Encyclopedia2GeoKG : un outil pour l'extraction d'informations et la génération de graphes de connaissances géo-historiques à partir d'articles encyclopédiques

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

*INSA Lyon, CNRS, Universite Claude Bernard Lyon 1,
LIRIS, UMR5205, 69621 Villeurbanne, France
**Universite Claude Bernard Lyon 1, CNRS, INSA Lyon,
LIRIS, UMR5205, 69622 Villeurbanne, France
prenom.nom@liris.cnrs.fr

Résumé. Les dictionnaires et encyclopédies anciennes, comme celle de Diderot et d'Alembert (1751-1772), renferment des connaissances, notamment géographiques, qui sont précieuses pour étudier leur évolution au cours des derniers siècles. Les milliers d'articles dans ces oeuvres nécessitent cependant des outils automatisés pour extraire des informations fiables et structurées. Notre outil s'appuie sur des modèles entraînés spécifiquement pour les textes anciens et permet de construire un graphe RDF représentant les entités géographiques et leurs relations spatiales. Un prototype [1] interactif de l'outil ainsi que les modèles [2] sont disponibles sur HuggingFace, le code est disponible sur gitlab [3].

1 Introduction

Les dictionnaires encyclopédiques anciens ont joué un rôle majeur dans la diffusion des savoirs, notamment au siècle des Lumières. L'Encyclopédie de Diderot et d'Alembert (EDdA, 1751-1772) en est l'exemple emblématique : plus de 74 000 articles, dont 15 000 consacrés à la géographie, offrant une source précieuse sur les représentations du monde au XVIII$^{\text{ème}}$ siècle. L'exploitation de ce corpus reste complexe : richesse lexicale, longueur des textes et spécificités du français ancien limitent les approches manuelles. De plus, les modèles de traitement automatique du langage, entraînés sur des données modernes, sont peu adaptés à une extraction fiable des connaissances. Au-delà de ces défis techniques et linguistiques, la construction automatique d'un graphe de connaissances géographiques offre un moyen innovant d'explorer ce corpus, en structurant l'information sous forme de triplets ⟨sujet, prédicat, objet⟩ interrogeables pour des analyses comparatives ou diachroniques.

Si des ontologies génériques comme FOAF, DBpedia ou Wikidata structurent de vastes connaissances, les domaines spécialisés — tels que la géographie historique — requièrent des

1. https://huggingface.co/spaces/GEODE/encyclopedia-to-geokg
2. https://huggingface.co/GEODE
3. https://gitlab.liris.cnrs.fr/ecoda/encyclopedia2geokg

modèles dédiés. Des ontologies comme GeoNames Ontology (Wick et al., 2007), SWEET (Rietveld et Hoekstra, 2015) ou GEO décrivent les entités géographiques (pays, villes, rivières, montagnes, etc.) et leurs propriétés spatiales, assurant l'interopérabilité entre sources. Enfin, certains modèles intègrent des dimensions spatiales et temporelles, tels que T-GK (Hogan et al., 2021) et SpatioTemporal RDF (Tachev et al., 2022), mais restent centrés sur des données contemporaines et bien structurées.

L'extraction d'information est essentielle pour peupler automatiquement des ontologies à partir de textes non ou faiblement structurés. Elle comprend la reconnaissance et la classification d'entités nommées ainsi que l'extraction de relations entre ces entités (Li et al., 2020). Brenon et al. (2022) et Joliveau et al. (2024) ont appliqué ces techniques à un corpus encyclopédique en français pour analyser l'évolution des discours géographiques : identification d'articles, extraction et classification d'entités, relations spatiales et coordonnées. Les informations extraites servent à la désambiguïsation toponymique et à la génération de cartes. Rawsthorne et al. (2021) vont plus loin avec un graphe de connaissances géographiques construit à partir d'un corpus d'instructions nautiques, fondé sur l'ontologie ATLANTIS et le fine-tuning de BERT pour l'extraction d'entités et de relations spatiales. Plus récemment, l'IA générative a ouvert la voie à l'extraction zero-shot ou few-shot via le prompt engineering (Moncla et Zeghidi, 2025).

Dans cet article, nous proposons une approche d'automatisation de la construction d'un graphe de connaissances géo-historiques à partir d'articles géographiques encyclopédiques. Nous avons développé une chaîne de traitements hybride combinant apprentissage supervisé et extraction/classification few-shot à l'aide de modèles de langage génératifs (LLMs). Cette approche est appliquée à la classification des articles, au typage des entités et à la détection des relations spatiales. Le présent article de démonstration introduit une version allégée de cette chaîne[4] : la démonstration en ligne, Encyclopedia2GeoKG, traite les articles individuellement, sans inclure l'étape de segmentation des textes décrivant plusieurs lieux ni le module complet de matching et d'entity linking, simplifiés ici pour des raisons techniques et de coût.

La suite de cet article est organisée comme suit. La section 2 détaille la modélisation de l'ontologie spatiale ainsi que les principales étapes d'extraction d'information et de peuplement du graphe de connaissances. La section 3 présente le prototype de démonstration accessible en ligne. Enfin, la section 4 conclut et ouvre sur les perspectives futures.

2 Construction du graphe de connaissances

Notre outil Encyclopedia2GeoKG repose sur une chaîne de traitements afin de construire automatiquement un graphe de connaissances pour un article donné. Comme les dictionnaires anciens traitent de nombreux domaines, nous nous sommes focalisés sur celui de la géographie, qui est bien représenté (e.g., dans EDdA, environ 20% d'articles géographiques (Vigier et al., 2022)). Une ontologie a été proposée pour représenter les différents concepts géographiques. Pour le peuplement, une chaîne de traitements composée de différentes étapes d'extraction d'informations à partir de l'article (e.g. type de lieu, entités nommées, relations spatiales) construit progressivement des triplets en se basant sur cette ontologie. Dans la suite, nous décrivons d'abord l'ontologie puis la chaîne de traitements.

4. Un article décrivant la chaîne de traitements complète utilisée pour construire le graphe de connaissances de l'ensemble du corpus a été soumis en article long à EGC. La référence sera ajoutée ici en cas d'acceptation.

2.1 Modélisation

Un article encyclopédique comme ceux de l'EDdA, 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[5]), 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 proposons une ontologie spatiale pour décrire la structure du graphe permettant de représenter les articles géographiques de l'EDdA. Elle définit et structure les concepts liés à l'information géographique tels que les lieux ou entités géographiques et les relations spatiales. La 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[6] (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).

2.2 Extraction d'information et peuplement du graphe

L'enchaînement des étapes dans la chaîne de traitement proposée est illustré dans la figure 1. Chacune des étapes décrites ci-après génère les triplets correspondant aux informations extraites et enrichit le graphe. Par ailleurs, pour l'ensemble des modèles présentés dans cet article, le détail des scores est disponible sur leurs pages HuggingFace[2] respectives.

2.2.1 Classification du type d'article

Notre outil Encyclopedia2GeoKG prend en entrée le texte d'un article de l'encyclopédie. Il vérifie d'abord que l'article est classé en géographie et décrit un lieu. Pour la classification en domaine, nous utilisons un modèle de classification entraîné sur l'EDdA par Brenon

5. `https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/8/2425`
6. Ontologie des unités de mesures, `http://www.ontology-of-units-of-measure.org/`

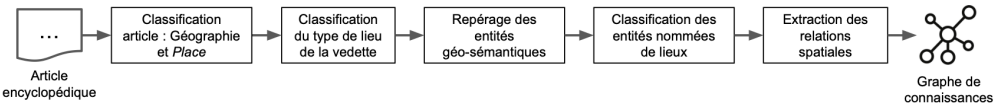

FIG. 1 – *Schéma de la chaîne de traitements de construction automatique du graphe de connaissances à partir d'un article encyclopédique.*

et al. (2022). La majorité des articles classés en géographie décrivent un lieu (e.g. Jean de Luz, S.[5]), mais certains décrivent des noms de peuples ou de communautés (e.g. SALYENS[7]) ainsi que des noms de concepts géographiques (e.g. LATITUDE[8]). La deuxième étape 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*. Pour cela, nous avons procédé au *fine-tuning* d'un modèle BERT pré-entraîné pour une tâche de classification de texte. Ce modèle atteint une f-mesure moyenne de 98%.

2.2.2 Classification des types de lieux

La première étape du traitement pour l'enrichissement du graphe consiste à identifier le type du lieu décrit par l'article selon la typologie définie dans l'ontologie. Pour cela, nous avons entraîné de manière supervisée un modèle de classification de texte basé sur BERT. Ce modèle atteint une f-mesure moyenne de 94%.

2.2.3 Repérage et classification des entités sémantiques

Les étapes de repérage et de classification des entités sémantiques s'intéressent aux informations contenues au sein des textes des articles (et plus aux vedettes). Un modèle BERT de classification de tokens est utilisé pour le repérage d'entités sémantiques (Moncla et Zeghidi, 2025). Il nous permet d'extraire les entités nommées de lieux, les expressions de relations spatiales, les coordonnées géographiques mais également les entités nommées de personnes et les dates. Un deuxième modèle de classification est appliqué sur les entités nommées de lieux (et leur contexte constitué des cinq mots qui précèdent et suivent l'entité) afin de les catégoriser selon leur type en suivant la typologie définie dans l'ontologie. Ce modèle a été entraîné à partir d'un jeu de données construit automatiquement avec `gpt4.1-mini` et obtient 84% de f-mesure moyenne.

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.

7. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/14/3429
8. https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/9/1466

2.3 Extraction des relations spatiales

Cette étape 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). Comme pour les étapes précédentes, nous avons entrainé un modèle BERT de classification. Pour cette tâche il obtient 92% de f-mesure moyenne.

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, 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. Les relations spatiales sont ajoutées dans le graphe en utilisant la notion de `rdf:Statement` afin de relier les différents éléments participant à la relation.

3 Démonstration

Un prototype interactif de Encyclopedia2GeoKG est disponible sur HuggingFace [1] et le code de la chaîne de traitements complète (incluant l'entraînement des modèles) est hébergé sur Gitlab[3]. Sur l'interface, il faut tout d'abord saisir le texte d'un article de lieu provenant d'un dictionnaire ancien. Prenons celui de Saint-Jean-Pied-de-Port mentionné dans l'EDdA, dont le texte est montré dans l'exemple 1. Notez que cet article fait partie des exemples proposés sur le prototype, et il est facile d'en récupérer d'autres sur le site de l'ARTFL[9].

(1) **Jean-pied-de-Port, S.**[10], (Géog.) ville de France en Gascogne, à une lieue des frontieres d'Espagne, autrefois capitale de la basse Navarre, avec une citadelle sur une hauteur [...] Elle est sur la Nive, à l'entrée d'un des passages des Pyrénées, à 8 lieues S. E. de Bayonne, 12 N. E. de Pampelune, 176 S. O. de Paris. Long. 16. 22. lat. 43. 8. (D. J.)

En cliquant sur le bouton `Run pipeline`, les différentes étapes décrites dans la section 2 sont exécutées. Comme illustré par la figure 2, plusieurs sorties sont textuelles (type et cardinalité de l'article, type du lieu). Saint-Jean-Pied-de-Port est bien reconnu comme un article de lieu, et plus précisément comme une ville. La visualisation de la reconnaissance des entités nommées, permet d'identifier les noms communs (NC) et les noms propres (NP) avec leur catégorie (e.g., "spatial", "person") ainsi que les relations.

9. `https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922`
10. `https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/8/2427`

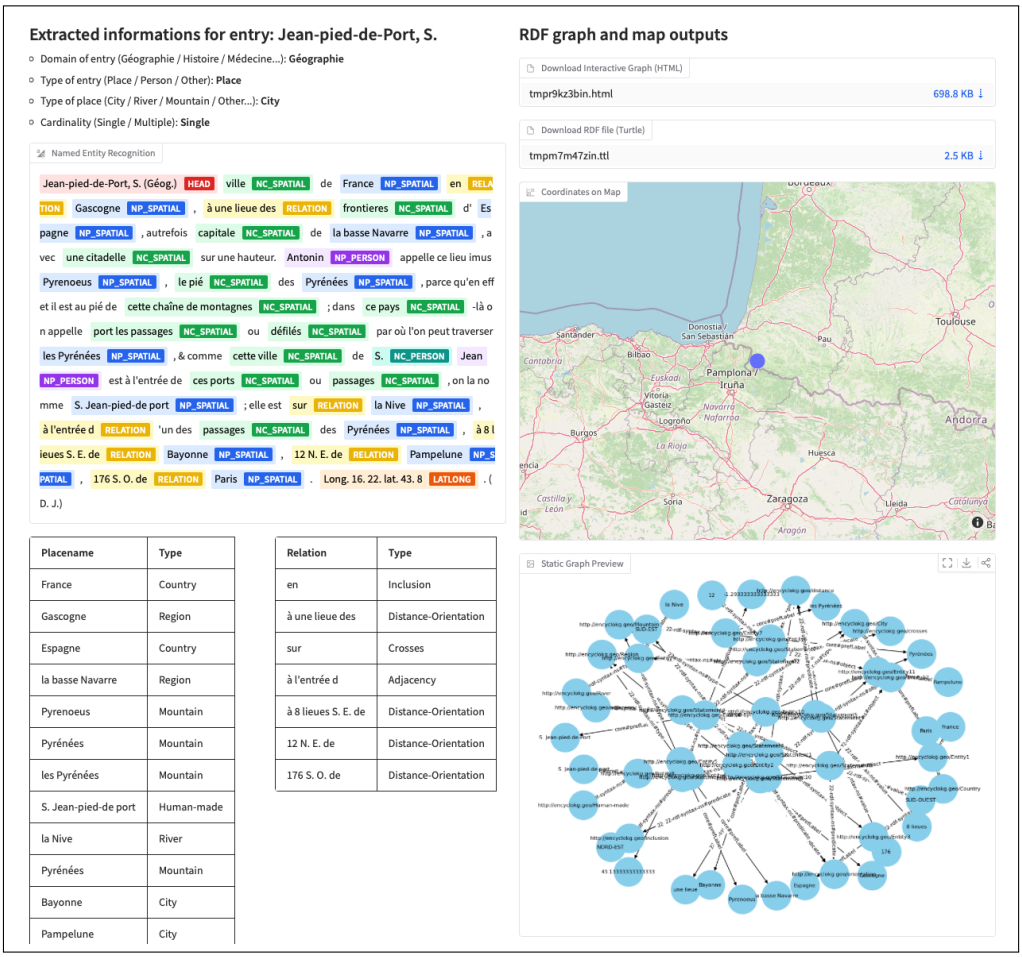

FIG. 2 – *Interface de visualisation des résultats*

La classification des noms propres spatiaux est affichée dans un tableau (e.g., Gascogne en `Region`). Les différentes relations spatiales sont aussi associées à leur type (e.g., *à une lieue des* classé en `DistanceOrientation`) et affichées dans un tableau.

À partir de ces informations, le graphe RDF est construit et le prototype permet de visualiser une version statique (peu lisible) et une version dynamique à télécharger. Cette dernière est montrée en figure 3. La relation entre Saint-Jean-Pied-de-Port et Bayonne est représentée par le `Statement4` (orientation, avec la valeur *Sud-Est*) et par le `Statement5` (distance, avec la valeur *8 lieues*). Enfin, quand des coordonnées géographiques sont détectées dans l'article, elles sont normalisées et positionnées sur une carte.

Lors de la démonstration, nous inviterons les participant·e·s à découvrir les connaissances sur leur commune ou leur région telles qu'on les partageait au XVIII^{ème} siècle à travers les encyclopédies et dictionnaires.

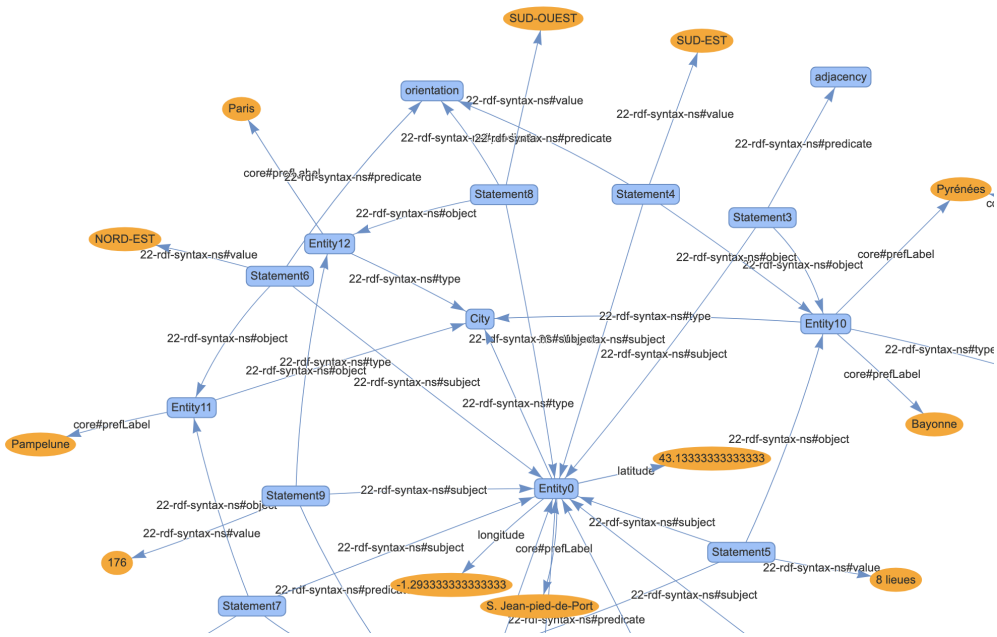

FIG. 3 – *Visualisation d'un extrait du graphe.*

4 Conclusion et perspectives

Dans cet article, nous proposons l'outil Encyclopedia2GeoKG qui permet de transformer un article géographique, tiré d'un dictionnaire ancien, en un graphe RDF. Notre chaîne de traitements automatisée permet d'extraire les informations telles que les entités nommées et leurs relations, puis de les classifier. Un prototype est accessible en ligne, ainsi que le code et les modèles.

L'outil ne traite actuellement qu'un seul article. Pour traiter une encyclopédie complète, il est nécessaire d'ajouter des étapes, notamment d'appariement. En effet un lieu mentionné dans un article peut déjà exister dans le graphe (vedette ou entité nommée dans un autre article). Même avec un processus de nettoyage des noms de lieux, les différences syntaxiques et sémantiques ne permettent pas de détecter toutes les correspondances, i.e., toutes les variantes de nom utilisées pour un même lieu. Un algorithme d'appariement sophistiqué est nécessaire pour détecter davantage de correspondances. Au niveau des sorties, le graphe produit serait difficilement exploitable visuellement, et le développement d'une interface dédiée permettrait à un public non-informaticien de découvrir ces connaissances. Enfin, l'outil pourrait proposer davantage d'interactions, par exemple une validation ou correction humaine.

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.

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.

Vigier, D., L. Moncla, I. Lefort, T. Joliveau, et K. Mcdonough (2022). Les articles de géographie dans le dictionnaire universel de trévoux et l'encyclopédie de diderot et d'alembert. *Langue française 214*(2), 59–80.

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

Summary

Ancient dictionaries and encyclopedias, such as Diderot and d'Alembert's Encyclopédie (1751–1772), contain knowledge—particularly geographical information—that is invaluable for studying how such knowledge has evolved over the past centuries. However, the thousands of articles in these works require automated tools to extract reliable and structured information. Our tool is based on models specifically trained on historical texts and it enables the construction of an RDF graph representing geographical entities and their spatial relationships. An interactive prototype[1], as well as the trained models[2], are available on Hugging Face, and the source code is available on GitLab[3].