Université Claude Bernard Lyon 1

Modélisation et enrichissement de données PFAS

Cahier des Charges

Étudiants :

Mohamed Hamed MOHAMED AHMED
Sidi NDAO
Dinel DEBIB

Sous l'encadrement de :

M. Fabien DUCHATEAU

Université Claude Bernard Lyon 1
Département informatique

Année 2024-2025

Introduction

1.1. Contexte

Les PFAS (Per- and polyFluoroAlkyl Substances), communément appelés "polluants éternels", sont des composés chimiques largement utilisés dans de nombreux produits industriels et de consommation courante[1]. Une équipe de journalistes européens a recensé en 2023 l'utilisation et les contaminations de PFAS à travers le projet Forever Pollution Project (une centaine de sources de données, 23000 sites contaminés et autant de sites présumés contaminés). La communauté scientifique s'est ensuite efforcée à documenter, mettre à jour et enrichir ces données de façon pérenne dans le projet PDH (PFAS Data Hub[2]).

Actuellement, les données sont stockées dans un format CSV basique, avec une visualisation cartographique simple. L'enrichissement de ces données permettrait une analyse plus approfondie par divers experts (toxicologues, sociologues, etc.). Des sources externes comme des bases de données de chimie contiennent des informations détaillées sur la composition chimique et la structure 3D des PFAS, tandis que l'INSEE fournit des données socio-démographiques et économiques par quartier, notamment via le projet Mongiris. Ce dernier est une base de données MongoDB qui consolide les données INSEE sur les 50000 IRIS (Ilots Regroupés pour l'Information Statistique) de France, permettant ainsi d'étudier l'impact des PFAS sur différentes populations à une échelle fine.

[1] Emma L. Schymanski et al., "Per- and polyfluoroalkyl substances (PFAS) in PubChem: 7 million and growing", Environmental Science & Technology, https://pubs.acs.org/doi/10.1021/acs.est.3c04855

[2] Projet PDH (PFAS Data Hub), "Base de données européenne sur les PFAS", https://pdh.cnrs.fr/fr/

1.2 Objectifs

Le projet vise à atteindre quatre objectifs principaux :

1. Modélisation des données.

Le format actuel (fichiers CSV) ne permet qu'une représentation tabulaire, favorise la redondance d'informations (et donc les incohérences), ne permet pas d'interroger facilement les données et limite l'enrichissement. Il est donc prévu de :

- Concevoir une représentation des données existantes en utilisant un modèle plus adapté sous forme de document (MongoDB)
- Utiliser le format GeoJSON pour les données spatiales
- Assurer la compatibilité avec les sources de données existantes

2. Migration des données.

La transformation des données d'un format tabulaire à un modèle orienté document nécessite de développer un script qui migre les données en respectant les contraintes existantes ainsi que celles du nouveau modèle . Il est donc prévu de :

- Transférer les données PFAS des fichiers CSV vers MongoDB
- Intégrer les données territoriales et démographiques de l'INSEE, en mettant à jour les données de Mongiris
- Valider l'intégrité des données après migration

3. Enrichissement automatique.

L'enrichissement des données PFAS avec des informations complémentaires est crucial pour une analyse approfondie. Les bases CompTox[3] et EcoTox[4] contiennent des informations précieuses sur la structure et la composition des PFAS qu'il faut exploiter de manière automatisée. Il est donc prévu de :

- Développer un algorithme d'interrogation des bases CompTox et EcoTox
- Établir des correspondances entre les identifiants des différentes sources
- Mettre en place un système de mise à jour régulière

4. Visualisation des données.

La représentation visuelle des données est essentielle pour comprendre la distribution et l'impact des PFAS sur les territoires. Une interface interactive permettra aux utilisateurs d'explorer et d'analyser les données de manière intuitive.
Il est donc prévu de :

- Créer une application web de visualisation cartographique
- Permettre l'exploration interactive des données
- Intégrer des filtres et des options d'analyse

[3] CompTox Chemicals Dashboard, https://comptox.epa.gov/dashboard
[4] EcoTox Knowledgebase, https://cfpub.epa.gov/ecotox/

Travail à réaliser

Notre travail sera divisé en 4 étapes, chacune portant sur l'un des objectifs mentionnés précédemment.

2.1. Phase de modélisation.

L'analyse approfondie des structures de données existantes est nécessaire pour concevoir un modèle optimal qui facilitera les analyses futures. Cette phase est cruciale car elle déterminera l'efficacité des traitements ultérieurs. Les documents MongoDB possèdent un schéma flexible, ce qui permet d'ajouter de nouvelles informations sur le long terme. Cette étape comprend l'analyse des structures existantes (CSV, CompTox, EcoTox), la conception d'un schéma MongoDB optimisé, et la validation du modèle avec des jeux de données test.

2.2. Phase d'intégration

La migration des données vers le nouveau système nécessite une attention particulière pour garantir l'intégrité et la cohérence des informations. Cette étape implique la transformation des données tout en préservant leurs relations et leur qualité. Les développements incluent la création des scripts de migration CSV vers MongoDB, la mise en place des connecteurs pour les données INSEE, ainsi que l'exécution des tests de validation et de performance.

2.3. Phase d'enrichissement

L'enrichissement automatique des données est un aspect central du projet, nécessitant des algorithmes robustes pour établir des correspondances fiables entre les différentes sources de données. La précision et la performance sont essentielles. Le travail se concentre sur l'implémentation de l'API d'interrogation CompTox/EcoTox, le développement des algorithmes de correspondance, et la mise en place d'un système de cache et d'actualisation performant.

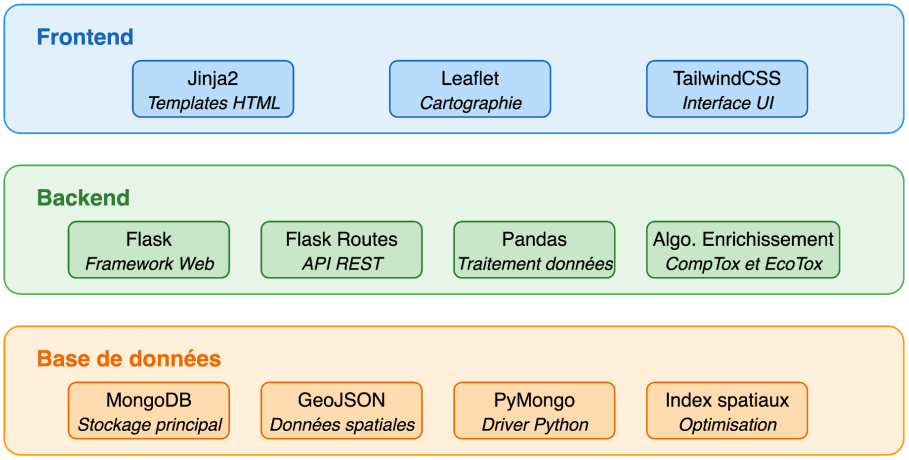
2.4. Phase de visualisation

L'interface utilisateur doit permettre une exploration intuitive des données tout en offrant des fonctionnalités d'analyse avancées. L'accent est mis sur la performance et l'ergonomie. Le développement comprend la création de l'interface web avec Flask, l'intégration des composants cartographiques Leaflet, et l'implémentation des fonctionnalités d'analyse et d'export des données.

Architecture technique du projet

La Figure 1 présente l'architecture technique du projet en trois couches : le Frontend avec Jinja2, TailwindCSS et Leaflet pour la visualisation cartographique ; le Backend géré par Flask avec Pandas et l'algorithme d'enrichissement pour CompTox/EcoTox ; et la Base de données utilisant MongoDB, GeoJSON et des index spatiaux pour les données géographiques.

Figure 1 : Architecture technique du projet PFA



Planning prévisionnel