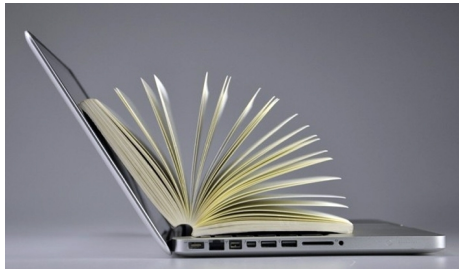
Sorbonne Nouvelle
université des cultures



Reconnaissance d'entités nommées

Ljudmila PETKOVIC

Introduction aux humanités numériques (L1HN001)
Mineure « Humanités numériques », licence Lettres
Paris, le 23 novembre 2023, année 2023-2024

Les projets en humanités numériques

Quatre principales étapes :

1. Acquisition d'un objet d’étude
2. Traitement d'un objet d’étude
3. Exploitation d'un objet d’étude
4. Publication des résultats

⚠ ***NB***:
Les frontières entre ces étapes sont très poreuses et l’ordre des opérations ne respecte pas toujours cette progression logique.

Les projets en humanités numériques

Quatre principales étapes :

1. Acquisition d'un objet d’étude
2. **Traitement d'un objet d’étude**
3. Exploitation d'un objet d’étude
4. Publication des résultats

⚠ ***NB***:
Les frontières entre ces étapes sont très poreuses et l’ordre des opérations ne respecte pas toujours cette progression logique.

Les projets en humanités numériques

Quatre principales étapes :

1. Acquisition d'un objet d’étude
2. **Traitement d'un objet d’étude**
3. **Exploitation d'un objet d’étude**
4. Publication des résultats

⚠ ***NB***:
Les frontières entre ces étapes sont très poreuses et l’ordre des opérations ne respecte pas toujours cette progression logique.

Entités nommées

Les entités nommées (EN) sont des types particuliers d'unités lexicales (groupes de formes) qui font référence à une entité du monde concret dans certains domaines spécifiques (humains, sociaux, politiques, économiques ou géographiques).

- associées aux noms propres et aux descriptions définies (p. ex. : *le chat noir*)
- trois catégories principales (dites angl. *coarse-grained*, « à gros grain ») :
  - noms de **personnes** : *Barack Obama*, *Hugo*...
  - noms de **lieux** : *rue de Rivoli*, *Mercure*...
  - noms d'**organisations** : *Sorbonne Nouvelle*, *ILPGA*...
- peuvent inclure : les fonctions de personnes (*le roi Henri IV*), les dates (*2023*), etc.

Une définition précise du nom propre ?

Du point de vue de la linguistique, la catégorie « nom propre » est difficile à définir, car il existe de nombreuses exceptions.

Les critères traditionnels :

- **forme** des mots – marque de majuscule (⚠️ *la gare de Montparnasse*)
- d'ordre **factuel** : la non traduction et l'absence de dictionnaires de la langue
- d'ordre **morphosyntaxique** : l'absence de déterminant et de flexion en théorie

  ⚠️ *la Seine*, *le Paris d'après-guerre* (emploi figuré), « J'ai acheté trois *Picassos* »...

Nom propre : La marque de la majuscule ?

- n’est **pas translinguistique**
  - usage différent d’une langue à l’autre, ex. : l’allemand
- n’est **pas valide en diachronie**
  - usage inexistant dans les corpus anciens
  - notion qui apparaît avec l'imprimerie : besoin de normaliser un corpus
- n’est **pas appréciable à l’oral** (aussi, besoin de normaliser)

Types d'ambiguïtés

- le **même nom** est utilisé pour **plusieurs entités**
  - *Paris* (France) et *Paris* (Texas)
- une **même entité** peut avoir **plusieurs noms**
  - *Paris*, *Lutèce*, *Lutetia Parisiorum*
- une entité qui désigner une entité en catégories différentes (**métonymie**)
  - *la Sorbonne* – une organisation et un lieu
  - « Le *prix Nobel de la littérature* s'est montré digne » – un prix et son lauréat·e

Difficultés à définir précisément une EN

les bornes / limites de l'EN

- *la rue de Strasbourg*
  i. une rue qui porte ce nom
  ii. une rue dont le nom n'est pas précisé, localisée à Strasbourg

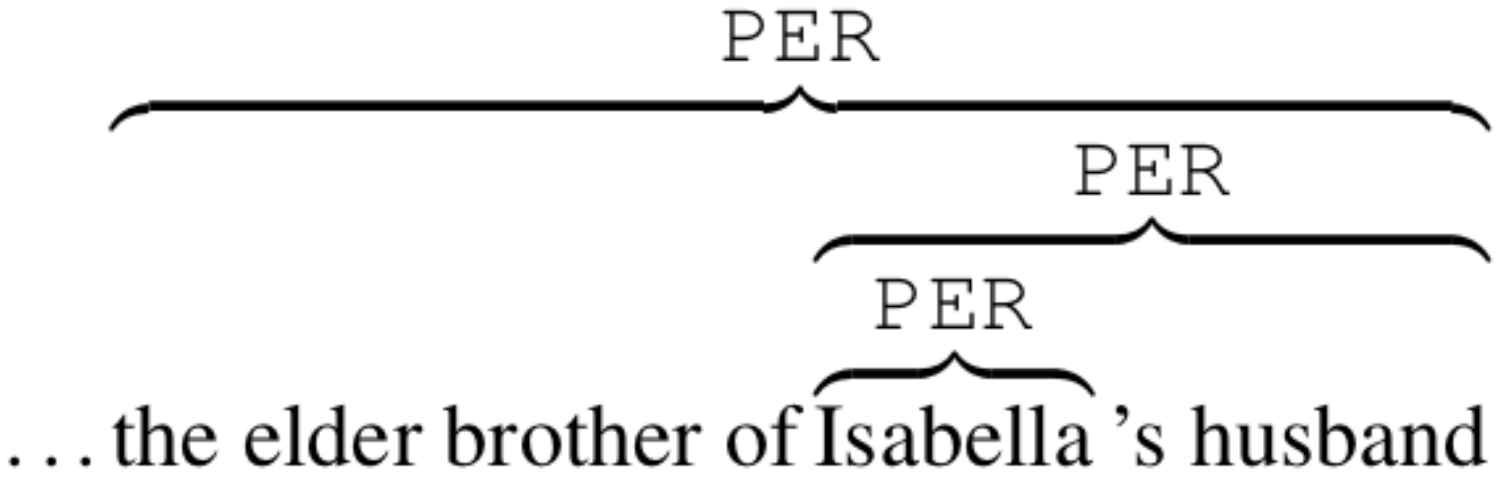
imbrication de l'annotation de plusieurs EN

- *le président de la France* : personne qui occupe la fonction + un pays (*France*)

un référent parfois flou, collectif ou historique

- *les côtes de la Guyane*, *le nord de l'Europe*, *La Bohême*

EN imbriquées



Annotation imbriquée d'une EN (Sims & Bamman, 2020).

Les campagnes d'évaluation de la REN en TAL

- The Lit Bank: an Annotated Dataset of Literary Entities (Bamman *et al.*, 2019)
  - exemples d'EN : *Tom Sawyer* ; certains noms communs (*le policier*)
  - consignes d'annotation avec les catégories retenues :
    - People (PER) : *Tom Sawyer*, *her daughter*
    - Facilities (FAC) : *the house*, *the kitchen*
    - Geo-political entities (GPE) : *London*, *the village*
    - Locations (LOC) : *the forest*, *the river*
    - Vehicles (VEH) : *the ship*, *the car*
    - Organizations (ORG) : *the army*, *the Church*
  - corpus littéraire de fiction (XVIII-XX s.) des auteurs anglophones (projet Gutenberg)

Caractéristiques les plus souvent utilisées pour la REN

- caractéristiques au niveau des mots
- listes d'entités : gazetiers (index géographiques, angl. *gazetteer*), dictionnaires...
- caractéristiques des documents et corpus

Caractéristiques les plus souvent utilisées pour la REN

- **caractéristiques au niveau des mots**
- listes d'entités : gazetiers (index géographiques, angl. *gazetteer*), dictionnaires...
- caractéristiques des documents et corpus

Caractéristiques au niveau des mots

- **casse** :
  - commence par une majuscule (*Sorbonne*)
  - le mot est tout en majuscules (*ILPGA*)
  - le mot est en majuscules et en minuscules (*iCampus*)
- **ponctuation** :
  - se termine par un point (*St.*)
  - présence d'un point interne (*I.B.M.*)
  - d'un apostrophe, trait d'union ou esperluette interne (*O'Connor*)
- **chiffre** : succession de chiffres (*2023*) / lettres (*XVIII*) / lettres + chiffres (*W3C*)
- **caractère** : marque possessive (*Esther's* family)
- **catégorie grammaticale** : - nom commun + nom propre (*madame Bovary*)

REN : types d'approches

Approches à base de méthodes symboliques

- reposent sur des **règles** élaborées par un expert et des dictionnaires (listes)

Approches guidées par les données (angl. *data-driven*) et l'apprentissage[1]

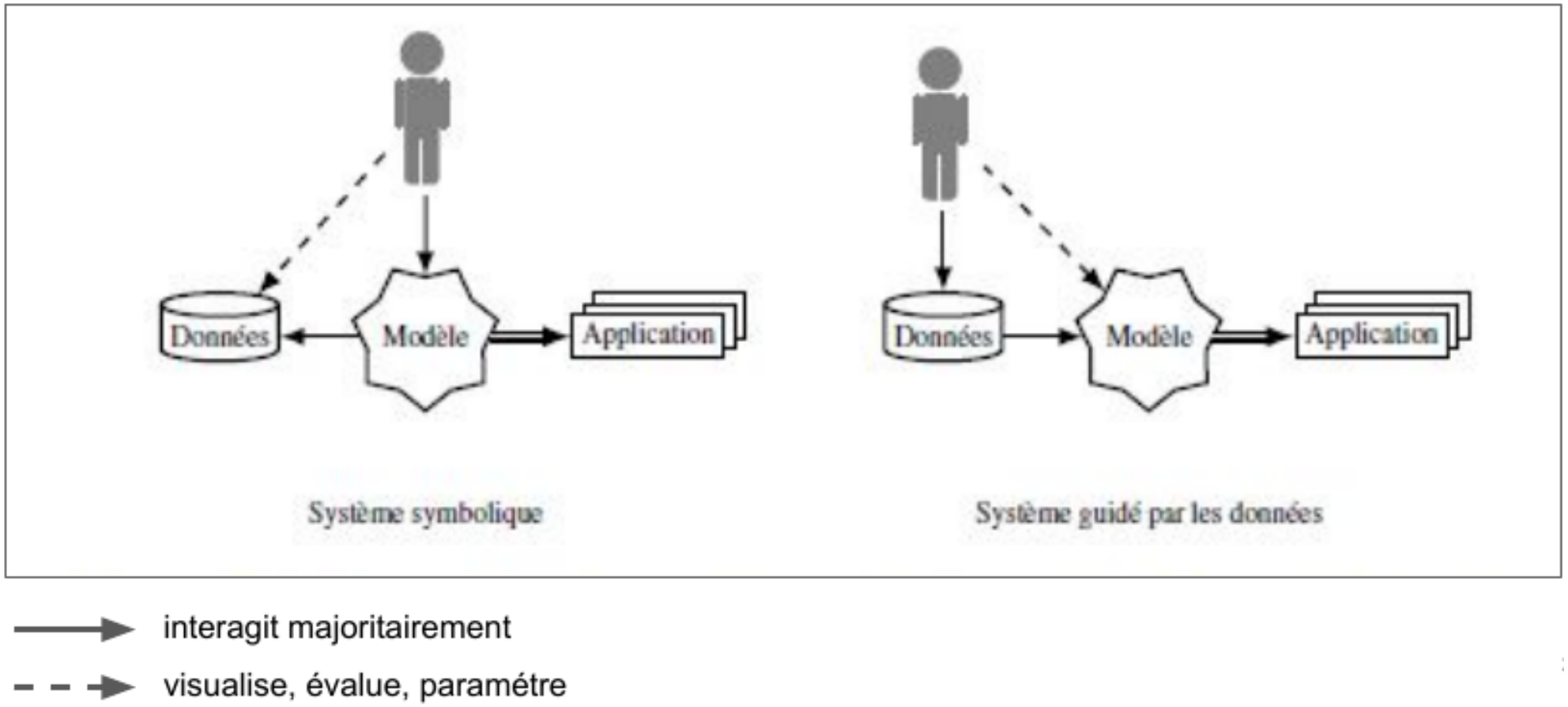
- au croisement des mathématiques, statistiques et sciences cognitives, cherchent à déterminer les paramètres d'un modèle à partir de données

  a) apprentissage automatique « classique » (non-, semi- et supervisé)
  b) apprentissage profond[2] avec les modèles de langue

[1] apprentissage machine / automatique (angl. *machine learning*), et par extension, les algorithmes d'apprentissage

[2] angl. *deep learning*

REN : types d'approches



Approche symbolique *vs.* approche guidée par les données (Nouvel *et al.*, 2015)

Classification

- en apprentissage machine, la REN est modélisée comme un problème de **classification**
- à partir de données, l'algorithme vise à déterminer de valeurs discrètes (catégories) à attribuer à une séquence de mots donnée en entrée

Annotation des EN : convention BIO

- B - *beginning* ; I - *inside* ; O - *outside* (d'un segment textuel)

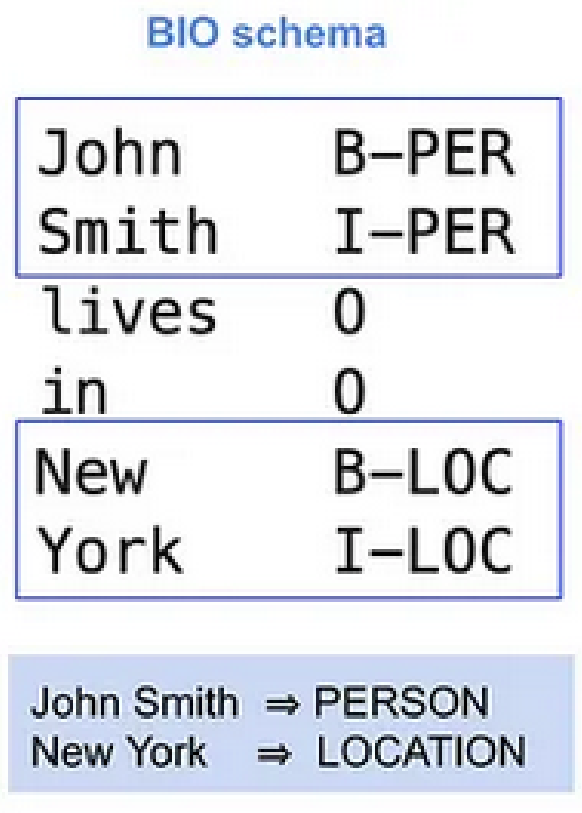


Schéma *BIO* d'annotation des EN (Kocaman, 2020).

Outils de REN

`spaCy`

- librairie du langage de programmation Python *open source*
- permet d'identifier les EN en plus de 73 langues
- modules : tokenisation, lemmatisation, étiquetage morphosyntaxique[1] etc.

[1] angl. *POS (part-of-speech) tagging*

EN des lieux

Une fois les EN spatialisées (géographiques) extraites sous un format spécifique, existe-t-il un moyen de les visualiser sur une carte géographique ?

Cartographier les EN

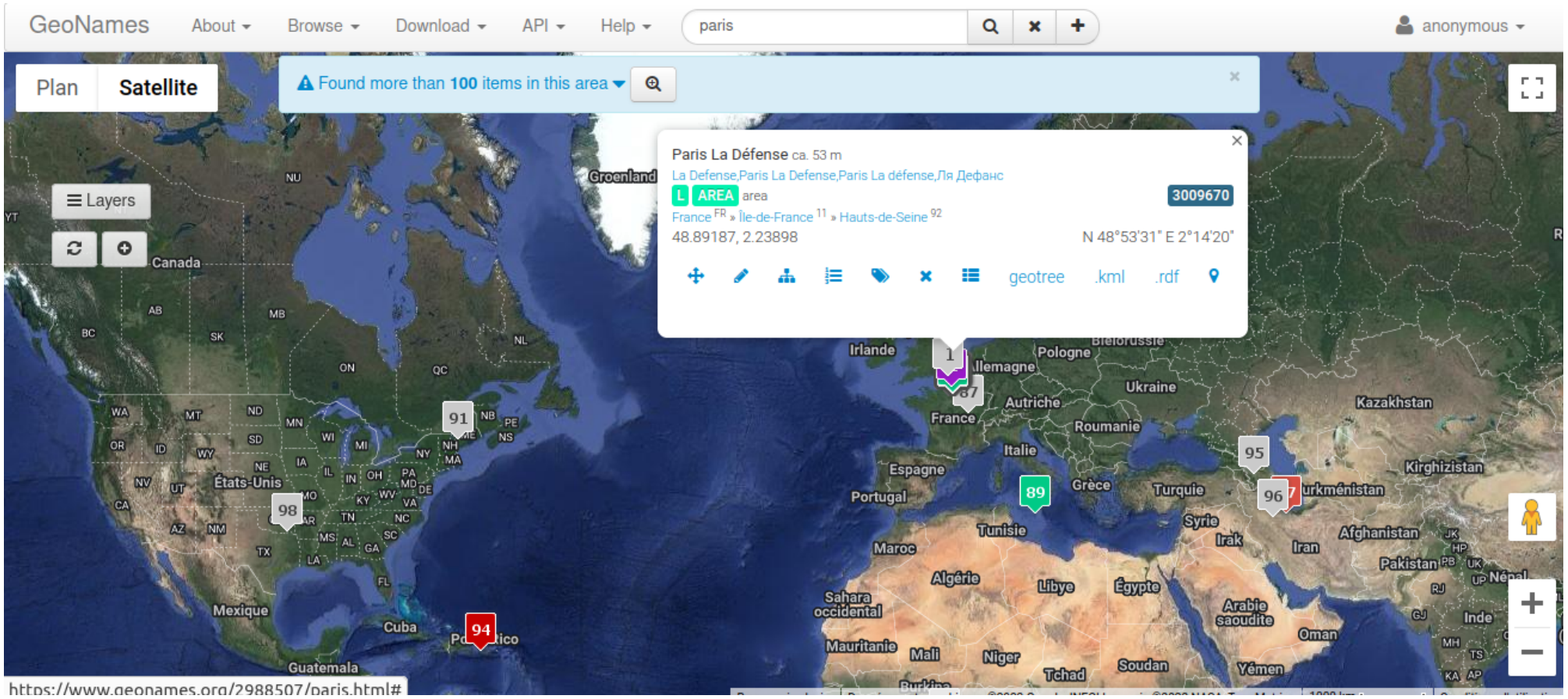
1. extraire les entités nommées de lieux dans un texte (OCRisé ou non)
2. récupérer leurs coordonnées géographiques
3. projeter les EN sur une carte

Pour que les EN des lieux soient visualisées sur une carte géographique, il faudrait les aligner avec un référentiel standard (p. ex. GeoNames).

GeoNames

- référentiel géographique : base de données géographiques consultable en ligne
- plus de 25 millions de noms géographiques et plus de 11 millions de noms uniques

GeoNames



EN *Paris* projetée sur la carte géographique avec le référentiel GeoNames.

Cartographier les EN *via* Pandore

Tanagra

- interface web pour géolocaliser et cartographier des EN de lieux dans les textes
- `spaCy` → GeoNames → projection sur une carte interactive Leaflet
- les nouveaux lieux peuvent également être ajoutés à la base de données
- possibilité de filtrer les résultats par métadonnées (titre ou date), et d'exporter les lieux cartographiés au format `.csv`

Collecte de corpus de Wikisource *via* Wikisource

1. Télécharger les sept premiers chapitres (`Chapitre I` , ..., `Chapitre VII`) de l'ouvrage *Le Tour du monde en quatre-vingts jours* de Jules Verne de Wikisource.
2. Sélectionner le format texte brut `.txt` (angl. *plain text*).
3. Compresser (zipper) les fichiers téléchargés (clique droit → `Compresser...`).

Collecte de corpus de Wikisource *via* Pandore (beta)

1. Aller sur `Chapitre I` de *Le Tour du monde en quatre-vingts jours*.
2. Copier le lien URL du premier chapitre `https://fr.wikisource.org/wiki/Le_Tour_du_monde_en_quatre-vingts_jours/Chapitre_1`.
3. Ouvrir le module de collecte de corpus de Wikisource sur Pandore.
4. Coller le lien URL de l'étape 2 dans le champ `À partir d'URL(s)`.
5. Ajouter encore 6 champs en cliquant sur `Ajouter URL`.
6. Coller le lien URL des 6 chapitres suivants (il suffit de remplacer le dernier élément de l'URL `Chapitre_1` avec `Chapitre_2` ... jusqu'au `Chapitre_7`).
7. Cliquer sur `Collecter corpus`.

Cartographier les EN avec Tanagra

8. Aller sur Tanagra, sélectionner le modèle `French_lg` de `spaCy`.
9. Téléverser le dossier compressé (zippé) dans le champ `Upload file(s)`.
10. Observer les EN cartographiées avec leurs fréquences d'utilisation : sur quel continent y a-t-il le plus grand nombre d'EN reconnues ?

Références

- **Bamman, D., Popat, S., & Shen, S.** (2019). An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 2138-2144) https://aclanthology.org/N19-1220/.
- **Brando, C.** (2022). « Les entités nommées · Module TAL - Master HN PSL » [*diapositives*].
- **Galleron, I.** (2021). « Introduction aux humanités numériques (L1HN001) [*diapositives en interne*].
- **Kocaman, V.** (2020). « Named Entity Recognition (NER) with BERT in Spark NLP » https://towardsdatascience.com/named-entity-recognition-ner-with-bert-in-spark-nlp-874df20d1d77

Références II

- **Nouvel, D., Ehrmann, M., & Rosset, S.** (2015). *Les entités nommées pour le traitement automatique des langues*. ISTE Group.
- **Sims, M., & Bamman, D.** (2020). Measuring information propagation in literary social networks. *arXiv preprint arXiv:2004.13980*.