Analyse et représentation d'un corpus littéraire, les récits d'esclaves

DOCUMENT TECHNIQUE

Octobre 2025 – Janvier 2026

Inty Gely
Samuel Rigal
Laura Deloffre

Introduction

Ce projet résulte d'une commande concrète de Madame Marie-Pierre Baduel, professeure agrégée d'anglais et doctorante au sein du laboratoire CAS (Culture Anglo-Saxonne). Sa thèse, intitulée « Réinterprétation des récits d'esclaves au travers du prisme des images du sang et de la couleur », s'intéresse à l'analyse et à la valorisation des récits d'esclaves à travers une approche visuelle et spatiale.

Dans ce cadre, le projet consiste en la conception et le développement d'une application web interactive permettant de retracer la vie des esclaves à partir de leurs récits. L'objectif principal est de donner une dimension concrète et visuelle aux recherches menées, en facilitant leur diffusion et leur compréhension. Une attention particulière est portée à la visualisation cartographique des trajets et des déplacements des esclaves tout au long de leur existence, élément central du projet.

Ce travail s'inscrit dans la continuité de trois projets antérieurs, qui ont posé les bases de la collecte, de la structuration et de la représentation des données historiques utilisées.

Tous les programmes utilisés pour ce projet sont sur le github : https://github.com/lauradelo/slave_narratives

Base de données

Nouvelle structure

Tout d'abord nous avons restructuré la base de données tel qui le MCD suivant :

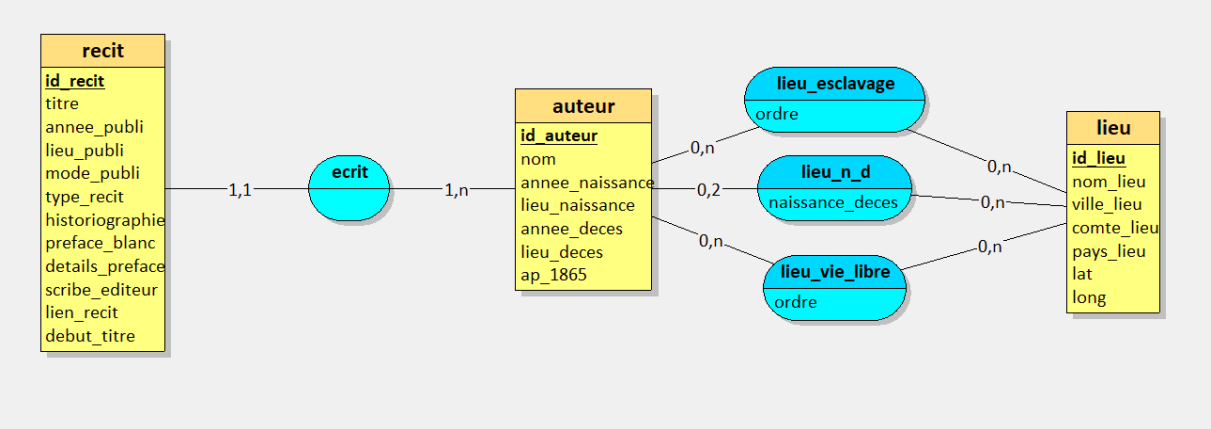

Figure 1 : Proposition du nouvel MCD

Nous avons ensuite créé la base de données dans la BDD MySQL avec le script ''creation_bdd.sql''.

```sql
USE resclaves;


-- Table auteur
CREATE TABLE IF NOT EXISTS auteur (
    id_auteur INT PRIMARY KEY,
    nom VARCHAR(255),
    annee_naissance VARCHAR(50),
    lieu_naissance VARCHAR(255),
    annee_deces VARCHAR(50),
    lieu_deces VARCHAR(255),
    ap_1865 TINYINT(1) DEFAULT 0,
    moyen_lib VARCHAR(255),
    origine_parents VARCHAR(255),
    militant_abolitionniste VARCHAR(255),
    particularites VARCHAR(255),
    plrs_recits TINYINT(1) DEFAULT 0,
    op_source VARCHAR(255);
);

-- Table lieu
CREATE TABLE IF NOT EXISTS lieu (
    id_lieu INT AUTO_INCREMENT PRIMARY KEY,
    nom_lieu VARCHAR(255) UNIQUE,
    ville_lieu VARCHAR(255),
    comte_lieu VARCHAR(255),
    pays_lieu VARCHAR(255)
);
```

Figure 2 : Extrait script de création de la bdd

Insertion des données

Pour insérer les données dans la base de données, le script Python *insertion_csv_bdd.py* a été utilisé.

Ce script permet de lire le fichier CSV *ap_1865.csv*, de nettoyer et normaliser les données, puis de les insérer dans la base de données MySQL.

Une attention particulière a été portée à la gestion des lieux, qui sont initialement stockés sous forme de texte dans une seule colonne du fichier CSV. Le script sépare ces informations afin de créer une table *lieu* dédiée, puis établit des relations avec les auteurs à l'aide de tables de liaison.

Cette méthode permet d'éviter les doublons de lieux dans la base de données, de normaliser les informations géographiques et de garantir une structure relationnelle cohérente. Les lieux de vie libre et les lieux liés à l'esclavage sont ainsi traités séparément, tout en partageant une table commune des lieux.

Nettoyage

Pour nettoyer et restructurer la base de données, nous avons utilisé plusieurs script Python au transfert des données depuis des tables intermédiaires vers les tables finales.

Ces scripts ont pour objectif de normaliser les données, de corriger certaines incohérences et de garantir une structure relationnelle cohérente au sein de la base.

Fonctionnement des script

Le script commence par se connecter à la base de données MySQL et récupère les données issues des tables tab_auteurs et tab_recits_v3. Ces tables contiennent des informations hétérogènes, parfois mal formatées ou incomplètes.

Un ensemble de fonctions de nettoyage est ensuite appliqué :

- Nettoyage des champs textuels : suppression des valeurs vides, des mentions telles que « *non spécifié* » ou « *n/a* », et homogénéisation des formats.
- Normalisation des identifiants auteurs : certains identifiants étant stockés sous forme alphanumérique (par exemple *A15*), le script extrait uniquement la partie numérique afin d'assurer la cohérence des clés primaires.
- Découpage des lieux : les lieux pouvant contenir plusieurs informations dans une seule cellule (séparées par des conjonctions ou des ponctuations), le script les divise en entités distinctes.

Gestion des lieux

Afin d'éviter les doublons, les lieux sont stockés dans une table dédiée (*lieu*). Un mécanisme de cache permet de vérifier si un lieu existe déjà avant de l'insérer, garantissant ainsi l'unicité des entrées. Les lieux sont ensuite reliés aux auteurs via des tables de liaison distinctes selon le contexte :

- lieux de vie libre
- lieux liés à l'esclavage

L'ordre d'apparition des lieux est conservé, ce qui permet de restituer la chronologie ou la progression géographique lorsque cela est pertinent.

Insertion des auteurs et des récits

Les auteurs sont insérés dans la table finale. Certaines informations textuelles sont également converties en valeurs booléennes afin d'améliorer leur exploitation (par exemple la présence de plusieurs récits), mais la plupart des informations ne concernant pas les lieux sont restées telles quelles.

Les récits sont ensuite insérés dans la table *recit*, avec leurs métadonnées associées (date et lieu de publication, type de récit, mode de publication, présence d'une préface, lien vers la ressource, etc.).

Chaque récit est relié à son auteur via une table de correspondance, assurant ainsi une relation claire entre auteurs et œuvres.

Résultat

Ce processus de nettoyage et de transfert permet d'obtenir une base de données :

- structurée selon les principes de la normalisation relationnelle,
- débarrassée des doublons et des valeurs incohérentes,
- prête à être exploitée pour des analyses, des visualisations ou une mise en ligne.

Analyses R

Clustering

Pour le clustering, le programme clustering.R a été utilisé, il a permis de rassembler les variables, et de construire deux acp, une groupé par variables socio-historiques, une autre groupé par données géographiques.

Chargement et préparation des données

Le programme commence par se connecter à la base de données MySQL afin de charger les tables nécessaires, notamment celles contenant les informations sur les récits, les auteurs et les localisations géographiques associées.

Ces différentes sources sont ensuite fusionnées afin d'obtenir un jeu de données unique reliant chaque récit à ses caractéristiques sociales, éditoriales et spatiales.

Un travail de nettoyage est appliqué en amont : les variables pertinentes sont sélectionnées et les valeurs manquantes sont remplacées par des catégories explicites, ce qui permet d'assurer la stabilité des traitements statistiques ultérieurs.

Clustering socio-historique

Les variables socio-historiques (type de récit, mode de publication, présence d'une préface blanche, engagement abolitionniste de l'auteur, multiplicité des récits) sont encodées sous forme de variables catégorielles puis transformées en variables numériques.

À partir de ces données, un clustering par k-means est réalisé. Cette approche permet de regrouper les récits selon des profils socio-historiques similaires.

Une analyse en composantes principales (ACP) est ensuite utilisée afin de projeter ces groupes dans un espace à deux dimensions, facilitant leur visualisation et leur interprétation.

Clustering géographique

En parallèle, un second clustering est effectué à partir des coordonnées géographiques associées aux récits. Les données spatiales sont converties en objets géographiques exploitables, puis un clustering est appliqué directement sur les coordonnées.

Cette analyse permet de faire émerger des regroupements spatiaux de récits, révélant des logiques géographiques de production ou de diffusion.

Diagramme alluvial

Le programme trajet_flux.R permet de construire un diagramme alluvial afin de représenter visuellement les trajectoires des auteurs à travers différents lieux et étapes de leur vie.

Objectif du diagramme

Le diagramme alluvial a pour objectif de montrer les enchaînements spatiaux entre :

- le lieu de naissance,

- le premier lieu d'esclavage,
- le premier lieu de vie libre,
- le lieu de décès.

Cette visualisation permet de synthétiser des parcours complexes et de faire apparaître des régularités ou des trajectoires dominantes.

Chargement et préparation des données

Le programme se connecte à la base de données MariaDB afin de charger les auteurs concernés, en filtrant uniquement ceux postérieurs à 1865.
Les données de lieux sont ensuite normalisées : les noms sont mis en minuscules, les accents supprimés et les différentes sources d'information géographique fusionnées afin d'obtenir une appellation exploitable et homogène.

Construction des trajectoires

Les trajectoires individuelles sont reconstruites en reliant plusieurs tables de la base :

- les lieux de naissance et de décès,
- le premier lieu d'esclavage,
- le premier lieu de vie libre.

Chaque auteur est ainsi associé à une suite ordonnée de lieux correspondant aux différentes étapes de son parcours. Les valeurs manquantes sont remplacées par une catégorie explicite (« *Inconnu* ») afin de conserver l'ensemble des trajectoires.

Réduction et synthèse des catégories

Afin de rendre la visualisation lisible, un mécanisme de réduction des catégories est appliqué.

Seuls les lieux les plus fréquents sont conservés, tandis que les autres sont regroupés dans une catégorie générique (« *Autres* »). Cette étape permet de limiter la complexité graphique tout en préservant les tendances principales.

Les trajectoires identiques sont ensuite comptabilisées, ce qui permet de pondérer le diagramme par le nombre d'auteurs concernés.

Visualisation alluviale

Le diagramme alluvial est généré à partir de ces trajectoires agrégées.

Chaque bande représente un flux d'auteurs entre les différentes étapes de leur vie, son épaisseur étant proportionnelle au nombre d'auteurs suivant cette trajectoire. Les strates correspondent aux lieux, et les flux permettent de visualiser les transitions entre eux.

Apport de la visualisation

Ce diagramme offre une lecture synthétique et intuitive des parcours géographiques des auteurs. Il met en évidence les lieux centraux, les trajectoires majoritaires ainsi que les transitions les plus fréquentes entre esclavage, vie libre et fin de vie.