Mise en situation 1 (E1)

Collecte, stockage et mise à disposition des données d’un projet IA

Les pesticides ont-ils un impact sur l’incidence du cancer en France ?

Formation Développeur en Intelligence Artificielle

RNCP 37827

Promotion 2024-2026

Maximilien Proust

**SOMMAIRE**

[**Introduction 3**](#_b1tudj17zite)

[Présentation du projet 3](#_uvnfaczeocre)

[Contexte du projet 3](#_sls243humeli)

[Spécifications techniques 3](#_t63q8irzdjyy)

[**Extraction des données 4**](#_d23rwq2amlcq)

[Données issues d’un service web (API) 4](#_mb6mm7ilc5sf)

[Données issues d’une page web 4](#_4pml5r6h2mtm)

[Données issues d’un fichier 4](#_65z12uqbkv8n)

[Données issues d’un base de données relationnelles 4](#_yx66up57silz)

[Données issues d’un système big data 4](#_2py97jnobm6x)

[**Requêtage des données 5**](#_nmatos6z06km)

[Requêtes de type SQL 5](#_72k15os7wtz)

[Requêtes depuis un système big data 5](#_9p42dsmf1g3n)

[**Agrégation des données 5**](#_ix8pmbbp4vmc)**Création de la base de données 6**

[Préparation des données 5](#_jpcgefr1vbic)

[Agrégation des données 5](#_mcudajlzlv4v)

[Jeu de données final 5](#_9qqswaxjqjfp)

[Modélisation des données 6](#_693t11557h5x)

[Modèle physique des données (MPD) 6](#_ik1bzpg09yj)

[Choix du système de gestion de la base données (SGBD) 6](#_o5tvt35s3hx6)

[Création de la base de données 6](#_efn1os9dj9vv)

[Import des données 6](#_44pm7hwukmr)

[Conformité RGPD 6](#_4gn8wrikwvwl)

[**Développement de l’API 7**](#_j0ml1cjlhqrr)

[Spécifications fonctionnelles et techniques 7](#_4jpdprf74k1b)

[Conception de l’architecture de l’API 7](#_6chkzwt6eflu)

[**Documentation 7**](#_7f423kg80qfx)

[**Perspectives et améliorations 7**](#_41tghgp87glx)

[**Conclusions 7**](#_8xwh17rpmr8z)

[**Annexes 8**](#_tezm7i4fli34)

Note : il est attendu un rapport de 2 à 5 pages pour ce livrable

# Introduction

## Présentation du projet

Le terme pesticide va désigner les substances actives qui vont contrôler ou éliminer des organismes indésirables. Ces organismes indésirables peuvent être des champignons, des animaux, des plantes ou encore des bactéries. Les pesticides vont être utilisés en majorité pour l’agriculture afin de protéger les cultures. Cependant, on va en retrouver également dans l’usage domestique en tout genre tel que le traitement aux mauvaises herbes, l’élimination de nuisibles ou encore les produits antiparasitaires[[1]](#footnote-1).

L’usage des pesticides de synthèse dans l’agriculture s’est développé de manière considérable depuis leur apparition dans les années 1940-1950. Si on prend en compte les données récentes, depuis les années 1990, la quantité de pesticides utilisée dans le monde a augmenté de 80%. En France, certaines régions vont utiliser énormément de pesticides, dont les régions du nord, du centre Val-de-Loire, de la Garonne ou encore du Rhône. Mais aucune région ne semble épargnée par l’utilisation des pesticides. Pourtant, en 2008, la France a mis en place le plan Ecophyto dont l’objectif était de réduire de 50% l’utilisation des pesticides d’ici à 2018. Force est de constater que ce plan fût un échec, l’utilisation des pesticides augmentant de 25% sur cette période[[2]](#footnote-2). L’échec de ces mesures montrent l’incapacité de l’état à agir sur l’utilisation des pesticides. Mais quelles peuvent être les conséquences pour la population française ? Il serait impossible de traiter l’ensemble des conséquences, je vais donc me concentrer sur les conséquences pour le cancer humain.

Chaque individu est constitué d’environ 50 000 milliards de cellules répartis en sous-ensembles structurés appelés tissus qui vont eux-mêmes former des organes. Au sein de chaque organe, les cellules vont avoir des fonctions précises qui peuvent être par exemple la multiplication cellulaire. C’est ce phénomène qui va nous intéresser. En effet, lorsque ce processus de multiplication va se dérégler, l’organisme va détruire les cellules dysfonctionnelles. Cependant, ce mécanisme de protection peut ne plus fonctionner et la cellule malformée va alors se multiplier encore et encore et va devenir une tumeur qui peut être bénigne ou maligne. Les tumeurs dites malignes sont ce que l’on va appeler le cancer[[3]](#footnote-3).

L’incidence du cancer en France ne cesse d’augmenter depuis 30 ans. En 2023, on estime à 430 000 le nombre de nouveaux cas en France. Le cancer va représenter la première cause de mortalité prématurée. En 2021, le nombre de décès imputable au cancer était de 162 000[[4]](#footnote-4). Le cancer, son diagnostique et son traitement sont donc un enjeu de santé public majeur. En effet, même si le taux de guérison progresse, il est important de s’attaquer au problème en amont plutôt qu’à la guérison de celui-ci.

Il existe beaucoup d’études montrant l’impact des pesticides dans le milieu professionnel mais pratiquement aucune sur l’impact des pesticides sur la population civile.

Quelques éléments pour introduire le projet.

Acteurs, objectifs fonctionnels et techniques

## Contexte du projet

Environnements et contraintes techniques, budget, organisation du travail et planification.

## Spécifications techniques

Les technologies et outils, les services externes, les exigences de programmation (langages), l’accessibilité (disponibilité, accès).

Proposer un schéma global du projet en partant des 5 sources et en terminant par l’API (diagramme des flux de données)

# Extraction des données

*L’extraction de données depuis un service web, une page web (scraping\*), un fichier de données, une base de données et un système big data\* en programmant le script\* adapté afin de pérenniser la collecte des données nécessaires au projet.*

Présentation des sources, des moyens techniques mis en oeuvre pour automatiser l’extraction (si besoin de gagner en place, un tableau synthétique pourrait convenir, donner plus de détails en annexe)

## Données issues d’un service web (API)

## Données issues d’une page web

Il est important de comprendre quel peut être l’impact des substances chimiques présents dans les polluants. Pour cela, je me suis appuyé sur le site web sagepesticides[[5]](#footnote-5). En effet, ce site web autorise le scrapping comme le mentionne son robot.txt qui affiche la mention user : \* allow : /. Pour l’extraction, j’ai utilisé un script python avec comme bibliothèques : requests, BeautifulSoup, json, re et logging. L’objectif d’utiliser la bibliothèque json est de permettre un enregistrement des données récupérées dans un fichier json et ainsi de pouvoir le rentrer dans une base Mongodb. Enfin, j’ai ajouté dans mon script un délai d’une à deux secondes afin de ne pas surcharger le serveur et éviter ainsi les éventuelles saturation serveur.

Au niveau des données récupérés, j’ai choisi de conserver tout ce qui pouvait avoir un intérêt pour la compréhension des molécules chimiques et leurs effets néfastes sur l’humain, puis sur l’environnement. En Annexe 1, vous pouvez retrouver une capture d’écran du site web complet d’une substance recherchée. Le premier bloc, va contenir des informations générales sur la substance. Ce qui va m’intéresser ici va être la description de la substance, indiquant notamment de façon générale l’utilisation de celle-ci. Pour récupérer cette information, j’ai simplement utilisé les balises html présentes.

## Données issues d’un fichier

## Données issues d’un base de données relationnelles

## Données issues d’un système big data

# Requêtage des données

*Extraction des données depuis un système de gestion de base de données et un système big data en appliquant le langage de requête propre au système afin de préparer la collecte des données nécessaires au projet.*

ksdfj<pl d

Préciser ici l’objet et les résultats des requêtes, soit depuis la source d’origine des données, soit depuis les bases développées pour les besoins du projet.

## Requêtes de type SQL

Objectif des requêtes et résultats obtenus (détails possible en annexe)

## Requêtes depuis un système big data

Objectif des requêtes et résultats obtenus (détails possible en annexe)

# Agrégation des données

*Agrégation de données issues de différentes sources en programmant, sous forme de script, la suppression des entrées corrompues et en programmant l’homogénéisation des formats des données afin de préparer le stockage du jeu de données final.*

## Préparation des données

Gestion des données partielles et/ou manquantes (si nécessaire et le cas échéant), nettoyage, suppression des données corrompues ni nécessaire, homogénéisation des formats de données, …

## Agrégation des données

Normalisation des différentes sources pour un jeu de données unique

## Jeu de données final

Résultats obtenus suite à l’agrégation

# Création de la base de données

Dans le respect du RGPD en élaborant les modèles conceptuels et physiques des données à partir des données préparées et en programmant leur import afin de stocker le jeu de données du projet.

## Modélisation des données

Méthode Merise et son formalisme

## Modèle physique des données (MPD)

## Choix du système de gestion de la base données (SGBD)

Adéquation entre le type de base et les contraintes du projet

## Création de la base de données

## Import des données

Alimentation de la base de données à partir de scripts d’import

## Conformité RGPD

Registre et procédure de conformité RGPD le cas échéant

# Développement de l’API

*Mettre à disposition le jeu de données en utilisant l’architecture REST pour permettre l’exploitation des données par les autres composants du projet.*

## Spécifications fonctionnelles et techniques

Analyses des spécifications fonctionnelles et techniques

Intégrer un schéma fonctionnel si possible.

## Conception de l’architecture de l’API

Conception de l’architecture de l’API, les points de terminaison, les règles d’accès, etc.

# Documentation

Rédiger une documentation technique portant sur :

* extraction des données depuis les bases de données : sélection, filtrage, jointure, optimisations, etc
* agrégation : dépendances, nettoyage, homogénéisation, etc
* création de la base : dépendances, commandes pour l’exécution des scripts, conformité RGPD, etc
* API : points de terminaison, règles d’authentification, etc

# Perspectives et améliorations

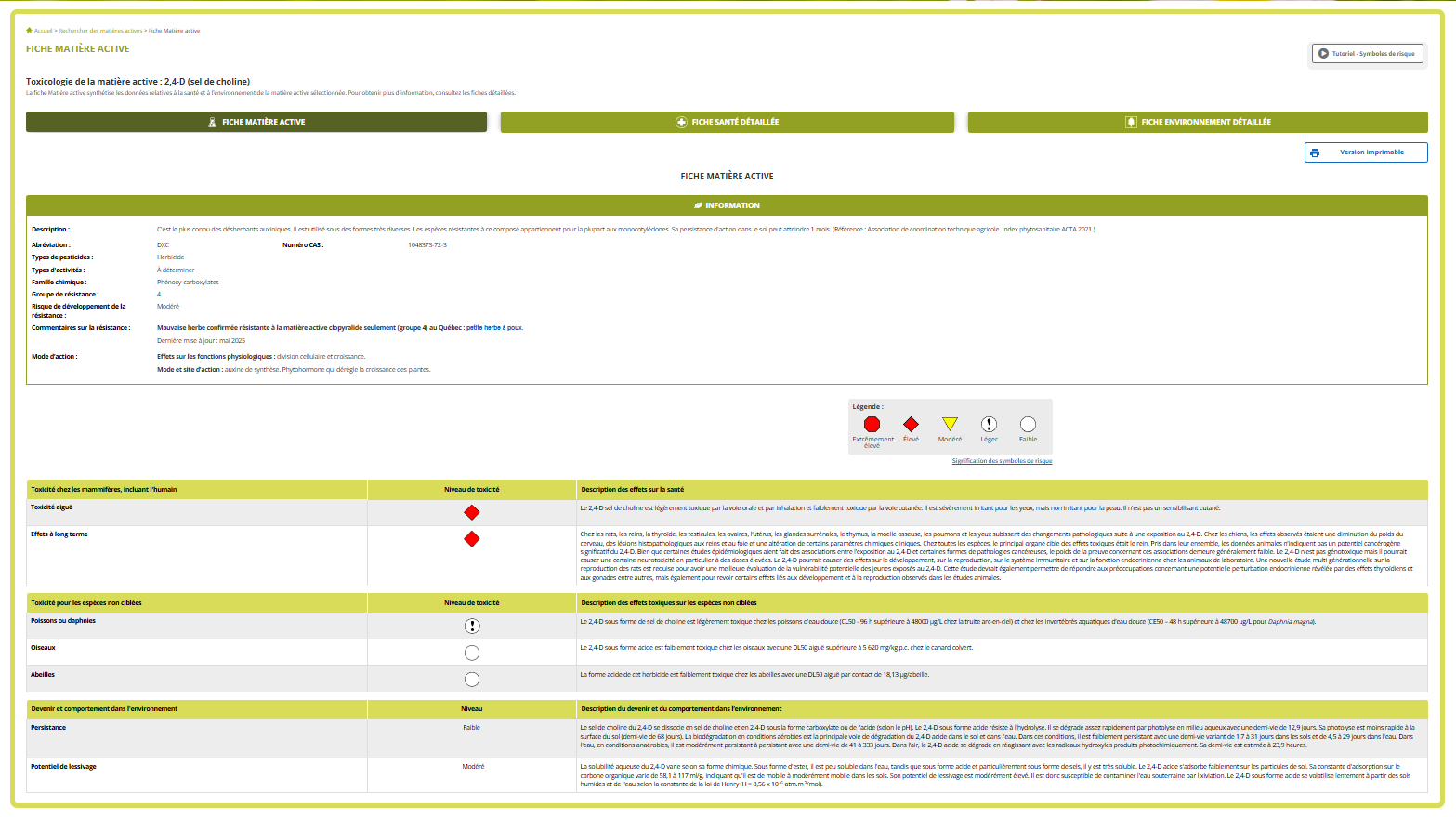
Déploiement et usage à venir de l’API développée

Les points possibles d’améliorations si il y en a

# Conclusions

# Annexes

Annexes possibles :

* Informations plus détaillées sur les sources de données
* Annexe 1 : Visualisation du résultat de la recherche d’une substance chimique sur le site web :  https://www.sagepesticides.qc.ca/
* 
* Figure 1 : Visualisation des données obtenues sur la recherche d'une substance
* Requêtes SQL et Big Data : exemples d’instructions et commentaires associés
* Règles d’agrégation : fournir ici les règles si complexes et nombreuses pour alléger le rapport
* Base de données : schéma de la BDD complet si étendue, schéma par domaine spécifique de la BDD, dictionnaires de données, modélisation (MCD, MLD, MPD), etc
* RGPD : registre de traitement des données personnelles, procédure de conformité, etc
* API : documentations

1. Santé gouv, « Pesticides ». [↑](#footnote-ref-1)
2. Boitas, « Etat des lieux de l’usage des pesticides en France ». [↑](#footnote-ref-2)
3. Fondation ARC, « Qu’est ce qu’un cancer ? » [↑](#footnote-ref-3)
4. Fondation ARC, « Le cancer en chiffres ». [↑](#footnote-ref-4)
5. « https://www.sagepesticides.qc.ca/ ». [↑](#footnote-ref-5)