Construction automatique d'un graphe de connaissances géo-historiques à partir de textes encyclopédiques anciens

Bin Yang*, Ludovic Moncla*
Fabien Duchateau** Frédérique Laforest*

*INSA Lyon, CNRS, Universite Claude Bernard Lyon 1,
LIRIS, UMR5205, 69621 Villeurbanne, France
prenom.nom@insa-lyon.fr
**Universite Claude Bernard Lyon 1, CNRS, INSA Lyon,
LIRIS, UMR5205, 69621 Villeurbanne, France
prenom.nom@univ-lyon1.fr

Résumé. Les encyclopédies anciennes, comme celle de Diderot et d'Alembert (1751-1772), offrent une ressource précieuse pour étudier l'évolution des savoirs géographiques, mais leur ampleur complique toute analyse manuelle. Cet article présente une méthode automatique de construction d'un graphe de connaissances géo-historiques à partir de ces textes. Nous proposons des ontologies spatiale et de provenance adaptées au corpus et introduisons un *gold standard* de 2 750 articles géographiques. Le pipeline combine apprentissage supervisé et grands modèles de langage pour la classification d'articles, le typage d'entités et l'extraction de relations spatiales. Les performances atteignent F1 = 92% pour les relations et F1 > 97% pour la classification, aboutissant à un graphe RDF de 35 000 entités et 46 000 relations. Ce travail ouvre la voie à l'analyse computationnelle des savoirs géographiques anciens. Données, modèles et code sont disponibles sur HuggingFace [1] et Gitlab [2].

1 Introduction

Les dictionnaires encyclopédiques anciens ont joué un rôle essentiel dans la diffusion des savoirs, notamment au siècle des Lumières. L'Encyclopédie de Diderot et d'Alembert (EDdA, 1751-1772) est emblématique de cette entreprise intellectuelle : elle rassemble environ 74 000 articles couvrant de nombreux domaines, dont la géographie (environ 15 000 articles). Ces textes offrent aujourd'hui une source précieuse pour les historiens et les linguistes, car ils témoignent des représentations du monde et des connaissances géographiques disponibles au XVIIIe siècle. Cependant, exploiter ce corpus est une tâche complexe. La richesse lexicale, la longueur des descriptions et les spécificités linguistiques du français ancien limitent fortement les approches manuelles. Par ailleurs, les modèles de traitement automatique du langage

1. https://huggingface.co/GEODE
2. https://gitlab.liris.cnrs.fr/ecoda/encyclopedia2geokg

(TAL) existants sont principalement entraînés sur des données contemporaines et ne sont pas directement adaptés à ce type de corpus pour une extraction des connaissances fiables.

Au-delà de ces défis techniques et linguistiques, la construction automatique d'un graphe de connaissances géographiques offre une nouvelle manière d'explorer le corpus : elle permet de structurer les informations sous forme de triplets ⟨sujet, prédicat, objet⟩, interrogeables et exploitables dans une perspective comparative ou diachronique. Notre objectif est donc d'automatiser la construction d'un graphe de connaissances géo-historiques à partir des articles géographiques de l'EDdA. Pour ce faire, nous avons conçu une chaîne de traitements complète qui s'appuie à la fois sur des modèles de classification affinés sur un échantillon de données du corpus et sur des grands modèles de langage (LLMs) pour l'extraction et la structuration des informations géographiques.

Les contributions de cet article sont les suivantes :

— La définition d'une ontologie spatiale et d'une ontologie de provenance adaptées aux textes encyclopédiques anciens ;
— La conception d'une chaîne de traitements hybride, associant apprentissage supervisé et classification/extraction *few-shot* par LLMs génératifs, appliquée à la classification d'articles, au typage d'entités et à la détection de relations spatiales ;
— La création et la mise à disposition du jeu de données annoté `GeoEDdA-TopoRel` permettant l'entraînement et l'évaluation des différents modèles de la chaîne de traitement ;
— La construction d'un graphe de connaissances géo-historiques RDF à grande échelle (plus de 35 000 entités et 46 000 relations), première ressource de ce type sur une encyclopédie historique.

La suite de l'article est organisée comme suit : la section 2 présente les travaux connexes. La section 3 décrit la modélisation ontologique du domaine. Nous présentons la méthodologie en section 4. Les jeux de données et les expérimentations sont présentés en section 5, avant de conclure en section 6.

2 État de l'art

Les ontologies ou graphes de connaissances sont des structures formelles destinées à représenter des concepts et des relations qui les unissent. Plusieurs ontologies génériques comme FOAF, 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.

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 (Wick et al., 2007), SWEET (Rietveld et Hoekstra, 2015) 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. Certains travaux ont exploré l'enrichissement de graphes de connaissances avec des dimensions temporelles et spatiales, tels que T-GK (Hogan et al., 2021) et SpatioTemporal RDF (Tachev

et al., 2022)), mais ceux-ci restent souvent centrés sur des données contemporaines et bien structurées.

L'extraction d'information est une étape essentielle afin de peupler automatiquement ces ontologies à partir d'informations extraites de textes non ou faiblement structurés. Les principales tâches incluent la reconnaissance et la classification d'entités nommées et l'extraction de relations entre ces entités (Li et al., 2020). Les approches classiques reposant sur des méthodes symboliques (e.g., expressions régulières, grammaires d'extraction) ont été progressivement supplantées par des méthodes statistiques puis neuronales. Les architectures basées sur les réseaux de neurones profonds et les modèles pré-entrainés (notamment BERT) et ses variantes multilingues ont permis des avancées significatives, notamment en contexte multilingue ou en faible supervision. Brenon et al. (2022) et Joliveau et al. (2024) se sont intéressés à un corpus encyclopédique en français avec l'objectif d'étudier les évolutions survenues dans les discours géographiques. Ce travail s'appuie sur des étapes d'extraction d'information (identification des articles géographiques, extraction et classification des entités nommées, des relations spatiales et des coordonnées géographiques). Les informations extraites sont utilisées pour la désambiguïsation des toponymes et la génération de cartes sans aller jusqu'à la construction d'un graphe de connaissances. Rawsthorne et al. (2021) proposent une approche allant jusqu'à la construction d'un graphe de connaissances géographiques à partir de l'identification d'entités imbriquées et de relations spatiales binaires appliquées à un corpus d'instructions nautiques pour la navigation côtière. Cette approche s'appuie sur l'ontologie ATLANTIS spécialement développée pour le domaine maritime et côtier et sur le fine-tuning de modèles BERT pour l'extraction des entités et des relations. Plus récemment, l'IA générative a introduit une nouvelle façon d'aborder l'extraction d'information, via l'utilisation dite *zero-shot* ou *few-shot* de LLMs pré-entraînés, fondées sur le *prompt engineering* (Moncla et Zeghidi, 2025).

En adaptant ces travaux, nous proposons une méthodologie combinant apprentissage supervisé avec *fine-tuning* de modèles encodeurs et *few-shot prompting* de LLMs génératifs afin de modéliser et peupler un graphe de connaissances pour l'EDdA.

3 Modélisation du graphe

Une encyclopédie se découpe en différents volumes (e.g., 17 volumes de texte pour EDdA), dans lesquels sont rédigés des articles. Chaque article possède une vedette (ou titre), et s'accompagne parfois d'une marque de domaine – qui n'est pas uniformisée (e.g., *Géog*, *Géog. anc.*, ou *Géogr.*). Les premiers mots d'un article de géographie précisent généralement le type de lieu (e.g., *petite ville de France* pour l'article sur Saint Jean de Luz [3]). Le reste de l'article peut inclure une description du lieu, des coordonnées géographiques, et surtout des relations par rapport à d'autres lieux (e.g., de distance, d'orientation).

Nous avons défini deux ontologies (provenance et spatiale) pour décrire la structure du graphe permettant de représenter les articles géographiques de l'EDdA. Ces ontologies sont disponibles [4] en RDF (Turtle) et en version graphique (HTML). Elles utilisent le préfixe `ekg` correspondant à l'URI `http://encyclokg.geo/`.

3. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/8/2425
4. https://gitlab.liris.cnrs.fr/ecoda/encyclopedia2geokg/-/tree/main/ontologies

3.1 Ontologie de provenance

L'ontologie de provenance a pour objectif de lier les informations extraites du texte à l'article, au volume et à l'encyclopédie d'où elles proviennent. Pour cela nous proposons un ensemble de classes et de prédicats adaptés à la structure de notre corpus :

— Classes : `Encyclopedy`, `Volume`, `Article` ;

— Prédicats : `articleOf`, `volumeOf`, `articleNumber`, `order`, `extractedFrom`.

Chaque information extraite (matérialisée par un triplet) est représentée sous forme d'un `rdf:Statement` (procédé de réification), qui est rattaché à son article source par le prédicat `extractedFrom`.

3.2 Ontologie spatiale

L'ontologie spatiale définit et structure les concepts liés à l'information géographique tels que les lieux ou entités géographiques et les relations spatiales. Une classe principale `Place` permet de représenter les lieux avec leurs caractéristiques (e.g., coordonnées géographiques, dimensions). De plus, elle se spécialise selon une typologie de lieux établie à partir des types les plus présents dans l'encyclopédie.

— Sous-classes : `Country`, `City`, `Region`, `Sea`, `River`, `Lake`, `Mountain`, `Island`, `HumanMade`, `Other` ;

— Prédicats : `latitude`, `longitude`, `surface`, `length`.

Les relations spatiales entre entités géographiques sont inspirées de celles de DE-9IM (Mark et Egenhofer, 1998) et se limitent aux prédicats suivants :

— `inclusion` (e.g., dans le duché, en Allemagne, au royaume de France) ;

— `adjacency`, qui dénote une forte proximité entre deux lieux (e.g., à côté de, proche de, sur la côte de) ;

— `orientation`, qui précise en plus un cardinal grâce au prédicat `rdf:value` (e.g., au sud de, au couchant de, au midi) ;

— `distance`, qui spécifie généralement une valeur et une unité[5] (e.g., à deux lieues de, à cinq milles de, à deux parasanges de). Actuellement les valeurs sont stockées sous forme de chaine, mais des solutions existent pour affiner (Lefrançois et Zimmermann, 2016) ;

— `movement` (e.g., se jette dans, prend sa source de, coule dans la mer) ;

— `crosses` (e.g., se situe sur la rivière, traverse la ville) ;

— `other` (e.g., entre le Liban et l'Antiliban).

La figure 1 illustre la représentation du triplet (Lyon, orientation, Alpes) dans un graphe, au moyen du *statement* S1. Ce dernier précise également le cardinal (*West*), ainsi que sa provenance (article A1, issu du volume V1 de l'encyclopédie EDdA).

4 Construction du graphe de connaissances

La chaîne de traitements proposée pour la construction automatique du graphe de connaissances (voir figure 2) se décompose en 5 principaux modules : le pré-traitement (étapes 1 à 3), la classification des articles décrivant un lieu (étape 4), le repérage des entités géo-sémantiques

5. Ontologie des unités de mesures, `http://www.ontology-of-units-of-measure.org/`

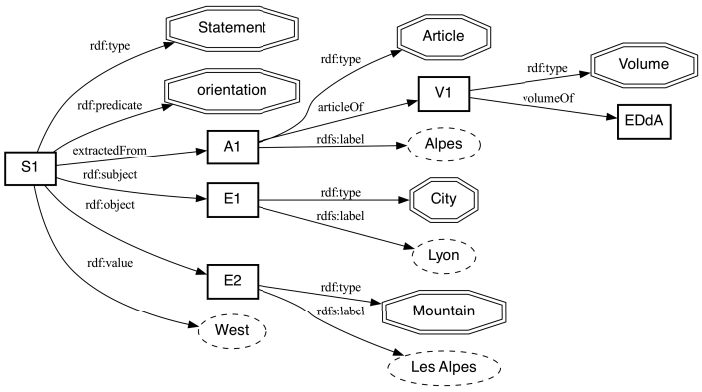

FIG. 1 – *Exemple (simplifié) d'une relation cardinale entre Lyon et les Alpes. Les classes des ontologies sont représentées avec un double octogone, les ressources instances sous forme de rectangle et les littéraux par un cercle en pointillés.*

(parmi lesquelles les entités nommées de lieux, les relations spatiales et les coordonnées géographiques) (étape 5), la classification des entités nommées de lieux (étape 6) et l'extraction des relations spatiales (étape 7). Chacun de ces modules génère des triplets qui viennent peupler le graphe au fur à mesure. Les LLM génératifs ne sont utilisés que lors préparation des données, mais ne sont plus nécessaires ensuite (sauf pour l'étape de segmentation).

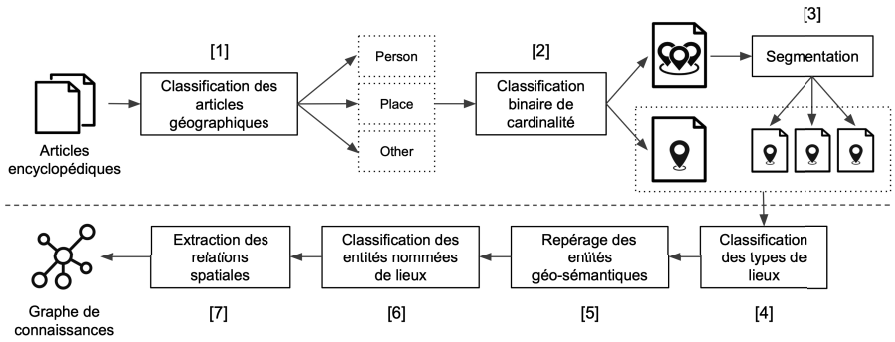

FIG. 2 – *Schéma de la chaîne de traitement de construction automatique du graphe de connaissances. Les numéros d'étape sont précisés entre crochets.*

4.1 Pré-traitement

Notre chaîne de traitements prend en entrée les articles de l'encyclopédie classés en géographie d'après un modèle de classification entraîné sur l'EDdA par Brenon et al. (2022). La majorité des articles classés en géographie décrivent un lieu (e.g. Jean de Luz, S.[3]), mais cer-

tains décrivent des noms de peuples ou de communautés (e.g. SALYENS[6]) ainsi que des noms de concepts géographiques (e.g. LATITUDE[7]). Dans le cadre de ce travail, nous nous intéressons uniquement aux articles décrivant les lieux, la première étape [1] est donc une étape de classification qui permet de distinguer les lieux, les peuples ou les concepts géographiques, respectivement les classes *Place*, *Person* et *Other*.

Les articles de lieux peuvent décrire un ou plusieurs lieux (voir les exemples (1) et (2)). Pour distinguer ces deux cas (voir section 5.2) nous avons ajouté une étape de classification supervisée (étape [2]).

(1) **MÉGARSUS, ou MAGARSUS**[8], (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.

(2) **SYCAE**[9], (Géog. anc.) nom **d'une ville** de la Cilicie, **& d'une ville** de la Thrace, selon Étienne le géographe. (D. J.)

La dernière étape du pré-traitement (étape [3]) a pour rôle de segmenter les articles décrivant plusieurs lieux. La grande diversité des formulations employées pour exprimer qu'une même vedette fait référence à 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'entraînement d'un modèle de classification supervisé performant. Pour contourner cette difficulté, nous proposons une approche *few-shot* fondée sur l'instruction de grands modèles de langage dont l'objectif est de générer de manière indépendante chacune des descriptions relevant d'un lieu unique. Les exemples (3) à (5) montrent les sorties obtenues pour l'article MÉGARSUS, ou MAGARSUS (voir exemple (1)).

(3) MÉGARSUS, ou MAGARSUS (Géog. anc.) **1°** une ville de Cilicie, près du fleuve Pyrame ;

(4) MÉGARSUS, ou MAGARSUS (Géog. anc.) **2°** une rivière de Scythie, selon Strabon ;

(5) MÉGARSUS, ou MAGARSUS (Géog. anc.) **3°** un fleuve de l'Inde, selon Denys le Périégète.

Ainsi, nous obtenons à la fin du pré-traitement des articles traitant d'un lieu unique et autant de noeuds sont créés dans le graphe.

4.2 Enrichissement du graphe

Cette partie décrit les étapes [4] à [7] de la chaîne de traitements.

4.2.1 Classification des types de lieux

L'étape [4] est la première étape du traitement pour l'enrichissement du graphe ; elle consiste à identifier le type de lieu d'un article en s'appuyant sur la typologie définie dans l'ontologie

6. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/14/3429
7. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/9/1466
8. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/10/1365
9. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/15/3468

spatiale (voir section 3.2). Cette étape fait de nouveau appel à un modèle de classification de texte entraîné de manière supervisée. L'exemple (3) sera classé en `City` et les exemples (4) et (5) en `River`. Cette étape génère les triplets correspondants et enrichit le graphe.

4.2.2 Repérage et classification des entités nommées

Les étapes [5] et [6] de repérage et de classification des entités nommées s'intéressent aux informations contenues au sein des textes des articles (et plus aux vedettes). Un modèle de repérage d'entités sémantiques (Moncla et Zeghidi, 2025) est utilisé, il nous permet en particulier d'extraire les entités nommées de lieux, les expressions de relations spatiales et les coordonnées géographiques.

De la même manière que précédemment pour les articles de lieux, nous souhaitons classer les entités nommées de lieux selon leur nature en s'appuyant sur les types prédéfinis dans l'ontologie spatiale. Le modèle entraîné attribue une des classes de l'ontologie à chaque entité nommée en fonction de son contexte constitué des cinq mots qui précèdent et suivent l'entité.

Pour éviter de créer des entités déjà existantes (e.g., celles issues des vedettes d'articles), un procédé d'appariement (*matching)* compare le nom des entités nommées de façon exacte (ainsi que leurs types) avec celui des entités existantes. En cas d'échec, les mesures de similarité Jaro-Winkler et Levenshtein normalisée, selon les travaux de Gali et al. (2016), effectuent une comparaison approximative avec un seuil élevé (95%, pour limiter les correspondances incorrectes). Si aucune correspondance n'est trouvée, une nouvelle URI est créée pour ce lieu et son noeud est ajouté dans le graphe, ainsi que le triplet associant ce noeud à son type de lieu.

Pour les coordonnées géographiques, nous appliquons un modèle *Transformers* de type *encoder-decoder* pour une tâche de génération *text-to-text* afin de normaliser les coordonnées extraites vers le format DMS (degré, minute, seconde) puis de les transformer en degrés décimaux. Les triplets indiquant latitude et longitude sont ajoutés dans le graphe et associés au noeud correspondant à la vedette.

4.2.3 Extraction des relations spatiales

Cette étape [7] consiste d'une part à classer les expressions de relations extraites à l'étape précédente en s'appuyant sur les classes prédéfinies de l'ontologie et d'autre part à identifier les entités impliquées dans la relation (Aurnague et al., 1997).

La tâche de classification est réalisée par un modèle supervisé entrainé sur le jeu de données `GeoEDdA-TopoRel` (décrit à la Section 5). Pour les relations spatiales, les étiquettes utilisées dans le jeu de données annoté reprennent la liste des prédicats définis dans l'ontologie spatiale à l'exception des prédicats `orientation` et `distance` qui sont regroupés en une seule classe. En effet, les relations spatiales exprimant la distance ou l'orientation sont très majoritairement utilisées conjointement dans les articles de l'encyclopédie (voir l'exemple (6)) et les expressions complètes sont regroupées sous la même étiquette `orientation-distance` par le modèle de repérage des entités. Afin de construire les triplets en suivant les spécifications de l'ontologie spatiale, nous appliquons des expressions régulières pour extraire et normaliser les valeurs et unités de distance et d'orientation.

(6) [...] à 4 lieues N. E. de Fontarabie, 4 S. O. de Bayonne, 174 S. O. de Paris. [...]

Une fois les expressions de relations spatiales identifiées et catégorisées, il est nécessaire de les relier aux entités sujet et objet correspondantes. L'approche retenue consiste à considérer que le sujet de la relation peut être soit la vedette de l'article, soit la dernière entité précédant la relation (en terme de position dans la phrase) tandis que dans la majorité des cas, son objet est clairement identifié comme étant l'entité suivante. 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 (e.g. City, inclusion, Region) et (Country, inclusion, Region). Pour cela, nous avons sélectionné les 500 premiers articles de l'EDdA (toutes les entités nommées ont déjà leur URI à cette étape), dans lesquels les expressions de relation ont été associées à leur sujet et leur objet par le modèle `gpt-4.1-mini`. À partir de cet échantillon, nous avons construit un ensemble de triplets et calculé la fréquence d'apparition de chaque type de triplet. Par exemple, `City, inclusion, Region` a une probabilité élevée tandis que celle de `Sea, inclusion, Region` sera faible. Cet ensemble sert de référence de probabilité des différents types de triplets.

5 Jeu de données et expérimentations

5.1 Création du jeu de données `GeoEDdA-TopoRel`

Afin d'entraîner et d'évaluer les différents modèles utilisés par la chaîne de traitements (voir section 4), nous avons construit un jeu de données *gold-standard* annoté. Ce jeu de données se compose de 2 750 articles annotées, extraits de l'Encyclopédie de Diderot et d'Alembert (données fournies par l'ARTFL [10]), dont 2 250 relèvent exclusivement du domaine de la géographie. `GeoEDdA-TopoRel` met à disposition un ensemble de variables exploitées à différentes étapes de notre pipeline de construction d'un graphe de connaissances, représenté à la figure 2. Chaque donnée comprend 10 champs :

- `volume` : numéro du volume de l'article dans le corpus Encyclopédie ARTFL ;
- `numero` : numéro de l'article à l'intérieur d'un volume (Encyclopédie ARTFL) ;
- `head` : intitulé de l'article ;
- `text` : texte intégral de l'article en format brut ;
- `entryType` : typologie de l'entrée géographique, avec trois modalités possibles : *Place*, *Person* ou *Other* ;
- `cardinality` : information permettant de distinguer les cas où l'article renvoie à un lieu unique (*Single*) ou à plusieurs lieux (*Multiple*) ;
- `placeType` : type du lieu décrit par l'article ;
- `placeNames` : liste des toponymes extraits, associés à leur position dans le texte et à leur type sémantique ;
- `spatialRelations` : relations spatiales identifiées dans le texte, associées à leur position et à leur type ;
- `segmentedDescriptions` : liste des sous-descriptions lorsque la cardinalité est *Multiple* ; champ vide dans le cas contraire.

Pour constituer le jeu de données, nous avons adopté une approche fondée sur la validation manuelle des prédictions d'un LLM génératif utilisé en *zero* ou *few-shot*. Plusieurs modèles ont été comparés selon les tâches. Pour la classification des entités nommées de lieux,

10. `https://encyclopedie.uchicago.edu`

les F-mesures obtenues sont : 87,6% avec `gpt-4.1-mini`, 78,7% avec `gpt4-turbo`, puis 56,7% et 37,3% avec `mixtral:8x7b` et `deepseek-r1:7b`. Pour la segmentation en sous-articles (lorsque plusieurs lieux sont décrits), les précisions atteignent 96,8% avec `gpt-4.1-mini`, 95,6% avec `gpt4-turbo`, puis 85,6% et 46,8% avec `llama3:70b` et `mixtral:8x7b`. Les résultats détaillés pour l'ensemble des tâches de construction du jeu de données sont disponibles sur le GitLab du projet[2].

5.2 Évaluation de la chaîne de traitements

En complément de l'évaluation des LLMs qui ont servi de base à la constitution du jeu de données, `GeoEDdA-TopoRel` a été utilisé pour l'entraînement et l'évaluation des modèles supervisés utilisés dans la chaîne de traitements. Le jeu de données est partitionné en 3 parties, 1 800 articles pour le jeu d'entraînement, 225 articles pour les jeux de test et autant pour la validation. La répartition des articles au sein de ces trois sous-ensembles a été faite de manière à respecter au maximum la distribution des différentes classes géographiques.

Pour l'ensemble des étapes (voir figure 2), nous avons fine-tuné des modèles encodeurs pré-entraînés (BERT) avec une couche de classification. Les scores obtenus sur le jeu de test sont présentés dans le tableau 1. Les évaluations détaillées de chacun des modèles sont disponibles sur HuggingFace[1].

TAB. 1 – *F-mesures moyennes pondérées des différents modèles entraînés (sur `GeoEDdA-TopoRel`) et utilisés dans la chaîne de traitements.*

Étape	**[1]**	**[2]**	**[4]**	**[6]**	**[7]**
F-mesure	98%	98%	94%	84%	92%

Pour la tâche de classification des entités nommées de lieux (étape [6], voir figure 2), nous proposons une solution hybride pour entraîner le modèle de classification. Cette approche repose sur une première étape *few-shot* utilisant un modèle génératif pour constituer un jeu d'entraînement non corrigé et utilisé pour entraîner un modèle supervisé[11]. Les résultats montrent que le modèle supervisé sur des données bruitées obtient des scores légèrement plus faibles que l'approche par LLM. Néanmoins, l'avantage de cette approche réside dans l'autonomie vis-à-vis de très gros modèles coûteux et dans la capacité à être ré-entraînée sur des données nettoyées. Nous avons également étudié l'impact de la taille du contexte autour de l'entité nommée. Les résultats du tableau 1 pour l'étape [6] sont donnés pour une taille de contexte de 5-gram.

5.3 Graphe complet

Le graphe résultant de la construction de l'EDdA contient 428 098 triplets : 1 encyclopédie, 17 volumes, 15 384 vedettes (autant que le nombre d'articles) dont 15 252 uniques.

L'étape de matching doit être évaluée sur le graphe complet (et pas sur le sous-ensemble `GeoEDdA-TopoRel`). Les articles contiennent 87 500 entités nommées. Le résultat de l'appariement est montré dans le tableau 2. Les entités sont matchées en priorité avec une égalité

11. Pour cette étape, seul le jeu de test a été corrigé à la main

stricte sur la vedette (54 033) puis sur une entité nommée (9 310). Les mesures de similarité ajoutent quelques milliers de correspondances. Enfin, 3 505 entités sont ambiguës (pointant vers 2 entités de même nom et de même type) et 16 594 n'ont pas de correspondance. Au final,

TAB. 2 – *Nombre de correspondances trouvées selon la stratégie de matching.*

Égalité		**Similarité**		**Ambiguïté**	**Sans**
vedette	entité nommée	vedette	entité nommée		**correspondance**
54 033	9 310	3 096	962	3 505	16 594

le graphe inclut 35 552 entités géographiques et 46 585 relations spatiales. Parmi les entités, 13 476 proviennent d'un article unique, 1 977 sont issues des 728 articles segmentés, 16 594 sont des entités nommées sans correspondance et 3 505 sont ambiguës.

Le tableau 3 précise le nombre d'entités pour chaque classe géographique. Les villes sont largement dominantes (48%), suivies des régions (15%) et des rivières (14%). Les types les moins représentés sont pays et mer (ce qui est cohérent avec leur nature) et lac (429, un nombre qui semble plutôt faible). Le tableau 4 liste le nombre de relations spatiales. La relation d'inclusion domine largement dans le graphe (53%) car les descriptions évoquent avant tout l'organisation hiérarchique des entités spatiales (e.g., chaque ville est située par rapport à son pays, et souvent dans sa région). Les relations d'orientation (11%) et de distance (10%) montrent l'importance de la localisation exprimée en fonction d'autres lieux. Les 3 000 relations *Other* nécessitent une analyse approfondie pour détecter de nouveaux types de relation.

TAB. 3 – *Nombre d'instances par classe géographique.*

Classe	**Nb. instances**
City	17 159
Region	5 167
River	4 930
Island	2 541
Other	1 752
Mountain	1 312
HumanMade	1 071
Country	599
Sea	592
Lake	429
TOTAL	35 552

TAB. 4 – *Nombre d'instances par relation spatiale.*

Relation	**Nb. instances**
inclusion	24 629
orientation	5 330
distance	4 640
adjacency	3 724
crosses	3 487
other	3 000
movement	1 775
TOTAL	46 585

Les tableaux 5 et 6 montrent les 5 pays et les 5 régions ayant le plus de relations spatiales (en tant que sujet ou objet d'un prédicat). La France est en première place, suivies par des pays voisins. Comme notre ontologie n'inclut pas de classe pour les continents, ceux-ci ont donc été reconnus comme des pays (e.g., l'Afrique dans le tableau 5). Au niveau des régions, l'Égypte et l'Irlande ont été classées de façon erronée.

Enfin, nous avons étudié la distribution des types de relations spatiales selon leur sujet. Certaines associations ne sont a priori pas possibles, comme un pays sujet d'un prédicat de mouvement. Une analyse approfondie est nécessaire pour identifier ces cas et les corriger.

TAB. 5 – *Top-5 des pays avec le plus de relations spatiales (sujet et objet).*

Pays	Nb. relations
France	3 208
Allemagne	2 454
Italie	2 350
Afrique	1 632
Angleterre	944

TAB. 6 – *Top-5 des régions avec le plus de relations spatiales (sujet et objet).*

Région	Nb. relations
Égypte	394
Irlande	364
Languedoc	334
Barbarie	258
Westphalie	256

6 Conclusion et perspectives

Cet article propose une chaîne de traitements automatisée pour transformer les données textuelles d'une encyclopédie (ici, celle de Diderot et d'Alembert) en un graphe de connaissances centré sur la géographie. Elle repose sur l'identification des articles du domaine ciblé, leur éventuelle segmentation s'ils contiennent plusieurs sous-articles, et leur typage selon notre ontologie spatiale. La chaîne de traitements permet ensuite d'extraire, de typer et de regrouper les entités nommées spatiales. Enfin, elle détecte et classifie les relations spatiales entre lieux. Le jeu de données expertisé `GeoEDdA-TopoRel` comporte 2 750 articles (dont 2 250 lieux) manuellement annotés pour être utilisés à chaque étape de notre pipeline. Le graphe EDdA, qui comprend plus de 15 000 articles de géographie, 35 000 lieux et 46 000 relations spatiales, peut être chargé dans un *triple-store* pour une analyse approfondie.

Ce travail offre de nombreuses perspectives. La chaîne de traitements peut être améliorée à différents niveaux : les distances (et leurs éventuelles unités) peuvent être représentées par une valeur et une unité, ce qui clarifie leur sémantique. Cela permettrait par ailleurs d'uniformiser les définitions des unités (e.g., une lieue possède 3 définitions avec des mesures différentes selon les périodes). Une comparaison contextuelle (via les relations spatiales) serait utile pour lever l'ambiguïté de certaines entités (e.g., les villes de Vienne en France et en Autriche). Une analyse des relations de la classe *Other* permettrait d'affiner l'ontologie spatiale, et d'autres types de relations (entre un lieu et un autre type, comme événement ou personne) enrichiraient le graphe. Pour promouvoir l'interopérabilité, nous envisageons de définir des correspondances vers d'autres ontologies spatiales, comme Geonames, ou plus générales comme DBpedia et Wikidata. Au niveau des données, la spécification de mappings est moins évidente de par l'évolution des représentations (e.g., la Savoie décrite dans EDdA n'a plus les mêmes caractéristiques – géographiques, organisationnelles – que la Savoie actuelle). Une autre perspective s'intéresse à la diachronie, c'est à dire à l'évolution des connaissances dans le temps, par exemple pour la comparaison de différentes encyclopédies ou de différentes éditions d'une même encyclopédie ou dictionnaire. Enfin, l'étude de dictionnaires publiés dans d'autres langues permettrait de faire émerger un ensemble de savoirs communs à une région donnée tout en identifiant les connaissances spécifiques à un territoire.

Références

Aurnague, M., L. Vieu, et A. Borillo (1997). *La représentation formelle des concepts spatiaux dans la langue*, pp. 69–102. Masson.

Brenon, A., L. Moncla, et K. McDonough (2022). Classifying encyclopedia articles : Comparing machine and deep learning methods and exploring their predictions. *Data & Knowledge Engineering 142*, 102098.

Gali, N., R. Mariescu-Istodor, et P. Fränti (2016). Similarity measures for title matching. In *2016 23rd International Conference on Pattern Recognition*, pp. 1548–1553. IEEE.

Hogan, A., 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, et A. Zimmermann (2021). Knowledge graphs. *ACM Computing Surveys 54*(4), 1–37.

Joliveau, T., L. Moncla, A. Taroni, D. Vigier, et K. Mcdonough (2024). A digital exploration of geographic knowledge in diderot and d'alembert's encyclopédie. In *30th International Conference on the History of Cartography (ICHC)*.

Lefrançois, M. et A. Zimmermann (2016). Supporting arbitrary custom datatypes in rdf and sparql. In *European Semantic Web Conference*, pp. 371–386. Springer.

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

Mark, D. et M. Egenhofer (1998). Modeling spatial relations between lines and regions : Combining formal mathematical models and human subjects testing. *Cartography and Geographic Information Systems 21*, 195–212.

Moncla, L. et H. Zeghidi (2025). Token and span classification for entity recognition in french historical encyclopedias. Technical Report arXiv preprint arXiv :2506.02872, LIRIS.

Rawsthorne, H. M., N. Abadie, E. Kergosien, C. Duchêne, et É. Saux (2021). Automatic nested spatial entity and spatial relation extraction from text for knowledge graph creation. In *Proceedings of the 12th International Conference on Geographic Information Science*.

Rietveld, L. et R. Hoekstra (2015). The linked data fragments approach : A low-cost solution for publishing and querying linked data. In *International Semantic Web Conference (ISWC)*.

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

Wick, M., T. Boutreux, et E. Nauer (2007). The geonames geographical database. Technical report, Geonames.

Summary

Ancient encyclopedias, such as Diderot and d'Alembert's Encyclopédie (1751–1772), are invaluable resources for studying the evolution of geographical knowledge, but their scale hinders manual analysis. This paper presents an automated method for building a geo-historical knowledge graph from these texts. We design spatial and provenance ontologies tailored to the corpus and introduce a gold-standard dataset of 2,750 geographical articles. The proposed pipeline combines supervised learning and large language models for article classification, entity typing, and spatial relation extraction. Our approach achieves strong results (F1 = 92% for relation classification, F1 > 97% for article classification), producing an RDF graph of over 35,000 entities and 46,000 spatial relations. All datasets, models, and code are available on HuggingFace[1] and Gitlab[2].