Construction automatique d'un graphe de connaissances géographiques à partir d'entrées encyclopédiques

Bin YANG

9 septembre 2025

Laboratoire d'InfoRmatique en Images et Systèmes d'information (INSA Lyon)

Sous la supervision de

Ludovic Moncla, INSA lyon

Fabien Duchateau, Université Claude Bernard Lyon 1

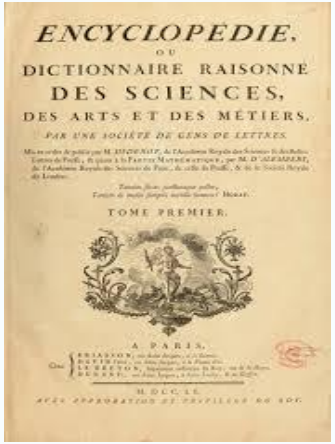
Frédérique Laforest, INSA Lyon

Contexte

- **Projet ECoDA** (2025-2026) : Étude sur les évolutions du savoir géographique dans les anciens dictionnaires (***Encyclopédie de Diderot et d'Alembert*** (1751-1772), ***Dictionnaires de Trevoux*** (1704-1771))

Objectifs

- **Extraction d'entités géographiques et de relations spatiales**
- **Construction du graphe de connaissances en RDF**


ENCYCLOPEDIE,
OU
DICTIONNAIRE RAISONNÉ
DES SCIENCES,
DES ARTS ET DES MÉTIERS,
PAR UNE SOCIÉTÉ DE GENS DE LETTRES.
TOME PREMIER.
A PARIS,

EDdA (1751-1772)

GRENOBLE, (*Géogr.*) ancienne ville de ***France***, avec un évêché suffragant de ***Vienne***, & un parlement érigé en 1493 par Louis XI. qui n'étoit encore que dauphin ; mais son pere ratifia cette érection deux ans après. *Grenoble* est sur ***l'Isere***, *à onze lieues S O. de* ***Chambéri***, *quarante-deux N. O. de* ***Turin***, *seize S. E. de* ***Vienne***, *cent vingt-quatre S. O. de* ***Paris***. *Long.* suivant Harris, *23d. 31'. 15".* suivant Cassini, *23d. 14'. 15". latit 45d. 11'.*

PLAN

1 Modélisation du graphe

2 Peuplement du graphe

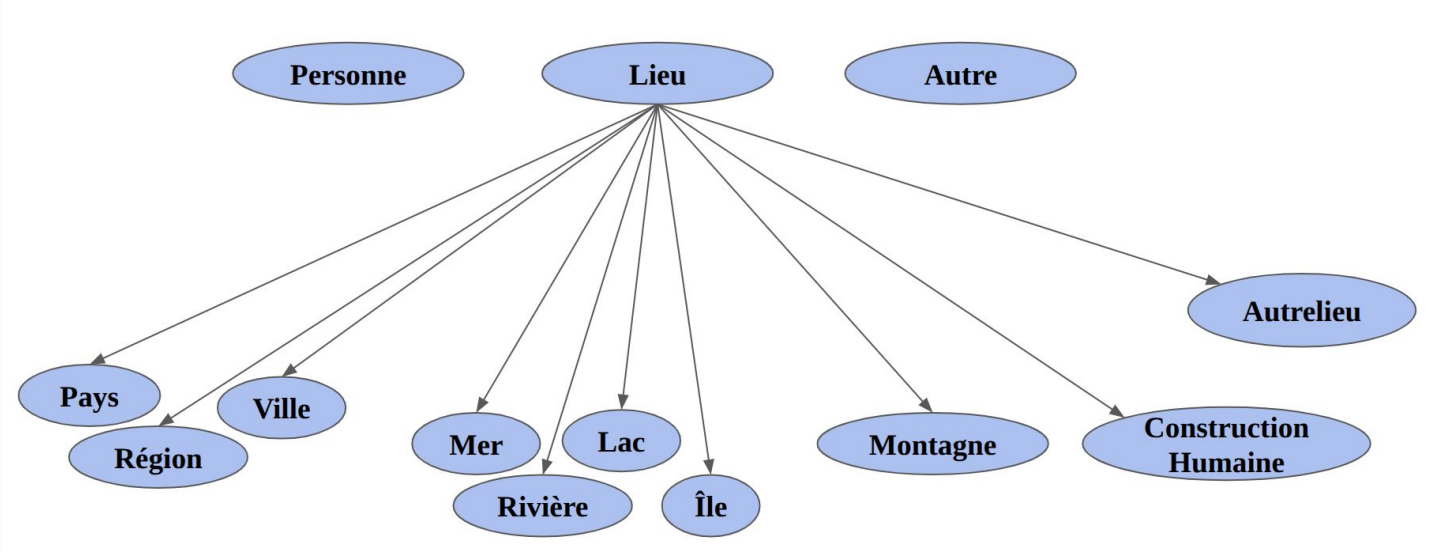
3 Évaluation du graphe

4 Conclusion et perspectives

1 Modélisation du graphe

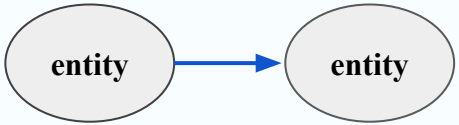
L'ontologie spatiale

- Modèle formel qui représente et structure des concepts géographiques et des relations spatiales



M. Wick, T. Boutreux, and E. Nauer (2007). *The Geonames geographical database*. http://www.geonames.org
R. Laurini (Lyon, France) & O. Kazar (Biskra, Algeria)(2016). *Geographic Ontologies: Survey and Challenges.* https://perso.liris.cnrs.fr/rlaurini/mri-pdf/5.pdf

Relations dans l'ontologie spatiale



Inclusion: "dans", "ville de"...
Adjacence: "à côté de", "est adjacent à"...
Orientation: "au sud de"...
Distance: "à 20 lieues de"...
Mouvement: "se jette dans"...
Crosses: "sur la rivière", "traverse la ville"...
Autre_relation: "par", "au dessous de"...

a_longueur: "2km"...
a_surface: "25 kilo mètres carés"...
a_longitude: "Long. 23.31."...
a_latitude: "Lat. 45.11."...

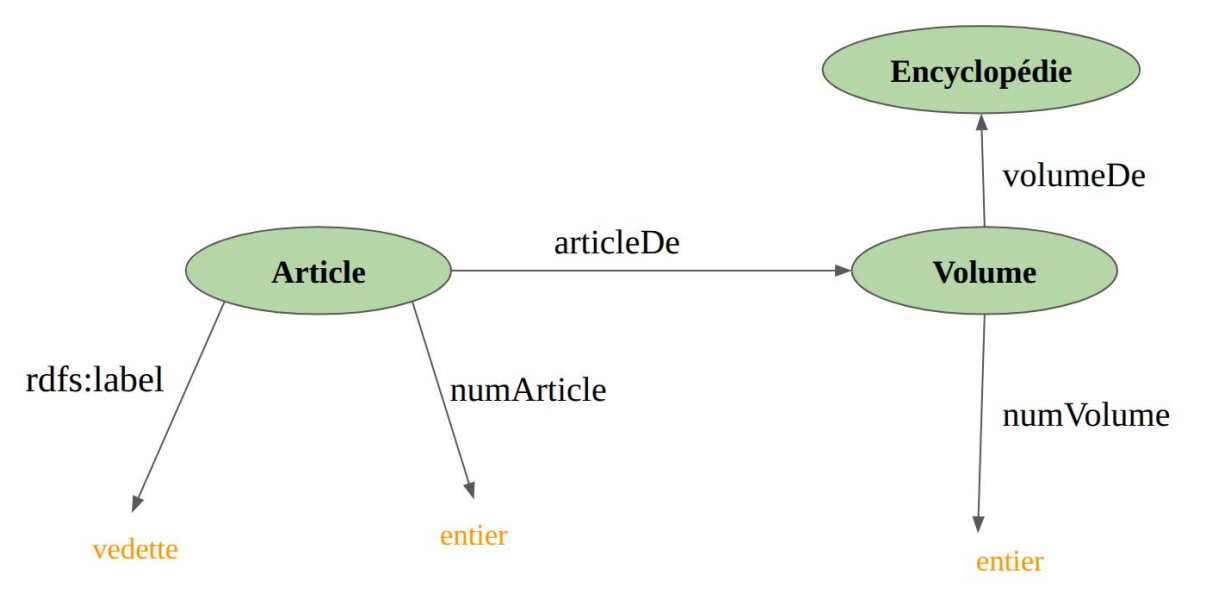
M. Wick, T. Boutreux, and E. Nauer (2007). *The Geonames geographical database*. http://www.geonames.org
R. Laurini (Lyon, France) & O. Kazar (Biskra, Algeria)(2016). *Geographic Ontologies: Survey and Challenges.* *https://perso.liris.cnrs.fr/rlaurini/mri-pdf/5.pdf*

L'ontologie de provenance

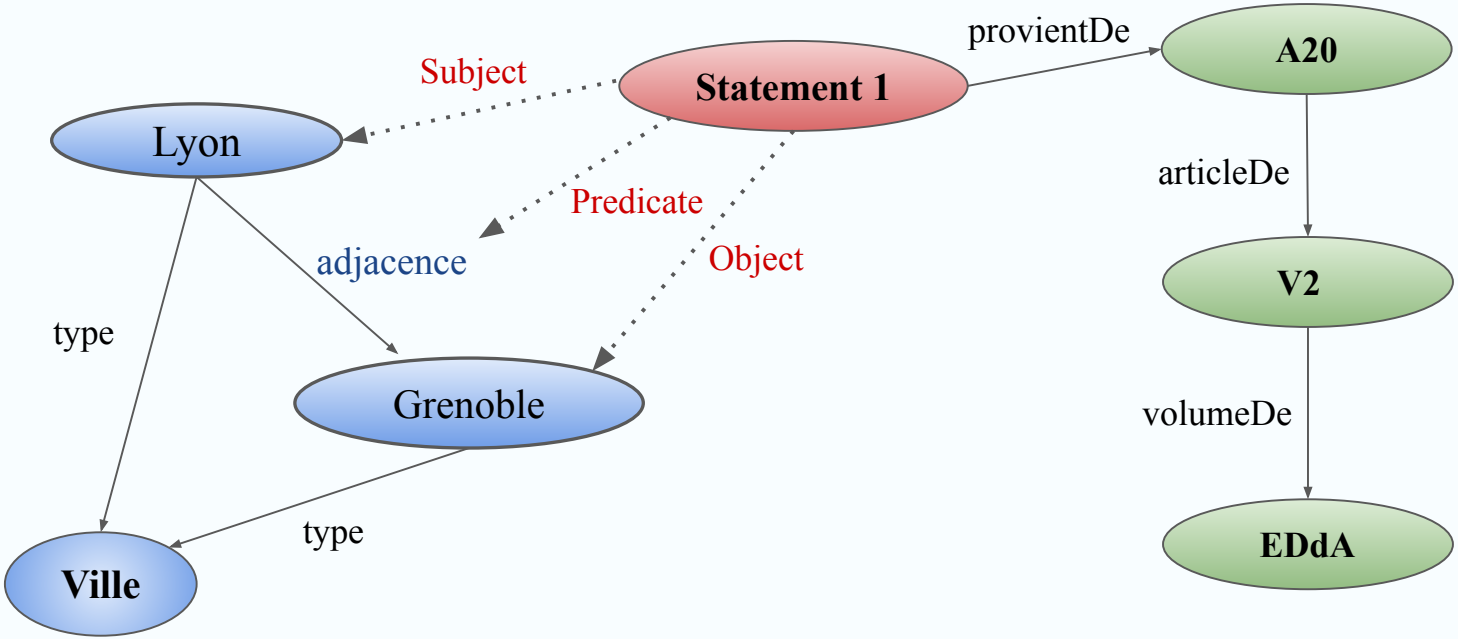
- Modèle formel qui réprésente d'où viennent des informations extraites

  Classes : Encyclopédie, Volume, Article

  Relations: articleDe, volumeDe, numArticle, numVolume



Exemple de graphe



2 Peuplement du graphe

- ➔ Prétraitement d’articles
- ➔ Extraction d’information

Illustration simplifiée du pipeline

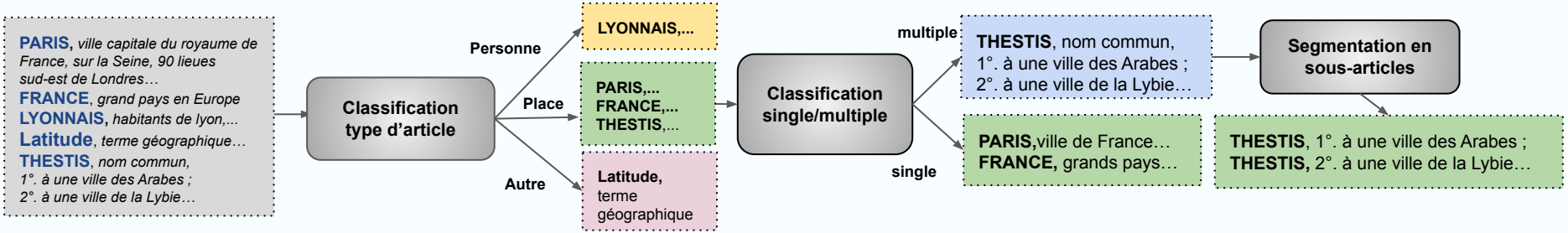


Illustration simplifiée du pipeline

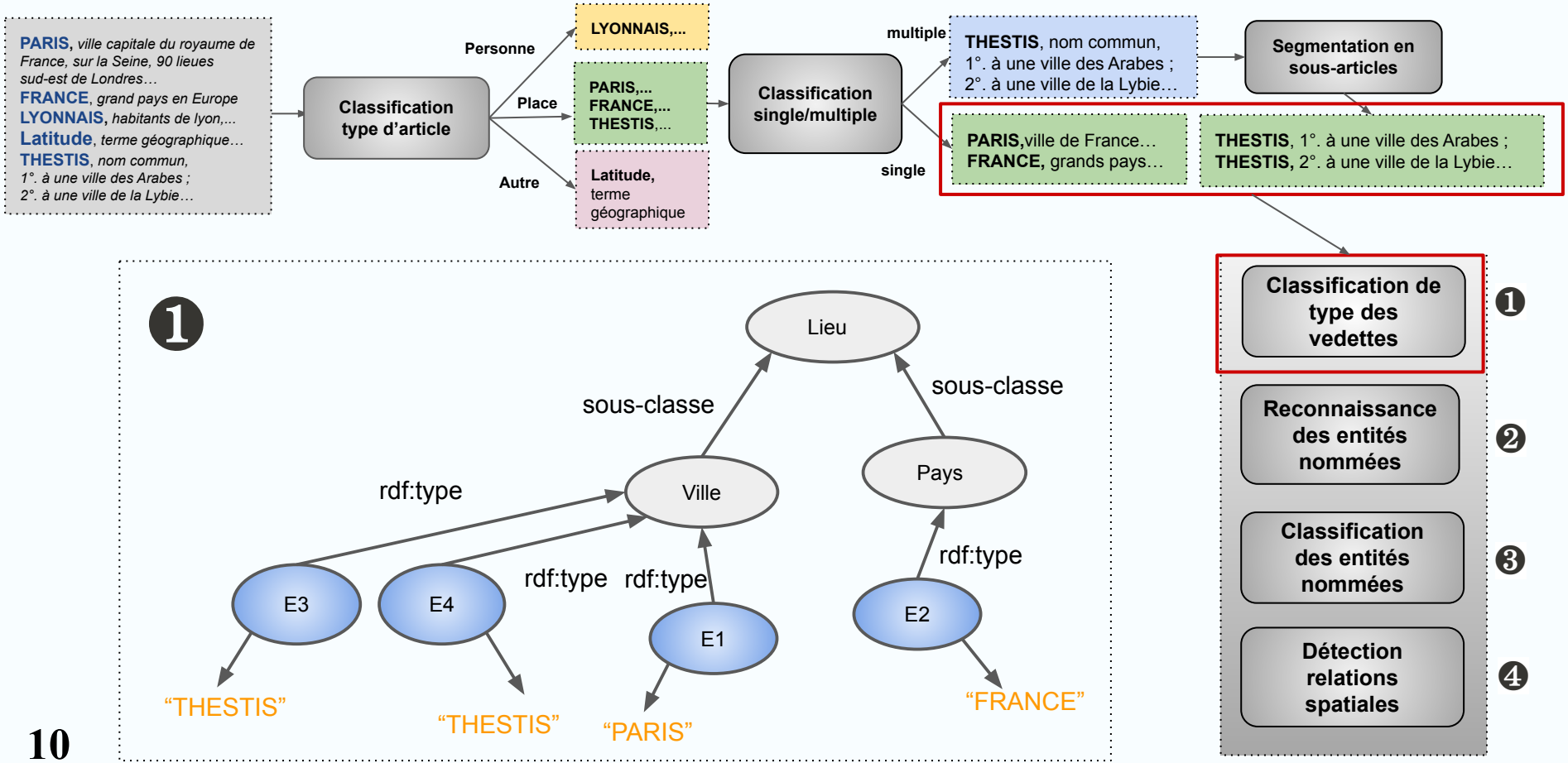
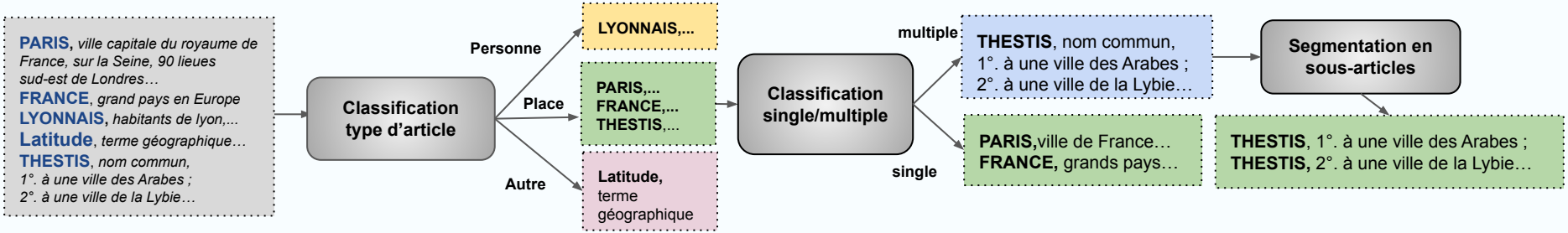
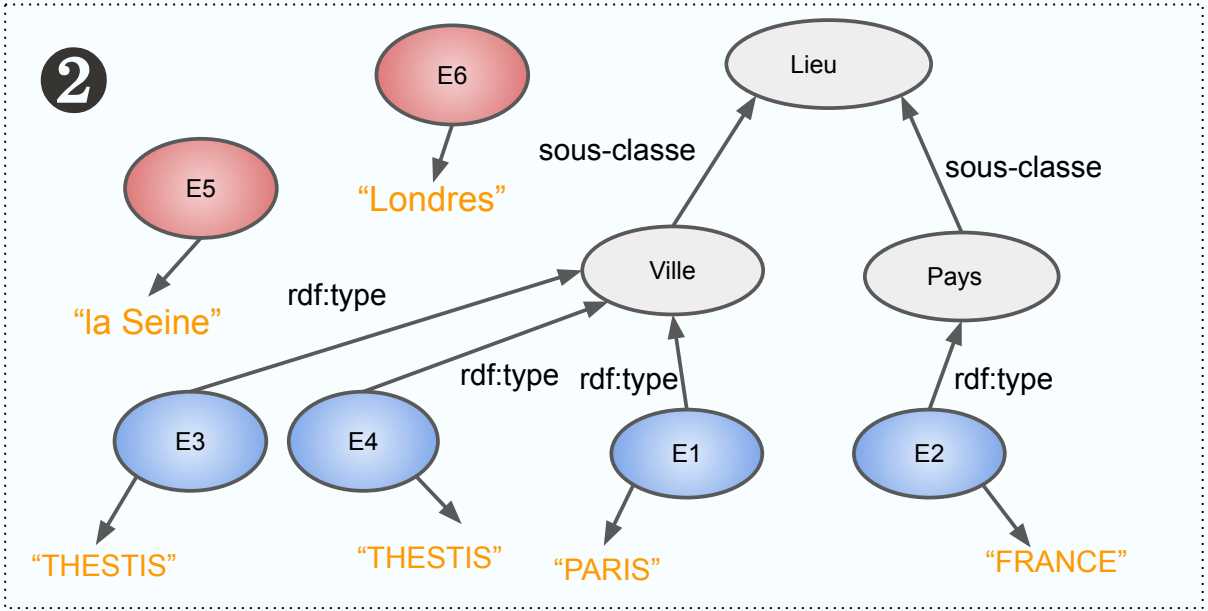


Illustration simplifiée du pipeline





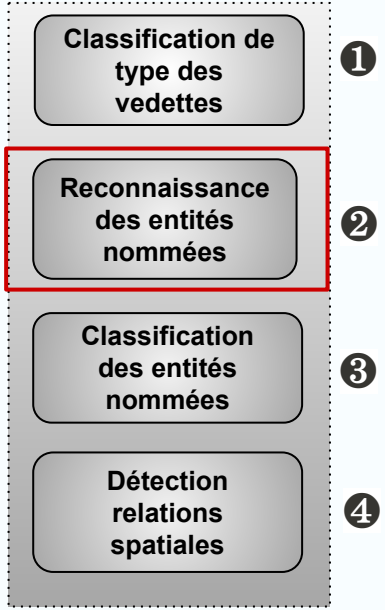
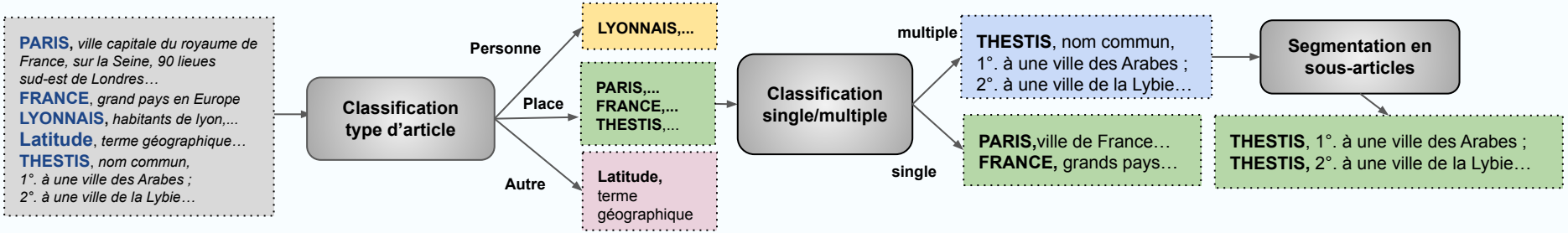
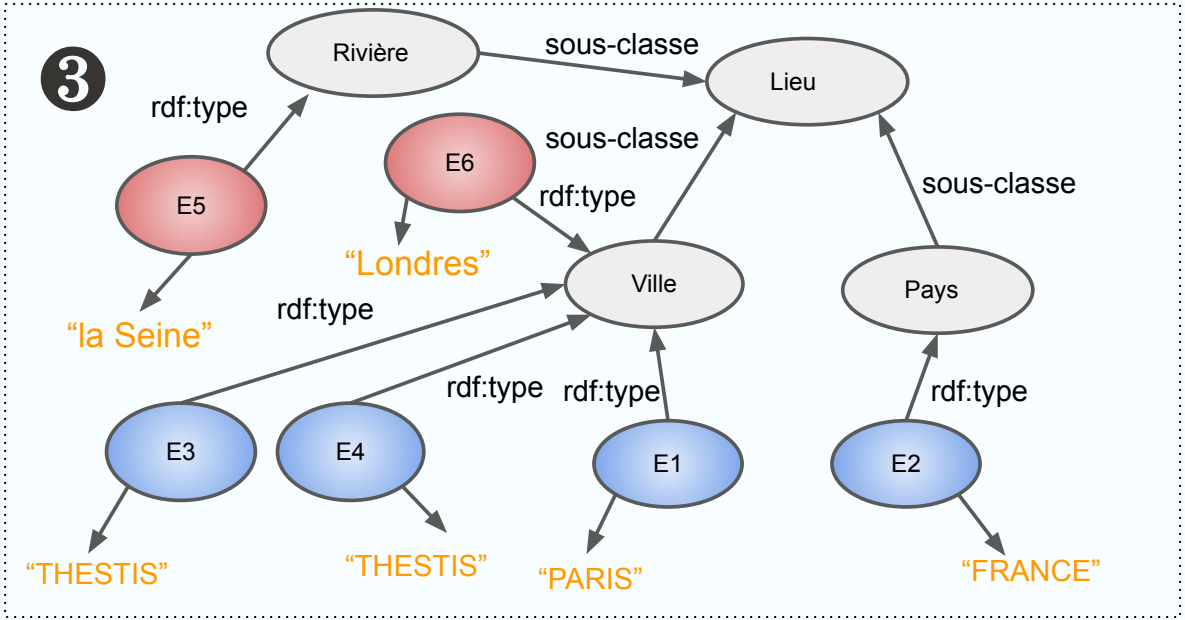


Illustration simplifiée du pipeline





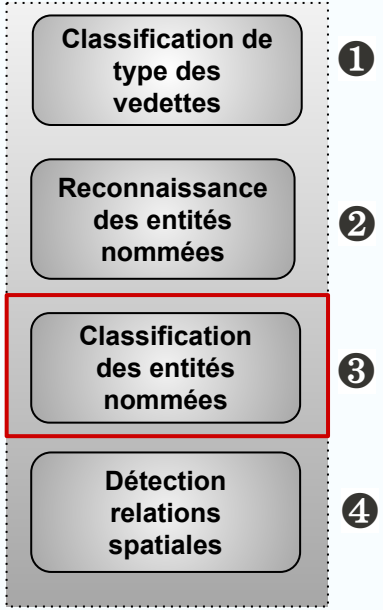
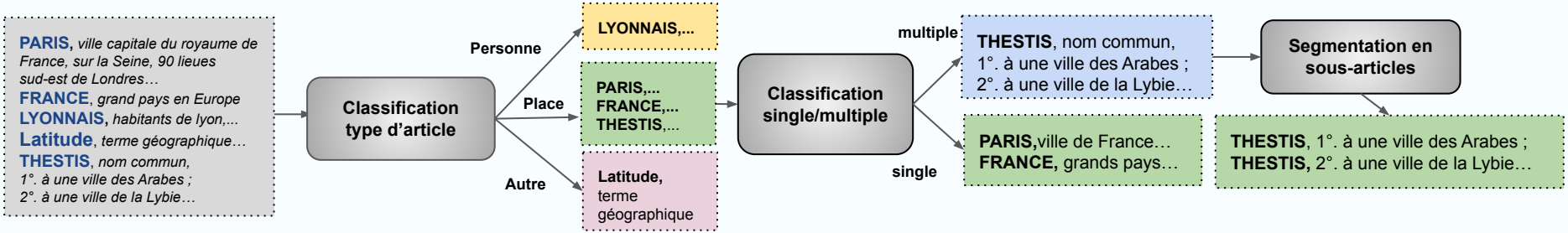
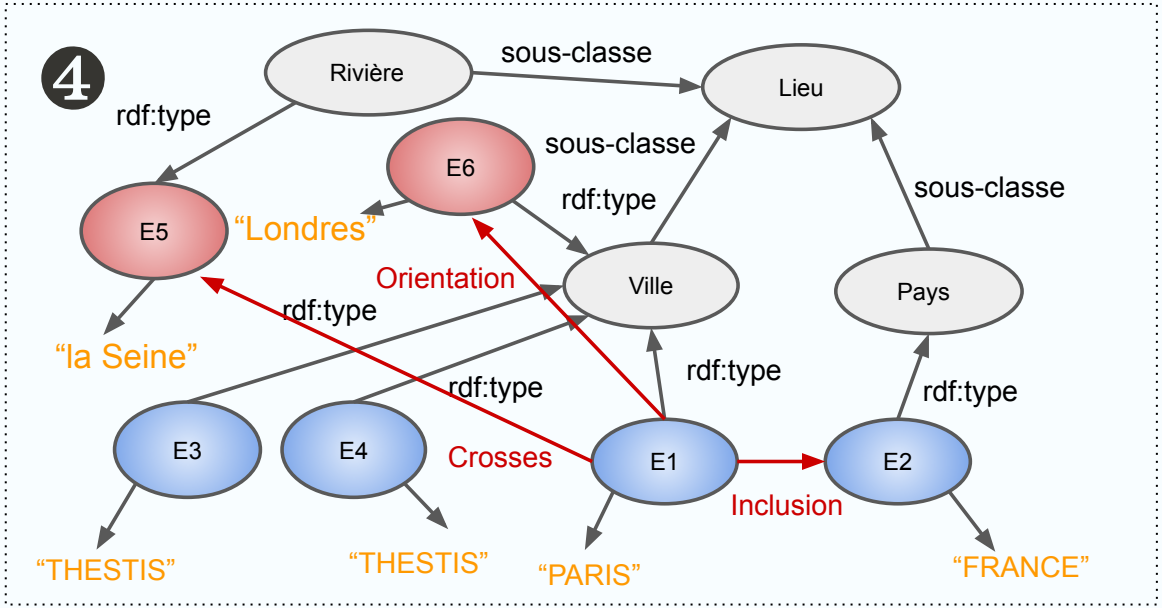
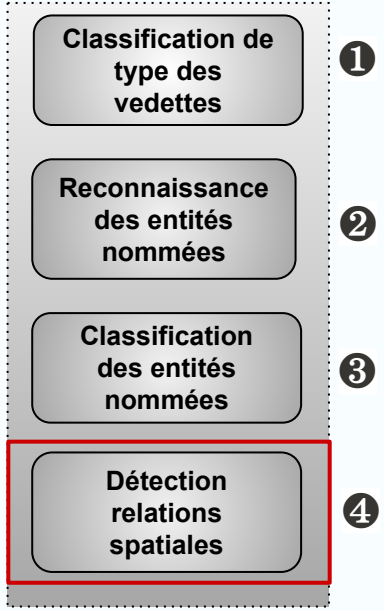


Illustration simplifiée du pipeline







Classification type d'article

Classification single/multiple

fine-tuning
Bert-base-multilingual-cased

Segmentation en sous-articles

few-shot prompting

Classification type de lieu des vedettes

Reconnaissance des entités nommées

Classification des entités nommées

Détection relations spatiales

Dataset GeoEDdA-TopoRel

Lieu (2250)

Personne(250)

Autre (250)

single (2000)

multiple (250)

1012 Ville	**59** Construction Humaine
263 Île	**51** Mer
188 Région	**51** Autrelieu
181 Rivière	**43** Pays
114 Montagne	**40** Lac

Dataset GeoEDdA-TopoRel: https://huggingface.co/datasets/GEODE/GeoEDdA-TopoRel

Classification type d'article

Précision: 97.04%
Rappel: 97.09%
F-mesure: 96.91%

Classification single/multiple

Précision: 97.75%
Rappel: 97.78%
F-mesure: 97.76%

Segmentation en sous-articles

mistral-7b : 75.37%
llama3-70b: 87.32%
gpt4-turbo: 93.27%
gpt4.1-mini: 89.19%

Classification type de lieu des vedettes

Dataset GeoEDdA-TopoRel

Lieu (2250)
Personne(250)
Autre (250)

single (2000)
multiple (250)

1012 Ville	**59** Construction Humaine
263 Île	**51** Mer
188 Région	**51** Autrelieu
181 Rivière	**43** Pays
114 Montagne	**40** Lac

Reconnaissance des entités nommées

Classification des entités nommées

Exécution sur l'ensemble de l'EDdA:

15384 articles "Géographie" ⟹ **14204** Lieu

13476 single ⋯⋯ **15453**

728 multiple ⟹ **1977** sous-articles

Détection relations spatiales

Dataset GeoEDdA-TopoRel: https://huggingface.co/datasets/GEODE/GeoEDdA-TopoRel

Classification type d'article

Classification single/multiple

Segmentation en sous-articles

Classification type de lieu des vedettes

Reconnaissance des entités nommées

Classification des entités nommées

Détection relations spatiales

bert-base-multilingual-cased-place-entry-classification

	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%**

Précision, Rappel et F-mesure de la classification de type des vedettes

Exemples d'erreur :

-"promontoire de l'île"

Île ✗ Autrelieu ✓

- "lieu des Pays-bas"

Pays ✗ Autrelieu ✓

- "temple"

Autrelieu ✗ Construction Humaine ✓

https://huggingface.co/GEODE/bert-base-multilingual-cased-place-entry-classification

Classification type d'article

Classification single/multiple

Segmentation en sous-articles

Camembert-base-edda-span-classification (GeoEDdA)

text string · lengths 22 560 84.6%	meta dict	tokens list · lengths 5 111 81.8%	spans list · lengths 8 16 29.2%

ILLESCAS, (Géog.) petite ville d'Espagne, dans la nouvelle Castille, à six lieues au sud de Madrid.	{
	"volume": 8,
	"head": "ILLESCAS",
	"author": "unsigned",
	"domain_article": "Géographie",
	"domain_paragraph": "Géographie",
	"article": 2637,
	"paragraph": 1
	}
		[
		{
		"text": "ILLESCAS",
		"start": 0,
		"end": 8,
		"id": 0,
		"ws": false
		},
		{
		"text": ",",
		"start": 8,
		"end": 9,
		"id": 1,
		"ws": true
		},
		{
		"text": "(",
			[
			"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

Exemple d'output du modèle

Classification type de lieu des vedettes

Reconnaissance des entités nommées

Classification des entités nommées

Détection relations spatiales

Moncla, L. and Zeghidi, H. (2025) *Token and Span Classification for Entity Recognition in French Historical Encyclopedias* https://doi.org/10.48550/arXiv.2506.02872

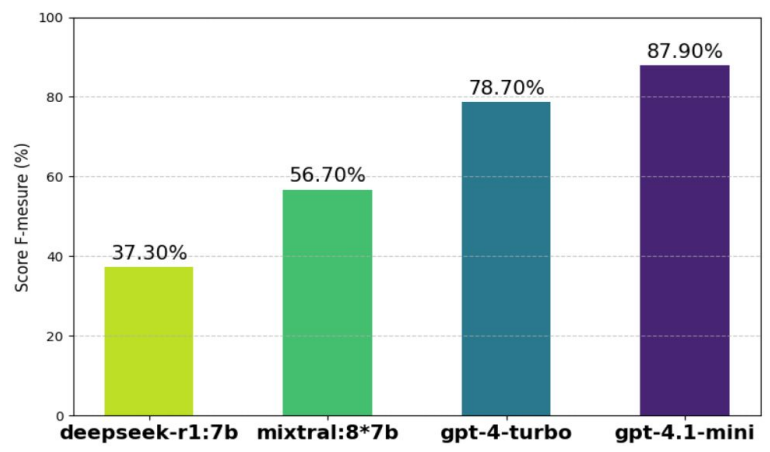
Dataset GeoEDdA: https://huggingface.co/datasets/GEODE/GeoEDdA-NER

Classification type d'article

Classification single/multiple

Segmentation en sous-articles

➢ *few-shot prompting*



F-mesure de la classification de type des entités par LLMs

➢ *Bert based fine-tuning*

	précision	rappel	f-mesure	weighted accuracy
4-gramme précédents	82.53%	82.10%	82.03%	82.10%
4-gramme précédents et suivants	83.99%	83.32%	83.28%	83.32%
5-gramme précédents et suivants	85.12%	84.85%	83.80%	**84.85%**
8-grammes précédents et suivants	85.16%	84.54%	84.49%	84.54%

Précision, Rappel et F-mesure de la classification de type des entités de N-grammes

Classification type de lieu des vedettes

Reconnaissance des entités nommées

Classification des entités nommées

Détection relations spatiales

Classification type d'article

Classification single/multiple

Segmentation en sous-articles

➢ **Bert-base-multilingual-cased-classification-relation**

	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%**

Précision, Rappel et F-mesure de classification de type de relations

Seule-distance: *"90 lieues de"*

Seule-Orientation: *"à l'est de"*

Cas mixed: *"trois lieues au sud de"*

Regex

distance ou/et Orientation

Classification type de lieu des vedettes

Reconnaissance des entités nommées

Classification des entités nommées

Détection relations spatiales

➢ **Relation-linking**

- Few-shot prompt
- Approche probabiliste

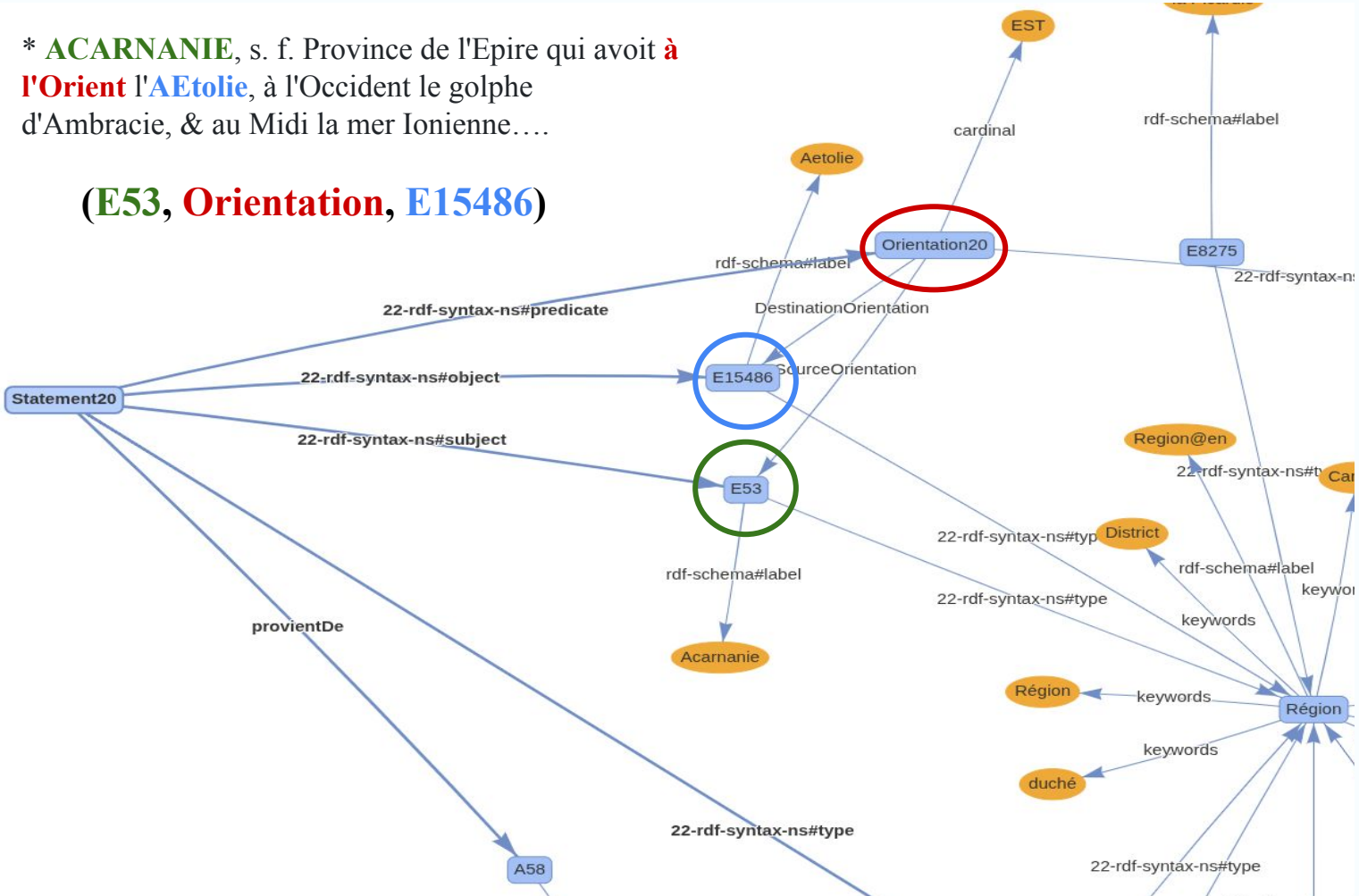
	nb_correct	nb_total	précision
llama3 :70b	106	124	85.48%
gpt-4-turbo	113	124	91.12%
gpt-4.1-mini	113	124	91.12%
approche probabiliste	115	124	92.74%

Précision de relation linking de LLMs et de l'approche probabiliste

Exemple du graphe produit

* **ACARNANIE**, s. f. Province de l'Epire qui avoit **à l'Orient** l'**AEtolie**, à l'Occident le golphe d'Ambracie, & au Midi la mer Ionienne….

(E53, Orientation, E15486)



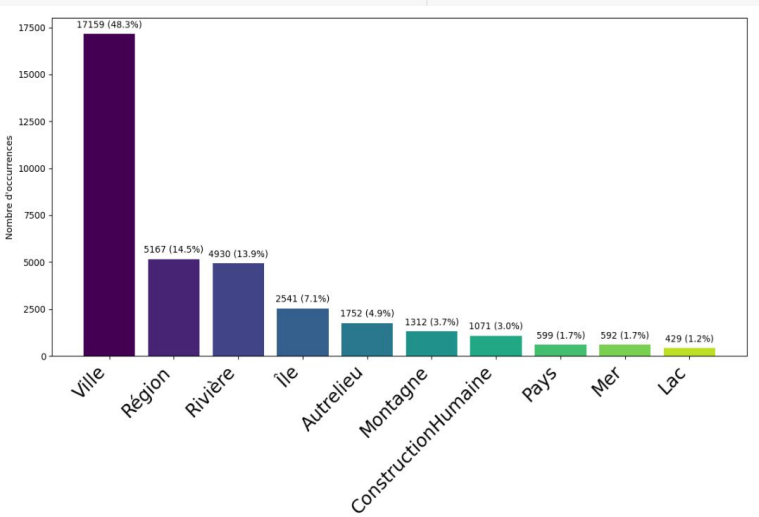
3 Évaluation du graphe

Description du graphe

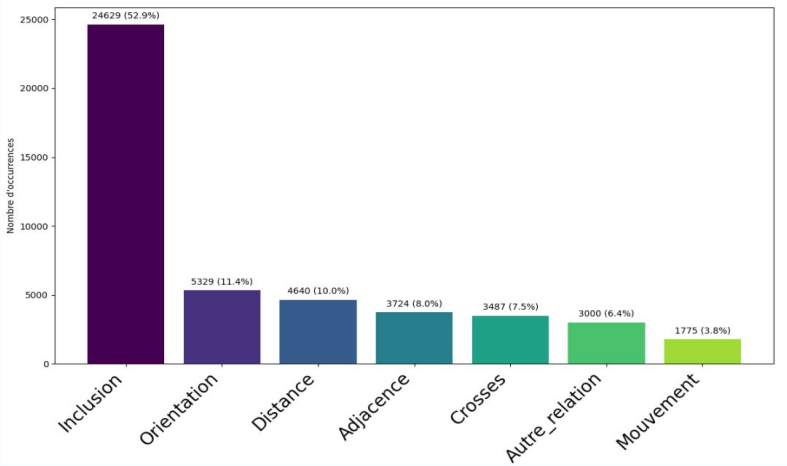
classe	nombre d'instances (ou sous-classes)
Encyclopédie	1
Volume	17
Article	15,453
Lieu	10
Entité (URI)	32,047
Relation (Statement)	46,584
Entité ambigüe	395

M. Perry and J. Herring (2021). *Ogc geosparql - a geographic query language for rdf data.*https://www.ogc.org/standards/geosparql

Distribution des types d'entités et des types de relations



Distribution des types d'entités



Distribution des types de relations

➢ **Entités dominantes:**

- Ville (48.3%)
- Région (14.5%)
- Rivière(13.9%)

≈ 80%

➢ **Relations dominantes:**

- Inclusion (52.9%)
- Orientation (11.4%)
- Distance (10.0%)
- Adjacence (8.0%)

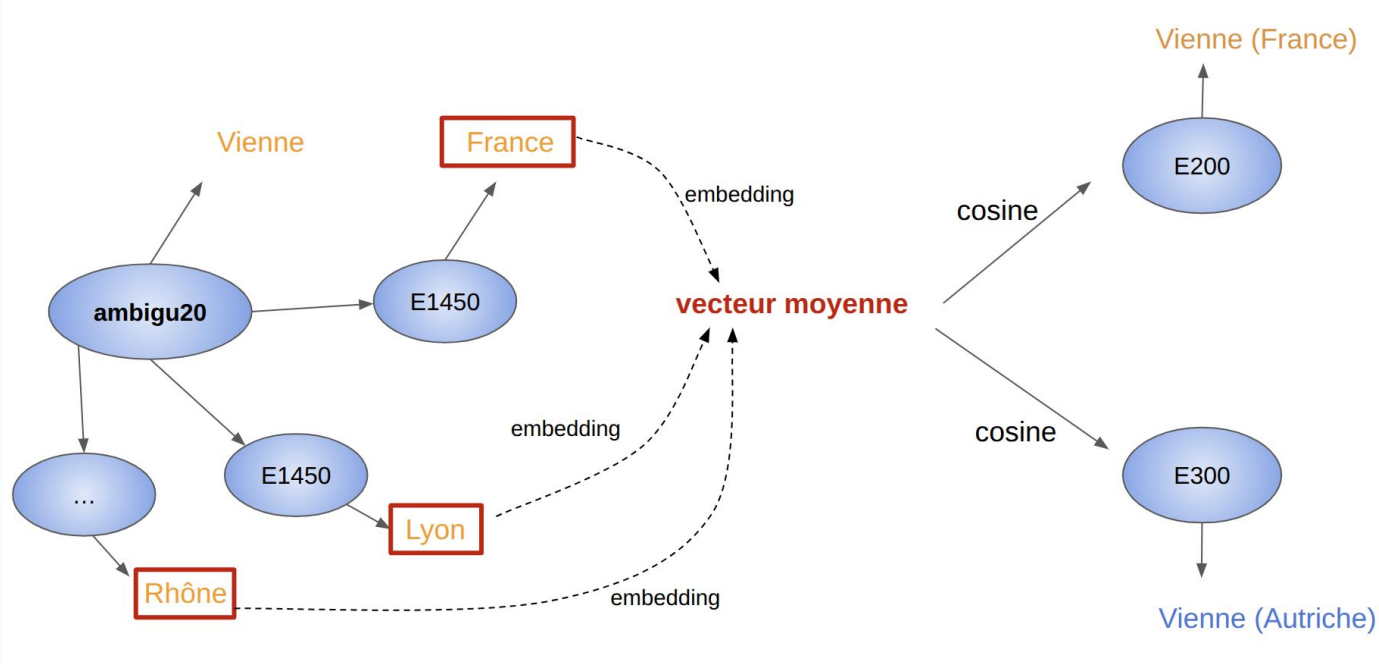
4 Conclusion et perspectives

Conclusion

- Définition de l'ontologie spatiale et l'ontologie de provenance dédiées
- Création du dataset *GeoEDdA-TopoRel*
- Entraînement de modèles de classification adaptés (type d'articles, single/multiple, type d'entités géographiques, type de relations spatiales)
- Une approche complète pour la construction de graphes de connaissances géographiques à partir de textes encyclopédiques
- Un pipeline automatique → support pour les travaux futurs en linguistique computationnelle, en histoire numérique et en sciences de l'information géographique ancienne

Perspectives

➢ **Désambiguisation des entités "ambigües"** : calcul de similarité sémantique des noeuds voisins



Perspectives

- Optimisation du modèle de classification de types des entités nommées
- Intégration d'un mécanisme de validation ontologique (e.g. un Pays ne peut pas être le sujet d'une relation de type "Mouvement")
- Application du pipeline à d'autres dictionnaires ou encyclopédies (e.g. Trévoux)
  - Étude diachronique sur les évolutions du savoir géographique
  - Comparaison entre dictionnaires ou entre éditions

Merci de votre attention et écoute!