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 déjà plusieurs études montrant l’impact des pesticides sur la survenue de divers cancers. Ces études sont très bien documentées dans la review « Cancer health effects of pesticides » qui va recenser les études sur le sujet. On s’aperçoit que les pesticides sont plus ou moins corrélés à l’apparition d’une multitude de cancers. Cependant, certaines études font débat et il est important de noter la difficulté d’associer la survenue d’un cancer à l’exposition aux pesticides[[5]](#footnote-5). L’objectif en mettant en corrélation des jeux de data autour de l’incidence du cancer et de l’achat de substances chimiques considérées comme cancérigènes, c’est d’observer s’il pourrait y avoir un rapport entre la quantité de pesticides utilisées dans certains départements et la survenue de cancers.

## Contexte du projet

L’objectif du projet est de réaliser un API qui va exposer les données sur un temps voulu pour savoir s’il pourrait y avoir un lien potentiel entre des départements utilisant beaucoup de pesticides et d’autres départements ou la quantité de pesticides est moindre. Pour cela, j’ai utilisé majoritairement du python dans un environnement virtuel avec quelques requêtes SQL et noSQL. A partir de python, j’ai utilisé le notebook pour visualiser mes données. J’ai pu utiliser plusieurs bibliothèques permettant la visualisation ou la mise à disposition des données comme FastAPI, Pandas, Plotly, SQLAlchemy. Mes données ont été stocké dans une base de données Postgres en local pour toutes les données nécessitant une base de données relationnelle. Les données Postgres ont été récupéré via une API pour certaines données et via des CSV pour la majeure partie. Pour les données en base de données non relationnelle, j’ai utilisé une base MongoDB. Les données de ma base non relationnelle sont issues du scrapping. Vous pouvez observer sur la figure 1 le schéma global du projet en visualisant les différentes sources de données utilisées.

En termes de planification, nous avions 6 semaines pour réaliser le projet.

J’ai découpé mes tâches en 3 parties. La première a été de trouver les différentes sources de données selon mes besoins sur les deux premières semaines. Une fois cette tâche accomplie, la suivante a été de regarder le contenu de mes sources, de les nettoyer pour les mettre en lien les unes avec les autres et cela sur une durée de 2 semaines également. Enfin, j’ai exposé ses données via une application FastAPI permettant de visualiser d’éventuels liens sur les deux dernières semaines. L’objectif était de rendre l’application agréable visuellement pour un utilisateur novice.

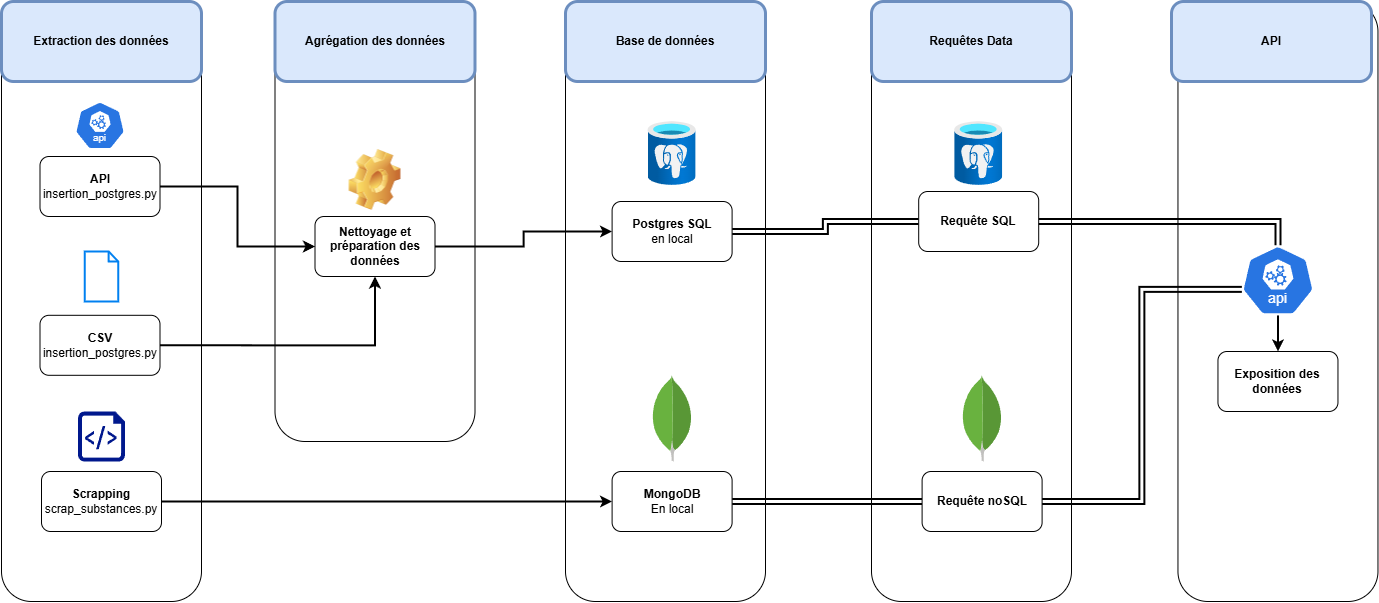


Figure 1: Schéma de l'extraction à la visualisation des données.

# 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.*

La première étape du projet a été de trouver des sources qui pourrait permettre d’effectuer un lien entre l’utilisation des pesticides en France et l’incidence des cancers. J’ai pu trouver plusieurs sources intéressantes que je vais présenter ci-dessous.

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

L’extraction des données est la première étape mise en œuvre

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

L’API (Application Programming Interface) est une interface qui va permettre à deux programmes de communiquer entre eux, notamment pour avoir accès à des données ou à des services en ligne. Ici, j’ai utilisé une API appelée Hubeau qui provient du ministère de la transition écologique. Cette source est une source officielle du gouvernement et la qualité du dataset est donc très bonne. De plus, l’API est librement accessible, ce qui est un atout pour le projet. Les données sont publiques et surtout régis par la licence Etalab. Cette licence signifie qu’on peut réutiliser les données y compris à des fins commerciales à condition de mentionner la source. Dans les conditions d’utilisation, on retrouve que quelques points de bonnes pratiques à suivre, tel que l’utilisation de filtre ou la compression des requêtes. Le lien de l’API est le suivant : <https://hubeau.eaufrance.fr/page/api-vente-et-achat-de-produits-phytopharmaceutiques>.

L’API permet de récupérer les informations par une méthode de pagination. Afin d’éviter une latence élevée, j’ai mis 20 000 lignes par page. En effet, dû à un mécanisme de pagination cumulative, plus il y a de page, plus les données vont être longues à télécharger. Les données récupérées sont des données d’achats de produits phytosanitaires et leurs substances. Je récupère donc 2 jeux de données, l’un pour l’achat de produits et l’autre pour l’achat de substances. J’effectue un premier préfiltre dans lequel je ne récupère que les données départementales.

Pour la base de données autour des produits phytosanitaires, j’effectue un nettoyage de celui-ci. En effet, après analyse dans un notebook, plusieurs éléments ressortent. Premièrement, il y a beaucoup de département avec des valeurs égales à 0. Je décide donc de les supprimer, n’ayant pas besoin de 0, représentant surement le total par département. De plus, je convertis les numéros de département dans un format string de façon à n’avoir que des lettres et non des chiffres pour mutualiser l’ensemble de nos données et notamment les départements corses 2A et 2B. Ensuite, je transforme les valeurs AMM (Autorisation de Mise sur le Marché) des produits phytosanitaires de façon à les rendre tous numérique, la aussi dans un souci de cohérence entre les différents jeux de données. Enfin, je reset l’index et j’y ajoute une colonne correspondant à cet index pour en faire ma clé primaire.

Pour le deuxième jeu de donnée autour des substances chimiques, ma première étape a été de récupéré uniquement les valeurs ou la classification était égale à CMR (Cancérigène, Mutagènes ou toxique pour la Reproduction) ou à T, T+, CMR (Nom des CMR avant 2018). En effet, ne travaillant que sur l’incidence des pesticides sur le cancer, les autres substances ne m’intéressent pas. Ensuite, j’ai appliqué les mêmes étapes que pour les produits chimiques à savoir de retirer les 0 des départements, les transformer en string et de transformer en format numérique l’AMM ainsi que d’ajouter comme clé primaire l’index. Avant l’insertion en table, les tables sont créées par SQLModel avec le format que j’ai choisi et que vous retrouverez en annexe 1. Nous allons donc retrouver les tables Produits\_vente et Substance\_cmr\_vente dans le schéma Pollution\_cancer. La table Produits\_vente après renommage va compter 7 colonnes. Dans l’ordre : id\_produits\_vente qui est la clé primaire, amm qui est une clé étrangère, annee qui est une clé étrangère, num\_departement qui est une clé étrangère, autorise\_jardin, quantite\_en\_kg, unite. La table Substance\_cmr\_vente va compter 10 colonnes : id\_substance qui est la clé primaire, amm (clé étrangère), annee (clé étrangère), classification\_mention, code\_cas, code\_substance, num\_departement (clé étrangère), fonction, nom\_substance et quantite\_en\_kg. Vous pouvez retrouver les détails dans la figure 2 ci-dessous.

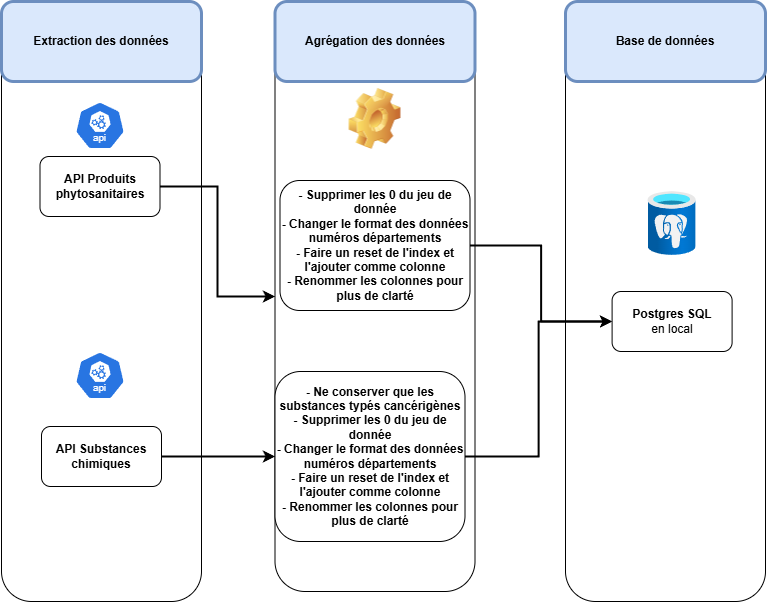


Figure 2: Extraction des données de l'API

## Données issues d’une page web

Le scrapping est une méthode permettant de récupérer des balises HTML notamment, avec Beautifoulsoup. Etant donné que le site web visé comporter des balises HTML intéressant à capturer, je me suis basé sur ce site web. Ce site web est sagepesticides.[[6]](#footnote-6). En effet, il est difficile de trouver des informations supplémentaires sur les molécules que l’on a récupéré avec l’API. Ce site web va répertorier les connaissances au sujet de certaines molécules. Le point négatif est le faible nombre de molécules présent qui est inférieur à 500. J’ai malgré tout choisi ce site web car il autorise la présence de robot comme l’indique le robot.txt qui affiche la mention user : \* allow : /. De plus, sur le site on ne retrouve aucune information interdisant la pratique du scrapping. 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 saturations.

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 2, 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. Ensuite, je vais récupérer les données des tableaux montrant un risque pour la biodiversité et la santé humaine et animale.

## Données issues des fichiers

# 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

# 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)

# 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

Annexe 1 : Code de création des tables en base de données relationnelles.

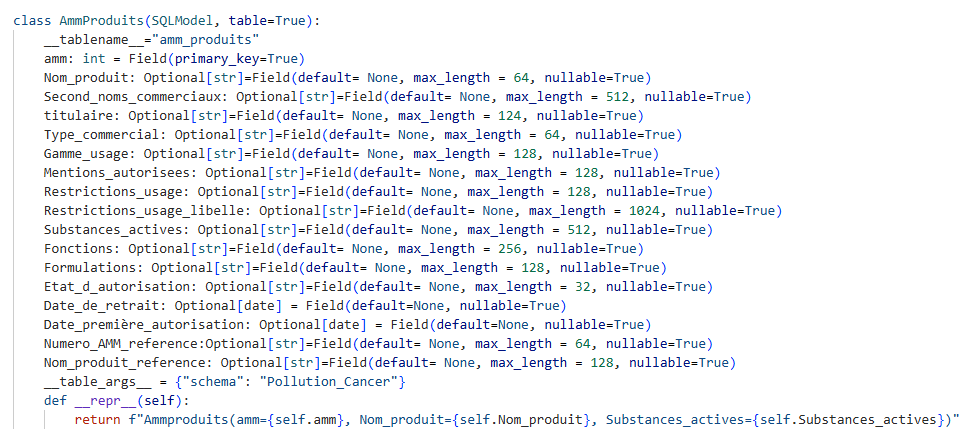




Suite de l’annexe 1 : Code de création des tables en base de données relationnelles.







Annexe 2 : Visualisation du résultat de la recherche d’une substance chimique sur le site web : <https://www.sagepesticides.qc.ca>

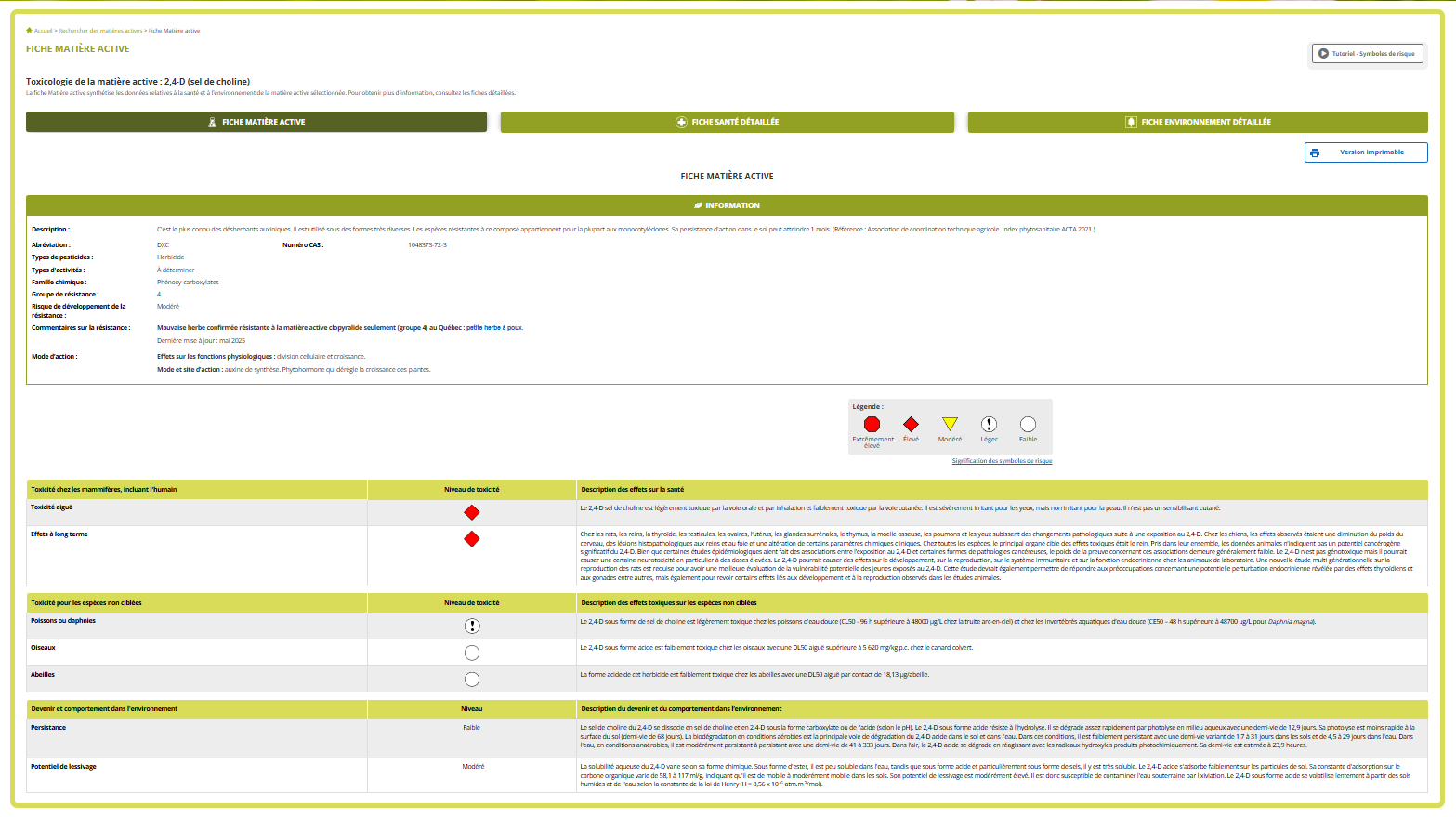


Figure 3: Visualisation des données obtenues sur la recherche d'une substance

Annexes possibles :

* Informations plus détaillées sur les sources de données
* 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. Bassil et al., « Cancer health effects of pesticides ». [↑](#footnote-ref-5)
6. « https://www.sagepesticides.qc.ca/ ». [↑](#footnote-ref-6)