



RAPPORT DE STAGE
MASTER 1 GÉOGRAPHIES NUMÉRIQUES MENTION GÉOMATIQUE
UNIVERSITÉ JEAN MONNET, SAINT-ÉTIENNE

**Une base de données géographiques pour
l'Encyclopédie de Diderot et d'Alembert**

Structure d'accueil :

CNRS UMR EVS – Université Jean Monnet, Saint-Étienne

CNRS UMR LIRIS – INSA Lyon

Auteur : GIMENEZ SARAIVA Matheus

Tutrice pédagogique : CUNTY Claire

Tuteurs professionnels : JOLIVEAU Thierry et MONCLA Ludovic

Stage effectué du 1^{er} Mars au 31 Juillet 2021

Promotion 2022-2023



Remerciements

Avant tout développement sur cette expérience professionnelle, je souhaite exprimer ma gratitude envers toutes les personnes qui ont rendu possible l'accomplissement de ce stage et qui ont joué un rôle essentiel dans la réalisation de cette expérience enrichissante.

Je remercie tout particulièrement Thierry Joliveau et Ludovic Moncla pour leur accompagnement, leur approche pédagogique et leur bienveillance tout au long de la durée de mon stage.

Merci à mes proches, famille et ami.e.s, qui m'ont entourée et soutenue durant la rédaction de ce mémoire, et avec lesquels les échanges sont toujours constructifs.

Table des matières

1.	Introduction	8
2.	Présentation de la structure de l'accueil et contexte du projet	10
3.	Matériels	11
3.1.	Les données utilisées	11
3.2.	Jupyter Notebook, VS Code et GitHub.....	11
3.3.	Librairies Python	11
3.4.	Outil géoweb.....	12
4.	Méthodes et outils	13
4.1.	Regex – Expressions régulières.....	15
4.2.	Création des enveloppes	17
4.2.1.	Enveloppes de l'encyclopédie	17
4.2.2.	Pays modernes et continents.....	18
4.3.	Perdido.....	19
4.4.	Jointure entre les deux jeux de données.....	21
4.5.	Calcul de la distance géodesique	21
4.5.1.	Arcpy – cursor	22
4.5.2.	Librairie GeoPandas, Geopy et Shapely	23
4.6.	Classification des articles	25
4.7.	Web Scraping pour corriger l'URL des articles	26
4.8.	Géocodage.....	27
4.9.	Jointure spatiale.....	30
4.10.	Priorité de désambiguïsation.....	30
4.11.	Traitement des données et mise en ligne des couches.....	32
4.12.	Experience Builder	33
5.	Évaluation des résultats	38
5.1.	Statistiques	38
5.2.	Analyse des articles par pays et continent	39
6.	Résultats et discussions	40
6.1.	Enveloppes EDdA	40
6.2.	Perdido.....	40
6.2.1.	Indicateur de la distance géodésique.....	40
6.2.2.	Analyse des 125 articles	46
6.2.3.	Analyse des articles par pays et continent.....	48
6.3.	Géocodage	49
7.	Application Experience Builder	51

8.	Conclusion	53
9.	Références bibliographiques	55
10.	Annexes	56
	Annexe A - Enveloppes EDdA	56
	Annexe B – Tableau des observations d’analyse des 125 articles	57

Table des illustrations

Figure 1. Ordre de réalisation des grands objectifs.	13
Figure 2. Pipeline des procédures réalisés pendant le stage Geode.....	14
Figure 3. Configuração das variáveis para criação de ferramenta no ArcGIS Pro.	22
Figure 4. Widget « Filtre » de colonnes type 1.....	34
Figure 5. Widget « Liste » pour filtrer la couche des articles selon l'enveloppe de correspondance.	35
Figure 6. Widget de « URL » pour représenter les pages ARTFL des articles.	35
Figure 7. Widget de coordonnées.	35
Figure 8. Widget de « édition » des articles.....	36
Figure 9. Diagramme en violon des variables seuil d'égalité et distance géodésique des tous les articles.	42
Figure 10. Histogramme de la variable seuil d'égalité de tous les articles.	43
Figure 11. Nuage des points entre les variables seuil d'égalité (axe X) et distance géodésique (axe Y) de tous les articles.....	43
Figure 12. Diagramme en violon des variables seuil d'égalité et distance géodésique des articles avec des coordonnées anciennes.	45
Figure 13. Histogramme de la variable seuil d'égalité des articles avec des coordonnées anciennes.	46
Figure 14. Nuage des points entre les variables seuil d'égalité (axe X) et distance géodésique (axe Y) des articles avec des coordonnées anciennes.....	46
Figure 15. Carte du nombre des correspondances par limites modernes.	49

Table de tableaux

Tableau 1. Liste des opérateurs de regex.....	16
Tableau 2. Classes de priorités de désambiguïsation.....	31
Tableau 3. Informations statistiques des variables distance géodésique et seuil d'égalité de tous les articles.....	41
Tableau 4. Informations statistiques des variables distance géodésique et seuil d'égalité des articles avec coordonnées anciennes.....	44
Tableau 5. Nombre d'articles par continents.....	49

1. Introduction

L'encyclopédie de Diderot et Alembert (simplifiée en EDdA) a été écrite et éditée, entre 1751 et 1772, par Denis Diderot et Jean Le Rond d'Alembert. Il faut également noter la participation très importante de Louis de Jaucourt, responsable pour la signature de presque 30% des articles de l'ouvrage. Au total elle contient 17 volumes, 71 818 articles, 18 000 pages et 21 700 000 mots. Néanmoins, les auteurs ont délibérément négligé l'aspect de la cartographie, comme l'a exprimé Diderot de la manière suivante :

«... il étoit nécessaire de s'en tenir à la seule connoissance géographique des villes qui fût scientifique, à la seule qui nous suffiroit pour construire de bonnes cartes des tems anciens, si nous l'avions, & qui suffira à la postérité pour construire de bonnes cartes de nos tems, si nous la lui transmettons... »

Diderot. Article « Encyclopédie ». Vol. V EDdA

Afin de compléter la dimension géographique de l'Encyclopédie, le Projet GEODE vise à comprendre les connaissances géographiques de l'époque où l'encyclopédie a été écrite. Pour ce faire une méthodologie informatique a été employée : numérisation de l'encyclopédie à partir de la Reconnaissance Optique de Caractères (OCR), extraction et analyse automatique du contenu textuel, et création de bases de données géographiques.

Une des solutions développées pour extraire des informations géographiques des articles est la librairie python Perdido, créée par Ludovic Moncla et disponible sur sa page GitHub (<https://github.com/GEODE-project/perdido-geoparsing-notebook>). Perdido est capable non seulement d'identifier les entités géographiques grâce à la reconnaissance d'entités nommées (Named Entity Recognition,NER), mais aussi de les géocoder, c'est-à-dire de fournir des coordonnées géographiques correspondantes.

L'extraction des entités géographiques est une étape importante, cependant il est nécessaire de vérifier si les coordonnées trouvées sont correctes. Dans cette optique il faut reconnaître que des erreurs peuvent résider dans la numérisation de l'encyclopédie, la géocodification, voire dans l'encyclopédie elle-même.

Dans le but d'enrichir le travail réalisé, il est essentiel d'établir un mécanisme permettant la participation des utilisateurs ayant une vaste connaissance géographique de

leurs pays ou du monde. Cela permettra l'accès aux informations obtenues ici, ainsi que la correction d'éventuelles erreurs identifiées.

Ainsi, une application géoweb constitue une alternative permettant aux utilisateurs de visualiser les coordonnées géographiques des articles de l'encyclopédie et de proposer des modifications. À partir de là, il est possible d'effectuer la désambiguïsation et de compléter de manière appropriée le travail réalisé par Diderot, Alembert et Jaucourt. La désambiguïsation est un processus visant à identifier de manière non ambiguë la localisation d'un lieu. Plusieurs types d'ambiguïtés existent et dans ce travail nous nous intéresserons notamment au fait qu'un même nom de lieu puisse être attribué à plusieurs lieux. Cela se produit à cause de deux facteurs. D'une part, parce que l'encyclopédie ne fournit pas suffisamment d'informations pour déterminer la localisation moderne de ces articles. D'autre part, parce qu'il existe une ambiguïté inhérente aux toponymes, c'est-à-dire que plusieurs lieux peuvent avoir le même nom.

La désambiguïsation peut être réalisée à différents niveaux. L'utilisateur a la possibilité de signaler des erreurs dans l'extraction des coordonnées (longitude ou latitude). De plus, il est possible d'indiquer les changements historiques dans les noms des pays. En outre, il y a un espace pour fournir des commentaires sur des cas spécifiques nécessitant des corrections.

Ce stage a pour objectif d'extraire les entités géographiques avec Perdido, de trouver une façon de vérifier si l'extraction a été bonne et de mettre en place une application géoweb sur Expérience Builder.

2. Présentation de la structure de l'accueil et contexte du projet

Le projet interdisciplinaire GEODE est développé par des chercheurs en linguistique, informatique et géographie des laboratoires ICAR, LIRIS et EVS. Ce projet est actuellement financé par le LabEX ASLAN et il s'agit de la continuation du projet GéoDISCO (2019-2020), financé par la MSH Lyon St-Etienne.

L'objectif scientifique du projet GEODE est d'analyser les discours géographiques dans les encyclopédies françaises. Les encyclopédies comprises dans le corpus sont : l'Encyclopédie de Diderot, d'Alembert et Jaucourt au XVIIIe siècle (1751-1772), la Grande Encyclopédie au XIXe siècle. (1886-1902), l'Encyclopædia Universalis ou de Wikipédia au XXIe siècle.

Les objectives de ce stage sont les suivantes :

- Terminer le travail d'extraction et de validation des coordonnées anciennes des articles et de caractérisation de ces lieux : type du lieu à extraire de l'article, nombre de caractères de l'article ;
- Mettre à jour et d'améliorer la maquette de géovisualisation de ces coordonnées sur ArcGIS Online et d'explorer des méthodes de géovisualisation Opensource ;
- Mettre au point une procédure d'extraction des autres lieux vedettes cités dans les articles disposant de coordonnées pour établir un indicateur de cohérence de ces coordonnées par un indicateur de distance ;
- Rechercher les coordonnées modernes de ces lieux par l'interrogation de gazetiers.

3. Matériels

3.1. Les données utilisées

Au cours du stage, des fichiers préalablement traités par Monsieur Joliveau et Monsieur Moncla ont été utilisés. Les deux ensembles de données couvraient les articles ayant subi un processus de ROC (reconnaissance optique de caractères), ce qui a permis de manipuler le contenu individuel de chaque article. Ces informations sont disponibles sur la plateforme ARTFL (<https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/>).

Dans le fichier mis à disposition par Monsieur Joliveau, une étape préliminaire d'extraction des informations géographiques anciennes de ces articles avait déjà été effectuée manuellement. Cela incluait l'obtention de la latitude (en degrés, minutes et secondes), de la longitude (en degrés, minutes et secondes) et la conversion en coordonnées modernes.

À l'aide de la bibliothèque Perdido, Monsieur Moncla a élargi ces ensembles de données en ajoutant des colonnes contenant le contenu textuel des articles ainsi que les coordonnées extraites automatiquement.

3.2. Jupyter Notebook, VS Code et GitHub

Les codes Python utilisés pendant le stage ont été écrits dans Jupyter Notebook ou Visual Studio Code. Au début, Jupyter Notebook était l'option la plus utilisée. Cependant, après avoir eu besoin de partager les codes avec M. Moncla, Visual Studio Code est devenu plus utilisé car il était nécessaire d'intégrer GitHub. À partir de cet outil, il a été possible de travailler en cloud avec les codes.

3.3. Bibliothèques Python

Les bibliothèques Python utilisés dans le développement de ce projet ont inclus:

- Perdido : pour l'extraction des entités géographiques et la réalisation du géocodage des articles ;
- Geopandas, Geopy et Shapely : utilisés pour la manipulation et l'analyse des ensembles de données contenant des informations géométriques (shapefiles), ainsi que pour l'exécution d'opérations impliquant la géométrie, y compris les calculs de distance géodésique ;

- FuzzyWuzzy : utilisée dans l'analyse de similarité entre les mots, notamment pour évaluer l'égalité entre les termes ;
- Selenium : utilisé pour la technique de WebScraping, dans le but de corriger les URL associées aux articles ;
- Seaborn et Matplotlib : utilisées pour la création de graphiques et de visualisations pour l'analyse et la présentation des résultats.

3.4. Outil géoweb

Le choix d'Experience Builder a été influencé par deux facteurs principaux. Premièrement, l'utilisation préalable d'ArcGIS Pro a rendu la communication entre les deux outils plus facile, car ils appartiennent tous deux à l'écosystème Esri. Deuxièmement, Experience Builder présente une interface de création d'applications web intuitive, avec des fonctionnalités de glisser-déposer, évitant ainsi la nécessité d'apprendre un nouveau langage de programmation.

4. Méthodes et outils

La **Figure 1** présente un schéma contenant l'ordre dans lequel les objectifs ont été réalisés. En outre, il est également possible d'observer les outils nécessaires à la réalisation de chaque étape.

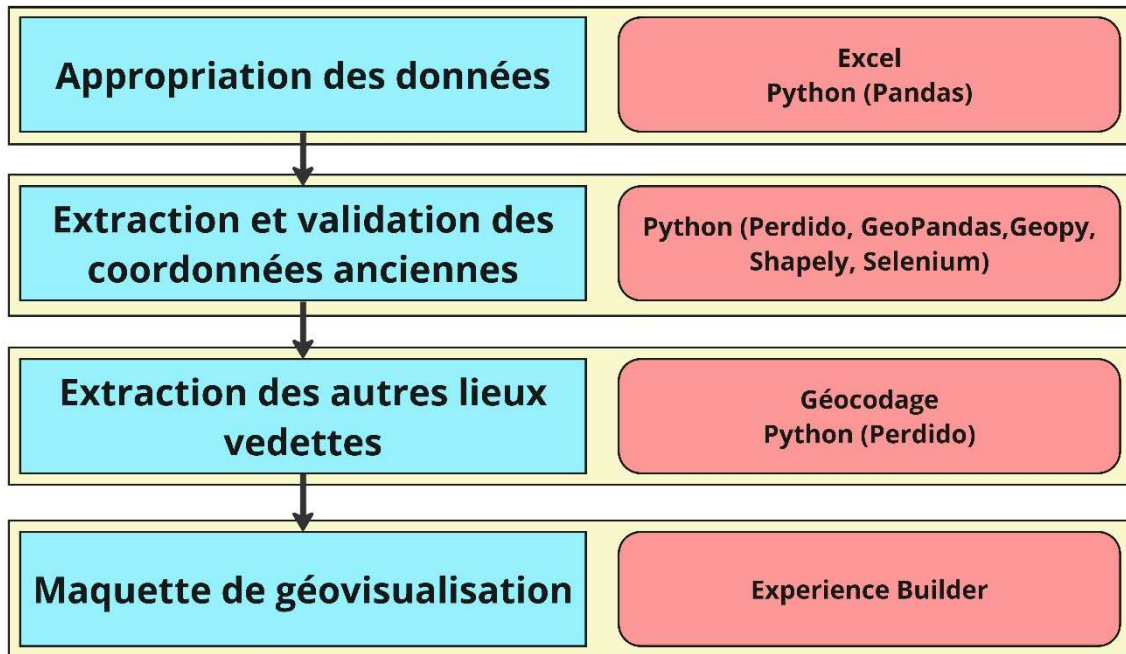


Figure 1. Ordre de réalisation des grands objectifs.

Dans la *Erro! Fonte de referênciã não encontrada.*, il est présenté le schéma complet de la réalisation de ce stage. Toutes les étapes sont représentées graphiquement dans l'ordre suivi et de plus détaillés que dans la figura précédente, en montrant le changement du nombre de lignes et colonnes des fichiers obtenus. Chaque étape sera expliquée dans les sections suivantes.

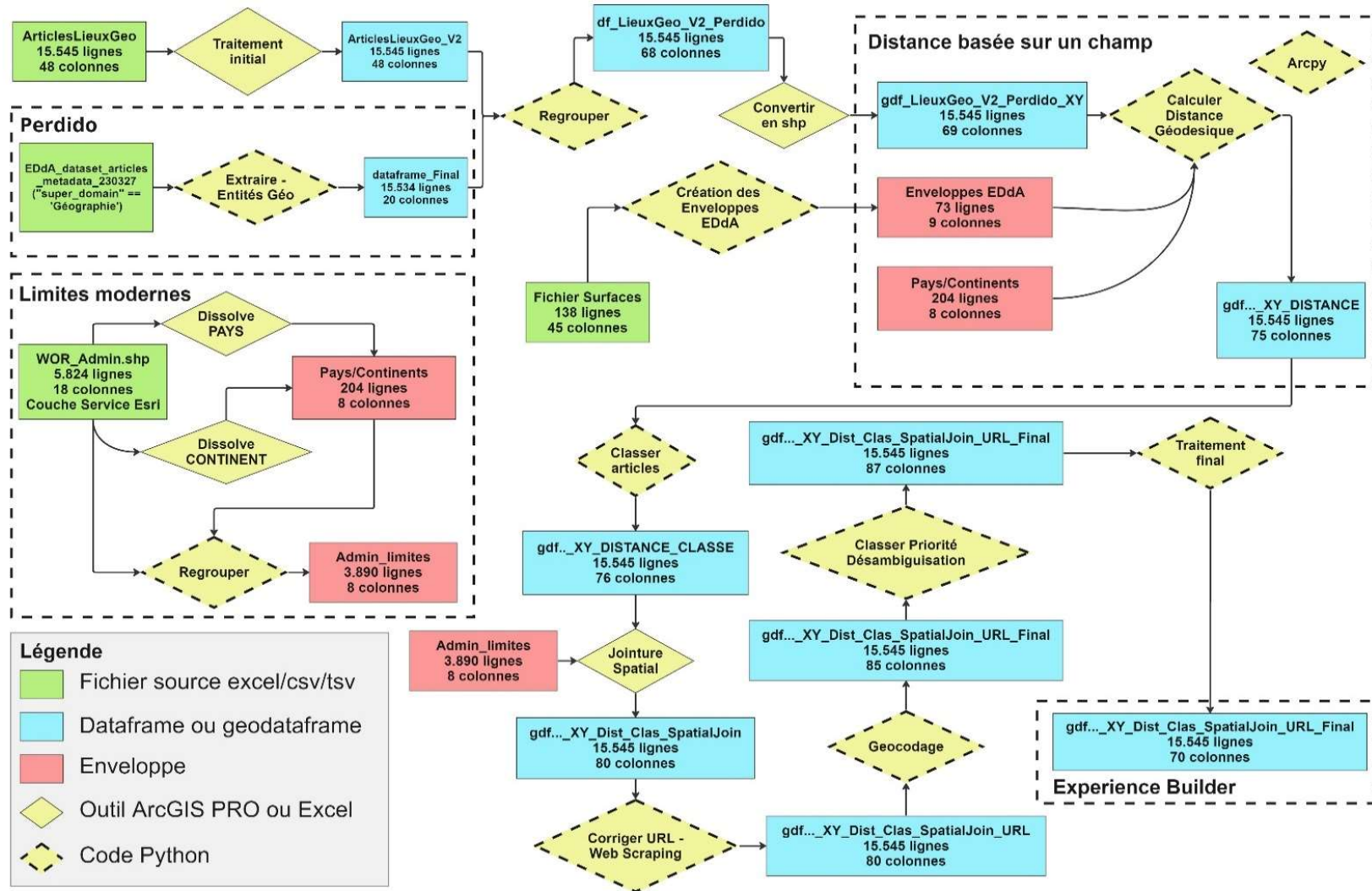


Figure 2. Pipeline des procédures réalisés pendant le stage Geode.

4.1. Regex – Expressions régulières

Les expressions régulières, ou les expressions rationnelles en français, telles que définies par la librairie Python « re », sont utilisées pour trouver des structures textuelles dans une chaîne de caractères. Cette librairie offre une plus grande variété de possibilités lorsqu'il s'agit de travailler avec des variables de type chaîne de caractères. À l'aide de celle-ci, il est possible d'extraire tous les nombres d'un texte, les dates, les adresses électroniques et tout autre type d'information, à condition de définir la structure à rechercher.

Cette structure est définie par la bibliothèque regex elle-même, qui propose plusieurs opérateurs. Ces opérateurs peuvent être combinés pour obtenir la structure d'intérêt. La liste des opérateurs utilisés est représentée dans le **Tableau 1**.

Opérateurs	Fonction
[]	Spécifiez un jeu de caractères. Par exemple : [a-z] pour extraire tous les caractères de « a » à « z » en minuscules.
\	Pour utiliser des caractères spéciaux qui fonctionnent comme des opérateurs regex. Ex : pour extraire le point final d'une phrase.
^	Recherche d'éléments au début d'une chaîne de caractères.
\$	Recherche des éléments à la fin d'une chaîne de caractères.
*	Recherche de zéro ou plusieurs répétitions d'un caractère ou d'une sous-chaîne.
+	Recherche d'une ou plusieurs occurrences d'un caractère ou d'une sous-chaîne.
?	Recherche de zéro ou d'une seule apparition d'un caractère ou d'une sous-chaîne.
	Recherche d'un caractère ou d'un autre.
{}	Définir un nombre spécifique de caractères.
[^]	Recherche d'un caractère autre que le caractère spécifié immédiatement après le ^.
()	Regrouper les règles et définir l'ordre d'application (comme en mathématiques).
.	Extraire n'importe quel caractère.
\d	Extraire n'importe quel chiffre.
\D	Extrait tout caractère qui n'est pas un chiffre.

<code>\w</code>	Extraire n'importe quel caractère alphanumérique.
<code>\W</code>	Extraire tout caractère qui n'est pas alphanumérique.
<code>\s</code>	Représenter un espace vide.
<code>\S</code>	Représenter qu'il ne s'agit pas d'un espace vide.

Tableau 1. Liste des opérateurs de regex.

Les opérateurs regex ont été utilisés pour extraire les coordonnées des articles classés comme « surface » par M. Joliveau. La création d'une structure pour extraire les coordonnées n'a pas été évidente, car tous les articles ne suivaient pas la même structure. Il y avait des cas où les coordonnées se trouvaient avant les mots « longitude » ou « latitude », et des cas où elles se trouvaient après ces mots. De plus, ces mots pouvaient être abrégés de différentes manières.

À partir de l'analyse des opérateurs et des structures possibles présentes, la structure suivante a été structurée pour extraire les coordonnées :

```
pattern = re.compile(r'(Long(?:itude)?\.|lat(?:itude)?\.)\D*([\d]+(?:\.\s?\d+)?(?:\s*\s*[\d]+(?:\.\s?\d+)?)*(?:\.\s?([\d\.])))?', re.IGNORECASE)
```

La variable "pattern" stocke la structure construite à l'aide des opérateurs regex. Voici ci-dessous une explication de chaque partie :

- `r` - est placé devant pour traiter cette structure comme une chaîne de caractères brute (« raw string »). Cela est nécessaire pour éviter les problèmes liés aux caractères spéciaux de Python ;
- `(Long(?:itude)?\.|lat(?:itude)?\.)` – la partie concernant les mots « longitude » et "latitude" ainsi que leurs éventuelles abréviations ;
- `\D*` - fait référence aux caractères non numériques qui peuvent être présents ou non après les mots longitude, latitude et leurs abréviations ;
- `[\d]+(?:\.\s?\d+)?` – fait référence à la capture de la partie numérique des coordonnées ;
- `(?:\s*\s*[\d]+(?:\.\s?\d+)?)*` - fait référence aux situations où les caractères numériques des coordonnées sont séparés par un trait d'union ;

- `(?:\.s?([\d\.]+)?)?` – fait référence à la partie décimale des coordonnées qui peut être présente ou non ;
- `re.IGNORECASE` – il ne fait pas partie de la structure, mais c'est un argument qui prend en compte à la fois les caractères en minuscules et en majuscules, c'est-à-dire qu'il n'est pas sensible à cet aspect.

Après l'extraction des coordonnées des surfaces à partir de la structure créée, il a été vérifié combien d'articles la structure n'a pas réussi à extraire les coordonnées. Cela était prévisible car cet outil présente des limites pour détecter toutes les différentes façons dont les auteurs ont écrit les articles. Ces articles ont eu leurs coordonnées extraites manuellement.

4.2. Création des enveloppes

L'étape de construction des enveloppes a joué un rôle crucial, car ces structures seraient utilisées dans la détermination subséquente d'indicateur de distance. Cet indicateur présenterait la distance entre les coordonnées décrites par l'encyclopédie et les enveloppes, ainsi que par rapport aux pays/continents avec leurs limites actuelles

Les enveloppes sont des polygones qui délimitent les frontières des nations. L'encyclopédie a fourni des coordonnées qui ont permis la création de rectangles. Dans le cas des pays et des continents contemporains, les limites en vigueur ont été prises en compte. La création des enveloppes, de même que celle des pays/continents, est décrite ci-dessous.

4.2.1. Enveloppes de l'encyclopédie

Le tableau Excel fourni par M. Joliveau contenait 139 articles classifiés comme surfaces. Parmi ce total, il a été possible d'obtenir 73 enveloppes, c'est-à-dire 73 rectangles formés par deux valeurs de longitude et deux valeurs de latitude. Les autres 66 articles ne contenaient pas suffisamment d'informations pour former des enveloppes.

Parmi ces 73 enveloppes, la méthodologie basée sur les expressions régulières (REGEX) s'est révélée efficace pour 60 d'entre elles. Les 13 autres avaient au moins une valeur de longitude ou de latitude extraite, et pour les compléter, il a été nécessaire d'extraire manuellement les valeurs.

Les valeurs des quatre points (rectangle) pour former les enveloppes ont été enregistrées dans un fichier Excel. L'outil Table de coordonnées vers polygone (Défense) d'ArcGIS Pro exige un ordre spécifique pour la disposition des sommets du rectangle.

Chaque rectangle nécessite quatre lignes, une pour chaque sommet. Cependant, l'ordre des lignes doit être le suivant : 1ère ligne - Longitude 1 et Latitude 1 ; 2ème ligne - Longitude 2 et Latitude 1 ; 3ème ligne - Longitude 2 et Latitude 2 ; et enfin 4ème ligne - Longitude 1 et Latitude 2. Cet ordre est extrêmement important, car s'il n'est pas respecté, les sommets sont échangés et l'enveloppe créée présente une forme de « papillon ».

4.2.2. Pays modernes et continents

Les enveloppes définies par l'encyclopédie ne contenaient pas tous les pays décrits par celle-ci. Afin de permettre davantage de possibilités pour l'indicateur de distance, une couche a été créée contenant les limites modernes des pays et des continents. Ainsi, même si un article d'un lieu appartenait à un pays qui n'existait pas dans la couche des enveloppes, la distance pouvait être calculée à partir de la couche des limites modernes.

Il a été nécessaire d'ajouter les continents car les villes des pays d'Amérique du Sud ou d'Afrique étaient souvent référencées comme « ville d'Amérique » ou « ville d'Afrique » au lieu de leur pays respectif.

La couche contenant les limites modernes a été trouvée sur le serveur ArcGIS (https://services.arcgis.com/P3ePLMYs2RVChkJx/arcgis/rest/services/WOR_Boundaries_2023/FeatureServer) et s'appelle « WOR_Admin ». Cette couche est initialement composée des subdivisions des pays et a été choisie car elle permettait une analyse non seulement des pays, mais aussi de leurs divisions telles qu'observées dans certains articles (« Alsace », « Auvergne », « Naples », « Bavière », « Westphalie », etc.). Dans cette couche, il y a 5 824 entités (états, régions, districts, provinces, émirats et autres types de subdivisions). Son système de coordonnées est WGS 1984 et elle contient des colonnes telles que Nom, Pays d'affiliation, Continent, Codes ISO, et autres.

Grâce aux colonnes « Pays d'affiliation » et « Continent », il a été possible de créer la couche des limites modernes en utilisant l'outil « Fusionner » basé sur ces colonnes. Ensuite, un traitement a été effectué avec la librairie « Geopandas » pour regrouper ces deux couches à l'aide de la commande « concat() ». Cette couche finale n'existe que

lorsqu'il est nécessaire de calculer l'indicateur de distance, ce qui est l'un des avantages de travailler en Python.

4.3. Perdido

L'objectif de cette étape du projet était d'extraire les noms des entités géographiques des articles de l'encyclopédie. Cette extraction permettra de mettre en correspondance les continents et les pays avec leurs coordonnées anciennes, en respectant les limites définies par l'encyclopédie lorsque possible, ainsi qu'avec les pays modernes dont les noms n'ont pas changé.

Cela est rendu possible grâce à la librairie Perdido, développée pour classer chaque mot et découvrir sa fonction grammaticale et ce qu'il représente. Au sein de ses classes, il existe une base de données contenant les noms des continents, pays, régions, villes, rivières, etc., ce qui permet de distinguer les entités géographiques des autres mots qui présentent la même classification, dans ce cas « NPr », nom propre.

En accord avec M. Joliveau et M. Moncla, il a été convenu d'extraire toutes les entités géographiques de chaque article. Cependant, seules les trois premières seraient accompagnées du mot qui les précède, car cela nous permettrait de déterminer s'il s'agit d'un pays, d'un royaume, d'une ville, d'une rivière, etc. Cette décision a été prise car les articles sur les lieux contiennent généralement les informations géographiques dans le premier paragraphe. Ainsi, nous réduirions le temps nécessaire pour exécuter le code vu qu'on limitait l'analyse.

Avant l'extraction de ces entités, il a été nécessaire de limiter le début et la fin du texte sur lesquels Perdido effectuerait l'extraction. La colonne « Head » et la librairie Python `etree` ont été nécessaires pour écrire la partie du code qui déterminerait le début du texte. La colonne « Head » contient le titre de l'article et permet de repérer celui-ci dans le contenu de chaque article. Les articles de l'Edda pouvaient être rédigés selon deux modèles, comme indiqué ci-dessous :

Modèle 1 - * ADENBOURG, ou ALDENBOURG, (Géog. mod.) ville d'Allemagne, cercle de Westphalie, Duché de Berg. Long. 25. lat. 51. 2.

Modèle 2 - * ADRIANE, s. f. ville de la Province de Cyreneen Afrique, ainsi nommée d'Adrien, Empereur.

Dans le modèle 1, la classe de l'article est placée entre parenthèses « (Géog. mod.) ». Avec Perdido, il est possible de produire une sortie au format XML-TEI, où toutes les informations sont encodées avec des éléments et des attributs XML. Ainsi, la position de la classe de l'article a été déterminé. Celle-ci est un attribut de type 'articleClass'. Avec la librairie etree, la position de la dernière parenthèse «) » a été extraite et ainsi définir le début du texte après ce caractère.

Dans les cas du type de modèle 2, cette méthodologie n'est pas possible, donc l'alternative choisie a été d'utiliser la colonne « Head ». Cette colonne contient le titre de l'article. En général, les premiers mots du texte d'un article sont le titre. Par conséquent, le code a été structuré de manière à ne considérer que les mots après le titre.

Pour déterminer la fin du texte, la méthodologie a été plus simple que pour le début. Grâce à la librairie sentence_splitter, il a été possible de séparer les phrases du paragraphe. Avec la connaissance préalable que les informations géographiques étaient écrites dans les premières phrases d'un article, un total de 255 caractères a été défini pour faire l'extraction. Pour atteindre cette valeur, les phrases séparées ont été réunies. Si l'article comportait en totalité moins de 255 caractères, tout le texte de l'article a été utilisé.

Comme mentionné précédemment, l'extraction des entités géographiques a été réalisée en utilisant Perdido. Pour configurer la librairie pour cette fonction, il a été nécessaire de définir la langue en français et la version du geoparser (responsable de l'obtention des entités géographiques) en tant que « Encyclopedie ». Avec ces paramètres définis, il a été possible de commencer l'extraction en utilisant la commande « named_entities ».

En plus de « named_entities », les commandes « start_offset » ont été utilisées pour obtenir la position de l'entité géographique et « pos » pour connaître la classe grammaticale d'un mot. Grâce à ces commandes, il a été possible d'obtenir le mot précédant l'entité géographique. Ainsi, les mots ont été itérés en sens inverse à partir de la position de l'entité et leurs caractéristiques ont été analysées. Si le mot était de type « N » (substantif) et non « P » (ponctuation), il était extrait. Une liste de mots plus liés à la géographie a été privilégiée, comprenant « ville », « village », « comté », « province », « pays », « contrée », « royaume » et « bourg ». Cependant, l'extraction aurait pu concerner d'autres mots, à condition qu'ils soient des substantifs.

Pendant l'extraction des entités, il a été nécessaire de travailler avec des fonctions pour aider à l'analyse textuelle : « upper() » pour travailler avec les mots en majuscules ; « strip() » pour supprimer les espaces blancs inutiles ; « replace() » pour supprimer les caractères spécifiques, par exemple « * », et « unidecode() » de la librairie unidecode pour supprimer les accents. Ces fonctions étaient importantes pour éviter les erreurs de correspondance dues aux accents ou aux différences de cas de lettres, étant donné que Python distingue les minuscules des majuscules.

4.4. Jointure entre les deux jeux de données

La fusion des deux ensembles de données était nécessaire pour ajouter les données d'extraction des entités géographiques de Perdido à l'autre ensemble de données. La librairie Pandas permet d'effectuer cette fusion à condition qu'il y ait une colonne commune. C'est pourquoi la colonne « id_article » a été créée, qui est la concaténation du volume, « v » et du numéro de l'article.

Avec la colonne « id_article » créée, la commande « merge() » de la librairie Pandas a été appliquée, ce qui a permis d'obtenir l'ensemble de données final qui serait utilisé pour les étapes restantes.

En plus, il a été nécessaire d'effectuer une analyse des articles exclusifs de chaque ensemble de données. Ces articles ont été traités manuellement.

4.5. Calcul de la distance géodesique

Le premier indicateur pour aider à la désambiguïsation des articles était la distance entre les coordonnées anciennes extraites de l'article dans l'encyclopédie et le point le plus proche de l'enveloppe construite à partir des coordonnées de EDdA et/ou du pays/continent moderne. Plus cette distance est réduite, plus la probabilité que la coordonnée ancienne extraite de l'Encyclopédie est forte.

Il est important de mettre en évidence que les points sans anciennes coordonnées, c'est-à-dire ceux pour lesquels l'Edda n'a pas décrit la latitude et la longitude, ou seulement une des deux, ont reçu les valeurs suivantes : Longitude = $-17,66277^\circ$ et Latitude = 0° . La longitude n'est pas zéro car auparavant, l'encyclopédie utilisait le méridien de l'île de Fer comme référence, et qui est arbitrairement placé à 20° à l'ouest du méridien de Paris. Aujourd'hui le méridien de Greenwich est utilisé. La conversion donne cette valeur de $-17,66277^\circ$.

Deux méthodologies ont été testées pour calculer la distance : arcpy et geopandas.

4.5.1. Arcpy – cursor

Le package arcpy a été la première méthodologie testée, car tout le travail était effectué dans le logiciel ArcGIS Pro. Ce logiciel permet d'ouvrir un onglet de l'IDE du bloc-notes Jupyter, ce qui facilite l'utilisation de Python dans le SIG.

Pour calculer la distance, il a été nécessaire d'utiliser les commandes de « cursor » d'arcpy. Cette commande permet d'accéder aux informations de la table attributaire d'une couche à l'aide de la méthode SearchCursor(). Il est également possible de mettre à jour les informations de la table attributaire en utilisant la méthode UpdateCursor().

Le calcul de la distance a été réalisé à l'aide de la méthode angleAndDistanceTo(). Pour ce faire, les coordonnées du point de l'article et les coordonnées de l'enveloppe ou du pays/continent ont été utilisées. La distance géodésique a été sélectionnée, et la raison en sera expliquée dans le prochain paragraphe.

Une fois le code terminé, il est possible de créer un outil qui exécute ses commandes. Il suffit de configurer les variables que l'utilisateur doit remplir, comme représenté dans la **Figure 3**. De plus, il est nécessaire d'importer le script du code au format *.py dans l'outil.

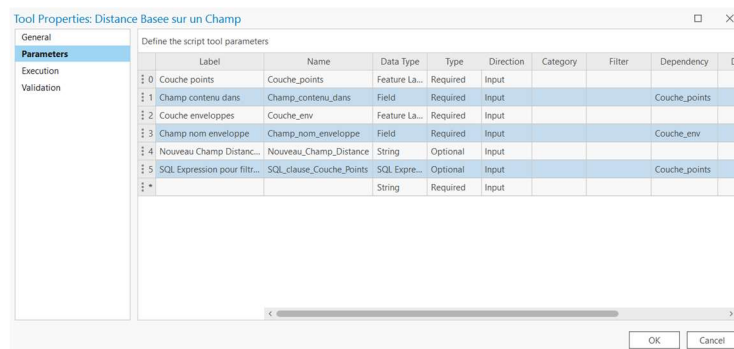


Figure 3. Configuration des variables pour la création de l'outil dans ArcGIS Pro.

Malgré les possibilités offertes par arcpy, il présente également des inconvénients. Le principal inconvénient est que lorsqu'il est nécessaire de créer une colonne, il faut utiliser un outil (arcpy.AddField_management()), ce qui rend l'exécution du code plus lente. En outre, l'utilisation de la géométrie d'une entité dans la table attributaire n'est pas évidente, car elle n'est pas visible directement. Pour l'obtenir, il faut utiliser la colonne « SHAPE@ ». C'est pour ces raisons qu'il a été décidé de choisir une autre alternative.

4.5.2. Librairie GeoPandas, Geopy et Shapely

En raison des difficultés rencontrées avec le package `arcpy`, nous avons recherché une alternative pour calculer la distance entre les points et leurs enveloppes respectives ainsi que leurs pays/continents. Il a été choisi d'utiliser les librairies `GeoPandas`, `Geopy` et `Shapely`, qui offrent les fonctionnalités nécessaires à notre analyse.

La librairie `GeoPandas` est une extension de la célèbre librairie `Pandas`, offrant des fonctionnalités supplémentaires dédiées à la manipulation de données géospatiales. On utilise `GeoPandas` pour lire des fichiers au format `shapefile` et les transformer en `Geodataframes`, qui sont des `Dataframes` auxquels une colonne de géométrie a été ajoutée. Ainsi, il est possible d'accéder aux informations contenues dans la table attributaire et ainsi déterminer la distance entre les points.

Le calcul de la distance a été divisé en quatre étapes : tout d'abord, on identifie le polygone correspondant au point ; ensuite, on définit le point de référence à l'intérieur de ce polygone ; puis, on prépare les coordonnées pour le calcul de la distance géodésique ; enfin, on effectue le calcul proprement dit. Il convient de noter que ce calcul a été restreint aux articles dans lesquels la colonne « `PLACE_1` » contenait le terme « ville ».

Dans la première étape, divisée en deux parties, il a été utilisé les couches d'enveloppes de l'EDdA et de limites modernes. Il a été choisi d'utiliser deux couches de polygones en raison du faible nombre d'enveloppes définies par l'encyclopédie, soit seulement 73. Comme la couche de limites modernes comportait 204 pays/continents, cela a augmenté la probabilité d'obtenir une correspondance.

La correspondance a été réalisée en utilisant deux méthodologies, la deuxième étant utilisée uniquement en cas d'échec de la première. Dans un premier temps, nous avons adopté une approche d'égalité complète, en utilisant les informations de la colonne « `NE_1` » de la couche d'articles pour sélectionner, si disponible, la même valeur présente dans la colonne « `LIEU` » de la couche d'enveloppes.

Pour la couche des limites modernes, la colonne « `NE_1` » a été comparée à la colonne « `NOM_FR` ». Afin d'éviter les erreurs, l'analyse dans les deux cas (enveloppes et limites modernes) a été effectuée en lettres majuscules en utilisant la méthode « `upper()` ».

Après avoir effectué la correspondance, les coordonnées du polygone filtré ont été extraites et seront utilisées à l'étape suivante. De plus, à cette étape, un paramètre d'égalité est généré, qui sera expliqué en détail dans la deuxième méthodologie. Lorsque la correspondance est trouvée, ce paramètre est attribué à la valeur 100 ; sinon, la deuxième méthodologie est exécutée.

La deuxième méthodologie repose sur l'utilisation de la librairie FuzzyWuzzy, qui permet de réaliser des correspondances entre des mots avec un pourcentage de similitude. Contrairement à l'approche précédente basée sur une égalité complète (100%), FuzzyWuzzy permet d'obtenir des correspondances avec une similarité moindre, par exemple, « Allelemagne » = « Allemagne », atteignant une équivalence de 90%.

Il est essentiel de mettre en évidence que ce pourcentage d'égalité varie en fonction du nombre de caractères d'un mot et de la tolérance aux caractères différents. Par exemple, pour le pays « France », un seul caractère différent peut avoir un impact significatif, tandis que pour le pays « Allemagne », qui comporte plus de caractères, cette différence peut être plus tolérée.

Cette méthodologie utilisant FuzzyWuzzy offre une plus grande flexibilité pour identifier des correspondances dans des situations où une égalité exacte n'est pas possible, améliorant ainsi l'efficacité du processus de correspondance entre les données. Cette flexibilité a été une solution pour les erreurs contenues dans l'encyclopédie.

Pour réaliser la correspondance en utilisant cette méthodologie, il a d'abord été nécessaire de générer une liste contenant toutes les correspondances possibles répondant à un critère minimum établi lors de tests, défini à 70%. Pour les enveloppes, cette liste était composée des valeurs de la colonne « LIEU », tandis que pour les limites modernes, nous avons ajouté une colonne supplémentaire, en plus de « NOM_FR », appelée « NOM_Hist ». Cette colonne a été créée, après de nombreux tests, pour stocker des noms qui ont subi des modifications au fil de l'histoire ou ont reçu différentes désignations par Esri. Quelques exemples présents dans cette colonne comprenaient : « Angleterre » (Royaume-Uni de Grande-Bretagne), « Arabie » (Émirats arabes unis), « Abyssinie » (Éthiopie) et « Nubie » (Soudan du Sud).

À partir de cette liste, il a été sélectionné la correspondance présentant le plus haut degré de similitude selon notre métrique de 70%. Ainsi, on récupère la géométrie du polygone correspondant, ainsi que le paramètre d'égalité obtenu.

Dans la deuxième étape, il a été utilisé les géométries du point de l'article et de son enveloppe ou de son pays/continent respectif. Si le polygone contenait le point de l'article, la distance serait considérée comme nulle et le calcul serait terminé. Sinon, on trouve un point de référence dans le polygone par rapport au point de l'article.

Le point de référence a été obtenu en utilisant la méthode « `nearest_point()` » de la librairie Shapely. Cette méthode réalise une analyse de voisinage entre deux géométries et renvoie le point le plus proche entre elles. Ainsi, les coordonnées du point du polygone le plus proche des coordonnées de l'article ont été obtenues. Dans la documentation de la méthode « `nearest_point()` », il était indiqué qu'il était possible de travailler avec des coordonnées en degrés décimaux. Par conséquent, le système de coordonnées WGS 1984 a été conservé.

Dans la troisième étape, les longitudes et latitudes des points ont été extraites en utilisant les méthodes « `get_x()` » et « `get_y()` » de la librairie Shapely. Cette séparation était nécessaire car le calcul de la distance exige un ordre spécifique des coordonnées, avec la latitude en premier et la longitude ensuite.

Avec les coordonnées dans l'ordre correct, il a été calculé la distance géodésique, en kilomètres, entre les deux points en utilisant la méthode « `GD()` » de la librairie Geopy.

La distance géodésique a été choisie comme mesure appropriée pour le calcul de la distance entre les points, en tenant compte de la courbure de la Terre. Ce choix est particulièrement pertinent lorsqu'on travaille avec des coordonnées géographiques en degrés décimaux, garantissant des résultats plus précis par rapport aux mesures de distance planes, telles que la distance euclidienne.

4.6. Classification des articles

L'étape de classification s'est avérée essentielle pour distinguer entre les articles faisant référence à des lieux et ceux qui n'avaient pas cette référence. La classification existante à travers la colonne « `edda_class` » n'était pas bonne.

La nouvelle classification des articles a donc été basée sur quatre catégories : « LIEU », « TERME », « RENVOI » et « PEUPLE ». Ces catégories ont été définies par M. Joliveau pour filtrer les articles se référant à des lieux réels, des mots explicatifs, des références à d'autres articles de l'encyclopédie et des mentions de personnes ou de peuples (démonymes).

L'analyse de la colonne « edda_class » a permis d'identifier certaines classes directement associées aux quatre catégories. En utilisant Python, il a été possible de créer une nouvelle colonne appelée « CLASSE », qui a reçu les classifications correctes pour ces catégories.

Cependant, la catégorie « RENVOI » a nécessité une analyse supplémentaire en fonction du nombre de mots dans chaque article. Si le nombre de mots, à l'exclusion de la partie faisant référence au titre, était inférieur à 7 mots, l'article serait classé dans cette catégorie. Par exemple, en considérant l'article sur Goyane (volume 7, numéro 2557) avec la description « GOYANE, (Géog.) Voyez Guiane. », la partie pertinente « Voyez Guiane » ne compte que 2 mots. Ainsi, 7 mots ont été établis comme critère après des analyses préalables pour distinguer ces types d'articles des autres.

Cette approche détaillée a permis d'effectuer une classification appropriée des articles dans les quatre catégories, assurant ainsi une plus grande précision et cohérence dans le processus.

4.7. Web Scraping pour corriger l'URL des articles

Cette étape était nécessaire en raison de problèmes de correspondance entre les URL des articles dans la feuille Excel et leurs contenus respectifs. Le Web Scraping a été utilisé pour corriger automatiquement ces divergences.

L'URL suit un modèle spécifique, qui est la concaténation de « <https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/> », suivi du numéro du volume de l'article et du numéro de l'article lui-même. Par exemple, pour l'article sur Paris, appartenant au volume 11 et au numéro 4606, l'URL est « <https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/navigate/11/4606> ».

Cependant, il peut arriver que la plateforme ARTFL apporte des modifications et que la partie de l'URL relative au numéro de l'article soit modifiée. Pour résoudre ce problème, un code a été développé en utilisant la librairie « Selenium » pour effectuer le Web Scraping et corriger les articles dont les URL ont été modifiées.

Le Web Scraping serait utilisé pour comparer la colonne « Head » avec le titre de la page ARTFL stockée dans la colonne « URLARTFL » de l'article.

Pour corriger les URLs des articles sur la plateforme ARTFL, le processus a été réalisé de manière itérative, en suivant une séquence spécifique. Initialement, l'URL du

premier article, sur les 15 545, a été corrigée. Ensuite, on est passé à l'article numéro 51 et on a corrigé son URL. Par la suite, on est passé à l'article numéro 101 et ainsi de suite, en corrigeant les articles à intervalles réguliers de 50.

Cette approche a permis de corriger les URLs de chaque 50ème article sur l'ensemble total des 15545 articles. Après avoir corrigé l'URL d'un article, la valeur du numéro de l'article contenu dans cette URL était stockée. Ensuite, cette valeur était comparée au numéro réel de l'article, et en fonction de cette comparaison, une variable appelée « offset » se voyait attribuer une valeur de zéro si les numéros étaient identiques, ou une valeur de différence (par exemple, -2, -1, 1 ou 2) sinon.

Cette variable a été stockée dans une feuille auxiliaire qui stockait aussi ID de l'article, volume de l'article, numéro de l'article, URL corrigée. Pour chaque intervalle d'articles (par exemple, de l'article 2 à l'article 50), la valeur de l'offset obtenue pour l'article 51 était ajoutée à la partie du numéro de l'article dans l'URL. Pour les articles de 52 à 100, l'offset obtenu pour l'article 101 était ajouté, et ainsi de suite.

Cependant, lors du passage d'un volume d'article à un autre, une analyse minutieuse a été nécessaire, car il a été constaté que l'offset était remis à zéro. Cette analyse attentive a permis de garantir la correction précise des URLs, en tenant compte des volumes des articles et en évitant les distorsions des valeurs de l'offset.

De cette manière, tous les articles ont été pris en compte dans la correction des URLs sans dépasser la limite d'accès à une page électronique, et en assurant l'intégrité du processus de Web Scraping pour la plateforme ARTFL.

4.8. Géocodage

Le géocodage, tout comme la distance géodésique, a été un autre indicateur choisi pour aider à la désambiguïsation des articles de l'encyclopédie. À partir du géocodage, il a été possible d'obtenir les coordonnées modernes des articles. Ainsi, permettant de comparer les anciennes coordonnées avec les coordonnées modernes. Une forte distance entre la localisation ancienne extraite de l'encyclopédie et sa localisation moderne peut conduire à suspecter une erreur. Il sera possible aussi de localiser avec des coordonnées modernes, les articles sans coordonnées dans l'Encyclopédie.

La librairie « Perdido » a été utilisée pour effectuer le géocodage, en étant configurée pour utiliser différents gazetiers, y compris Nominatim (OpenStreetMap),

GeoNames, World Historical Gazetteer et IGN. Parmi ces options, le gazetier choisi a été Nominatim, car il a présenté des résultats plus complets et cohérents après des tests comparatifs.

Le géocodage a été effectué en utilisant la colonne « Head » et les colonnes des entités géographiques (« NE_1 », « NE_2 » et « NE_3 »). Cependant, un traitement spécial de la colonne « Head » a été nécessaire en raison de la présence de listes de valeurs, d'articles en fin de nom et d'inversions dans l'ordre des mots. Sans ce traitement, la requête à Nominatim aurait eu moins de chances de réussir selon les tests effectués.

Lorsqu'une liste de valeurs était présente dans la colonne « Head », chaque élément était séparé et consulté individuellement dans le Nominatim. Par exemple, pour le titre d'article « ACADIE ou ACCADIE », deux requêtes seraient effectuées, une pour « ACADIE » et une autre pour « ACCADIE ».

Lorsque des articles (L', l', Le, le, La, la, Les, les) étaient trouvés à la fin des noms, ils étaient retirés pour un géocodage correcte. Par exemple, « ALEXANDRIN L' » serait traité comme « ALEXANDRIN ».

Enfin, les cas où les noms de lieux commençaient par « Saint- » ont été traités. Dans l'encyclopédie, « Saint- » était placé à la fin, et ce mot pouvait être écrit de différentes manières, telles que « Saint- », « (Saint-) », « Sant- », « St. », « ST », « Sainte- », etc. Pour tous ces cas, « Saint- » a été placé au début du terme. Par exemple, « AIGNAN (Saint) » a été traité comme « Saint-AIGNAN ».

Après le traitement de la colonne « Head », le géocodage a été effectué en suivant une règle structurée pour la construction des requêtes. Cette règle a été élaborée dans le but de commencer les requêtes de manière plus spécifique et de progresser vers des requêtes plus simples. La première requête consistait à concaténer la colonne « Head » ou l'un des éléments de sa liste avec les trois colonnes d'entités géographiques. Si aucun résultat n'était obtenu, la requête était répétée en supprimant la troisième entité géographique, et ainsi de suite, jusqu'à ce qu'une requête soit effectuée uniquement avec la colonne « Head ».

Ce processus a permis d'aborder de manière progressive les différentes possibilités d'écriture et de formatage des données, ce qui a abouti à un géocodage plus complet et précis. Cela a été rendu possible car des tests préliminaires ont montré que consulter « Lyon, France » était plus précis que simplement « Lyon », par exemple.

Pour éviter les restrictions liées à la limite de requêtes quotidiennes de Nominatim, le géocodage a été appliqué exclusivement aux articles correspondant aux villes de France, identifiés par la présence du terme « ville » dans la colonne « PLACE_1 » et « France » dans la colonne « NE_1 ». Ce filtrage a également été appliqué aux résultats du géocodage, en considérant que, selon la documentation, les villes ont le paramètre « type » égal à « administrative ». Cette approche garantissait que le géocodage soit réalisé de manière plus efficace et ciblée sur les informations géospatiales pertinentes pour le contexte de la recherche.

Les résultats ont été enregistrés dans un fichier shapefile contenant la table attributaire avec les colonnes suivantes : l’ID de l'article, la requête effectuée, le paramètre auxiliaire, la longitude, la latitude et le nom du lieu. Dans la colonne paramètre auxiliaire appelée « req_type », elle servait à déterminer quel type de requête a été effectuée au format numérique. Voici les valeurs possibles :

- 1 - requête effectuée avec les colonnes « Head », « NE_1 », « NE_2 » et « NE_3 » ;
- 2 - requête effectuée avec les colonnes « Head », « NE_1 » et « NE_2 » ;
- 3 - requête effectuée avec les colonnes « Head » et « NE_1 » ;
- 4 - requête effectuée uniquement avec la colonne « Head ».

Cette structure de colonnes a permis de représenter les résultats géocodés sous forme de points sur la carte et d'établir la communication avec l'ensemble de données d'origine grâce à la colonne ID de l'article qui serait utile pour l'application géoweb.

En plus du nouveau shapefile créé, les résultats ont également été stockés dans l'ensemble de données d'origine. La colonne « nom_geocod » a enregistré la requête effectuée qui a donné des résultats, et si aucun résultat n'était obtenu, elle stockait simplement « - ». Les colonnes suivantes ont stocké le nombre de résultats pour chaque type de requête :

- « Head_seul » - nombre de résultats de la requête avec seulement la colonne « Head » ;
- « Head_1 » - nombre de résultats de la requête contenant les colonnes « Head » et « NE_1 » ;

- « Head_1_2 » - nombre de résultats de la requête contenant les colonnes « Head », « NE_1 » et « NE_2 » ;
- « Head_1_2_3 » - nombre de résultats de la requête contenant les colonnes « Head », « NE_1 », « NE_2 » et « NE_3 ».

Ainsi, le géocodage des articles a été réalisé de manière exhaustive, surmontant les défis linguistiques et de mise en forme, afin d'obtenir des coordonnées géographiques précises et à jour, contribuant de manière significative à la désambiguïsation des informations géospatiales dans l'encyclopédie.

4.9. Jointure spatiale

À cette étape, l'objectif était de réaliser une jointure spatiale entre les articles et une couche contenant les trois niveaux modernes de division administrative (continent, pays et subdivision). L'objectif était d'enrichir le jeu de données original avec des colonnes enregistrant le nom et le type de chaque division géographique. Par exemple, pour la France, la colonne « Type_adm » stockerait « région, pays, continent » et la colonne « Nom » stockerait « Auvergne-Rhône-Alpes, France, Europe ».

Cependant, l'analyse de cette jointure a été interrompue car il n'était pas possible de garantir que les entités géographiques définies par l'encyclopédie correspondaient exactement aux mêmes niveaux de division géographique utilisés dans la couche externe. De plus, le manque de normalisation dans la description des lieux dans l'encyclopédie et l'extraction limitée à seulement trois entités avec des mots associés rendaient probable l'omission d'informations géographiques importantes.

4.10. Priorité de désambiguïsation

Pour faciliter la priorisation des articles à corriger, une colonne appelée « Urgence » a été ajoutée pour stocker la classe de priorité pour la désambiguïsation. La classe de chaque article a été déterminée en tenant compte de divers paramètres, notamment les coordonnées anciennes, la correspondance avec le pays/continent moderne, la distance jusqu'au pays correspondant et la distance par rapport aux points de géocodage.

La distance par rapport aux points de géocodage était le seul paramètre à calculer, car les autres avaient déjà été obtenus. Ce calcul impliquait l'utilisation des coordonnées de l'article et des coordonnées de chaque résultat de géocodage pour déterminer la

distance géodésique en kilomètres, à l'aide de la méthode GD() de la librairie Shapely, comme expliqué dans la section 4.5.2. Pour les articles avec plusieurs résultats de géocodage, la distance finale a été calculée comme la moyenne de ces distances.

Les classes sont représentées dans le **Tableau 2**.

Classe	Coordonnées anciennes	Correspondance - pays/continent moderne	Distance au pays/continent moderne (km)	Geocodage	Distance aux points geocodés (km)
1	Non	Non	-	Non	-
2	Non	Oui	-	Non	-
3	Non	Oui	-	Oui	-
4	Oui	Non	-	Non	-
5	Oui	Oui	≥ 500	Non	-
6	Oui	Oui	≥ 500	Oui	-
7	Oui	Oui	≥ 100 et < 500	Non	-
8	Oui	Oui	≥ 100 et < 500	Oui	-
9	Oui	Oui	> 0 et < 100	Non	-
10	Oui	Oui	> 0 et < 100	Non	-
11	Oui	Oui	0	Oui	0
12	Oui	Oui	0	Oui	≥ 30
13	Oui	Oui	0	Oui	< 30

Tableau 2. Classes de priorités de désambiguïsation.

Ainsi, la colonne « Urgence » a aidé à prioriser et à organiser les corrections des articles, en prenant en compte plusieurs critères géospatiaux, permettant ainsi de diriger les efforts vers les articles avec une priorité plus forte pour la désambiguïsation.

4.11. Traitement des données et mise en ligne des couches

Une étape de traitement des données a été réalisée dans la table d'attributs du fichier final obtenu. En utilisant les librairie Pandas et Geopandas, les colonnes non nécessaires ont été supprimées, le type de données (numérique ou textuel) des colonnes a été ajusté, les colonnes ont été renommées et les valeurs des données ont été mises à jour si nécessaire.

Après ce traitement, le fichier shapefile final a été converti en une couche web à l'aide d'ArcGIS Pro. La conversion a permis de rendre la couche disponible sur ArcGIS Online (AGOL), où il est possible d'y accéder et d'effectuer des modifications dans la table d'attributs.

Dans la table d'attributs, les colonnes suivantes ont été ajoutées, que les utilisateurs rempliraient dans l'application géoweb pour aider à la désambiguïsation des données : « Correcteur », « Transcription des Coordonnées EDdA », « Vraie LongDegréEDda », « Vraie LongMinuteEDda », « Vraie LongSecondeEDda », « Vraie LatDegréEDda », « Vraie LatMinuteEDda », « Vraie LatSecondeEDda », « Identification entre noms ancienne et moderne », « Nom moderne », « Pays moderne », « LongModern », « Latmodern », « Remarques » et « Statut ».

De plus, AGOL offre la possibilité de créer des restrictions ou des listes de valeurs autorisées pour certaines colonnes. Cette série d'ajustements et de fonctionnalités dans AGOL garantit un environnement plus contrôlé et structuré pour la désambiguïsation des données, permettant une meilleure interaction et collaboration entre les utilisateurs dans le processus de correction et d'enrichissement de la base d'informations géospatiales. Les restrictions suivantes ont été appliquées :

- Transcription des Coordonnées d'EDdA – liste avec les options : « Correcte » et « Erronée » ;
- Vraie LongDegréEDda – seulement entiers entre -150 e 150 ;
- Vraie LatDegréEDda – seulement entiers entre -90 e 90 ;
- Vraie LongMinuteEDda, Vraie LongSecondeEDda, Vraie LatMinuteEDda, Vraie LatSecondeEDda – seulement entiers entre 0 e 60 ;

- Identification entre noms ancienne et moderne – liste avec les options : Nom moderne identique, Nom moderne différent, Nom moderne incertain e Nom moderne introuvable ;
- Statut – liste avec les options : ‘NON-VERIFIÉ’ e ‘VERIFIÉ’.

4.12. Experience Builder

Après la préparation des fichiers nécessaires, a commencé la phase finale de l'application géoweb, dont l'objectif est de permettre à l'utilisateur de vérifier et corriger les coordonnées et de fournir des informations mises à jour sur les articles, contribuant ainsi à leur désambiguïsation.

L'application géoweb a été conçue avec trois pages principales : la page « Introduction », la page « Édition » et la page « Carte ». Sur la page d'introduction, des explications sur les cartes et les outils disponibles pour aider l'utilisateur sont fournies. De plus, il existe trois sous-pages qui détaillent le mode d'utilisation : « Filtres » explique les filtres configurés par les widgets et autres interactions, « Colonnes » détaille les colonnes pouvant être modifiées par l'utilisateur, et « Classes » présente les classes de priorité de désambiguïsation telles que décrites dans la section 4.10.

La page d'édition est l'espace où l'utilisateur peut effectuer des modifications et ajouter des informations, elle contient deux cartes interactives. La première carte affiche les articles de l'encyclopédie avec deux symbologies différentes, ainsi que les enveloppes de l'encyclopédie. La deuxième carte, quant à elle, présente à la fois les articles de l'encyclopédie et les résultats du géocodage pour la France. C'est dans cette carte que l'utilisateur peut éditer les informations.

Sur la page de la carte, l'utilisateur dispose d'une fonctionnalité similaire à un SIG (Système d'Information Géographique). La première carte de la page d'édition est présentée à plus grande échelle, tandis qu'en bas se trouve la table attributaire des articles de l'encyclopédie.

Pour rendre l'application dynamique et interactive, plusieurs widgets ont été ajoutés et des interactions entre les cartes ont été configurées dans Experience Builder. Parmi ces interactions, on peut souligner la facilité de visualisation la page d'édition, qui permet de filtrer un article sélectionné sur la première carte ainsi que ses résultats de

géocodage sur la deuxième carte, grâce à un déclencheur (Trigger) établissant une connexion via la colonne ID de l'article.

D'autres filtres et actions ont été configurés à partir de widgets disponibles sur Experience Builder, nécessitant seulement des connaissances de base sur le travail avec les tables d'attributs. Deux types principaux de widgets ont été utilisés pour filtrer les couches des cartes en fonction des colonnes.

Le premier d'entre eux est le widget « filtre », qui permet à l'utilisateur de filtrer une colonne en fournissant ou en sélectionnant des valeurs d'intérêt. Sa configuration est similaire à la sélection des attributs dans une table attributaire, ne nécessitant que la définition de la colonne à filtrer et les critères de filtrage, y compris la possibilité de combiner des colonnes à l'aide des méthodes ET ou OU. Ce widget a été largement utilisé sur les deux pages de l'application. Sur la page « Édition », il a été utilisé pour filtrer les colonnes du titre de l'article et du pays/continent moderne. Sur la page « Carte », il a été utilisé pour filtrer les colonnes « Urgence de désambiguïsation », « Distance au pays/continent moderne » et « Statut ». La **Figure 4** illustre un exemple de ce widget.

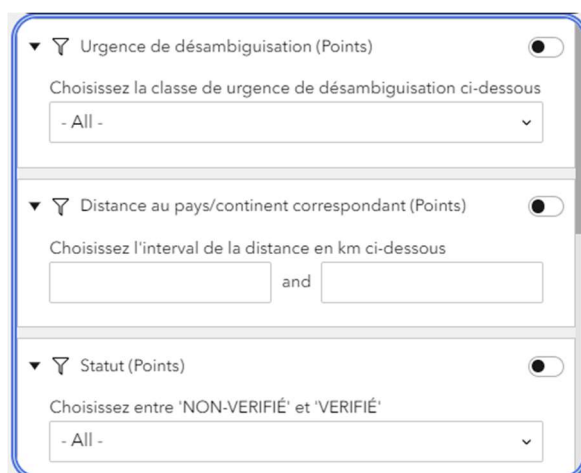


Figure 4. Widget « Filtre » de colonnes type 1.

Le deuxième widget est appelé « Liste » et offre également la fonctionnalité de filtrer une colonne. Cependant, la différence est qu'il présente toutes les valeurs présentes dans la colonne configurée sous forme de liste. Pour appliquer le filtre, l'utilisateur n'a qu'à cliquer sur la valeur d'intérêt. Ce widget s'est avéré particulièrement utile pour représenter les enveloppes de l'encyclopédie de manière plus visuelle que le widget « Filtre » mentionné précédemment. La **Figure 5** illustre ce widget.



Figure 5. Widget « Liste » pour filtrer la couche des articles selon l'enveloppe de correspondance.

En plus des widgets de filtres, quatre autres types de widgets ont été utilisés. Le widget « Embed » a été utilisé pour permettre l'accès à l'URL de chaque article sur le site ARTFL. Ce widget est dynamique, permettant la configuration d'une colonne contenant les URL à afficher. Ainsi, en cliquant sur un article spécifique, le widget affiche exclusivement l'URL correspondant à cet article. Par conséquent, il est possible d'afficher le contenu de l'article de manière plus visuelle que dans une fenêtre pop-up ou un tableau d'attributs. La **Figure 6**, illustre ce widget.



Figure 6. Widget de « URL » pour représenter les pages ARTFL des articles.

Les widgets de coordonnées et d'édition ont pour objectif commun de contribuer à la désambiguïsation des données. Le widget de coordonnées offre à l'utilisateur deux options après la définition de la carte de référence : afficher les coordonnées en déplaçant la souris ou fixer les coordonnées du point où il a cliqué. Ces coordonnées peuvent être utilisées par l'utilisateur pour remplir directement les colonnes « Longmodern » et « Latmodern » dans le widget d'édition. Dans toute l'application, la convention de présenter d'abord la longitude, puis la latitude a été suivie, ce qui s'applique également à ce widget. La **Figure 7** illustre ce widget.



Figure 7. Widget de coordonnées.

La configuration du widget d'édition nécessite la définition non seulement de la couche d'intérêt, mais également des colonnes de cette couche qui seront autorisées à être éditées par l'utilisateur. Ces colonnes sont les mêmes que celles mentionnées à la section 4.11. De plus, il est possible de donner à l'utilisateur l'option de déplacer le point, ce qui entraîne la modification des coordonnées. Cependant, cette fonctionnalité peut produire

des erreurs, c'est pourquoi la création des colonnes de coordonnées modernes a été établie. Ainsi, les coordonnées d'origine de l'encyclopédie sont préservées, tout en permettant d'ajouter et de connaître les coordonnées modernes. La **Figure 8** illustre ce widget.

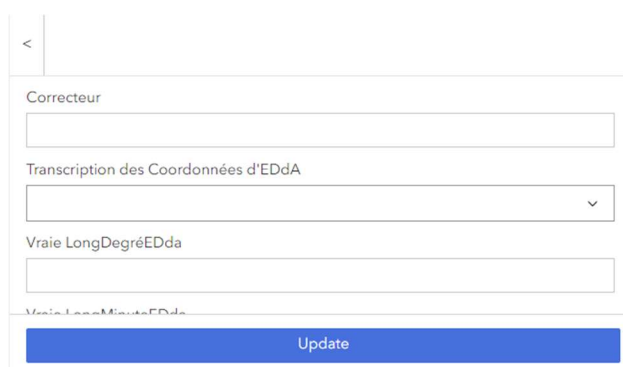


Figure 8. Widget de « édition » des articles.

Enfin, le widget de table d'attributs a été utilisé uniquement sur la page « Carte » et pour sa configuration, il suffit de fournir la couche de référence.

En plus des widgets et des interactions, une attention particulière a été portée à la symbologie et aux colonnes des couches. Ces configurations doivent être effectuées sur le Web Map avant de commencer à utiliser l'Experience Builder.

Sur le Web Map, il a été utilisé deux types de symbologie, un pour chaque indicateur. Pour l'indicateur de distance par rapport au pays/continent correspondant, il a été choisi des cercles de tailles proportionnelles. Les classes ont été définies manuellement de manière à mieux représenter les points. Les 10 classes utilisées étaient les suivantes : 0 km, 1 - 10 km, 11 - 50 km, 51 - 100 km, 101 - 200 km, 201 - 500 km, 501 - 1 000 km, 1 001 - 5 000 km, 5 001 - 10 000 km et 10 001 - 15 425 km.

Pour l'indicateur de classe de priorité de désambiguïsation, il a été adopté une symbologie d'échelle de couleurs et de tailles. Les classes les plus prioritaires ont été représentées avec du rouge foncé et avec des symboles plus grands, tandis que les classes moins prioritaires se sont rapprochées du jaune clair et ont eu des symboles plus petits.

Les résultats de géocodage ont été représentés par des triangles de couleur bleu foncé pour éviter toute confusion avec les points des articles, qui sont représentés par des cercles.

La dernière chose à régler sur Web Map était les colonnes qui ont été renommées ou supprimées (si pas nécessaires) dans le but d'aider l'utilisateur à comprendre la table attributaire. Par exemple, lors de l'étape de réalisation du ArcGIS Pro, le nom d'une

colonne était limité à 10 caractères, ce qui obligeait à abrégé plusieurs colonnes. Par exemple, la colonne « D_Pays_NE1 » est devenue « Distance Pays/Continent et contenu colonne NE1 - en km ». De cette manière, les fenêtres pop-ups, configurées sur Experience Builder, sont plus compréhensibles pour l'utilisateur.

Avec toutes les configurations terminées et l'application prête, l'accès public à l'application a été autorisé. Elle appartient au serveur ESRI et sera disponible tant que l'université y aura accès.

5. Évaluation des résultats

5.1. Statistiques

Pour évaluer la méthodologie employée, trois outils ont été utilisés : un tableau avec des informations statistiques, des graphiques et une analyse minutieuse d'une fraction des articles.

Le tableau et les graphiques seront réalisés en fonction des variables de distance géodésique et du critère de similarité Fuzzy Wuzzy par rapport aux enveloppes EDdA et aux limites modernes. Les deux ont été réalisés à partir des bibliothèques Matplotlib et Seaborn de Python.

Ces outils ont été appliqués initialement à tous les articles, puis ultérieurement à un sous-ensemble ne comprenant que les articles ayant des coordonnées anciennes.

Dans le tableau, les informations suivantes seront incluses : le décompte des valeurs non nulles, la moyenne, l'écart-type, la valeur minimale, la valeur maximale et les quartiles.

Les graphiques utilisés visent à rendre l'analyse plus visuelle et à faciliter la compréhension des résultats. Les graphiques suivants ont été utilisés :

- Diagramme en violon, qui présente les mêmes informations qu'un diagramme en boîte (boxplot) avec en plus la distribution des résultats ;
- Nuage de points, qui permet de visualiser les combinaisons trouvées entre les deux variables, avec le critère d'égalité Fuzzy Wuzzy sur l'axe X et la distance sur l'axe Y ;
- Histogramme, qui aide à comprendre la distribution des résultats de la variable critère d'égalité Fuzzy Wuzzy, en complément du graphique en violon.

En plus de ces graphiques, une carte sera confectionnée pour indiquer combien de correspondances ont été obtenues par pays ou continent. Cela permettra d'évaluer les endroits où l'encyclopédie présente le plus d'informations.

L'analyse minutieuse a été réalisée sur 125 articles, répartis en 5 catégories de 25 articles chacune. Dans cette analyse, le facteur qui a empêché le calcul de la distance ou si le calcul de la distance a été correctement effectué sera déterminé en relation aux limites

modernes. Les cinq catégories considérées avaient en commun la classe « LIEU ». Elles sont énumérées ci-dessous :

- Articles avec des coordonnées définies par l'encyclopédie et une distance égale à zéro ;
- Articles avec des coordonnées définies par l'encyclopédie et une distance supérieure à zéro et inférieure ou égale à 100 km ;
- Articles avec des coordonnées définies par l'encyclopédie et une distance supérieure à 100 km ;
- Articles avec des coordonnées définies par l'encyclopédie et sans distance calculée;
- Articles sans coordonnées.

La division des catégories a été réalisée à l'aide de la librairie Pandalas. Le choix des 25 articles de chaque catégorie a été effectué de manière aléatoire, rendu possible grâce à la méthode « sample() » de Pandalas.

5.2. Analyse des articles par pays et continent

Dans le but d'identifier les pays ou les continents avec la plus grande quantité d'informations décrites dans l'encyclopédie, un filtre a été appliqué en fonction de l'indicateur de distance géodésique. Le but de ce filtre était de sélectionner les articles pour lesquels la méthodologie de distance s'est avérée plus précise.

Le filtre consistait à sélectionner les articles avec une distance géodésique inférieure ou égale à 100 km et ayant un critère d'égalité supérieur ou égal à 90. Ces articles ont été cartographiés pour une analyse visuelle, permettant de comparer les régions du monde avec des informations plus détaillées dans l'encyclopédie.

6. Résultats et discussions

Les résultats obtenus pendant la réalisation de ce stage peuvent être divisés en quatre parties distinctes : la création des enveloppes, l'extraction d'informations à l'aide de la librairie Perdido, le géocodage également en utilisant Perdido, et enfin l'application finale développée dans l'Experience Builder. Il est important de souligner que les trois premières parties ont été conçues pour converger vers la dernière partie. Chacune de ces parties sera détaillée dans les sections suivantes.

6.1. Enveloppes EDdA

La méthodologie utilisée s'est avérée capable d'extraire les informations nécessaires pour créer les enveloppes EDdA, même face aux divers modèles utilisés par l'encyclopédie. Ainsi, il a été capable de créer un fichier shapefile pour stocker les 73 enveloppes, sous forme de rectangle, de l'encyclopédie.

Les enveloppes obtenues sont présentées à l'Annexe A – Enveloppes EDdA.

6.2. Perdido

6.2.1. Indicateur de la distance géodésique

L'indicateur de distance géodésique a été calculé pour les articles contenant le mot « ville » dans la colonne « PLACE_1 », ce qui a entraîné une réduction du nombre total d'articles de 15 545 à 6 541.

Parmi ces 6 541 articles, la méthodologie utilisée a permis la correspondance avec les enveloppes pour 3 352 articles et avec les limites modernes des pays ou des continents pour 5 178 articles. Cette différence était attendue, car la couche de limites modernes contenait trois fois plus d'informations, avec 204 limites contre seulement 73 enveloppes, ce qui augmentait la probabilité de correspondance.

Malgré le niveau élevé de correspondance exacte, observé par le critère d'égalité, comme le montrent les trois quartiles égaux à 100 dans le **Tableau 3**, en comparant les valeurs de distance, on note que le deuxième quartile, la médiane, était différent. Alors que pour les enveloppes, 50 % des distances sont inférieures à 111,32 km, pour les limites modernes, cette valeur augmente à 945,22 km. Cette proximité avec les enveloppes peut s'expliquer par la présence de pays plus connus, pour lesquels l'encyclopédie a fourni les coordonnées anciennes.

Dans le cas des limites modernes, il y a eu plus de correspondances. Cependant, si l'encyclopédie n'avait pas fourni les coordonnées anciennes, les articles se verraient attribuer une longitude de $-17,6627^\circ$ (en raison du changement de méridien de référence) et une latitude de 0° . Par conséquent, la distance pourrait augmenter considérablement, par exemple pour Lisbonne, une ville située sur la côte ouest de l'Europe, la distance entre un point avec ces coordonnées et Lisbonne approcherait les 4400 km. Ainsi, pour les points plus à l'est de l'Europe, la distance serait encore plus grande.

Cette analyse se concentre sur l'Europe, car c'est le continent qui compte le plus grand nombre d'articles, comme ce sera discuté ultérieurement.

Variable	Distance NE_1 au Enveloppe	Seuil Fuzzy Wuzzy NE_1 et Enveloppe	Distance NE_1 au Pays/Continent	Seuil Fuzzy Wuzzy NE_1 et Pays/Continent
Nb. non nul	3 352	3 352	5 178	5 178
Moyenne	2 531,39	97,01	2 618,71	96,17
Écart type	2 966,30	8,15	2 898,08	8,89
Minimum	0,00	70	0,00	70
25% (Q1)	0,00	100	0,00	100
50% (Q2)	111,32	100	945,22	100
75% (Q3)	5 571,83	100	5 014,47	100
Maximum	15 810,16	100	15 425,07	100

Tableau 3. Informations statistiques des variables distance géodésique et seuil d'égalité de tous les articles.

Dans les diagrammes en violon présentés dans la **Figure 9**, on observe une concentration des articles dans des correspondances exactes, mise en évidence par la région la plus dense du graphique, avec des critères d'égalité compris entre 95 et 100.

En ce qui concerne la distance, on observe deux régions principales de concentration. La première est proche de zéro, ce qui indique que ces articles se trouvent à l'intérieur de leurs polygones de correspondance. La deuxième région varie : pour les enveloppes, elle se situe entre 2 500 et 7 000 km, tandis que pour les limites modernes, elle se situe entre 2 000 et 8 500 km. Cela confirme ce qui a été expliqué précédemment, démontrant que les limites modernes offrent plus d'options et, par conséquent, la distance tend à être plus grande lorsque les coordonnées ne sont pas disponibles.

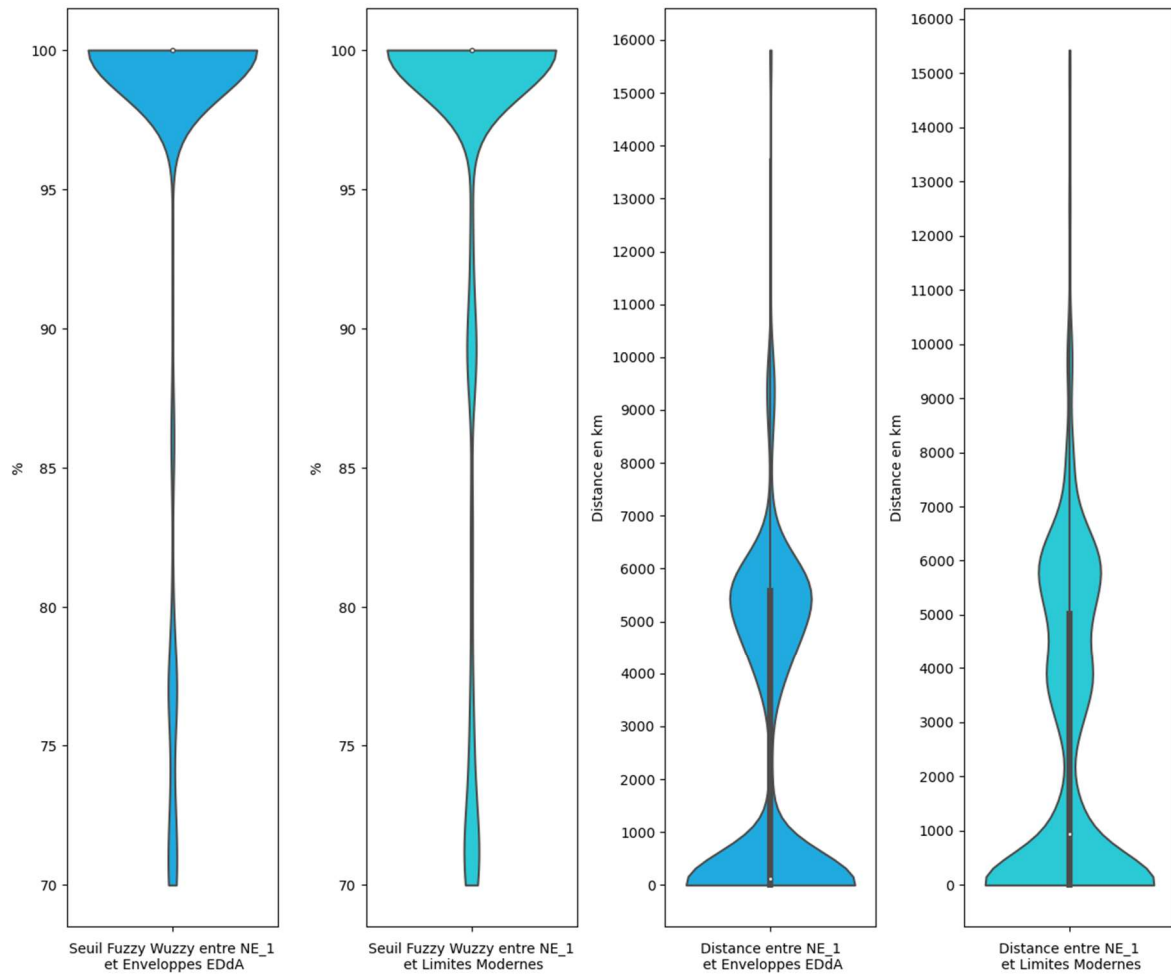


Figure 9. Diagramme en violon des variables seuil d'égalité et distance géodésique des tous les articles.

Dans la **Figure 10**, l'histogramme du critère d'égalité Fuzzy Wuzzy est présenté, ce qui confirme ce qui a été discuté précédemment, mettant en évidence qu'une grande partie des articles a obtenu des correspondances exactes. Pour les enveloppes, environ 3000 articles ont atteint un critère d'égalité de 100, tandis qu'un petit sous-ensemble seulement a obtenu un critère d'égalité proche du minimum défini à 70.

Dans le cas des limites modernes, les résultats sont similaires. Environ 4000 articles ont présenté une correspondance exacte, tandis qu'on peut également observer une petite concentration de correspondances entre 90 et 95. De plus, une autre partie des articles a présenté des valeurs proches de la limite minimale, variant entre 70 et 75.

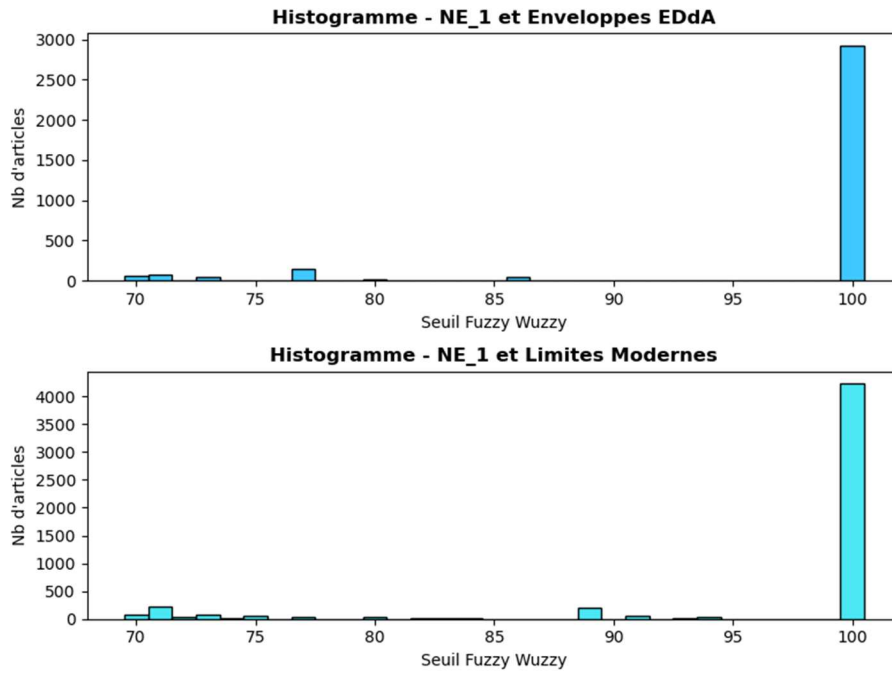


Figure 10. Histogramme de la variable seuil d'égalité de tous les articles.

Dans les graphiques de nuage des points présentés dans la **Figure 11**, il est possible d'observer la relation entre les deux variables, le critère d'égalité et la distance géodésique. Il est évident que les limites modernes offrent plus de possibilités de correspondance, car leur graphique présente une plus grande densité de points et une plus grande variété de combinaisons entre les variables.

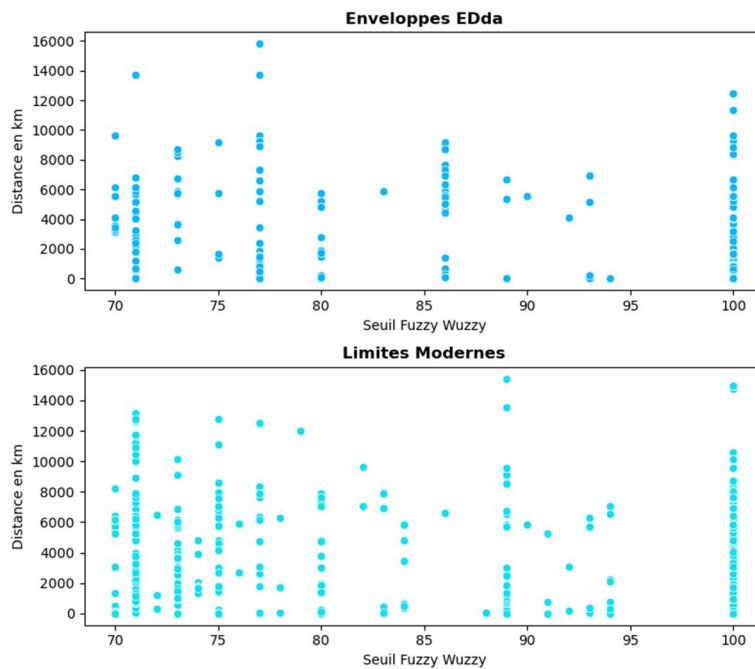


Figure 11. Nuage des points entre les variables seuil d'égalité (axe X) et distance géodésique (axe Y) de tous les articles.

Dans le **Tableau 4**, il est présenté les articles avec des coordonnées définies par l'encyclopédie, auxquels la méthodologie a été appliquée. On peut constater l'importance de la présence de coordonnées anciennes. Avec le filtre des articles contenant des coordonnées définies par l'encyclopédie, on constate qu'au moins 75 % des articles présentent une faible distance ou sont inclus dans les polygones de correspondance (distance égale à zéro).

De plus, la différence par rapport à la moyenne des points a été considérablement réduite. Auparavant, l'écart type était proche de 3000 km, maintenant cette valeur a été réduite à 1244,02 km pour les enveloppes et 1499,27 km pour les limites modernes. Une plus grande dispersion est attendue pour les limites modernes en raison de la plus grande possibilité de correspondances que cette couche fournit.

Encore une fois, le même schéma a été observé pour le critère d'égalité, avec une correspondance exacte se produisant dans la majorité des articles, représentée par les trois quartiles égaux à 100.

Variable	Distance NE_1 au Enveloppe	Seuil Fuzzy Wuzzy NE_1 et Enveloppe	Distance NE_1 au Pays/Continent	Seuil Fuzzy Wuzzy NE_1 et Pays/Continent
Nb. non nul	1 777	1 777	2 616	2 616
Moyenne	306,85	97,59	402,54	97,26
Écart type	1 244,02	7,27	1 499,27	7,65
Minimum	0,00	70	0,00	70
25% (Q1)	0,00	100	0,00	100
50% (Q2)	0,00	100	0,00	100
75% (Q3)	0,00	100	20,58	100
Maximum	12 496,92	100	15 425,07	100

Tableau 4. Informations statistiques des variables distance géodésique et seuil d'égalité des articles avec coordonnées anciennes.

Le diagramme en violon présenté dans la **Figure 12**, concernant les articles avec des coordonnées définies par l'encyclopédie, a montré une modification notable par rapport au diagramme précédent.

Auparavant, il y avait deux régions de concentration pour la variable de distance, mais maintenant on observe seulement une seule région proche de zéro. Ce résultat souligne l'importance des coordonnées anciennes, car en les fournissant, la probabilité de relier le point défini par ces coordonnées avec les informations textuelles géographiques décrites par l'encyclopédie augmente. Par conséquent, la méthodologie employée, qui consiste à extraire les entités géographiques et à calculer la distance, devient plus précise.

En ce qui concerne le critère d'égalité, on observe le même schéma de concentration principale proche de la valeur 100, indiquant une correspondance exacte.

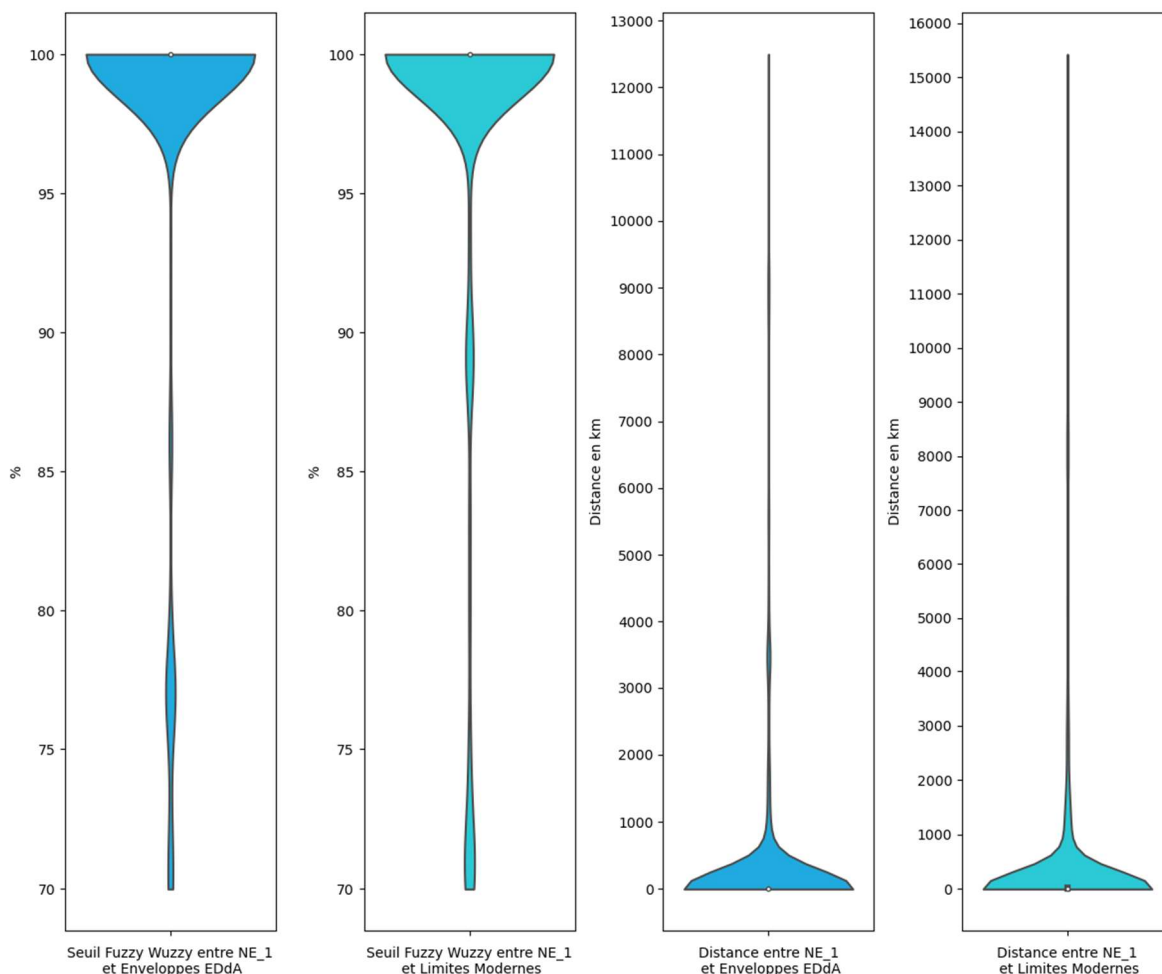


Figure 12. Diagramme en violon des variables seuil d'égalité et distance géodésique des articles avec des coordonnées anciennes.

L'histogramme présenté dans la **Figure 13** montre la même distribution que l'histogramme de tous les articles (**Figure 10**), mettant en évidence une grande proportion de correspondances exactes.

Dans les graphiques nuage de points des articles avec des coordonnées définies par l'encyclopédie (**Figure 14**), on observe un schéma similaire au graphique de dispersion de tous les articles (**Figure 11**). Les limites modernes présentent plus de combinaisons de valeurs entre les variables critère d'égalité et distance géodésique en raison du plus grand nombre de polygones, ce qui augmente la probabilité de correspondances.

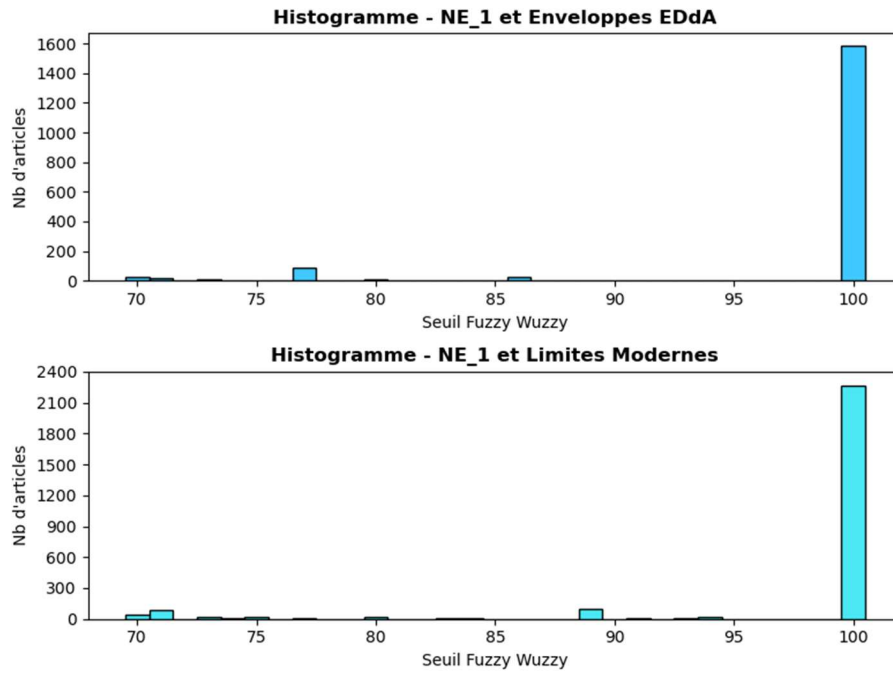


Figure 13. Histogramme de la variable seuil d'égalité des articles avec des coordonnées anciennes.

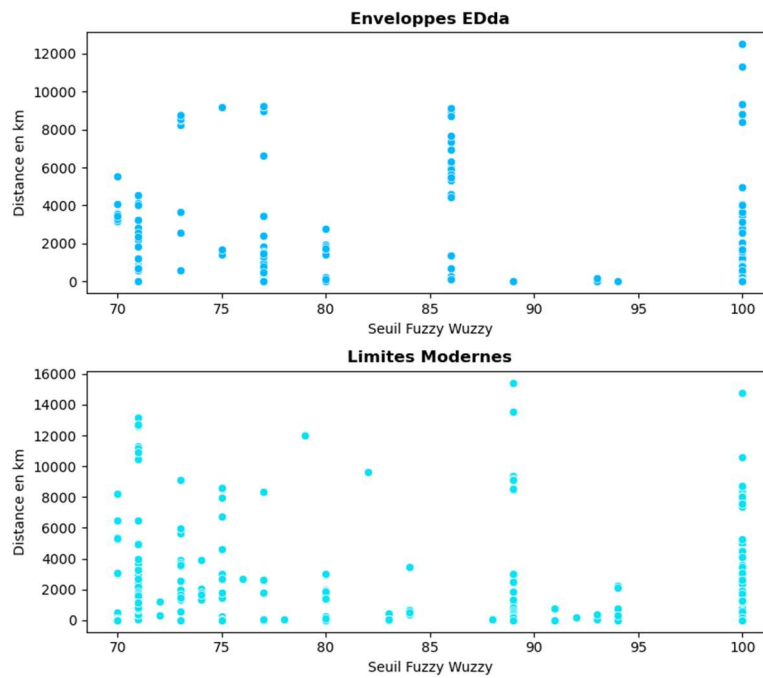


Figure 14. Nuage des points entre les variables seuil d'égalité (axe X) et distance géodésique (axe Y) des articles avec des coordonnées anciennes.

6.2.2. Analyse des 125 articles

Les résultats de l'analyse des 125 articles, divisés en 5 catégories de 25 articles chacune, seront discutés ci-dessous.

Dans la première catégorie, qui comprend les articles avec des coordonnées anciennes et une distance égale à zéro, il a été constaté que les valeurs égales à zéro correspondaient à des points situés à l'intérieur du polygone de correspondance, indiquant ainsi que la méthodologie était efficace. Cependant, certains articles ont été décrits comme des villes de leurs continents, sans mention de leurs pays, soulignant la nécessité d'améliorations pour déterminer le pays de l'article et augmenter la précision géographique.

Pour la deuxième catégorie, qui inclut les articles avec des coordonnées anciennes et une distance supérieure à zéro mais inférieure à 100 km, il a été identifié que les principaux facteurs causant cette distance étaient les coordonnées décrites dans l'encyclopédie, les changements territoriaux au fil de l'histoire et le manque de précision pour les pays du continent américain. Généralement, les coordonnées de l'encyclopédie nécessitaient des ajustements d'environ $0,5^\circ$ de longitude ou de latitude. Les changements territoriaux ont été observés dans les cas suivants : entre l'Allemagne et le Danemark, l'Italie et la Slovénie, et la France et la Belgique. De plus, le manque de précision pour les pays du continent américain s'est produit dans des cas tels que l'article "PARILLA, Santa", qui aurait dû être au Pérou, mais était au Panama. D'autres cas similaires se sont produits avec les articles "Isabelle" et "HALPO, ou HALAPO".

Dans la troisième catégorie, qui comprend les articles avec des coordonnées anciennes et une distance supérieure à 100 km, deux facteurs liés à la méthodologie ont été identifiés comme étant à l'origine de la distance élevée. Le premier d'entre eux était la présence d'erreurs de correspondance dues au Fuzzy Wuzzy. Le deuxième facteur était l'extraction des entités géographiques, la méthodologie n'ayant pas été en mesure de comprendre comment l'encyclopédie a décrit l'article. De plus, d'autres facteurs ont contribué au manque de précision, tels que l'absence d'une des coordonnées, une latitude incorrecte et des changements territoriaux au fil de l'histoire.

Pour la quatrième catégorie, qui englobe les articles avec des coordonnées anciennes mais sans distance, plusieurs facteurs ont été identifiés comme empêchant le calcul de la distance. Cela incluait des articles qui ne faisaient pas référence à des villes, des articles dans lesquels le mot « ville » était stocké dans les entités « PLACE_2 » ou « PLACE_3 », des correspondances inférieures à 70 % pour les pays ou les continents modernes, différents mots utilisés pour décrire les villes (bourg, village, capitale, comté, etc.), des informations géographiques situées après la position d'extraction des entités et

des erreurs d'écriture dans l'encyclopédie. Ces erreurs sont liées à la façon dont l'encyclopédie a décrit l'article.

Dans la cinquième catégorie, qui contient les articles sans coordonnées, plusieurs facteurs ont été observés comme empêchant le calcul de la distance. Sur les 25 articles, 15 n'étaient pas des villes, ce qui a rendu le calcul de la distance impossible. De plus, dans 3 articles, le manque de coordonnées a entraîné une distance très élevée par rapport au polygone de correspondance. Cinq autres articles n'ont pas eu leur distance calculée en raison de correspondances inférieures à 70 %. Dans les 2 articles restants, l'impossibilité de calculer la distance est due à l'extraction des coordonnées. Pour un article, le mot « ville » a été stocké uniquement dans la colonne « PLACE_2 ». Pour le dernier article, l'encyclopédie a utilisé le mot « bourg » juste avant le mot « France ».

Ainsi, on observe que la méthodologie utilisée s'est avérée efficace dans la première catégorie, mais nécessite des améliorations pour mieux extraire les entités géographiques et aborder les diverses possibilités. Le tableau avec les observations détaillées pour chacun des 125 articles analysés est disponible en Annexe B - Tableau des observations d'analyse des 125 articles.

6.2.3. Analyse des articles par pays et continent

L'analyse des pays ou continents avec le plus grand nombre de villes décrites dans l'encyclopédie a été réalisée en considérant la méthodologie de distance géodésique. Il a été sélectionné les résultats les plus précis (distance inférieure ou égale à 100 km et critère d'égalité supérieur ou égal à 90).

Les résultats sont présentés sur la carte (**Figure 15**) et dans le **Tableau 5**. Sur la carte, seulement les pays contenant des informations sont représentés, sinon il peut avoir des cas que seulement les continents ont été cartographiés.

Il a été constaté que l'Europe possède la plus grande quantité d'informations, en total 1 587 villes, avec notamment 518 villes pour la France et 366 villes pour l'Allemagne. Cependant, aucun résultat n'est trouvé pour la Belgique. Cette absence peut être attribuée aux changements territoriaux et aux variations de nomenclature au fil de l'histoire de ce pays. En raison de sa vaste superficie, la Russie affiche une densité urbaine de moins de 1 ville par 1 000 000 km², avec seulement 5 villes répertoriées dans l'encyclopédie. De même, le Luxembourg, en raison de sa petite taille, présente une densité urbaine plus grande (≤ 1 025 villes) avec seulement 2 villes répertoriées.

Il est important de souligner que les articles situés dans les pays d'Amérique du Sud et d'Amérique du Nord ont encore été décrits comme des villes de ces continents, et non de leurs pays spécifiques. Cela s'explique par la période historique pendant laquelle l'encyclopédie a été rédigée, lorsque moins d'informations étaient disponibles sur le nouveau continent et surtout car les pays modernes n'existaient pas encore. La même situation se produit sur le continent africain.

En revanche, en Asie, il a été constaté que l'encyclopédie décrit des villes de pays spécifiques, telles que la Chine (19 villes), le Japon (5 villes) et l'Inde (4 villes).

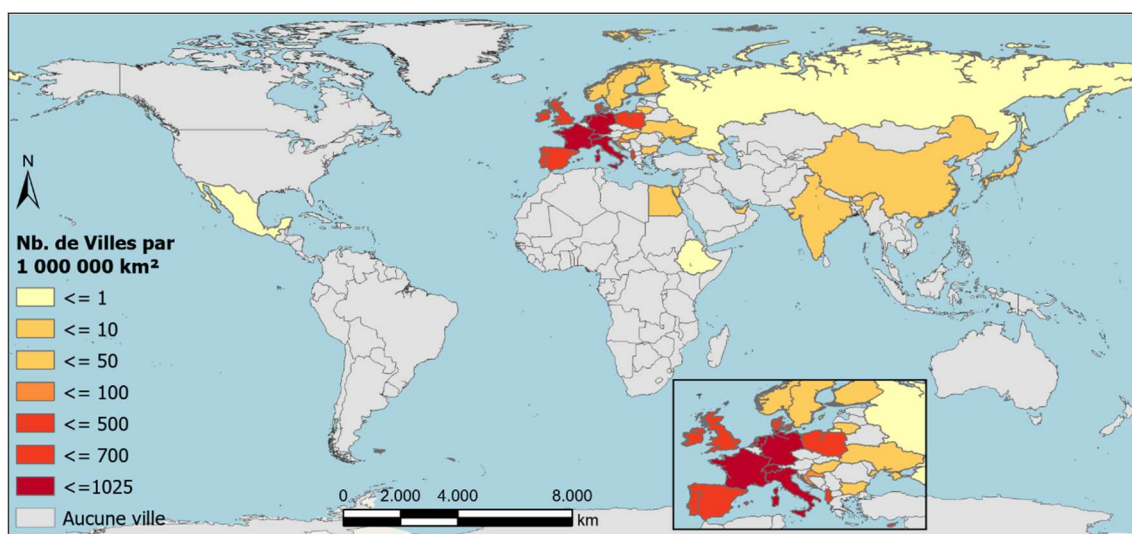


Figure 15. Carte du nombre des correspondances par limites modernes.

Continent	Nombre d'articles
Afrique	91
Amérique	196
Asie	157
Europe	1 587

Tableau 5. Nombre d'articles par continents.

6.3. Géocodage

Le géocodage a été appliquée exclusivement aux articles identifiés comme villes de France, totalisant 890 enregistrements parmi les 1335 articles où la colonne « NE_1 » contient « France » et la colonne « PLACE_1 » contient « ville ». Sur les 890 articles, le géocodage a donné des résultats pour 648 d'entre eux, représentant 72,8% du total. La couche de géocodage résultante contient 1241 points avec des coordonnées modernes extraites de la base de données de Nominatim.

Il est important de souligner que le nombre de points (1241) est supérieur au nombre d'articles (648) en raison de la possibilité qu'un même article ait plusieurs résultats de géocodage. Le maximum a été de 19 résultats pour un seul article.

La méthodologie utilisée a été considérée comme satisfaisante sur la base de ces résultats. Cependant, des améliorations ont été identifiées, notamment en ce qui concerne les problèmes de correspondance entre le terme utilisé par Nominatim et le mot extrait de l'encyclopédie, stocké dans la colonne « PLACE_1 ».

En cas de non-correspondance entre les deux mots, les résultats peuvent être limités. Par exemple, l'article « MONT-SAINT-MICHEL, sur mer », une ville de France, a été interprété par Nominatim comme une attraction touristique, et dans la base de données d'OpenStreetMap, il ne possède pas le paramètre « type » égal à « administratif ». En conséquence, un lieu du même nom a été trouvé, mais situé au Canada.

Une solution possible pour résoudre ce problème serait d'utiliser une boîte de délimitation pour limiter les résultats au pays d'intérêt. Cependant, cela ne résoudrait pas le problème de Nominatim interprétant erronément un article.

Une solution complémentaire serait d'utiliser les coordonnées anciennes des articles, le cas échéant, pour vérifier les résultats obtenus par le géocodage. Si les résultats se trouvent dans un rayon de distance prédéfini des coordonnées anciennes, ils seraient considérés comme valides, offrant à l'utilisateur plus d'options pour déterminer les coordonnées modernes de l'article.

7. Application Experience Builder

L'application développée dans Experience Builder a atteint les objectifs initialement fixés et est publiquement accessible via le lien : <https://experience.arcgis.com/experience/a1833d64dc904f3f9408d4b59008b720/>. Il est important de noter que l'utilisation sur un grand écran offre un espace de visualisation plus large des cartes, facilitant ainsi son utilisation.

Avec cette application, les utilisateurs peuvent visualiser les points et les enveloppes des articles avec des coordonnées anciennes. Pour les points sans coordonnées anciennes, ils ont été représentés avec des valeurs standardisées de longitude (-17,6627°, ce n'est pas 0° en raison de la différence de méridien de référence) et de latitude (0°).

La présence de deux cartes permet la séparation des fonctions disponibles. La carte « Les articles de l'Encyclopédie de Diderot et d'Alembert liés à la géographie » a pour objectif la visualisation des points, en utilisant deux symbologies différentes pour mettre en évidence les points les plus prioritaires. En revanche, la carte « Validation des articles de l'EDdA avec l'aide du geocodage de Nominatim » permet l'édition des articles.

De plus, l'application permet aux utilisateurs d'appliquer divers filtres à l'aide des widgets disponibles sur Experience Builder, rendant ainsi son utilisation interactive et dynamique. Le widget d'édition permet de compléter les informations géographiques des articles tout en conservant les données d'origine de l'encyclopédie et en ajoutant de nouvelles informations.

Malgré les objectifs atteints, certaines améliorations peuvent être apportées. La fonction de prévisualisation dans Experience Builder doit être améliorée par Esri, car les résultats affichés ne correspondent pas toujours à la version finale après la publication, nécessitant des ajustements supplémentaires. À plusieurs reprises, la taille des widgets ou la taille de la police des titres a été définie, mais lors de la publication, ils ont été déformés ou leur visualisation était compromise.

De plus, certains widgets n'ont pas de paramètres courants dans un SIG, tels que masquer les étiquettes ou les tables attributaires, ce qui limite le développement selon les besoins spécifiques.

Esri propose une version pour les développeurs d'Experience Builder, permettant la création de nouveaux widgets avec des fonctionnalités personnalisées. Cependant,

l'utilisation de cette version nécessite des connaissances en JavaScript, TypeScript et CSS, qu'il n'était pas possible d'acquérir dans le délai du stage.

8. Conclusion

Tout au long de ce stage, il y a eu d'importants défis notamment en raison de la diversité des formes et du manque de normes dans les descriptions des textes de l'encyclopédie. L'existence de plusieurs représentations pour un même concept a compliqué l'application d'une méthodologie unique pour l'extraction des informations géographiques. Par exemple, pour désigner le terme « ville », les termes « bourg », « village », « capitale », ou « comté » peuvent être utilisés.

Un autre point important était la question du changement de noms de villes, de pays et d'autres entités au fil de l'histoire, ce qui a également eu un impact sur la précision et la cohérence des résultats obtenus.

Pour surmonter ces défis, il a été nécessaire d'adapter la méthodologie et d'utiliser des techniques telles que REGEX pour extraire plus efficacement les informations des enveloppes géographiques.

Au cours du développement de l'application dans l'Experience Builder, il a été remarqué que certaines fonctionnalités des widgets disponibles ne répondaient pas entièrement aux besoins du projet. En conséquence, nous avons identifié la nécessité d'apprendre d'autres langages, tels que JavaScript, TypeScript et CSS, pour créer des widgets plus appropriés et personnalisés pour l'application.

Durant le développement de l'application dans Experience Builder, il a été constaté que certaines fonctionnalités des widgets disponibles ne répondaient pas complètement aux exigences spécifiques du projet. Cependant, pour mettre en œuvre des fonctionnalités plus personnalisées et améliorées, une connaissance des langages de programmation JavaScript, TypeScript et CSS serait nécessaire. Malgré les obstacles rencontrés, les résultats obtenus étaient satisfaisants. L'application développée a permis la visualisation des points et des enveloppes des articles avec des coordonnées anciennes, ainsi que l'édition et le complètement des informations géographiques des articles.

La méthodologie utilisée pour géocoder les articles à partir de la librairie Perdido s'est révélée prometteuse et peut être appliquée à d'autres pays, élargissant ainsi son potentiel d'utilisation au-delà des villes. La flexibilité de l'approche permet également son extension à d'autres entités géographiques.

Sur la base de cette expérience, il est évident que l'étude et l'application de méthodologies pour l'extraction d'informations géographiques à partir de textes historiques, tels que l'Encyclopédie de Diderot et d'Alembert, revêtent une importance significative en complétant l'aspect géographique de ces documents. Grâce à ces méthodologies, il est possible de récupérer et d'enrichir les connaissances sur la géographie historique, ainsi que d'explorer leur applicabilité dans les recherches et les études impliquant l'évolution des lieux et des territoires au fil du temps.

9. Références bibliographiques

Perdido. **Perdido documentation**. Disponible sur : <https://github.com/GEODE-project/perdido-geoparsing-notebook>. [Consulté le 8 mars 2023].

GeoPandas. **Geopandas documentation**. Disponible sur : <https://geopandas.org/en/stable/>. [Consulté le 10 mai 2023].

Geopy. **Geopy documentation**. Disponible sur : <https://geopy.readthedocs.io/en/latest/#>. [Consulté le 20 mai 2023].

Shapely. **Shapely documentation**. Disponible sur : <https://shapely.readthedocs.io/en/stable/index.html>. [Consulté le 25 mai 2023].

Selenium. **Selenium documentation**. Disponible sur : <https://selenium-python.readthedocs.io/index.html>. [Consulté le 18 juin 2023].

Nominatim. **Map Features OpenStreetMap**. Disponible sur : https://wiki.openstreetmap.org/wiki/Map_features. [Consulté le 20 juin 2023].

Fuzzy Wuzzy. **Fuzzy Wuzzy documentation**. Disponible sur : <https://pypi.org/project/fuzzywuzzy/>. [Consulté le 20 juin 2023].

Experience Builder. **What is ArcGIS Experience Builder?**. Disponible sur : <https://doc.arcgis.com/en/experience-builder/latest/get-started/what-is-arcgis-experience-builder.htm>. [Consulté le 24 juin 2023].

Experience Builder. **Widgets**. Disponible sur : <https://doc.arcgis.com/en/experience-builder/latest/configure-widgets/widgets-overview.htm>. [Consulté le 24 juin 2023].

Pandas. **Pandas documentation**. Disponible sur : <https://pandas.pydata.org/docs/>. [Consulté le 10 juillet 2023].

Seaborn. **Seaborn documentation**. Disponible sur : <https://seaborn.pydata.org/>. [Consulté le 11 juillet 2023].

Matplotlib. **Matplotlib documentation**. Disponible sur : <https://matplotlib.org/stable/index.html>. [Consulté le 11 juillet 2023].

10. Annexes

Annexe A - Enveloppes EDdA

Les 73 enveloppes EDdA extraites de la feuille de calcul de surface sont présentées ci-dessous. Les titres sont le même que stockés dans la colonne Head de chaque article.

Enveloppes EDdA		
ABISSINIE	FEMEREN ou FEMERN	MINORQUE
ABLAÏ	FLORIDE	MOLDAVIE, Moldavia
ABRUZZE	FORMOSE	NICARIA ou Nicarie
ACADIE ou ACCADIE	FRANCE	NORWEGUE
ADERBIJAN	FREESLAND	NUBIE
AFRIQUE	FUNEN ou FUYNEN	ONÉGA lac d'
ALBANIE	GHILAN	PANAY
ALLEMAGNE	GINS-ENG	PERSE, la
ALSACE	GOTHLAND, (l'île de)	PHILIPPINES, les
ANCONE, (La Marche d')	GROENLAND, (le)	Philippines, les nouvelles, ou les îles de Palaos
ANDALOUSIE	Guinée, (la Nouvelle)	POLOGNE
Ange (Saint)	HAINAN	PONT-EUXIN
ANGLETERRE	HONGRIE	Provinces-Unies
ANTILLES	HURONS lac des	SARDAIGNE, la
ARABIE	IRLANDE	Scyros
BORNO ou BOURNOU	Isles du Cap-verd, les	SCYTHES
CANARIES (les îles)	JAPON, le	SI-FAN
CASPIENNE (la mer)	JAVA l'île de	Sonde, îles de la
DALHACA ou DALACA	KAMTSCHATKA	SUEDE
DAUPHINE	KOPING	Suisse
ESPAGNE	LABRADOR, Estotilandia	Tartares ou Tatars
ETHIOPIE	LALAND, Lalandia	Terres australes
EUROPE	LEEWIN, la terre de	USBECKS
FALSTER	LÉMAN, le lac	
FARTACH	LYCAONIE, Lycaonia	

Annexe B – Tableau des observations d’analyse des 125 articles

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
15v3873	15	3873	Tafilet	Afrique	0,00	d0	Precision de continent.
2v4717	2	4717	CAKET	Asie	0,00	d0	Precision de continent.
10v2965	10	2965	MONTEREAU-FAUT-YONNE	France	0,00	d0	-
4v727	4	727	CONZA	Italie	0,00	d0	-
9v650	9	650	KILLALLOW	Irlande	0,00	d0	-
8v3814	8	3814	ISLEB	Allemagne	0,00	d0	-
9v772	9	772	KORSOÉ ou KORSOR	Danemark	0,00	d0	-
4v1000	4	1000	CORIA	Espagne	0,00	d0	-
12v919	12	919	PEGNAFIEL	Espagne	0,00	d0	-
1v2003	1	2003	AMADAN	Asie	0,00	d0	Il y a une ville, proche qu'il s'appelle Hamedan.
9v1640	9	1640	LECCO	Italie	0,00	d0	-
7v2566	7	2566	GRACCHURIS	Espagne	0,00	d0	il s'agit d'une ville ancienne.
1v2590	1	2590	ANDERNACH	Allemagne	0,00	d0	-
1v1946	1	1946	ALTENA, ou ALTENAW	Allemagne	0,00	d0	-
13v214	13	214	PONT-ORSON	France	0,00	d0	-
14v1480	14	1480	RHEIMS ou REIMS	France	0,00	d0	-
2v120	2	120	BADENWEILER	Allemagne	0,00	d0	-
11v3110	11	3110	OSMA	Espagne	0,00	d0	-
14v2326	14	2326	ROSTOCK	Allemagne	0,00	d0	-
10v2997	10	2997	MONTPENSIER	France	0,00	d0	-
3v951	3	951	CHAROUX	France	0,00	d0	-
1v3474	1	3474	AQUINO	Italie	0,00	d0	-

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
10v977	10	977	Mataram	Asie	0,00	d0	Precision de continent.
10v2999	10	2999	MONT-RÉAL	Espagne	0,00	d0	-
16v3898	16	3898	TURIN	Italie	0,00	d0	-
10v3272	10	3272	MOTRIL	Espagne	35,04	dm100	Long: 1° dif; Lat: -0,5° dif.
12v2785	12	2785	PINHEL	Portugal	36,62	dm100	Long: 1° dif; Lat: -0,1° dif.
12v2878	12	2878	PIRANO	Italie	23,34	dm100	Il appartient à Slovenie maintenant. Long: 0,5° dif; Lat: 0,3° dif.
9v2604	9	2604	LINCOLN	Angleterre	60,06	dm100	Long: 2,5° dif.
17v697	17	697	VICO-AQUENSE	Italie	8,42	dm100	Long: -0,2° dif.
11v216	11	216	NASSARI ou NAUSARI	Inde	19,20	dm100	Long: -0,7 ° dif.
4v696	4	696	CONVERSANO	Italie	13,07	dm100	Lat: 0,2° dif.
11v4604	11	4604	PARILLA, Santa	Amérique Méridionale	47,93	dm100	Il était censé d'être au Pérou, il est au Panama. Il faut réduire une dizaine de degrés.
4v2616	4	2616	CROTONE	Italie	27,49	dm100	Long: 0,3° dif; Lat: 0,1 ° dif.
9v2904	9	2904	LLIVIA	Espagne	2,41	dm100	Lat: 0,3° dif.
14v4855	14	4855	SEGEBERG	Danemark	67,06	dm100	Il appartient à Allemagne.
16v3285	16	3285	TRIESTE	Italie	32,37	dm100	Long: 0,4° dif.; Lat: 0,2° dif.
15v4540	15	4540	Tatah ou Tata	Inde	4,17	dm100	Proche de la côte ouest de l'Inde
1v4606	1	4606	ATHENES	Grèce	0,91	dm100	Long: 0,5° dif. Lat: 0,1 ° dif.
8v1248	8	1248	HOCHSTET	Allemagne	63,82	dm100	Long: 1,8° dif; Lat: -0,3° dif.
16v4623	16	4623	VEGEL, VEGER, & BEGÈ ou BEGER	Espagne	23,13	dm100	Long: -0,2° dif; Lat: -0,2 ° dif.
14v2795	14	2795	RYE	Angleterre	6,14	dm100	Long: 0,04° dif; Lat: -0,1° dif.

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
8v3748	8	3748	Isabelle	Amérique	16,82	dm100	Proche de la côte est de la République Dominicaine.
1v180	1	180	ABER-YSWITH	Angleterre	18,16	dm100	Long: -0,3° dif; Lat: -0,1 ° dif.
8v216	8	216	HALPO, ou HALAPO	Amérique	78,97	dm100	Proche de la côte ouest de Mexique.
12v2039	12	2039	PHILIPPEVILLE	France	24,49	dm100	Il appartient à Belgique.
17v1337	17	1337	UMBRIATICO	Italie	6,81	dm100	Long: 0,3° dif; Lat: 0,1 ° dif.
2v5711	2	5711	CARLOWITZ	Hongrie	80,85	dm100	Proche du sud de l'Hongrie.
17v2212	17	2212	WESTERVICK	Suède	50,88	dm100	Long: 1° dif; Lat: 0,2° dif.
15v4404	15	4404	TARIFFE	Espagne	7,42	dm100	Long: 0,4° dif; Lat: -0,2° dif.
2v406	2	406	BAMFE	Amérique Septentrionale	2 020,98	dM100	Erreur de correspondance. Ecosse septentrionale > Amérique Septentrionale.
8v2643	8	2643	ILLOCK	Nubie	3 923,04	dM100	Erreur d'extraction des entités géographiques. Il est une ville de basse-Hongrie.
10v678	10	678	MARQUEFAVE	France	691,10	dM100	Lat: -7° dif.
1v1944	1	1944	ALTEMBOURG	Tanzanie	5 332,25	dM100	Erreur de correspondance. Transylvanie en Roumanie > Tanzanie.
2v1536	2	1536	Belgrade	Tchéquie	930,97	dM100	Erreur d'extraction des entités géographiques. Il est une ville de la Roumanie.
1v3149	1	3149	APHIOM-KARAHISSART	Tchéquie	1 560,07	dM100	Erreur de correspondance. Turquie > Tchéquie.
7v1440	7	1440	Gallipoli	Tchéquie	1 184,38	dM100	Erreur de correspondance. Turquie > Tchéquie.

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
5v243	5	243	DORAR	Maurice	8 340,97	dM100	Erreur d'extraction des entités géographiques. Il est une ville de la France.
9v3892	9	3892	MAGHIAN	Arabie	1 166,15	dM100	Erreur de correspondance. L'encyclopédie ne précise pas quelle Arabie, la correspondance a été fait avec Émirats Arabes Unis.
2v1320	2	1320	BAUSK	Irlande	1 967,46	dM100	Erreur de correspondance. Curlande en ancienne territoire polonais (actuellement Lettonie) > Irlande.
12v895	12	895	PEDIR	Inde	13 556,89	dM100	Probable erreur des coordonnées de l'encyclopédie.
16v2280	16	2280	TOULOUBAN	Inde	341,52	dM100	Probable erreur des coordonnées de l'encyclopédie. La ville Multan, ainsi le fleuve avec le meme nom, utilisée comme référence est au Pakistan.
15v3026	15	3026	SUDBURY	Angleterre	1 547,37	dM100	Erreur d'extraction de la latitude de l'encyclopédie. La latitude correcte est 52° positives et n'est pas 52° négatives.
16v3827	16	3827	TULN	Allemagne	195,10	dM100	Territoire allemand a changé. Actuellement il est Autriche. Lat: 0,4° dif.
13v2294	13	2294	PUNTA-DEL-GUDA	Saint-Marin	3 081,22	dM100	Erreur de correspondance. Île de Saint-Michel en Açores > Saint-Marin.
1v3934	1	3934	ARLON	Pays-Bas	111,94	dM100	Territoire du Pays-Bas a changé, maintenant cette ville appartient à la Belgique.

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
9v604	9	604	KHAIBAR	Arabie	168,83	dM100	Erreur de correspondance. L'encyclopédie ne précise pas quelle Arabie, la correspondance a été fait avec Émirats Arabes Unis. Cette ville est dans l'Arabie saoudite.
16v21	16	21	TÉCOANTEPEQUE	Amérique Septentrionale	3 037,87	dM100	Il manque la longitude.
16v1946	16	1946	Tolu	Amérique Méridionale	2 374,80	dM100	Il manque la latitude.
3v2796	3	2796	COKENHAUSEN	Suède	412,62	dM100	Territoire suédois a changé. Actuellement cette ville est dans la Lettonie.
11v672	11	672	NEWCASTLE	Angleterre	251,49	dM100	Long: 4° dif.
1v1648	1	1648	ALEP	Serbie	1 472,80	dM100	Erreur de correspondance. Syrie > Serbie. La couche utilisé contient République arabe syrienne, ce que cause une grande différence en relation à Syrie.
13v3794	13	3794	RAVENNE	Italie	260,89	dM100	Long: 5° dif.
4v1942	4	1942	COURTRAI	Autriche	657,74	dM100	Erreur d'extraction des entités géographiques. L'entité devrait être Pays-Bas autrichiens. Maintenant elle est dans la Belgique. Lat: 1° dif.
3v1709	3	1709	CHIMERA	Tchéquie	981,74	dM100	Erreur d'extraction des entités géographiques. Cette ville est censé d'être en Albanie, elle était une ville forte de la Turquie.

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
11v3735	11	3735	Palais			sans_dist	Il ne s'agit pas d'une ville. C'est une place forte.
8v3811	8	3811	Isles du Vent			sans_dist	Il ne s'agit pas d'une ville. C'est une île.
14v3853	14	3853	SATALIE			sans_dist	Il s'agit d'une ville. Cependant la méthodologie a extraite des entités qui ne font pas référence à localisation.
5v794	5	794	DURHAM			sans_dist	Colonne "PLACE_1" n'est pas "ville". Il s'agit d'une capitale de province anglaise.
10v1138	10	1138	MAYO ou MAY			sans_dist	Colonne "PLACE_1" n'est pas "ville". Il s'agit d'un comté.
1v4760	1	4760	AVA			sans_dist	Il ne s'agit pas d'une ville. C'est un royaume de l'Asie.
8v3706	8	3706	IRLANDE			sans_dist	L'encyclopédie a commencé à parlé des informations géographiques dans le deuxième paragraphe. Les entités extraites ne font pas référence à la localisation.
14v1951	14	1951	ROCHE-SUR-YON			sans_dist	Colonne "PLACE_1" n'est pas "ville". Il s'agit d'un bourg.
10v2361	10	2361	MIOLANS			sans_dist	Il ne s'agit pas d'une ville. C'est une forteresse.
16v4747	16	4747	VENAFRE			sans_dist	Colonne "PLACE_1" n'est pas "ville". Cependant, colonne "PLACE_2" est "ville".

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
14v3718	14	3718	SARBRUCK			sans_dist	Aucune correspondance a été possible avec un seuil d'égalité plus grand que 70%.
11v69	11	69	NAHARUAN			sans_dist	Aucune correspondance a été possible avec un seuil d'égalité plus grand que 70%.
13v534	13	534	PORTO-SANTO			sans_dist	Il ne s'agit pas d'une ville. C'est une île.
7v2752	7	2752	GRATZ			sans_dist	Colonne "PLACE_1" n'est pas "ville". Cependant, colonne "PLACE_2" est "ville".
16v1941	16	1941	Tolna			sans_dist	Colonne "PLACE_1" n'est pas "ville". Il s'agit d'un comté.
5v2752	5	2752	ENTREVAUX			sans_dist	Aucune correspondance a été possible avec un seuil d'égalité plus grand que 70%. Il s'agit d'une ville de France, mais il faut améliorer l'extraction des entités.
14v4287	14	4287	SCALG			sans_dist	Aucune correspondance a été possible avec un seuil d'égalité plus grand que 70%.
10v3238	10	3238	MOSKOW			sans_dist	L'encyclopédie a commencé a parlé des informations géographiques dans le deuxième paragraphe. Les entités extraites ne font pas référence à la localisation.

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
7v2092	7	2092	Gibraltar			sans_dist	La méthodologie n'a pas réussi a déconsidérer la partie du titre. La première entité devrait être la NE_2.
3v1463	3	1463	CHERZO			sans_dist	Il ne s'agit pas d'une ville. C'est une île.
10v733	10	733	MARSICO-NUOVO			sans_dist	Colonne "PLACE_1" n'est pas "ville". Cependant, colonne "PLACE_2" est "ville". Il faut améliorer l'extraction des entités, la méthodologie a extrait comme "NE_1" un autre nom donné à ce ville.
10v2989	10	2989	MONT-MEDI			sans_dist	Colonne "PLACE_1" n'est pas "ville". Cependant, colonne "PLACE_2" est "ville". Il faut améliorer l'extraction des entités, la méthodologie a extrait comme "NE_1" un autre nom donné à ce ville.
1v3467	1	3467	AQUILA			sans_dist	Encyclopédie a écrit "ville d'italie" et l'extraction n'était pas bonne.
1v671	1	671	ACRE			sans_dist	Il s'agit d'une ville, mais cette information était stockée dans la colonne "NE_3". L'encyclopédie a écrit des autres noms pour cette ville. Ces noms ont été considérés comme entités.

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
13v3432	13	3432	Rambert-le-joux			sans_dist	Colonne "PLACE_1" n'est pas "ville". Avant l'entité "France" l'encyclopédie a utilisé le mot "bourg", mais avant il y a le mot "ville". Il faut améliorer l'extraction des entités.
13v1372	13	1372	PREUILLY	France	3 804,80	sans_coord	Il manque les coordonnées. C'est pour ça que la distance est si grande.
11v3801	11	3801	PALESTINE			sans_coord	Il ne s'agit pas d'une ville. C'est un pays.
2v6332	2	6332	CECERIGO ou CERIGOTTO			sans_coord	Il ne s'agit pas d'une ville. C'est une île.
10v1085	10	1085	MAUGES les, ou le pays de Mauges			sans_coord	Il ne s'agit pas d'une ville. C'est une contrée.
2v2483	2	2483	BODANETZ			sans_coord	Aucune correspondance a été possible avec un seuil d'égalité plus grand que 70%.
10v407	10	407	Marche, la			sans_coord	Il ne s'agit pas d'une ville. C'est une province.
15v1321	15	1321	SOCHACZOW			sans_coord	Colonne "PLACE_1" n'est pas "ville". Cependant, colonne "PLACE_2" est "ville".
11v478	11	478	Némausus			sans_coord	Il ne s'agit pas d'une ville. C'est une fontaine à ville de Nismes.
15v4230	15	4230	TANITICUMOSTIUM			sans_coord	Il ne s'agit pas d'une ville. C'est un nom d'embouchure du Nil.
15v1284	15	1284	SMENUS			sans_coord	Il ne s'agit pas d'une ville. C'est un fleuve.

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
15v3168	15	3168	SUNIQUES, les			sans_coord	Il ne s'agit pas d'une ville. C'est un peuple de la Germanie.
9v1345	9	1345	LAPPA			sans_coord	Aucune correspondance a été possible avec un seuil d'égalité plus grand que 70%.
9v876	9	876	LABOURD (le)			sans_coord	Il ne s'agit pas d'une ville. C'est une contrée.
9v3625	9	3625	LYNCESTES			sans_coord	Il ne s'agit pas d'une ville. C'est un peuple de la Macédoine.
2v89	2	89	BACLAN			sans_coord	Il ne s'agit pas d'une ville. C'est un pays de la Perse.
15v2399	15	2399	STARACHINO	Tchéquie	6 193,76	sans_coord	Erreur de correspondance, Turquie > Tchèque. Cependant il s'agit d'une ville de Macédoine.
13v166	13	166	PONT-DE-ROYAN			sans_coord	Colonne "PLACE_1" n'est pas "ville". Avant l'entité "France" l'encyclopédie a utilisé le mot "bourg", mais avant il y a le mot "ville". Il faut améliorer l'extraction des entités.
17v2220	17	2220	Westminster, salle de			sans_coord	Il ne s'agit pas d'une ville. C'est une salle en Angleterre.
16v4745	16	4745	VEMPSUM	Italie	5 014,47	sans_coord	Il manque les coordonnées. C'est pour ça que la distance est si grande.
10v93	10	93	MANDRIA			sans_coord	Il ne s'agit pas d'une ville. C'est une île.

id_article	volume	numero	Head	pays_match	Distance au pays	Catégorie	Description
2v141	2	141	BAGE-LE-CHATEAU			sans_coord	Aucune correspondance a été possible avec un seuil d'égalité plus grand que 70%.
9v1831	9	1831	LEPETHYMNUS ou LEPETHYMUS			sans_coord	Il ne s'agit pas d'une ville. C'est une montagne.
1v2254	1	2254	AMIUAM			sans_coord	Il ne s'agit pas d'une ville. C'est une île.
2v2485	2	2485	BODENBURG			sans_coord	Aucune correspondance a été possible avec un seuil d'égalité plus grand que 70%.
13v2379	13	2379	PYDNA			sans_coord	Aucune correspondance a été possible avec un seuil d'égalité plus grand que 70%.