Automatic Construction of a Geo-Historical Knowledge Graph from Early Modern Encyclopedic Texts

Bin Yang[1], Ludovic Moncla[1,*], Fabien Duchateau[2,*] and Frédérique Laforest[1]

[1]INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69621 Villeurbanne, France
[2]Universite Claude Bernard Lyon 1, CNRS, INSA Lyon, LIRIS, UMR5205, 69621 Villeurbanne, France

Abstract

Early modern encyclopedias, such as Diderot and d'Alembert's (1751–1772), offer a valuable resource for studying the evolution of geographical knowledge, yet their sheer scale complicates manual analysis. This paper presents an automated method for constructing a geo-historical knowledge graph from these texts. We propose spatial and provenance ontologies tailored to the corpus and introduce a *gold standard* of 2,750 geographical articles. The pipeline combines supervised learning and Large Language Models (LLMs) for article classification, entity typing, and spatial relation extraction. Performance reaches F1 = 92% for relations and F1 > 97% for classification, resulting in an RDF graph of 35,000 entities and 46,000 relations. This work paves the way for the computational analysis of early geographical knowledge. Data, models, and code are available on HuggingFace[1] and Gitlab[2].

Keywords

information extraction, knowledge graph, early modern encyclopedias

1. Introduction

Early modern encyclopedic dictionaries played an essential role in the diffusion of knowledge, particularly during the Age of Enlightenment. Diderot and d'Alembert's *Encyclopédie* (EDdA, 1751–1772) is emblematic of this intellectual endeavor: it comprises approximately 74,000 articles covering numerous domains, including geography (about 15,000 articles). Today, these texts offer a valuable source for historians and linguists, as they bear witness to the representations of the world and geographical knowledge available in the 18th century.

However, exploiting this corpus is a complex task. The lexical richness, the length of descriptions, and the linguistic specificities of early modern French severely limit manual approaches [1]. Furthermore, existing Natural Language Processing (NLP) models are primarily trained on contemporary data and are not directly adapted to this type of corpus for reliable knowledge extraction.

Beyond these technical and linguistic challenges, the automatic construction of a geographical knowledge graph offers a new way to explore the corpus: it allows for the structuring of information in the form of triples ⟨subject, predicate, object⟩, which are queryable and exploitable from a comparative or diachronic perspective. Our objective is therefore to automate the construction of a geo-historical knowledge graph from the geographical articles of the EDdA. To do this, we designed a complete processing pipeline relying both on classification models fine-tuned on a data sample from the corpus and on Large Language Models (LLMs) for the extraction and structuring of geographical information.

The contributions of this paper are as follows:

- The definition of an ontology tailored to early modern encyclopedic texts;
- The design of a hybrid pipeline, combining supervised learning and generative *few-shot* LLM classification/extraction, applied to article classification, entity typing, and spatial relation detection;

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

GeoExT 2026: Fourth International Workshop on Geographic Information Extraction from Texts at ECIR 2026, April 2, 2026, Delft, The Netherlands

*Corresponding author.
ludovic.moncla@insa-lyon.fr (L. Moncla); fabien.duchateau@univ-lyon1.fr (F. Duchateau); frederique.laforest@insa-lyon.fr (F. Laforest)
0000-0002-1590-9546 (L. Moncla); 0000-0001-6803-917X (F. Duchateau); 0000-0002-9421-8566 (F. Laforest)

- The creation and release of the `GeoEDdA-TopoRel` annotated dataset, enabling the training and evaluation of the various models in the pipeline;
- The construction of a large-scale RDF geo-historical knowledge graph (over 35,000 entities and 46,000 relations), the first resource of this kind based on a historical encyclopedia.

The remainder of this paper is organized as follows: Section 2 presents related work. Section 3 describes the ontological modelling of the domain. We present the methodology in Section 4. Datasets and experiments are presented in Section 5, before concluding in Section 6.

2. State of the Art

Ontologies or knowledge graphs are formal structures designed to represent concepts and the relations uniting them. Several generic ontologies, such as FOAF, DBpedia, or Wikidata, have enabled the structuring of vast sets of knowledge across varied domains. Yet, when dealing with specialized domains, such as historical geography, more adapted models are required.

Modeling geographical entities in knowledge graphs allows for the formalization of geographical object types (countries, cities, rivers, mountains, etc.) and their spatial properties. Ontologies such as GeoNames Ontology [2], SWEET [3], or GEO (Geographic Ontology) have been developed to structure geographical information and facilitate interoperability between heterogeneous sources. In particular, these models allow for the explicit instantiation of geographical entities according to their nature, position, spatial hierarchy, and relationships with other objects. Some works have explored enriching knowledge graphs with temporal and spatial dimensions [4] but these often remain focused on contemporary and well-structured data.

Several projects focus on linking articles to an existing knowledge base. For instance, the Brazilian Historical-Biographical Dictionary contains around 7,800 articles, whose titles have been linked to Wikidata [5]. Articles from six German encyclopedias have been linked to Wikipedia, with a focus on word sense disambiguation [6]. The *frances* tool is dedicated to text mining techniques, for instance topic modelling, sentiment analysis or text summarization. It has been applied on the Britannica encyclopedia for cleaning and enriching information for end-users, but not for building a knowledge graph [7].

Information extraction is essential for populating ontologies from unstructured text, transitioning from symbolic methods to deep neural networks and pre-trained models like BERT [8]. Recently, generative AI and Large Language Models (LLMs) have shifted the paradigm toward zero-shot or few-shot extraction via prompt engineering [9]. Research on French encyclopedic corpora has utilized IE for toponym disambiguation and mapping, though without full knowledge graph integration [10, 11]. Conversely, [12] extended IE to construct a maritime knowledge graph using the ATLANTIS ontology, employing fine-tuned BERT models to extract nested entities and spatial relations from nautical instructions.

By adapting these works, we propose a methodology combining supervised learning with the *fine-tuning* of encoder models and *few-shot prompting* of generative LLMs in order to model and populate a knowledge graph for the EDdA.

3. Graph Modeling

An encyclopedia is divided into different volumes (e.g., 17 text volumes for the EDdA), in which articles are written. Each article has a headword (or title) and is sometimes accompanied by a domain marker, which is not standardized (e.g., *Géog*, *Géog. anc.*, or *Géogr.*). The first words of a geography article generally specify the type of place (e.g., *a city in South Holland* for the article on Delft[1]). The remainder of the article may include a description of the place, geographical coordinates, and other places with their relations to the headword (e.g., distance, orientation).

[1] https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/4/3906

The provenance aspect in our ontology aims to link information extracted from the text to the article, volume, and encyclopedia from which it originates. The Heritage Text Ontology enables the representation of encyclopedic information, but it lacks links between articles and some spatial classes and relationships [7]. To this end, we propose a set of classes and predicates tailored to our corpus:

- Classes: `Encyclopedy`, `Volume`, `Article`;
- Predicates: `articleOf`, `volumeOf`, `articleNumber`, `order`, `extractedFrom`.

Each piece of extracted information is represented in the form of an `rdf:Statement` (reification process), which is attached to its source article via the `extractedFrom` predicate.

Dealing with the spatial aspect, a main class `Place` allows for the representation of places with their characteristics (e.g., geographical coordinates, dimensions). Furthermore, it is specialized according to a place typology established based on the types most present in the encyclopedias.

- Subclasses: `Country`, `City`, `Region`, `Sea`, `River`, `Lake`, `Mountain`, `Island`, `HumanMade`, `Other`;
- Predicates: `latitude`, `longitude`, `surface`, `length`.

Spatial relations between geographical entities are inspired by those of DE-9IM [13]:

- `inclusion` (e.g., in the duchy, in Germany, in the Kingdom of France);
- `adjacency`, which denotes strong proximity between two places (e.g., next to, on the coast of);
- `orientation`, which further specifies a cardinal direction using the `rdf:value` predicate (e.g., to the south of, to the west [*au couchant*] of, to the south [*au midi*]);
- `distance`, which generally specifies a value and a unit[2] (e.g., two leagues from, five miles from, two parasangs from). Currently, values are stored as strings, but solutions exist to refine this [14];
- `movement` (e.g., flows into, originates from, flows into the sea);
- `crosses` (e.g., is located on the river, crosses the city);
- `other` (e.g., between Lebanon and Anti-Lebanon).

Our ontology is available[3] in RDF (Turtle) and graphical versions (HTML). It uses the prefix `ekg` corresponding to the URI `http://encyclokg.geo/`. Figure 1 illustrates the representation of the triplet (Lyon, orientation, Alps) in a graph, by means of statement S1. The latter also specifies the cardinal direction (*West*), as well as its provenance (article A1, from volume V1 of the EDdA encyclopedia).

4. Construction

In this section, we describe our pipeline for automatically constructing a knowledge graph (see Figure 2). It consists of five main modules: pre-processing (steps 1 to 3), classification of articles describing a place (step 4), identification of geo-semantic entities, including place named entities, spatial relations, and geographical coordinates (step 5), classification of place named entities (step 6), and extraction of spatial relations (step 7). Each of these modules generates triples that populate the graph incrementally. Generative LLMs are only used during data preparation but are not required afterward (except for the segmentation step).

4.1. Pre-processing

Our pipeline takes as input encyclopedia articles classified under geography according to a classification model trained on the EDdA [10]. The majority of articles classified as geography describe a place (e.g., Delft[1]), but a few describe names of peoples or communities (e.g., SALYENS[4]) as well as names of

[2]Ontology of units of measure, http://www.ontology-of-units-of-measure.org/

[3]https://gitlab.liris.cnrs.fr/ecoda/encyclopedia2geokg/-/tree/main/ontologies

[4]https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/14/3429

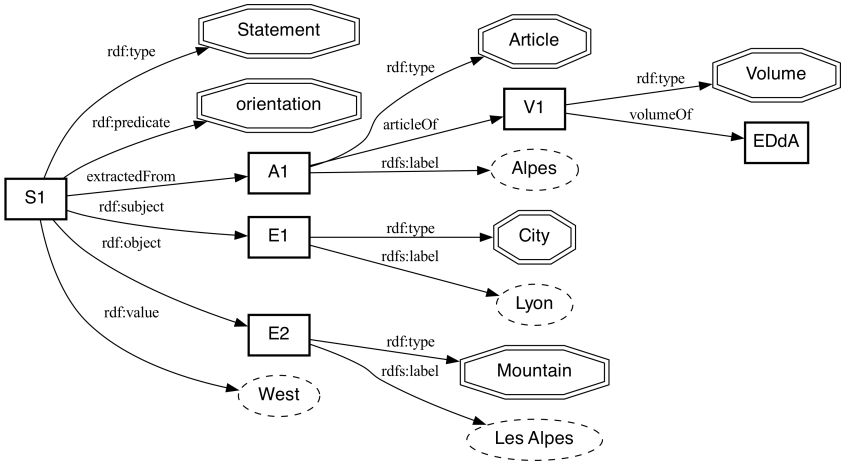

Figure 1: Simplified example of a cardinal relation between Lyon and the Alps. Ontology classes are represented by a double octagon, instance resources as rectangles, and literals by a dotted circle.

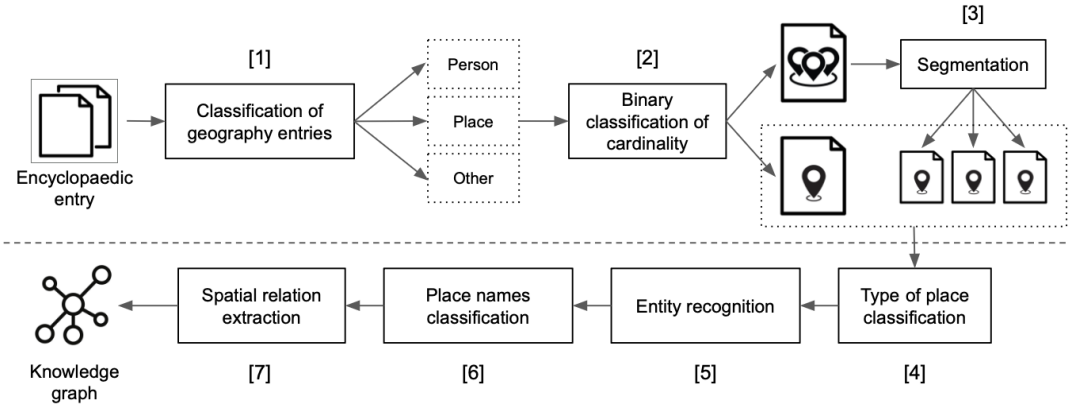

Figure 2: Diagram of the automatic knowledge graph construction pipeline. Step numbers are indicated in brackets.

geographical concepts (e.g., LATITUDE[5]). In the context of this work, we are only interested in articles describing places; thus the first step [1] allows for distinguishing places, peoples, or geographical concepts, specifically the classes *Place*, *Person*, and *Other*.

Place articles may describe one or multiple places (see examples (1) and (2)). To distinguish these two cases (see section 5.2), we added a supervised classification step (step [2]).

(1) **MÉGARSUS, or MAGARSUS**[6], (Anc. Geog.) **1°** a city in Cilicia, near the Pyramos river; **2°** a river in Scythia, according to Strabo; **3°** a river in India, according to Dionysius Periegetes.

(2) **SYCAE**[7], (Anc. Geog.) name **of a city** in Cilicia, **& of a city** in Thrace, according to Stephanus the Geographer. (D. J.)

The final step of pre-processing (step [3]) aims at segmenting articles describing multiple places. The great diversity of formulations used to express that a single headword refers to multiple places – such as enumerations, the use of the symbol "&", or expressions like "there is yet another city" – makes it difficult to train a high-performing supervised classification model. To overcome this difficulty, we propose a *few-shot* approach based on instructing large language models, with the objective of

[5]https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/9/1466
[6]https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/10/1365
[7]https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/15/3468

independently generating each description related to a unique place. Examples (3) to (5) show the outputs obtained for the article MÉGARSUS, or MAGARSUS (see example (1)).

(3) MÉGARSUS, or MAGARSUS (Anc. Geog.) **1**° a city in Cilicia, near the Pyramos river;

(4) MÉGARSUS, or MAGARSUS (Anc. Geog.) **2**° a river in Scythia, according to Strabo;

(5) MÉGARSUS, or MAGARSUS (Anc. Geog.) **3**° a river in India, according to Dionysius Periegetes.

Thus, at the end of pre-processing, we obtain articles dealing with a single place, and as many corresponding nodes are created in the graph.

4.2. Graph Enrichment

This part describes steps [4] to [7] of the pipeline.

4.2.1. Classification of Place Types

Step [4] is the first processing step for graph enrichment; it consists of identifying the place type of an article based on the typology defined in our ontology (see section 3). This step uses a text classification model trained in a supervised manner. Example (3) will be classified as `City` and examples (4) and (5) as `River`. This step generates the corresponding triples and enriches the graph.

4.2.2. Named Entity Recognition and Classification

Steps [5] and [6], for named entity recognition and classification, focus on information contained within the article texts (and no longer on the headwords). A semantic entity recognition model [15] allows us to extract place named entities, spatial relation expressions, and geographical coordinates.

Similarly to place articles, we classify named entities describing places according to their nature based on the predefined spatial types in our ontology. The trained model assigns one of the ontology classes to each named entity based on its context, consisting of the five words preceding and following the entity.

To avoid creating already existing entities (e.g., those derived from article headwords), a matching process compares the label of the named entities exactly (as well as their types) with that of existing entities. In case of failure, we rely on normalized Jaro-Winkler and Levenshtein similarity measures to perform an approximate comparison with a high threshold of 95%, to limit incorrect matches [16]. If no match is found, a new URI is created for this place and its node is added to the graph, along with the triple associating this node to its place type.

For geographical coordinates, we apply an *encoder-decoder* type *Transformers* model for a *text-to-text* generation task in order to normalize the extracted coordinates to the DMS (degree, minute, second) format and then transform them into decimal degrees. The triples indicating latitude and longitude are added to the graph and associated with the node corresponding to the headword.

4.2.3. Extraction of Spatial Relations

Step [7] consists, on the one hand, of classifying the relation expressions extracted in the previous step based on the predefined classes of the ontology and, on the other hand, of identifying the entities involved in the relation [17].

The classification task is performed by a supervised model trained on the `GeoEDdA-TopoRel` dataset (described in Section 5). For spatial relations, the labels used in the annotated dataset adopt the list of spatial predicates defined in our ontology, with the exception of the `orientation` and `distance` predicates, which are grouped into a single class. Indeed, spatial relations expressing distance or orientation are mainly used jointly in the encyclopedia articles (see example (6)), and the complete expressions are grouped under the same label `orientation-distance` by the entity recognition model. In order to construct the corresponding triples, we apply regular expressions to extract and normalize the values and units of distance and orientation.

(6) [...] at 4 leagues N. E. of Fontarabie, 4 S. W. of Bayonne, 174 S. W. of Paris. [...]

Once the spatial relation expressions are identified and categorized, it is necessary to link them to the corresponding subject and object entities. The adopted approach consists of considering that the subject of the relation can be either the article headword or the last entity preceding the relation (in terms of position in the sentence), whereas in the majority of cases, its object is clearly identified as being the following entity. Depending on the relation type and the object type, we compare the probabilities associated with each of the two candidate subjects in terms of the types of formed triples (e.g. <City, inclusion, Region> and <Country, inclusion, Region>). To define these probabilities, we selected the first 500 articles of the EDdA (all named entities already have their URI at this stage), in which the relation expressions were associated with their subject and object by the `gpt-4.1-mini` model. From this manually verified sample, we built a set of triples and calculated the frequency of appearance of each triple type. For example, <City, inclusion, Region> has a high probability while the one for <Sea, inclusion, Region> is low.

5. Datasets and Experiments

5.1. Creation of the `GeoEDdA-TopoRel` Dataset

To train and evaluate the various models used in the processing pipeline (see Section 4), we constructed an annotated *gold-standard* dataset. This dataset consists of 2,750 annotated articles extracted from Diderot and d'Alembert's *Encyclopédie* (data provided by ARTFL[8]), 2,250 of which fall exclusively within the domain of geography. `GeoEDdA-TopoRel` provides a set of variables utilized at different stages of our knowledge graph construction pipeline, represented in Figure 2. Each data point includes 10 fields:

- `volume`: volume number of the article in the ARTFL *Encyclopédie* corpus;
- `numero`: article number within a volume (ARTFL *Encyclopédie*);
- `head`: article headword (title);
- `text`: full text of the article in raw format;
- `entryType`: typology of the geographical entry, with three possible modalities: *Place*, *Person*, or *Other*;
- `cardinality`: information distinguishing cases where the article refers to a single place (*Single*) or multiple places (*Multiple*);
- `placeType`: type of the place described by the article;
- `placeNames`: list of extracted toponyms, associated with their position in the text and their semantic type;
- `spatialRelations`: spatial relations identified in the text, associated with their position and type;
- `segmentedDescriptions`: list of sub-descriptions when the cardinality is *Multiple*; empty field otherwise.

To compile the dataset, we adopted an approach based on the manual validation of predictions from a generative LLM used in a *zero-* or *few-shot* setting. Several models were compared depending on the task. For the classification of named place entities, the F-measures obtained are: 87.6% with `gpt-4.1-mini`, 78.7% with `gpt4-turbo`, followed by 56.7% and 37.3% with `mixtral:8x7b` and `deepseek-r1:7b`. For the segmentation into sub-articles (when multiple places are described), precision scores reach 96.8% with `gpt-4.1-mini`, 95.6% with `gpt4-turbo`, followed by 85.6% and 46.8% with `llama3:70b` and `mixtral:8x7b`. Detailed results for all dataset construction tasks are available on the project's GitLab[2].

[8] https://encyclopedie.uchicago.edu

5.2. Pipeline Evaluation

In addition to evaluating the LLMs that served as the basis for creating the dataset, GeoEDdA-TopoRel was used for training and evaluating the supervised models used in the processing pipeline. The dataset is partitioned into three parts: 1,800 articles for the training set, 225 articles for the test set, and the same number for validation. The distribution of articles within these three subsets was performed to respect the distribution of the different geographical classes as closely as possible.

For all steps (see Figure 2), we fine-tuned pre-trained encoder models (BERT) with a classification layer. The scores obtained on the test set are presented in Table 1. Detailed evaluations for each model are available on HuggingFace[1].

Table 1
Weighted average F-measures of the different models trained (on GeoEDdA-TopoRel) and used in the processing pipeline.

Step	**[1]**	**[2]**	**[4]**	**[6]**	**[7]**
F-measure	98%	98%	94%	84%	92%

For the classification of named place entities task (step [6], see Figure 2), we propose a hybrid solution to train the classification model. This approach relies on an initial *few-shot* step using a generative model to compile an uncorrected training set used to train a supervised model[9]. The results show that the model supervised on noisy data obtains slightly lower scores than the LLM approach. Nevertheless, the advantage of this approach lies in its independence from very large, costly models and in its capacity to be re-trained on cleaned data. We also studied the impact of the context size around the named entity. The results in Table 1 for step [6] are given for a 5-gram context size.

5.3. Full Graph

The graph resulting from the processing of the EDdA contains 428,098 triplets: 1 encyclopedia, 17 volumes, 15,384 headwords (matching the number of articles), of which 15,252 are unique.

The matching step must be evaluated on the full graph (and not on the GeoEDdA-TopoRel subset). The articles contain 87,500 named entities. The result of the matching is shown in Table 2. Entities are matched primarily via strict equality on the headword (54,033), then on a named entity (9,310). Similarity measures add a few thousand matches. Finally, 3,505 entities are ambiguous (pointing to 2 entities of the same name and type) and 16,594 have no match.

Table 2
Number of matches found according to the matching strategy.

Equality		**Similarity**		**Ambiguity**	**No match**
headword	named entity	headword	named entity		
54,033	9,310	3,096	962	3,505	16,594

At the end of the matching step, the graph includes 35,552 geographical entities and 46,585 spatial relations. Among the entities, 13,476 originate from a single article, 1,977 come from the 728 segmented articles, 16,594 are named entities without a match, and 3,505 are ambiguous.

Table 3 details the number of entities for each geographical class. Cities are largely dominant (48%), followed by regions (15%) and rivers (14%). The least represented types are countries and seas (which is consistent with their nature) and lakes (429, a number that seems rather low). Table 4 lists the number of spatial relations. The inclusion relation largely dominates the graph (53%) because descriptions primarily evoke the hierarchical organization of spatial entities (e.g., every city is located relative to its country, and often within its region). Orientation (11%) and distance (10%) relations demonstrate

[9]For this step, only the test set was manually corrected

the importance of localization expressed in relation to other places. The 3,000 *Other* relations require in-depth analysis to detect new relation types.

Table 3
Number of instances per geographical class.

Class	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

Table 4
Number of instances per spatial relation.

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

Tables 5 and 6 show the top 5 countries and regions with the most spatial relations (as either the subject or object of a predicate). France is in first place, followed by neighboring countries. As our ontology does not include a class for continents, these were recognized as countries (e.g., Africa in Table 5). Regarding regions, Egypt and Ireland were misclassified.

Table 5
Top-5 countries with the most spatial relations.

Country	Nb. relations
France	3,208
Germany	2,454
Italy	2,350
Africa	1,632
England	944

Table 6
Top-5 regions with the most spatial relations.

Region	Nb. relations
Egypt	394
Ireland	364
Languedoc	334
Barbary	258
Westphalia	256

Finally, we studied the distribution of spatial relation types according to their subject. Several associations are not *a priori* possible, such as a country being the subject of a movement predicate. An in-depth analysis is necessary to identify these cases and correct them.

6. Conclusion and Perspectives

This paper proposes an automated pipeline to transform textual data from an encyclopedia into a geography-centered knowledge graph. The expert-validated dataset `GeoEDdA-TopoRel` comprises 2,750 articles (including 2,250 places) manually annotated for use at each stage of our pipeline. The application of our pipeline to the Diderot and d'Alembert's encyclopedia resulted in the EDdA graph, which includes over 15,000 geography articles, 35,000 places, and 46,000 spatial relations. It can be loaded into a *triple-store* for in-depth analysis.

This work opens up numerous perspectives. The pipeline can be improved at various levels: distances (and their potential units) could be represented by a value and a unit, thereby clarifying their semantics. A contextual comparison (via spatial relations) would be useful to disambiguate certain entities (e.g., the cities of Vienne in France and Vienna in Austria). To promote interoperability, we plan to define alignments with other spatial ontologies, such as GeoNames, or more general ones like DBpedia and Wikidata. Another perspective focuses on diachrony, i.e., the evolution of knowledge over time, for example, for the comparison of different encyclopedias or different editions of the same encyclopedia or dictionary.

Acknowledgments

The authors thank the Computer Science Federation of Lyon (FIL) of the CNRS for financially supporting the ECoDA research project.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini in order to: grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Hinzmann, J. Röttgermann, A. Klee, M. Steffes, C. Schöch, The French enlightenment novel as a graph? potentials and challenges in the construction of a knowledge network, in: 6th International Conference on Graphs and Networks in the Humanities, volume 10, 2022, p. 9.

[2] M. Wick, T. Boutreux, E. Nauer, The GeoNames geographical database, Technical Report, Geonames, 2007.

[3] R. G. Raskin, M. J. Pan, Knowledge representation in the semantic web for Earth and environmental terminology (SWEET), Computers & Geosciences 31 (2005) 1119–1125.

[4] H. M. Rawsthorne, Creation of geospatial knowledge graphs from heterogeneous sources, Theses, Université Gustave Eiffel, 2024. URL: https://theses.hal.science/tel-04599846.

[5] V. de Paiva, A. Rademaker, Towards a brazilian history knowledge graph, arXiv preprint arXiv:2403.19856 (2024).

[6] T. Hagen, F. Jannidis, A. Witt, Word sense alignment and disambiguation for historical encyclopedias, in: Graphs and Networks in the Humanities 2022. Technologies, Models, Analyses, and Visualizations, Graphen & Netzwerke; AG des Verbandes Digital Humanities, 2022, p. 7. URL: https://encycnet.github.io/.

[7] R. Filgueira, frances: a deep learning nlp and text mining web tool to unlock historical digital collections: a case study on the Encyclopaedia Britannica, in: 2022 IEEE 18th International Conference on e-Science (e-Science), IEEE, 2022, pp. 246–255. URL: https://zenodo.org/records/13919115.

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

[9] Y. Lairgi, L. Moncla, R. Cazabet, K. Benabdeslem, P. Cléau, itext2kg: Incremental knowledge graphs construction using large language models, in: International Conference on Web Information Systems Engineering, Springer, 2024, pp. 214–229.

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

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

[12] H. M. Rawsthorne, N. Abadie, E. Kergosien, C. Duchêne, É. Saux, 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, 2021, pp. 21–30.

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

[14] M. Lefrançois, A. Zimmermann, Supporting arbitrary custom datatypes in RDF and SPARQL, in: European Semantic Web Conference, Springer, 2016, pp. 371–386.

[15] L. Moncla, H. Zeghidi, Token and Span Classification for Entity Recognition in French Historical Encyclopedias, Technical Report arXiv preprint arXiv:2506.02872, LIRIS, 2025. URL: https://arxiv.org/pdf/2506.02872.
[16] N. Gali, R. Mariescu-Istodor, P. Fränti, Similarity measures for title matching, in: 2016 23rd International Conference on Pattern Recognition, IEEE, 2016, pp. 1548–1553.
[17] M. Aurnague, L. Vieu, A. Borillo, La représentation formelle des concepts spatiaux dans la langue, Masson, 1997, pp. 69–102. URL: https://arxiv.org/pdf/1003.4894.