43. 32.

2° CAKET, (Géog.) château d'Asie, dépendant du roi de Perse, près du Caucase. Long. 63. 50. lat. 43. 32. """

Nous avons expérimenté les segmentations automatiques sur 4 LLMs : `mistral-7b`, `llama3-70b`, `gpt-4-turbo`, `gpt-4.1-mini` sur les 250 articles en "multiple" du da-

taset GeoEDdA-TopoRel. Nous avons ensuite fait une première évaluation en utilisant la mesure Rouge-L, et une deuxième évaluation à la main (Vrai/Faux). La mesure Rouge-L permet de calculer la similarité entre deux textes en comparant la séquence commune la plus longue, elle est souvent utilisée pour évaluer la traduction automatique et le résumé de texte [1].

Le tableau 6 présente les résultats obtenus avec les deux méthodes d'évaluation. Celles-ci mettent en évidence que les grands modèles GPT obtiennent de meilleures performances sur la tâche de segmentation. Lors de la première évaluation avec Rouge-L, le modèle gpt-4.1-mini a affiché une précision légèrement inférieure à celle de gpt-4-turbo, en raison de sa plus grande flexibilité dans la reformulation des phrases.

Vu les performances des 4 LLMs, nous avons validé finalement le gpt-4.1-mini pour segmenter ces 728 articles précédemment classifiés en "multiple", ce qui a produit 1977 sous-articles "single".

| | Score Rouge-L | Précision(Vrai/Faux) |
|---|---|---|
| mistral :7b | 75.37% | 46.80% |
| llama3 :70b | 87.32% | 85.60% |
| gpt-4-turbo | 93.27% | 95.60% |
| gpt-4.1-mini | 89.19% | 96.80% |

TABLE 6 – Evaluation de la segmentation des articles multiples

### 5.2 Classification selon le type de lieu des vedettes

Cette section a pour objectif de classifier les vedettes des articles "Place" selon leur type. Cette expérimentation s'est appuyée sur le sous-ensemble décrit dans le tableau 3. Les annotations ont été faites en respectant les principes suivants :

- **Pays** : grands pays, royaume, empire, état, etc.
- **Région** : petit pays, région, contrée, province, cercle, duché, etc.
- **Ville** : ville, bourg, village, etc.
- **Île** : île, presqu'île, etc.
- **Rivière** : rivière, fleuve,etc.
- **Montagne** : montagne, vallée, etc.
- **Mer** : mer, golphe, baie, etc.
- **Construction Humaine** : port, château, forteresse, abbaye, etc.
- **Lac** : lac, étang, marais, etc.

— **Autrelieu** : promontoire, cas, rivage, désert, etc.

À partir de ces annotations, un modèle de classification selon le type de lieu a ensuite été entraîné, en utilisant la même méthode de fine-tuning sur le modèle de "Bert-Base-multilingual-cased". La figure 5 présente l'évaluation des résultats de cet entraînement ainsi que la matrice de confusion des classes.

La figure 6 présente la distribution des types de lieu de toutes les 15,453 vedettes de l'EDdA prédites par ce modèle. Nous pouvons constater que la classe "Ville" représente près de 60% du total des catégories, tandis que les classes "Rivière" et "Région" se placent respectivement en deuxième et troisième position.

| | **Précision** | **Rappel** | **F-mesure** |
|---|---|---|---|
| Ville | 90.91% | 100% | 95.24% |
| Île | 96.30% | 96.30% | 96.30% |
| Région | 89.66% | 92.86% | 91.23% |
| Rivière | 96.56% | 100% | 98.25% |
| Montagne | 100% | 95.45% | 97.67% |
| Construction Humaine | 90.00% | 100% | 94.74% |
| Autrelieu | 88.89% | 66.67% | 76.19% |
| Mer | 100% | 91.67% | 94.74% |
| Lac | 100% | 88.89% | 94.12% |
| Pays | 100% | 92.31% | 96.00% |
| average | **94.63%** | **94.50%** | **94.37%** |

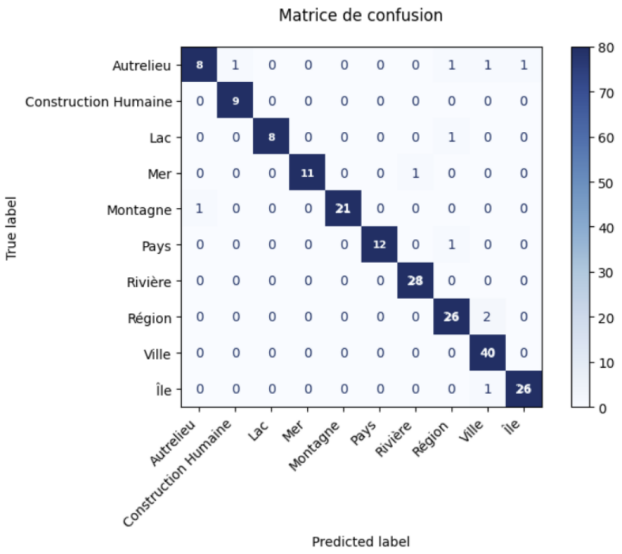


FIGURE 5 – Évaluation du modèle de classification selon le type des vedettes

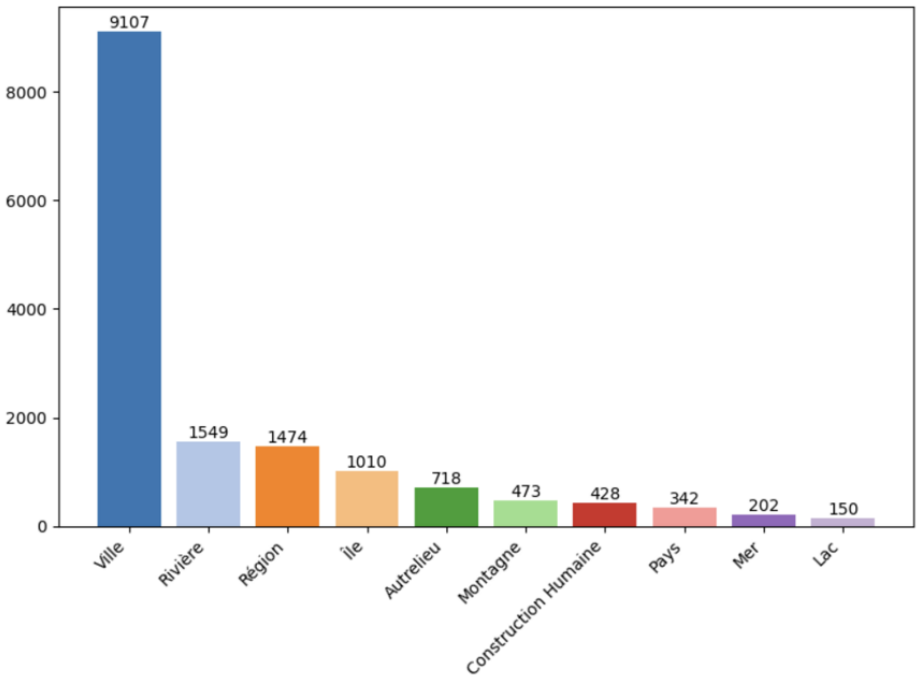


FIGURE 6 – Distribution des types de vedette de l'EDdA

## 5.3 Reconnaissance des entités nommées

Concernant l'étape de reconnaissance des entités nommées géographiques et des expressions de relations spatiales, nous nous appuyons sur les travaux précédents de l'équipe [10]. Un modèle de reconnaissance des entités nommées [17] a été entraîné en utilisant le jeu de données GeoEDdA [18] [9]. Il s'agit d'un modèle de classification de tokens qui attribue une catégorie à chaque token en sortant un span qui contient ses informations de positions dans la phrase, son étiquette prédit, etc. Les différentes catégories incluent les entités nommées (lieux, personnes, et autres), les entités nominales (lieux et personnes), les entités étendues (lieux, personnes et autre), les relations spatiales, les coordonnées géographiques, les vedettes et les marqueurs de domaine).

En particulier, nous nous sommes intéressés aux catégories "NP-spatial" et "NC-spatial". L'étiquette NP-spatial (pour Nom Propre spatial) désigne les entités nommées correspondant à des toponymes ou à des lieux identifiables de manière unique. Il s'agit par exemple de Paris, de la France, des Alpes ou encore de la Méditerranée.

À l'inverse, l'étiquette NC-spatial (pour Nom Commun spatial) regroupe les termes

17. `https://huggingface.co/GEODE/camembert-base-edda-span-classification`
18. `https://huggingface.co/datasets/GEODE/GeoEDdA`

de type générique qui renvoient à des catégories ou à des classes de lieux. On y trouve par exemple des mots comme ville, rivière, palais ou montagne. Ces termes ne désignent pas directement une entité spécifique mais caractérisent un type ou une nature d'objet géographique, ce qui permet d'indiquer la fonction ou le statut d'un lieu mentionné dans le texte.

Le modèle camembert-base-edda-span-classification [19] a été entrainé en prenant des textes avec des étiquettes BIO [20]. Ce modèle permet d'analyser automatiquement chaque token en lui attribuant une catégorie d'étiquette.

Le tableau 7 montre les résultats de sortie du modèle en format JSON avec l'entrée de la phrase : "ILLESCAS, (Géog.) petite ville d'Espagne, dans la nouvelle Castille, à six lieues au sud de Madrid."

Après avoir mis en application ce modèle sur les 15,453 articles de l'EDdA, nous avons obtenu 87,500 spans de "NP-Spatial" et 38,829 spans de "Relation".

| |
|---|
| {"text": "ILLESCAS","start": 0,"end": 8,"token_start": 0,"token_end": 0,"label": "Head"} |
| {"text": "Géog.","start": 11,"end": 16,"token_start": 3,"token_end": 4,"label": "Domain-mark"} |
| {"text": "petite ville","start": 18,"end": 30,"token_start": 6,"token_end": 7,"label": "NC-Spatial"} |
| {"text": "petite ville d'Espagne","start": 18,"end": 40,"token_start": 6,"token_end": 9,"label":"ENE-Spatial"} |
| {"text": "petite ville d'Espagne, dans la nouvelle Castille","start": 18,"end": 67,"token_start": 6,"token_end": 14,"label": "ENE-Spatial"} |
| {"text": "Espagne","start": 33,"end": 40,"token_start": 9,"token_end": 9,"label": "NP-Spatial"} |
| {"text": "dans","start": 42,"end": 46,"token_start": 11,"token_end": 11,"label": "Relation"} |
| {"text": "la nouvelle Castille","start": 47,"end": 67,"token_start": 12,"token_end":14,"label":"NP-Spatial"} |
| {"text": ", à six lieues au sud de","start": 67,"end": 91,"token_start": 15,"token_end": 21,"label":"Relation"} |
| {"text": "Madrid","start": 92,"end": 98,"token_start": 22,"token_end": 22,"label": "NP-Spatial"} |

TABLE 7 – Extrait de l'output du modèle camembert-base-edda-span-classification

19. https://huggingface.co/GEODE/camembert-base-edda-span-classification
20. https://plmlatex.math.cnrs.fr/project/686cfed859b0b6430d0378b2

## 5.4 Classification des entités nommées selon le type de lieu

Dans cette section, nous avons exploré deux approches pour classifier les entités nommées de lieux identifiées du texte dans la section précédente selon leur type de lieu : few-shot prompting (4.4.1) et bert-based supervised classification (4.4.2).

### 5.4.1 Few-shot prompting

Étant donnés le texte d'origine et les spans de types "NP-Spatial", un LLM est guidé par un prompt pour attribuer à chaque span l'une des classes parmi les dix types définis dans notre ontologie spatiale (voir figure 2).

Nous avons expérimenté avec 4 LLMs : **Deepseek-r1 : 7b**, **Mixtral : 8*7b**, **gpt-4-turbo** et **gpt-4.1-mini** sur le jeu de test du dataset GeoEDdA-TopoRel. Le tableau 8 présente les résultats de l'évaluation de ces 4 LLMs avec cette approche de few-shot prompting.

| | Précision | Rappel | F-mesure |
|---|---|---|---|
| gpt-4.1-mini | 86.90% | 91.10% | **87.60%** |
| gpt-4-turbo | 77.40% | 81.20% | 78.70% |
| mixtral : 8*7b | 58.50% | 61.50% | 56.70% |
| deepseek-r1 : 7b | 39.10% | 46.40% | 37.30% |

TABLE 8 – Evaluation sur 4 LLMs en classification de types des entités nommées

Les résultats montrent une nette supériorité de gpt-4.1-mini, qui atteint la meilleure F-mesure (87,6 %) grâce à un bon équilibre entre précision et rappel, suivi par gpt-4-turbo avec des performances correctes mais sensiblement inférieures (78,7 %). En comparaison, les modèles open-source mixtral :8x7b et surtout deepseek-r1 :7b obtiennent des scores beaucoup plus faibles (56,7 % et 37,3 %), révélant leurs limites pour la classification des entités via simple prompting et soulignant la nécessité, pour ces modèles, d'un affinage ou d'un entraînement spécialisé pour rivaliser avec les modèles GPT-4.

### 5.4.2 Bert-based classification supervisée

En fonction des résultats ci-dessus, nous avons fait annoter automatiquement le type de lieu de tous les spans "NP-spatial" dans les 2,000 articles “single” du GeoEDdA-

TopoRel par le modèle gpt-4.1-mini. Ces annotations vont servir de jeu d'entraînement pour un modèle supervisé de classification des types d'entités nommées.

Contrairement à une annotation manuelle réalisée par des experts, ce procédé génère un corpus bruité, c'est-à-dire que les étiquettes produites peuvent contenir des erreurs ou des imprécisions liées aux limites du modèle génératif. Néanmoins, cette approche permettrait de disposer rapidement d'un grand volume de données annotées, ce qui ouvre la possibilité de tester l'efficacité d'un apprentissage supervisé même à partir de labels imparfaits. L'hypothèse sous-jacente est que, malgré le bruit, les régularités statistiques capturées par un modèle basé sur BERT peuvent-elles suffire à généraliser et à apprendre des représentations utiles pour la classification du type de lieu associé aux entités nommées ?

Nous avons exploré différentes stratégies pour représenter le contexte d'apparition d'une entité. Pour capturer cette information, nous avons extrait des fenêtres de contexte de tailles variables autour de chaque entité cible, en testant des séquences de 4-grammes, 5-grammes et 8-grammes. L'intuition est qu'une petite fenêtre peut fournir un signal précis mais limité (par exemple, un déterminant ou un adjectif directement lié à l'entité), tandis qu'une fenêtre plus large intègre des indices syntaxiques et sémantiques supplémentaires (par exemple, des verbes ou des compléments qui précisent la fonction du lieu dans la phrase). Cette expérimentation vise à évaluer quelle granularité de contexte maximise la performance du modèle de classification tout en limitant l'introduction de bruit contextuel.

Le tableau 9 montre les différentes tailles de contexte de l'entité "la nouvelle Castille" dans le texte : *ILLESCAS, (Géog.) petite ville d'Espagne, dans la nouvelle Castille, à six lieues au sud de Madrid.*

| | contexte |
|---|---|
| 4-grammes précédents | "petite ville d'Espagne, dans la nouvelle Castille" |
| 4-grammes précédents et suivants | "petite ville d'Espagne, dans la nouvelle Castille, à six lieues au" |
| 5-grammes précédents et suivants | "(Géog.), ville d'Espagne, dans la nouvelle Castille, à six lieues au sud" |
| 8-grammes précédents et suivants | "ILLESCAS, (Géog.) petite ville d'Espagne, dans la nouvelle Castille, à six lieues au sud de Madrid." |

TABLE 9 – Exemple de différentes tailles de contexte pour "la nouvelle Castille"

Le tableau 10 présente les évaluations sur les résultats d'entraînement avec des contextes de différentes tailles de n-grammes. Les résultats mettent en évidence l'influence de la taille de la fenêtre de contexte sur les performances du modèle de

classification des types d'entités nommées. Nous pouvons constater une tendance générale à l'amélioration des métriques lorsque la fenêtre de contexte s'élargit. Par exemple, la précision augmente progressivement avec la taille du contexte, ce qui suggère que le modèle bénéficie d'informations supplémentaires pour mieux identifier le type d'une entité géographique.

L'ajout des mots situés après l'entité, c'est-à-dire l'utilisation d'une fenêtre symétrique comprenant à la fois les termes précédents et suivants, améliore systématiquement les scores par rapport à une fenêtre ne considérant que les 4-grammes précédents. Cela indique que les informations postérieures à l'entité contiennent souvent des indices importants pour la désambiguïsation des types de lieux.

En revanche, l'augmentation de la taille de la fenêtre au-delà de cinq termes de part et d'autre de l'entité apporte un gain relativement faible. Le passage de 5-grammes à 8-grammes précédents et suivants n'améliore les métriques que de manière marginale, ce qui suggère qu'une fenêtre de taille moyenne est suffisante pour capturer l'essentiel du contexte pertinent, et qu'un contexte plus large peut introduire du bruit ou des informations redondantes.

Enfin, ces résultats montrent également la robustesse du modèle face à un jeu d'entraînement bruité, généré automatiquement par un LLM. Malgré la présence potentielle d'erreurs dans les annotations, le modèle parvient à apprendre des représentations fiables et à produire des performances stables. Ces observations permettent de conclure que dans notre cas il est possible de tirer parti d'annotations automatiques pour entraîner des modèles supervisés, tout en choisissant une taille de contexte adaptée pour maximiser la performance.

Nous avons finalement validé le modèle de classification de type des entités nommées de 5-grammes qui est disponible ici [21] sur Huggingface.

| | Précision | Rappel | F-mesure | accuracy average |
|---|---|---|---|---|
| 4-grammes précédents | 82.53% | 82.10% | 82.03% | 82.10% |
| 4-grammes précédents et suivants | 83.99% | 83.32% | 83.28% | 83.32% |
| 5-grammes précédents et suivants | 85.12% | 84.85% | 83.80% | 84.73% |
| 8-grammes précédents et suivants | 85.16% | 84.54% | 84.49% | 84.54% |

TABLE 10 – Evaluation de l'impact de différentes tailles de contexte

La figure 7 présente les précisions, rappels et F-mesures pour chaque classe du modèle de classification utilisant une fenêtre de 5-grammes, ainsi que la matrice de confusion

21. https://huggingface.co/GEODE/bert-base-multilingual-cased-classification-ner

associée. Dans l'ensemble, le modèle montre de bonnes performances, avec des scores élevés pour la majorité des classes, ce qui indique que la taille de contexte de 5-grammes permet de capturer efficacement les informations nécessaires à la distinction des types d'entités spatiales. En revanche, selon la matrice de confusion, les erreurs se produisent principalement entre certaines paires de classes proches sémantiquement, notamment "Pays-Région", "Région-Ville" et "Autre-Ville", ce qui reflète la difficulté inhérente à distinguer des entités dont les frontières sémantiques peuvent être floues. Ces résultats suggèrent que, bien que le modèle généralise bien, des confusions restent inévitables pour des catégories proches ou ambigues.

| | **Précision** | **Rappel** | **F-mesure** |
|---|---|---|---|
| Ville | 86.81% | 85.87% | 86.34% |
| Île | 76.07% | 81.65% | 78.76% |
| Région | 89.40% | 84.90% | 87.09% |
| Rivière | 83.21% | 87.20% | 85.16% |
| Montagne | 85.00% | 72.86% | 78.46% |
| Construction Humaine | 60.00% | 85.71% | 70.59% |
| Autrelieu | 78.79% | 54.17% | 64.20% |
| Mer | 89.66% | 91.23% | 90.43% |
| Lac | 88.89% | 72.73% | 80.00% |
| Pays | 79.60% | 95.81% | 86.96% |
| average | **85.12%** | **84.85%** | **84.73%** |

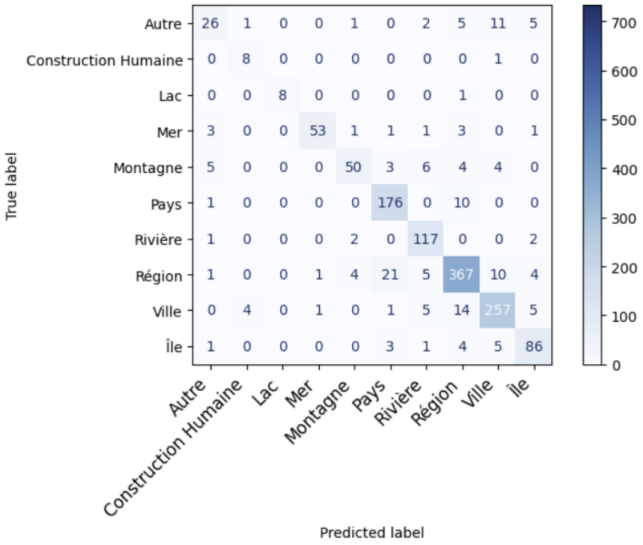


FIGURE 7 – Évaluation sur le modèle de classification de 5-grammes (à gauche) et sa matrice de confusion (à droite)

### 5.4.3 Normalisation des vedettes et des entités nommées

Les formulations d'écriture des vedettes par les auteurs de l'EDdA (e.g articles définis, tirets, virgules, "ou", préfixes, alias) étant très variées, une étape de normalisation est importante et nécessaire, notamment pour l'étape de matching entre vedettes et entités nommées identifiées dans le texte. Pour faciliter le matching qui consiste à vérifier si une entité nommée identifiée dans le texte existe déjà comme vedette, nous avons choisi de normaliser les vedettes en les passant en minuscules et en enlevant les accents, les tirets. En cas de plusieurs nominations, notre principe est de choisir un nom prioritaire comme "pref-label" et les autres sont considérés comme des alias.

La figure 8 présente quelques exemples de normalisation des vedettes.

| **Head** | **pref-label** | **alias1** | **alias2** | **alias3** |
|---|---|---|---|---|
| ABUYO ou ABUYA | abuyo | abuya | | |
| AIGNAN (Saint) | saint aignan | st. aignant | s. aigant | |
| AMIRANTE isles de l' | amirante | isles de l'amirante | | |
| Andalousie (la nouvelle) | nouvelle andalousie | | | |
| Marthe,Sainte, ou Sierra Néveda | sainte marthe | st. marthe | s. marthe | sierra neveda |

FIGURE 8 – Exemples de normalisation des vedettes

### 5.4.4 Entity Matching

Cette section a pour objectif d'associer les entités nommées extraites du texte aux vedettes existantes. Chaque vedette étant identifiée par une URI unique (par exemple, "E1", "E2"), toutes les entités nommées dans le texte seront comparées aux vedettes en fonction de leur nom et de leur type. Si aucune correspondance n'est trouvée, l'entité nommée est associée à une nouvelle URI et ajoutée comme nouveau nœud dans le graphe.

La figure 9 illustre le processus algorithmique de matching pour toute entité nom-

mée [22]. Un matching est valide uniquement si le nom et le type correspondent avec ceux d'une vedette. Toutefois, certains cas ambigus peuvent survenir lorsque plusieurs vedettes partagent le même nom et le même type (par exemple : la ville de Vienne en France et celle en Autriche). Dans de telles situations, l'entité nommée se verra attribuer une URI temporaire (comme "Ambigue1", "Ambigue2") afin de signaler cette ambiguïté.

Afin de pallier les erreurs issues de l'OCR (par exemple, des variantes comme "Italie" vs "Itali") et les alias d'entités nommées (comme "calix" et "calis"), nous avons intégré une étape de similarity-matching au cas où une entité n'aurait pas de correspondance stricte.

En suivant les résultats de l'étude de différentes mesures de similarité textuelle appliquées à la tâche de mise en correspondance de titres faite par [3], nous avons retenu les mesures de Jaro-Winkler [23] et Levenshtein [24], en définissant un seuil de similarité permettant de valider une correspondance lors de l'étape de similarité-match.

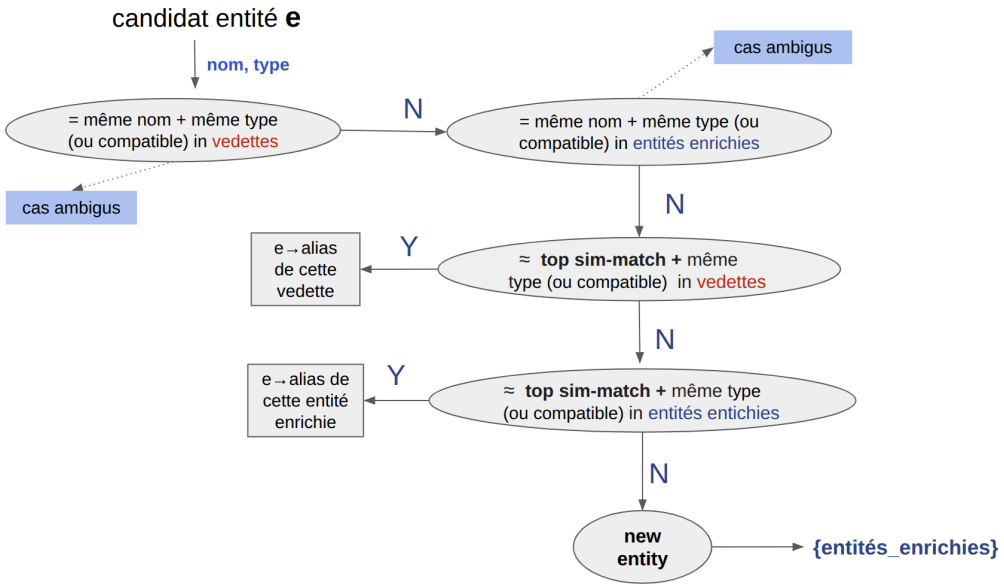


FIGURE 9 – Algorithme de matching entre entités nommées et vedettes

Dans le processus de matching, nous avons introduit des groupes de types compatibles (par exemple : "Pays, Région, Ville", ou "Mer, Lac, Rivière, Autre") afin d'atténuer l'impact des erreurs de prédiction du type des entités du modèle entraîné dans la section 5.4.2. Cette stratégie permet de tolérer certaines confusions typo-

22. https://gitlab.liris.cnrs.fr/lmoncla/stage-bin/-/tree/main/Entity_matching?ref_type=heads
23. https://fr.wikipedia.org/wiki/Distance_de_Jaro-Winkler
24. https://fr.wikipedia.org/wiki/Distance_de_Levenshtein

logiques sans multiplier inutilement les entités distinctes, tout en régularisant les correspondances entre entités proches sur le plan sémantique ou géographique.

La figure 11 présente les résultats de matching de tous les 87,500 spans "NP-Spatial" de l'EDdA par type de matching. Nous constatons que la majorité des entités correspond à un matching strict avec les vedettes (54033 occurrences, 5946 entités uniques), ce qui montre que beaucoup d'entités sont déjà présentes comme vedettes. Un nombre plus faible d'entités est identifié via le matching strict avec les entités enrichies (9310 occurrences, 2473 uniques), tandis que la recherche par similarité permet de récupérer un petit nombre de correspondances supplémentaires, indiquant la présence de variantes ou d'erreurs mineures dans le texte.

Les entités classées comme ambigues (3505 occurrences, 395 uniques) représentent des cas où l'algorithme a détecté plusieurs correspondances possibles (exemple de la ville de "Vienne"), illustrant les limites de désambiguïsation en plusieurs correspondances. Enfin, les entités sans correspondance (16594 occurrences, 15412 uniques) correspondent à de nouvelles entités, qui devront être ajoutées comme nouveaux nœuds dans le graphe de connaissances. Ces résultats montrent que l'algorithme est capable d'identifier efficacement les entités existantes tout en signalant les nouveautés ou les cas ambigus, préparant ainsi la construction d'un graphe enrichi et cohérent.

| match-case | occurences totales | nombre unique (déduplication) |
|---|---|---|
| strict match (vedette) | 54033 | 5946 |
| strict match (entité enrichie) | 9310 | 2473 |
| similarity match (vedette) | 3096 | 2558 |
| similarity match (entité enrichie) | 962 | 962 |
| ambigu | 3505 | 395 |
| no match (nouvelle entité) | 16594 | 15412 |
| total | 87500 | 27449 |

TABLE 11 – Evaluation sur les modèles de différentes tailles de contexte

## 5.5 Détection des relations spatiales

Cette section vise à détecter les relations spatiales entre les entités nommées identifiées et classifiées dans la section précédente. Ce processus s'appuie sur les spans de type "Relation" reconnus par le modèle "camembert-base-edda-span-classification" et se déroule en deux étapes : classification selon le type de relation (5.5.1) et Relation-Entity linking (5.5.2).

### 5.5.1 Classification selon le type de relation

Nous avons fait annoter tous les spans "Relation" dans les 2000 articles par le modèle gpt-4.1-mini selon leur type, avec la méthode de few-shot prompting. Le LLM est guidé à prédire la catégorie de relation parmi celles dans l'ontologie de peovenance, étant donnés le span de relation (souvent des prépositions) ainsi que le span suivant (NP-spatial ou NC-spatial).

Avec un contrôle sur la qualité des annotations générées en corrigeant des erreurs, nous avons fine-tuned le modèle "bert-base-multilingual-cased" à partir de ces annotations.

Concernant les catégories de relations, nous avons choisi de regrouper "Distance" et "Orientation" en **"Distance-Orientation"** car il existe souvent des cas binaires comme "trois lieues au nord de Lyon", "environs deux lieues au midi d'Amiens". Dans la suite, nous avons développé un algorithme en utilisant des expressions régulières qui permet d'identifier la distance et/ou le cardinal dans une relation classifiée en "Distance-Orientation".

La figure 12 illustre la distribution des types de relations dans les 2000 articles "single" du jeu de données GeoEDdA-TopoRel, tandis que la figure 10 présente les précisions, rappels et F-mesures obtenus par type de relation pour le modèle entraîné, ainsi que la matrice de confusion correspondante. Les résultats montrent que le modèle atteint une excellente performance globale, avec une F-mesure moyenne de 94,69%, et parvient à bien distinguer la grande majorité des relations spatiales. Ce modèle est disponible ici [25] sur Huggingface.

Cependant, certaines confusions apparaissent. Le type "Crosses" est celui qui obtient la précision la plus faible, avec des erreurs de classification récurrentes vers le type "Adjacence". Cette confusion s'explique par la nature parfois ambigue de la relation entre un lieu et une entité linéaire ou de surface (rivière, côte, mer). Ainsi, des exemples annotés en Adjacence tels que « le long des côtes occidentales d'Italie », « le long de la rivière Margus », « sur le golfe de Forth » ou encore « sur le bord d'un lac » ont été prédits à tort comme Crosses.

La matrice de confusion révèle également cinq erreurs entre Adjacence et Distance-Orientation, pour des cas comme « au-delà du Gange », « au-dessous de Bray », « en-deçà du Gange », « au bas du mont Capitolin » et « un peu au-dessous de

25. https://huggingface.co/GEODE/bert-base-multilingual-cased-classification-relation

la ville ». Ces confusions trouvent leur origine dans le jeu d'entraînement, où des relations qui contiennent des expressions de "Adjacence" sont en réalité annotées comme Distance-Orientation car elles comportent explicitement une indication de distance ou d'orientation. Par exemple, des formulations comme « à environ 5 lieues au-dessous de Vienne » ou « à trois lieues au-dessous d'Amiens » ont été annotés en Distance-Orientation. Ainsi, le modèle tend à se tromper pour prédire des expressions comme "au dessus de", "au bas de", "en-deça de".

| | train | validation | test |
|---|---|---|---|
| Inclusion | 1319 | 131 | 156 |
| Distance-Orientation | 1065 | 163 | 115 |
| Adjacence | 498 | 59 | 75 |
| Crosses | 397 | 50 | 29 |
| Mouvement | 184 | 15 | 35 |
| Autre-relation | 195 | 30 | 42 |

TABLE 12 – Distribution des types de relations dans les 2000 articles "single" du GeoEDdA-TopoRel

| | **Précision** | **Rappel** | **F-mesure** |
|---|---|---|---|
| Inclusion | 98.06% | 91.18% | 97.74% |
| Distance-Orientation | 93.44% | 99.13% | 96.20% |
| Adjacence | 95.71% | 85.90% | 90.54% |
| Crosses | 76.45% | 100% | 86.67% |
| Mouvement | 96.88% | 91.18% | 93.94% |
| Autre-relation | 97.43% | 88.37% | 92.68% |
| | **95.09%** | **94.69%** | **94.69%** |

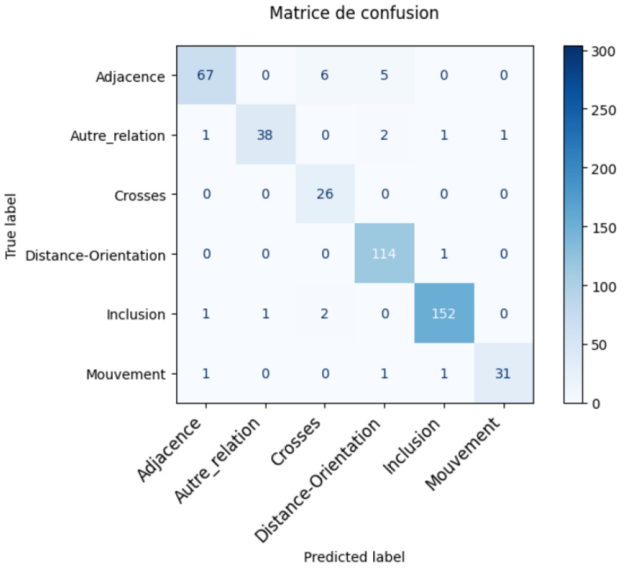


FIGURE 10 – Évaluation du modèle de classification selon le type de relation

### 5.5.2 Relation-Entity linking

Dans cette section, nous cherchons à associer chaque relation détectée à son sujet et son objet afin de former des triplets de type Sujet-Prédicat-Objet. Pour cela, nous avons exploré deux approches : le few-shot prompting et une méthode basée sur la probabilité conditionnelle, calculée à partir de la fréquence d'apparition des triplets.

#### Méthode de few-shot prompting

En fonction des spans séquentiels, un LLM est guidé via un prompt pour identifier le bon sujet et le bon objet à partir des spans précédents et suivants. Deux attributs supplémentaires, 's' (sujet) et 'o' (objet), sont ainsi ajoutés aux spans de relation.

La figure 11 montre l'entrée (texte et tous les spans) et la sortie (spans de relations enrichis) de notre instruction de prompt. Notez que l'entité E1 (que le LLM a considéré comme sujet pour les 2 relations spatiales) se réfère ici à la vedette ILLESCAS.

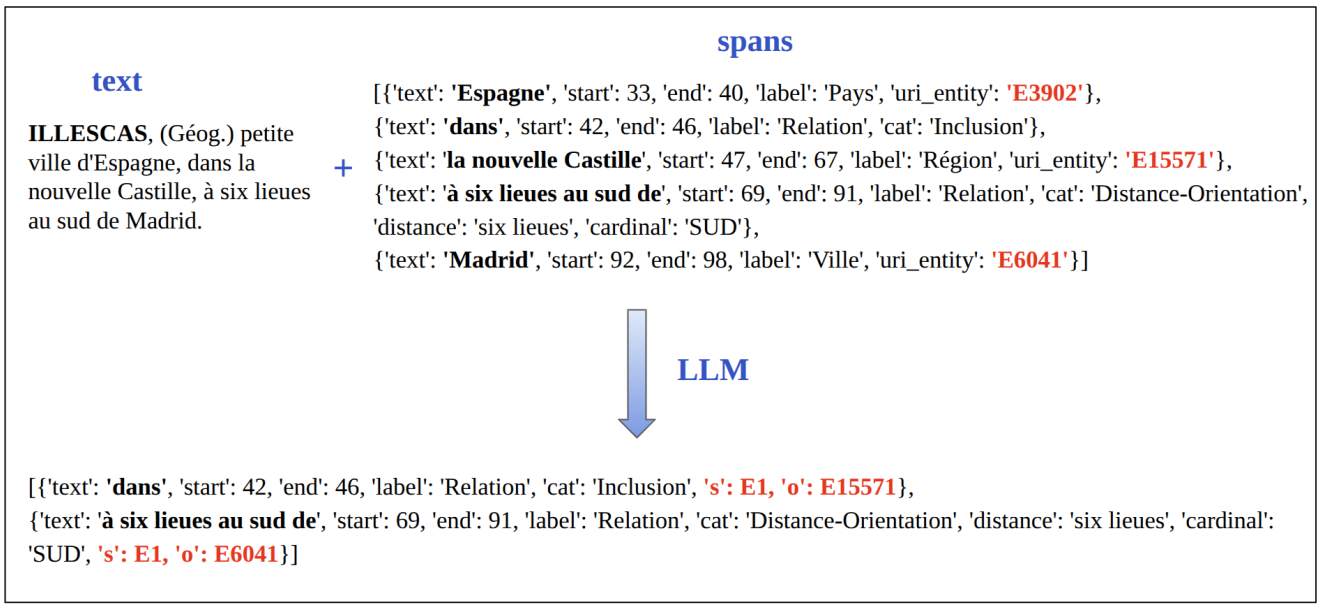


FIGURE 11 – Input et output de la méthode de few-shot prompting

#### Approche probabiliste

Étant donné que les spans sont identifiés de manière séquentielle, tout span de type "Relation" sera associé à son sujet qui correspond soit à la vedette de l'article, soit à la dernière entité nommée précédemment identifiée. Par exemple, dans la figure 12, la relation R2 ("dans") possède deux sujets candidats : E1 ("ILLESCAS") et E2 ("Es-

pagne"), tandis que son objet est clairement identifié comme étant l'entité suivante, E3. En fonction du type de la relation et de celui de l'objet, nous comparons les probabilités associées à chacun des deux sujets candidats en termes de types de triplets formés : (Ville, Inclusion, Région) et (Pays, Inclusion, Région). En effet, certains type de triplets sont peu fréquents voire irréalistes (d'un point de vue géographique).

Pour cela, nous avons sélectionné les 500 premiers articles de l'EDdA dans l'ordre alphabétique (toutes les entités nommées ont déjà leur URI après l'étape de Entity-matching), dans lesquels les spans de type "Relation" ont été associés à leurs sujets et objets par le modèle gpt-4.1-mini. À partir de cet échantillon, nous avons construit un sac de triplets et calculé la fréquence d'apparition de chaque type de triplet. Ce sac va servir de référence de probabilité des différents types de triplet.

La figure 13 présente les 20 types de triplets les plus fréquents dans cette référence de probabilité.

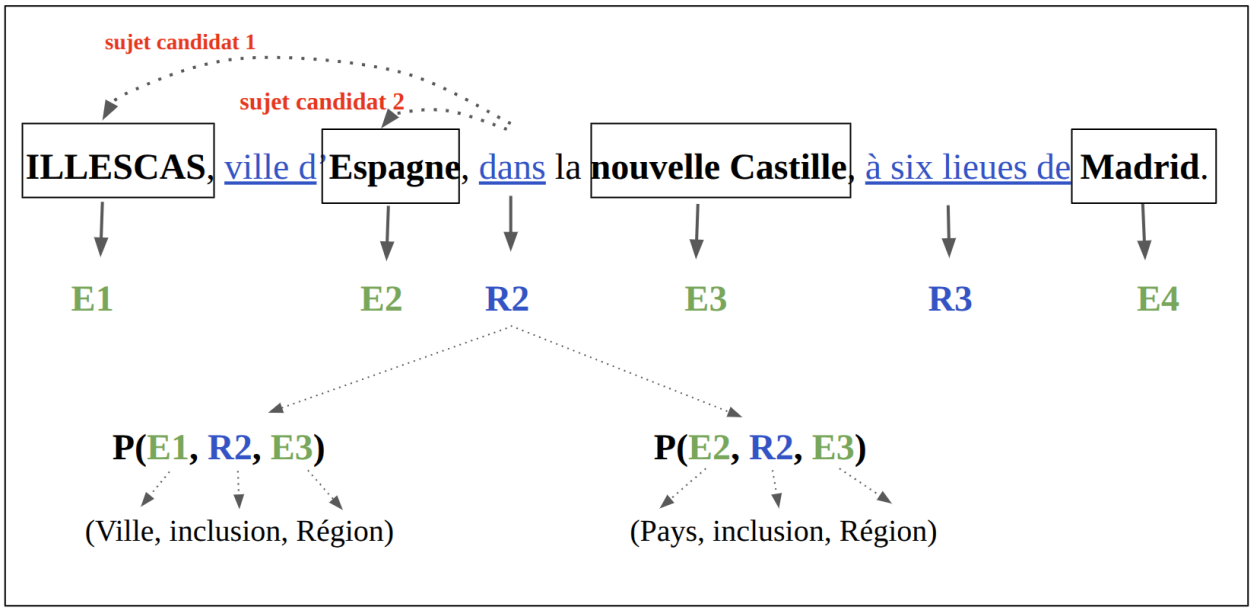


FIGURE 12 – Illustration de l'approche probabiliste

Pour comparer les performances de ces deux méthodes de relation-entity linking, nous avons fait une évaluation sur les spans de "Relation" dans les 100 premiers articles de l'EDdA. Le tableau 13 présente les résultats de l'évaluation de ces deux méthodes sur les 124 spans de "Relations".

Ces deux approches présentent chacune leurs limites. La méthode de few-shot prompting manque de stabilité dans ses sorties : même lorsqu'elle est strictement contrainte

| type_sujet | type_relation | type_objet | count | type_sujet | type_relation | type_objet | count |
|---|---|---|---|---|---|---|---|
| Ville | Inclusion | Région | 200 | Rivière | Inclusion | Région | 12 |
| Ville | Inclusion | Pays | 145 | Ville | Distance-Orientation | Ville | 12 |
| Ville | Crosses | Rivière | 82 | Rivière | Mouvement | Ville | 11 |
| Ville | Inclusion | Ville | 33 | Ville | Crosses | Mer | 11 |
| Région | Distance-Orientation | Région | 23 | Pays | Inclusion | Pays | 10 |
| Ville | Adjacence | Rivière | 20 | Ville | Adjacence | Région | 8 |
| Rivière | Mouvement | Rivière | 18 | Région | Inclusion | Pays | 7 |
| Région | Inclusion | Région | 17 | Ville | Adjacence | Mer | 7 |
| Rivière | Mouvement | Région | 16 | Île | Inclusion | Mer | 7 |
| Ville | Adjacence | Ville | 14 | Ville | Adjacence | Pays | 7 |

FIGURE 13 – Les 20 types de triplets les plus fréquents de l'échantillon de 500 articles

à ne renvoyer que l'URI du sujet et de l'objet, elle ne respecte pas toujours cette consigne. De son côté, l'approche probabiliste repose entièrement sur la qualité de la séquence des spans ; si le span suivant celui d'une relation n'est pas un bon candidat pour un syntagme nominal spatial (NP-spatial), cela conduit à une erreur de chaînage.

Par exemple, dans la phrase : « ABANA, rivière de Syrie qui se jette dans la mer de ce nom, après avoir arrosé les murs de Damas du côté du Midi. »

L'expression « la mer de ce nom » n'étant pas reconnue comme un NP-spatial, les LLMs ont pris par erreur le sujet initial (ABANA) comme objet de la relation « se jette dans », produisant ainsi un triplet incorrect (ABANA, Mouvement, ABANA). À l'inverse, l'approche probabiliste a retenu « Damas », annoté comme NP-spatial, comme objet (incorrect) de cette même relation.

| | nb_correct | nb_total | précision |
|---|---|---|---|
| llama3 :70b | 106 | 124 | 85.48% |
| gpt-4-turbo | 113 | 124 | 91.12% |
| gpt-4.1-mini | 113 | 124 | 91.12% |
| approche probabiliste | 115 | 124 | 92.74% |

TABLE 13 – Résultats de l'évaluation du prompt engineering et de l'approche probabiliste pour le relation-linking

## 5.6 Construction du graphe en RDF

Les relations associées aux entités via leurs URI sont intégrées dans un graphe construit à l'aide de la bibliothèque rdflib [26], en s'appuyant sur l'ontologie spatiale ainsi que sur l'ontologie de provenance. Plus spécifiquement, pour traiter des expressions introductives au début de l'article telles que « ville de », « rivière de » ou « duché de », qui décrivent la relation d'inclusion entre la vedette de l'article et sa région ou son pays, nous avons ajouté un algorithme dédié.

Celui-ci permet de créer une relation de type « Inclusion » lorsque le premier NP-spatial identifié correspond à un pays ou à une région.

Dans le graphe produit, chaque triplet (sujet, prédicat, objet) est représenté sous forme de statement, lequel est relié à l'article dont il est issu. La figure 14 illustre, par exemple, le statement2 de type de relation de mouvement correspondant au triplet (E1, Mouvement, E15454). Ce statement2 de type "Mouvement" a son sujet "E1" dont le nom est "A" et le type est "Rivière", son objet "E15454" dont le nom est "Fontaines" et le type est "Ville". La figure 15 illustre le statement20 de type de relation d'orientation correspondant au triplet (E53, Orientation, E15486). En particulier, ce statement a son prédicat "Orientation20" qui est un noeud intermédiaire implémenté car il introduit l'attribut de cardinal de cette relation : "EST".

26. `https://rdflib.readthedocs.io`

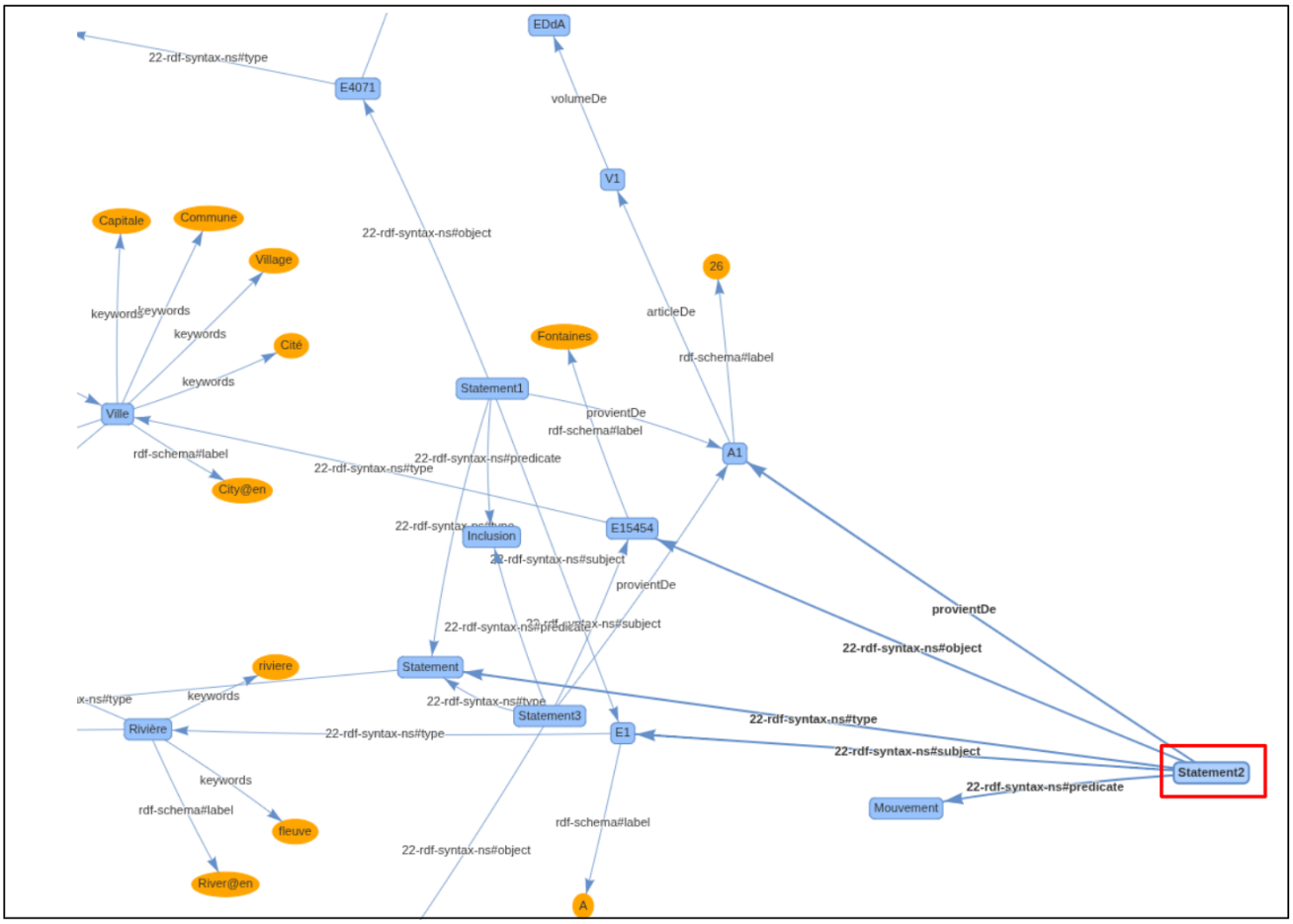


FIGURE 14 – L'illustration du statement 2 de type de relation de mouvement

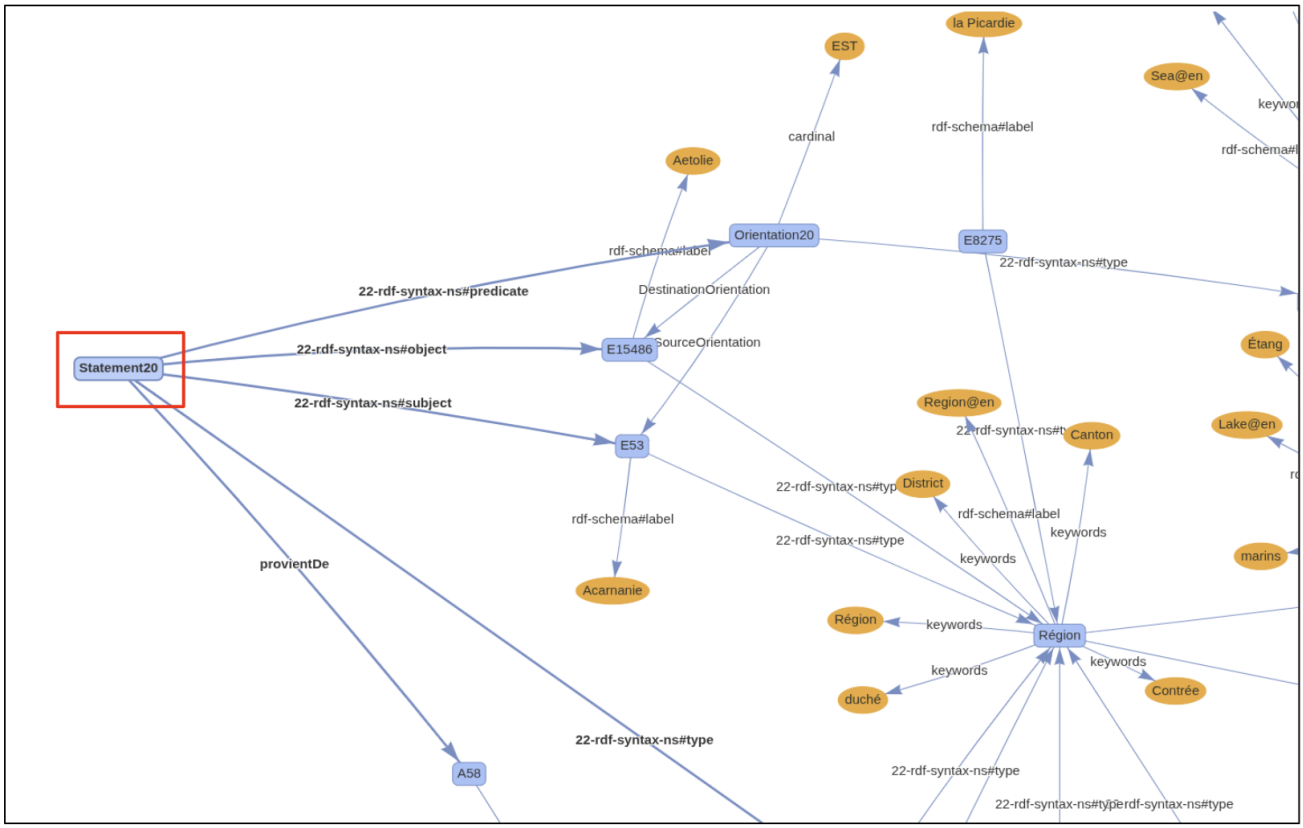


FIGURE 15 – L'illustration du statement20 de type de relation d'orientation

# Chapitre 6 - Évaluation du graphe

# 6 Évaluation du graphe

Cette section propose une double évaluation du graphe généré, à la fois sur les plans quantitatif et qualitatif, afin de mesurer la pertinence et la couverture des connaissances extraites à partir des textes sources.

## 6.1 Analyse quantitative

Dans un premier temps, nous avons procédé à une analyse statistique du graphe afin de mieux comprendre sa structure et les informations qu'il contient. Nous avons notamment comptabilisé le nombre d'entités nommées identifiées pour chaque type ainsi que le nombre de relations spatiales extraites pour chaque catégorie.

A l'aide des requêtes SPARQL sur notre graphe final (fichier en format ttl), nous avons comptabilisé 15,453 instances appartenant à la classe "Article", 17 instances à "Volume" et 10 instances à "Place" ; ces nombres sont conformes à la base de données de l'EDdA. De plus, nous avons 46584 statements, c'est-à-dire 46584 relations spatiales et 32047 URIs d'entité dans le graphe final.

Concernant la répartition des entités nommées et des relations, la figure 16 illustre la répartition des types d'entités nommées présentes dans le graphe, mettant en évidence les classes dominantes et les éventuels déséquilibres. On observe une prédominance marquée des entités de type Ville, représentant plus de 52% du total (17,159 occurrences), suivies des Régions (5,167 occurrences) et des Rivières (4,930 occurrences). Ces trois catégories couvrent à elles seules près de 80% des entités détectées, soulignant l'importance accordée aux divisions administratives et aux éléments hydrographiques majeurs dans les textes analysés. À l'inverse, les entités telles que Lac, Mer, ou Île sont moins fréquentes, bien qu'elles participent également à la richesse géographique du corpus.

De même, la figure 17 présente la distribution des types de relations spatiales détectées, ce qui permet d'observer les relations les plus fréquentes ainsi que celles plus marginales. La relation d'Inclusion domine largement avec 24 629 occurrences, soit près de 55% du total. Cela traduit une forte structuration hiérarchique de l'espace dans les textes, où les entités sont souvent situées à l'intérieur d'autres (par exemple, une ville dans une région ou un pays). Les relations d'Orientation (5,329) et de Distance (4,640) apparaissent également fréquemment, traduisant des descriptions plus

relatives de la localisation. Les relations comme Crosses, Adjacence ou Mouvement, bien que moins présentes, montrent une diversité dans les types d'interactions spatiales prises en compte, renforçant l'intérêt sémantique du graphe de connaissances construit.

Par ailleurs, nous avons comptabilisé un total de 3,505 nœuds dont l'URI est marqué comme Ambigu, reflétant les cas où l'attribution d'une entité unique n'a pas pu être résolue de manière certaine. Ce nombre correspond aux résultats obtenus lors de la phase de Entity Matching (section 5.4.4), où ces entités ont été temporairement annotées avec un identifiant ambigu faute de correspondance fiable dans la base de référence.

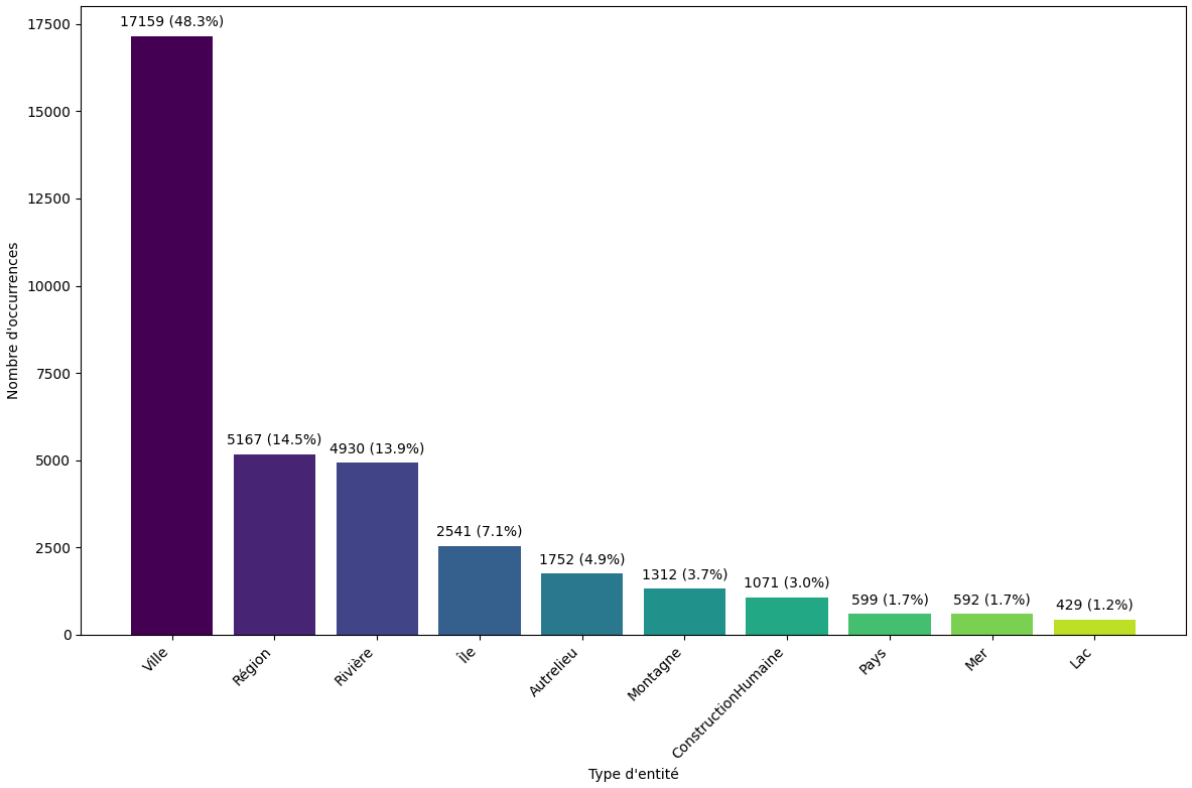


FIGURE 16 – Répartition des types d'entités géographiques de l'EDdA

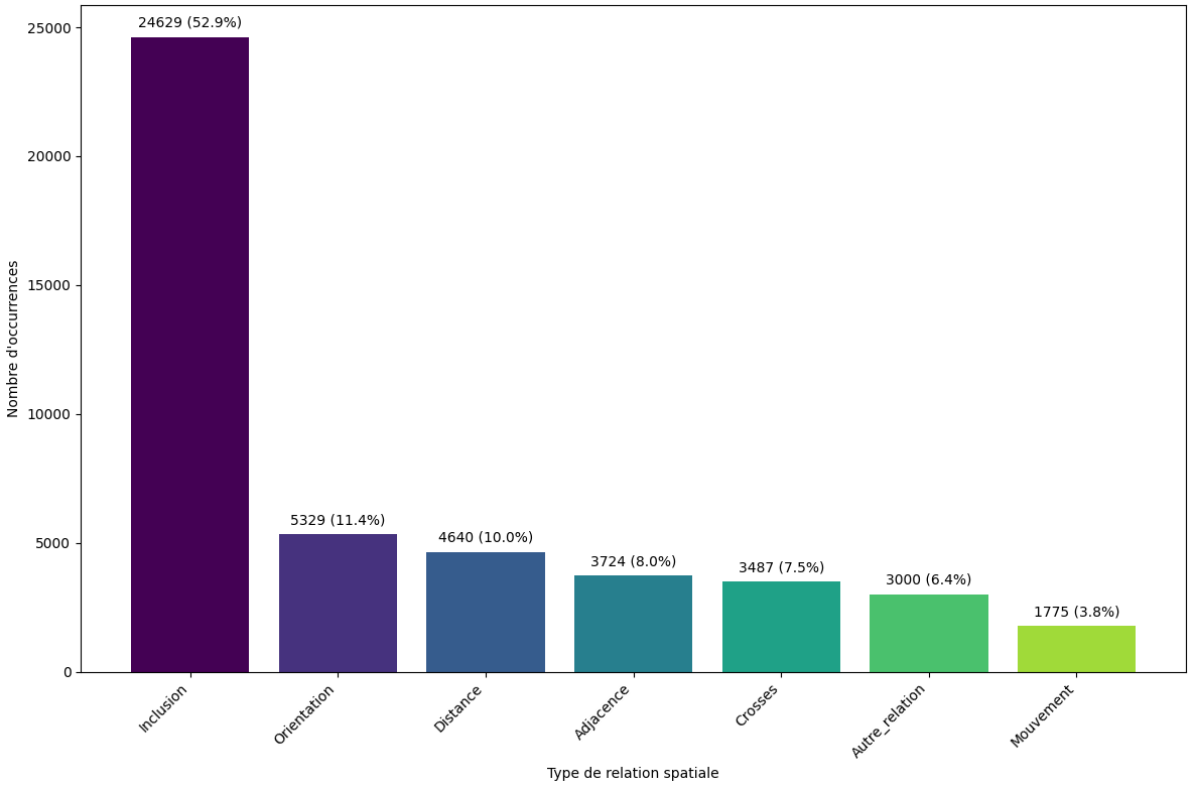


FIGURE 17 – Répartition des types de relations de l'EDdA

## 6.2 Analyse qualitative

Dans un second temps, nous avons procédé à une évaluation qualitative de la pertinence des relations au sein du graphe. Avant tout, un indicateur analysé est le nombre de nœuds isolés, c'est-à-dire des entités qui ne possèdent aucune relation topologique ou directionnelle avec d'autres nœuds. Au total, 11,428 entités, soit environ 30% de l'ensemble des nœuds, se retrouvent dans cette situation. Ce phénomène peut s'expliquer par plusieurs facteurs : tout d'abord, des erreurs dans la reconnaissance des entités nommées peuvent conduire à des entités mal identifiées ou non reconnues, empêchant ainsi leur mise en relation. Ensuite, certaines relations spatiales peuvent ne pas avoir été extraites, soit en raison de limites dans le modèle d'extraction, soit à cause de variations linguistiques complexes dans les textes. Enfin, une explication possible est que ces entités se trouvent dans des paragraphes qui consistent plutôt à décrire les contextes historiques et où il n'y a pas de relations spatiales associées. Notre approche consistant à traiter l'intégralité des textes, certains articles assez longs sont peu pertinents pour notre objectif et peuvent introduire du bruit : des mentions non pertinentes ou ambigues peuvent contribuer à l'isolement de certaines

entités. Ces résultats soulignent l'importance d'un pré-traitement rigoureux et d'un raffinement des règles d'extraction pour améliorer la connectivité du graphe et la fiabilité des relations spatiales modélisées.

La figure 18 présente les fréquences de type de relations entre toutes paires de classes d'entités. On observe une forte prédominance des relations entre Ville et Ville, avec 9594 occurrences et Ville et Région, avec 8757 occurrences ou encore Ville et Pays, avec 6742 occurrences. Cela traduit une tendance marquée des textes à situer les entités locales dans un cadre administratif plus large, révélant ainsi la hiérarchisation spatiale classique des descriptions géographiques. Ensuite, un bon nombre de relations entre Rivière et Ville, Rivière et Région ou encore Rivière et Pays soulignent le rôle important et central des cours d'eau dans les localisations administratives. La traversée d'une rivière ou la proximité d'un cours d'eau constitue un repère spatial significatif dans les descriptions anciennes, souvent liées à des enjeux économiques ou stratégiques.

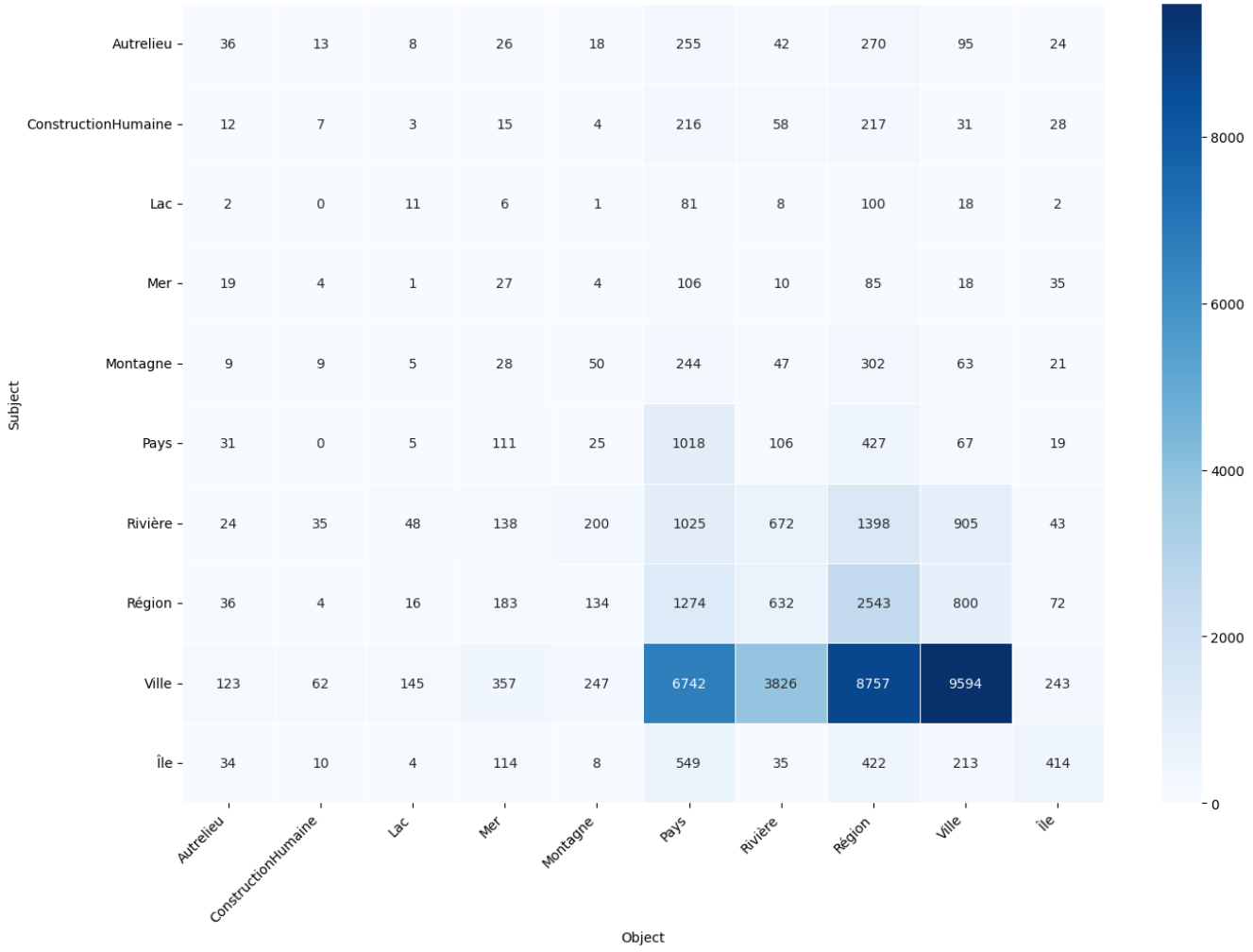


FIGURE 18 – Matrice des relations entre différents types d'entités

Malgré une extraction en grande partie cohérente, nous pouvons constater certaines anomalies sémantiques en analysant l'adéquation entre relations et entités. La figure 19 illustre la distribution des types de relations selon le type d'entité du sujet. Cette figure confirme la prédominance de la relation d'inclusion dans tous les cas. La position secondaire de la relation "Mouvement" dans la distribution de la classe "Rivière" s'explique naturellement par les propriétés intrinsèques d'une rivière. Néanmoins, il y a quelques incohérences entre les relations et les sujets d'entités. Par exemple, des triplets comme (Pays, Mouvement, Région) ou (Région, Mouvement, Ville) sont sémantiquement aberrants : un pays, en tant qu'entité géopolitique statique, ne peut pas être l'agent d'un mouvement spatial. De telles erreurs indiquent que certaines relations ont été appliquées sans validation des contraintes de type, c'est-à-dire sans vérifier que la relation est sémantiquement plausible entre les catégories d'entités concernées.

Ces incohérences proviennent soit d'une prédiction incorrecte des types d'entités, soit d'une mauvaise identification des types de relations (par exemple à cause d'une formulation spéciale ou métaphorique), soit encore d'un lien erroné établi entre une relation et une entité. Cette observation confirme l'absence d'un contrôle typologique lors de l'association des entités avec les relations.

Ainsi, il serait pertinent d'envisager l'introduction d'un mécanisme de validation ontologique a-posteriori, basé sur des règles de compatibilité entre types d'entités et relations (par exemple, seules des entités mobiles comme des rivières ou lacs peuvent être sujettes à une relation de Mouvement).

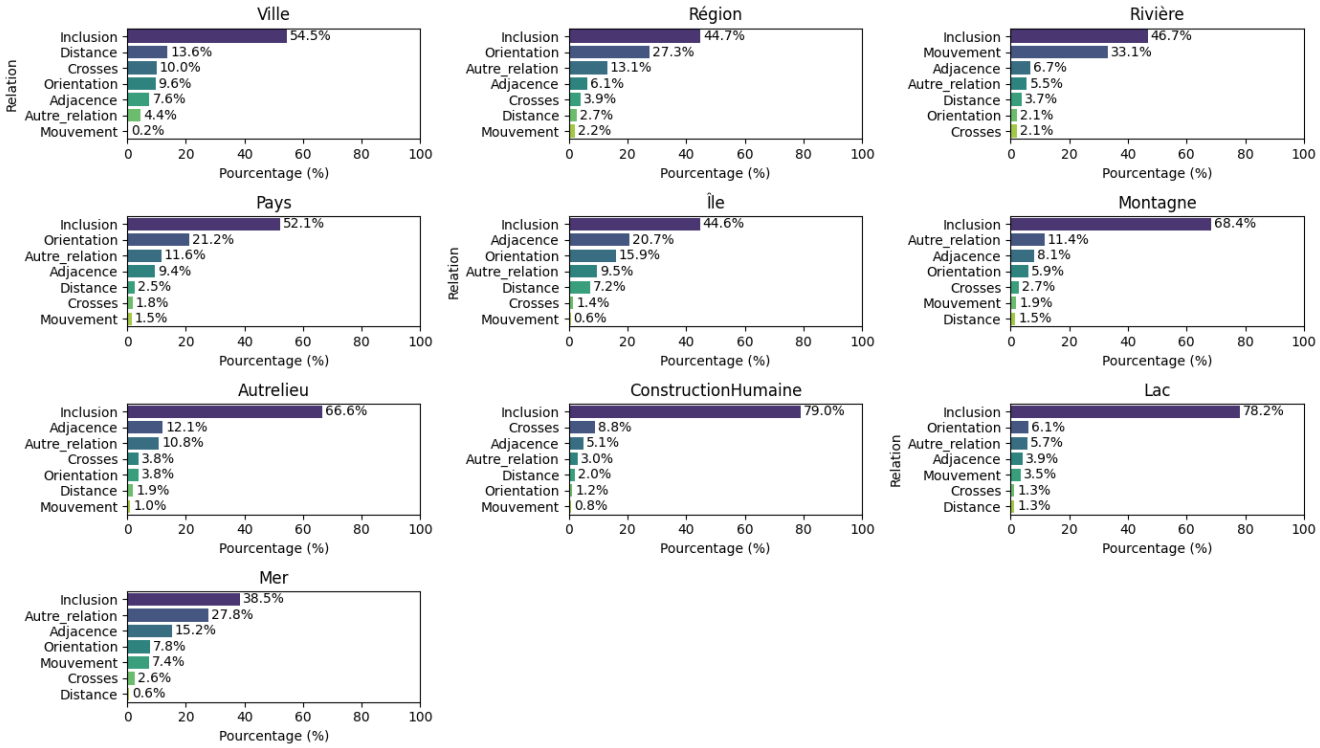


FIGURE 19 – Distribution des types de relations par sujet entité

# Chapitre 7 - Conclusion

# 7 Conclusion

## 7.1 Conclusion générale

Ce mémoire a présenté une approche complète et innovante pour la construction automatique d'un graphe de connaissances géo-historiques à partir des articles de l'EDdA. Cette démarche s'est structurée autour de trois phases majeures, qui combinent modélisation, extraction automatique et validation des résultats, offrant ainsi un cadre méthodologique rigoureux pour le traitement de corpus historiques complexes. L'ensemble du pipeline, accompagné des notebooks d'entraînement et scripts utilisés, est décrit en détail et disponible dans le fichier README.md [27].

Dans un premier temps, la modélisation du graphe, avec deux ontologies dédiées, a permis de formaliser la représentation des connaissances géo-historiques sous une forme interopérable et conforme aux standards du web sémantique, notamment à travers l'adoption du standard RDF. Cette étape a été cruciale pour garantir la compatibilité du graphe avec d'autres ressources ouvertes et faciliter sa réutilisation dans différents contextes scientifiques.

La phase suivante, centrée sur le peuplement automatique du graphe, a mobilisé des techniques avancées de traitement automatique du langage naturel. En particulier, nous avons combiné des approches de reconnaissance d'entités nommées, d'extraction de relations spatiales, ainsi que des méthodes de classification supervisée et de fine-tuning de modèles de langage de grande taille. L'utilisation de prompt engineering s'est révélée efficace pour adapter les capacités des LLM à la spécificité des textes anciens, tout en garantissant une extraction précise et contextualisée des informations. Ces techniques, conjuguées à la définition d'un pipeline de traitement, ont permis d'automatiser l'enrichissement du graphe avec des données fiables et structurées.

Par ailleurs, j'ai construit et mis à disposition un jeu de données standardisé nommé datasets GeoEDdA-TopoRel, qui constitue une ressource réutilisable et ouvre la voie à des travaux futurs en linguistique computationnelle, en histoire numérique et en sciences de l'information géographique.

Enfin, le développement de plusieurs modèles de classification dédiés – pour identifier le type d'article, le type des entités nommées et le type de relation spatiale – illustre la richesse et la modularité de notre pipeline. Cette architecture permet d'adapter

27. `https://gitlab.liris.cnrs.fr/lmoncla/stage-bin`

facilement l'outil à d'autres corpus ou d'intégrer des modules complémentaires pour approfondir l'analyse sémantique.

En conclusion, notre travail propose un cadre méthodologique et technique robuste pour la structuration et l'analyse des savoirs géographiques historiques en construisant des graphes de connaissances. Il constitue ainsi un pont entre les humanités numériques, l'histoire et le TAL, en offrant aux chercheurs un outil performant pour explorer, interroger et valoriser les données extraites. Nous espérons que ce cadre facilitera le développement de nouveaux projets interdisciplinaires, contribuant à enrichir la compréhension du patrimoine culturel et scientifique ancien.

Ce stage m'a permis d'acquérir de nouvelles compétences à la croisée du traitement automatique des langues et du web sémantique. J'ai approfondi mes connaissances sur les standards RDF et les principes de construction d'un graphe de connaissances, tout en mettant en pratique mes acquis en apprentissage automatique à travers l'entraînement de modèles de classification et de reconnaissance d'entités nommées. Cette double dimension, à la fois théorique et appliquée, m'a donné une vision plus complète de la chaîne allant de l'extraction automatique de l'information à sa structuration dans un format exploitable.

Les principales difficultés rencontrées ont concerné la gestion de données textuelles complexes et la diversité des relations spatiales à modéliser, qui nécessitent à la fois des choix méthodologiques et une réflexion sur la représentation des connaissances. Ces défis m'ont poussé à explorer différentes approches, à tester plusieurs architectures de modèles et à ajuster les pipelines de traitement, renforçant ainsi ma capacité d'adaptation et mon sens critique face aux résultats.

Enfin, ce stage a confirmé mon intérêt pour la recherche appliquée en traitement automatique du langage et en intelligence artificielle. Il s'inscrit pleinement dans mon parcours de master, en lien direct avec mes cours en apprentissage automatique et en génération de texte. Cette expérience m'encourage à poursuivre dans la voie du développement de méthodes intelligentes pour l'analyse et la valorisation de données textuelles, que ce soit dans le cadre d'une future thèse ou d'un poste d'ingénieur de recherche en NLP et web sémantique.

## 7.2 Limites et Perspectives

Bien que notre pipeline constitue une avancée significative pour la structuration automatique des savoirs géographiques issus de textes anciens, plusieurs limites subsistent et ouvrent des perspectives d'amélioration.

Premièrement, le module de reconnaissance des entités nommées présente des lacunes dans le traitement de cas particuliers, notamment les expressions telles que « ville de même nom », « duché de ce même nom » ou « rivière de même nom ». Ces expressions ne sont pas détectées correctement par les modèles actuels. Une piste d'amélioration serait d'y intégrer des règles fondées sur des expressions régulières, capables de capter ces expressions syntaxiques spécifiques.

Deuxièmement, le modèle de classification du type de lieu de 5-grammes pour les entités nommées repose sur une classification globale au niveau du texte. Cette approche pourrait être améliorée en adoptant des stratégies de classification token-level, par exemple, en insérant des tokens spéciaux autour des entités nommées (comme [CLS] ... [SEP]) et en fine-tunant un modèle de langage pré-entraîné sur cette tâche spécifique. Une telle adaptation permettrait probablement une meilleure prise en compte du contexte immédiat de l'entité et donc une amélioration des performances.

Troisièmement, notre pipeline ne prend pas en charge les entités ambigues qui sont rerpésentées avec des URI temporaires (par exemple : Ambigu20). Pour remédier à cette limitation, une piste envisageable consiste à exploiter la structure du graphe lui-même. En analysant les entités voisines directement connectées à une entité ambigue, il est possible de calculer la moyenne des similarités sémantiques vectorielles entre ces voisines et les différentes entités candidates. Cette approche permettrait d'estimer la probabilité de correspondance à une entité spécifique. Par exemple, considérons une entité ambigue de type ville, nommée « Vienne », reliée dans le graphe aux entités « Lyon » et « France ». En calculant les similarités sémantiques entre ces entités voisines, on peut identifier quelle entité de nom « Vienne » est plus proche contextuellement. Cela permet donc de désambiguïser l'entité de manière plus fiable en s'appuyant sur son contexte relationnel dans le graphe.

Quatrièmement, l'architecture actuelle du pipeline repose sur des modules indépendants, nécessitant des étapes manuelles intermédiaires. Une évolution naturelle de notre travail serait de développer une chaîne de traitement unifiée, automatisée de bout en bout, capable de générer directement un graphe RDF complet à partir d'un simple fichier source textuel.

Enfin, notre approche repose actuellement sur un parcours exhaustif de l'ensemble du texte de chaque article pour identifier les entités et les relations spatiales. Cette stratégie, bien que complète, introduit du bruit dans le graphe final. En effet, certaines entités erronées sont extraites (par exemple des mots latins interprétés à tort comme des toponymes), plusieurs nœuds isolés apparaissent sans relations contextuelles, et des relations incorrectes sont inférées. Pour réduire ces erreurs, une stratégie alternative consisterait à se concentrer sur les premières phrases de chaque article, souvent plus structurées et informatives du point de vue géographique. Ce filtrage amont pourrait permettre de renforcer la précision de l'extraction, en privilégiant les segments textuels les plus pertinents, et ainsi améliorer la qualité globale du graphe généré.

Ces axes d'amélioration ouvrent des perspectives de recherche intéressantes, tant sur le plan technique que méthodologique, pour enrichir et généraliser notre approche à d'autres corpus historiques ou encyclopédiques.

# Références


[1] A. Celikyilmaz, E. Clark, and J. Gao. Evaluation of text generation : A survey. *arXiv preprint arXiv :2006.14799*, 2021. Version 2, 18 May 2021.

[2] M. Gaio and L. Moncla. Extended named entity recognition using finite-state transducers : An application to place names. In *Proceedings of GEOProcessing 2017 : The Ninth International Conference on Advanced Geographic Information Systems, Applications, and Services*, pages 15–20, Nice, France, Mar. 2017. IARIA. Published March 19, 2017.

[3] N. Gali, R. Mariescu-Istodor, and P. Fränti. Similarity measures for title matching. In *Proceedings of the 15th IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1260–1266. IEEE, 2015.

[4] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A. Polleres, E. Prud'hommeaux, J. F. Sequeda, and A. Zimmermann. Knowledge graphs. *ACM Computing Surveys*, 54(4) :1–37, 2021.

[5] P. Kordjamshidi, M.-F. Moens, and D. Roth. Learning to interpret spatial language. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1338–1348, 2011.

[6] J. Li, A. Sun, J. Han, and C. Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1) :50–70, 2020.

[7] J. Liu, H. Liu, X. Chen, X. Guo, W. Guo, X. Zhu, and Q. Zhao. The construction of knowledge graph towards multi-source geospatial data. *Unknown Journal or Conference*, 202X. Zhengzhou, China.

[8] D. Maurel, N. Friburger, J.-Y. Antoine, I. Eshkol-Taravella, and D. Nouvel. Cascades de transducteurs autour de la reconnaissance des entités nommées [CasEN : a transducer cascade to recognize French named entities]. *Traitement Automatique des Langues*, 52(1) :69–96, 2011.

[9] L. Moncla, D. Vigier, and K. Mcdonough. Geoedda : A gold standard dataset for geo-semantic annotation of diderot & d'alembert's encyclopédie. In *Second International Workshop on Geographic Information Extraction from Texts (GeoExT) to be held at the 46th European Conference on Information Retrieval (ECIR 2024)*, 2024.

[10] L. Moncla and H. Zeghidi. Token and span classification for entity recognition in french historical encyclopedias. In *Proceedings of the 1st Workshop on*

*Natural Language Processing for Digital Humanities (NLP4DH)*, pages 11–21, Gothenburg, Sweden, 2021. Association for Computational Linguistics.

[11] M. Perry and J. Herring. Ogc geosparql - a geographic query language for rdf data. `https://www.ogc.org/standards/geosparql`, 2011. Open Geospatial Consortium.

[12] H. M. Rawsthorne, N. Abadie, E. Kergosien, C. Duchêne, and É. Saux. Automatic nested spatial entity and spatial relation extraction from text for knowledge graph creation : A baseline approach and a benchmark dataset. In *Proceedings of the 12th International Conference on Geographic Information Science (GIScience 2021)*. Leibniz International Proceedings in Informatics (LIPIcs), 2021.

[13] L. Rietveld and R. Hoekstra. The linked data fragments approach : A low-cost solution for publishing and querying linked data. In *Proceedings of the 14th International Semantic Web Conference (ISWC)*, 2015.

[14] B. Tachev, T. Ferranti, and D. Fensel. A survey on spatio-temporal knowledge graphs. *Semantic Web*, 13(1) :65–99, 2022.

[15] M. Wick, T. Boutreux, and E. Nauer. The geonames geographical database. *Available at http ://www.geonames.org*, 2007.

## Résumé

Ce travail présente un pipeline dédié à la construction d'un graphe de connaissances géo-historiques à partir des articles de l'Encyclopédie de Diderot et d'Alembert("EDdA" dans ce qui suit), une ressource précieuse du XVIIIe siècle. L'enjeu de ce travail est d'extraire, modéliser et structurer les informations géographiques contenues dans ces textes encyclopédiques, et de les représenter sous forme de graphes RDF interrogeables, utilisables dans le cadre de recherches en histoire ou en humanités numériques.

La méthodologie développée repose sur trois grandes étapes : la modélisation du graphe, le peuplement automatique du graphe à partir des textes, puis l'évaluation du graphe généré, tant sur le plan quantitatif que qualitatif.

L'approche technique mobilise plusieurs outils et méthodes : la reconnaissance d'entités nommées (NER) pour identifier les toponymes, l'extraction de relations spatiales (comme l'inclusion, l'orientation ou la distance) à l'aide de techniques de traitement automatique des langues (TAL), ainsi que des méthodes d'apprentissage automatique adaptées aux spécificités linguistiques du français du XVIIIe siècle. Ces composantes permettent de construire un graphe sémantique structuré, intégrant les relations spatiales entre lieux tout en tenant compte du contexte lexical, de la dimension géographique et de la traçabilité des informations extraites.

Ce travail ouvre des perspectives prometteuses pour l'analyse des savoirs géographiques anciens. Il facilite notamment l'étude diachronique des représentations spatiales, le croisement d'informations entre articles, ainsi que l'exploration des évolutions lexicales dans la manière de concevoir et de nommer les lieux au fil du temps.