AgroParisTech
Talents d'une planète soutenable

MÉMOIRE DE STAGE

ANALYSE ET MODÉLISATION DES DONNÉES TEXTUELLES POUR LA COMPRÉHENSION DE LA TRAJECTOIRE DE PRISE EN COMPTE DES ENJEUX ENVIRONNEMENTAUX DANS LES PROCESSUS DÉCISIONNELS

Cas pratique d'application NLP sur les avis de l'Autorité environnementale

---

Linh ĐINH

Certificat de Spécialité IODAA d'AgroParisTech

Mars - Août 2025

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude à ma tutrice de stage, Madame Cécile Blatrix, pour son accompagnement attentif et généreux à de nombreux niveaux. Son soutien ne s'est pas limité à ce stage, mais s'est étendu à d'autres aspects de mon parcours. Ainsi, c'est grâce à elle que j'ai découvert le programme Certificat de spécialité IODAA.

Je remercie également mon encadrant académique, Monsieur Vincent Guigue, pour son aide précieuse, ses conseils avisés et pour m'avoir mis en relation avec Monsieur Stéphane Dervaux (que je remercie également chaleureusement) afin d'obtenir l'accès au serveur, étape indispensable qui a permis la réalisation des parties les plus cruciales de ce projet.

J'adresse aussi ma profonde gratitude à Monsieur Pierre Leibovici et à Monsieur Rémi Labed, de *Disclose*, pour leur disponibilité, leur soutien bienveillant et pour m'avoir permis d'accéder à la précieuse base de données `data.disclose`.

Ma reconnaissance va aussi à l'ensemble des enseignants du programme IODAA, qui m'ont offert une base solide pour réussir ce parcours exigeant. Un remerciement particulier aux responsables de la formation, Monsieur Antoine Cornuéjol, Madame Christine Martin et Madame Sarah Cohen Boulakia qui ont accepté ma candidature et m'ont donné la chance de vivre une expérience académique et humaine d'une grande richesse.

Je n'oublie pas mes collègues, notamment Nicolas Lenglet et Samuel Pasquier, avec qui les échanges et moments de partage ont été aussi stimulants qu'enrichissants. Je tiens également à remercier mon frère et mes ami.es chers, qui m'ont accompagné dans les moments de joie comme dans les difficultés.

Enfin, une pensée toute particulière à Pierre Wan-Fat. Même si tu n'as pas été présent à mes côtés durant cette période, tout ce que j'ai pu accomplir aujourd'hui, je le dois en grande partie à toi. Au plus profond de mon cœur, tu as été et tu resteras toujours ma famille.

Abstract

This internship report, conducted within the OBSERVANCE project, explores how the recommendations of the French Environmental Authority (Ae) are considered in environmental impact assessments. To address the large and heterogeneous volume of documents generated in these processes, an automated NLP-based pipeline was designed and implemented. The system covers PDF conversion, text cleaning, metadata and content extraction, classification, database structuring, vector indexing, and deployment in a Retrieval-Augmented Generation (RAG) application. Results show near-perfect performance in simple metadata extraction (dates, pagination) and recommendation detection, as well as promising classification results with GPT-4. A first interactive prototype was deployed with Streamlit, confirming the feasibility of large-scale automated analysis while highlighting challenges related to data quality, document heterogeneity, and the stability of open-source models.

Résumé

Ce rapport de stage, réalisé dans le cadre du projet OBSERVANCE, analyse la manière dont les recommandations de l'Autorité environnementale (Ae) sont prises en compte dans les évaluations d'impact environnemental. Pour traiter le volume important et hétérogène de documents générés par ces procédures, un pipeline automatisé basé sur le traitement automatique des langues (TAL) a été conçu et implémenté. Le système couvre la conversion des PDF, le nettoyage des textes, l'extraction de métadonnées et de contenus pertinents, la classification, la structuration en base de données, l'indexation vectorielle ainsi que le déploiement dans une application de type Retrieval-Augmented Generation (RAG). Les résultats montrent des performances quasi parfaites pour l'extraction de métadonnées simples (dates, pagination) et la détection des recommandations, ainsi que des résultats prometteurs pour la classification avec GPT-4. Un premier prototype interactif a été déployé avec Streamlit, confirmant la faisabilité d'une analyse automatisée à grande échelle, tout en soulignant les défis liés à la qualité des données, à l'hétérogénéité documentaire et à la stabilité des modèles open source.

Table des matières

Encadrement et environnement

Stage de fin d'études effectué du 01/03/2025 au 31/08/2025, au Département Sciences Economiques, Sociales et de Gestion (SESG), UFR Gestion du Vivant et Stratégies Patrimoniales, AgroParisTech.
22 place de l'Agronomie - CS 20040 - 91123 Palaiseau Cedex
- Directrice de stage : Cécile BLATRIX - Professeure en Science Politique, Présidente du Département Sciences Economiques, Sociales et de Gestion, AgroParisTech – UFR Gestion du Vivant et Stratégies Patrimoniales/ UMR Printemps
- Référent académique de stage : Vincent GUIGUE – Professeur en Informatique,
- Localisation du bureau : Bâtiment E, 5e étage, E5.236, AgroParisTech
- Téléphone de la directrice du stage : (+33) 6 46 29 63 43
- Téléphone de la stagiaire : (+33) 6 76 75 08 68

Description de l'établissement d'accueil, et du service d'accueil

Le Département Sciences Économiques, Sociales et de Gestion (SESG) est l'un des 5 Départements que compte AgroParisTech. Organisé en 8 UFR implantées sur Palaiseau mais aussi Montpellier, Nancy et Clermont-Ferrand, ce département joue un rôle clé dans la formation et la recherche en sciences humaines appliquées. Il regroupe des disciplines comme Agriculture comparée, Droit, Economie, Gestion, Sociologie, Science politique, avec pour mission de dispenser des connaissances théoriques, méthodologiques et appliquées nécessaires aux ingénieurs dans le domaine de l'environnement et des ressources naturelles.

L'Unité de Formation et Recherche (UFR) « Gestion du Vivant et Stratégies Patrimoniales » (GVSP) est une composante du SESG. Elle est dédiée à l'étude des politiques publiques et des stratégies patrimoniales dans des contextes complexes, tels que les crises environnementales. L'UFR adopte une approche interdisciplinaire, alliant sciences du vivant et sciences sociales, et est en lien étroit avec l'UMR Printemps (UMR 8085, UVSQ-CNRS-AgroParisTech), laboratoire auquel sont rattachés les deux enseignants-chercheurs en science politique (Bruno Villalba et Cécile Blatrix), où sont menées des recherches sur la gouvernance écologique et les mobilisations environnementales.

Le stage s'est déroulé au sein de l'UFR GVSP, dans le cadre d'un projet de recherche piloté par Cécile Blatrix.

Position dans l'organigramme de l'entreprise et du projet

La stagiaire est intégrée au Département SESG d'AgroParisTech, au sein de l'UFR « Gestion du Vivant et Stratégies Patrimoniales », sous la direction de Cécile Blatrix. Elle contribue directement au projet de recherche OBSERVANCE.

Le projet OBSERVANCE est financé par l'APR DRITT d'AgroParisTech et s'inscrit dans deux axes stratégiques de l'établissement :

– l'accompagnement des transitions dans la gestion des ressources naturelles et la préservation/restauration de l'environnement ;

– l'analyse de données complexes et multi-échelles pour répondre aux grands défis contemporains.

Le projet OBSERVANCE est un projet conçu spécifiquement pour opérationnaliser les résultats de PEGASE en élaborant le prototype d'une analyse automatisée de la prise en compte des avis de l'autorité environnementale. Il repose essentiellement sur le financement d'un stage de 6 mois sur cette mission. La stagiaire est donc en position de chargée de projet en lien étroit avec la responsable scientifique

Échéancier des travaux avec les dates absolues

Période	**Activités**
01 mars – 31 mars	Démarrage du stage : lecture des documents, prise en main des outils, exploration du corpus
01 avril – 10 juin	Développement et test des modules de prétraitement, extraction de texte et appariement
10 juin – 21 juin	Rédaction du rapport intermédiaire
26 juin	Soutenance intermédiaire
30 juin – 20 août	Campagne d'expérimentations, tests croisés sur de nouveaux jeux de données, amélioration des scores de performance
21 – 31 août	Rédaction finale du rapport, préparation à la soutenance
9 septembre	Soutenance finale

I. Introduction

1. Présentation du projet OBSERVANCE

En France, la loi du 10 juillet 1976 relative à la protection de la nature[1] a introduit l'obligation de prendre en compte les incidences environnementales des projets les plus susceptibles d'affecter le milieu naturel. Au niveau européen, la directive EIA de 1985 a posé les bases de l'évaluation environnementale pour les grands projets, avant que la directive SEA de 2001 n'étende cette exigence aux plans et programmes. Cette évolution a conduit la France à adapter son cadre juridique afin de garantir l'indépendance de l'examen environnemental.

C'est dans ce contexte qu'a été créée en 2009 l'Autorité environnementale (Ae), au sein du Conseil général de l'environnement et du développement durable (CGEDD), sous la tutelle du ministère de l'Écologie. La réforme de 2016, avec le décret n° 2016-519[2], a ensuite institué les Missions régionales d'Autorité environnementale (MRAe) pour assurer l'évaluation au niveau régional. Depuis lors, l'Ae et les MRAe jouent un rôle central en garantissant des évaluations environnementales à la fois indépendantes, rigoureuses et objectives.

En 2018, a été lancé le projet de recherche PEGASE, dirigé par Cécile Blatrix et Nathalie Frascaria-Lacoste. Financé dans le cadre du programme ITTECOP (Infrastructures de Transports Terrestres, Écosystèmes et Paysages) et soutenu par le ministère de la Transition écologique ainsi que par l'ADEME, ce projet avait pour objectif d'examiner la gouvernance de l'évaluation environnementale en France à la suite de la réforme de 2016. Il s'attachait en particulier à analyser le rôle de l'Autorité environnementale (Ae) et de ses missions régionales (MRAe), ainsi que leur capacité à renforcer la prise en compte des enjeux de biodiversité, de paysage et de qualité de l'air dans les projets, plans et programmes soumis à évaluation. L'un de ses volets majeurs consistait à étudier la manière dont les avis de l'Ae étaient intégrés dans le processus décisionnel, notamment à travers les mémoires en réponse des maîtres d'ouvrage.

Le projet s'est achevé en 2023 avec la publication d'un rapport[3] . Forts des résultats obtenus, ses responsables ont souhaité poursuivre cette dynamique et poser les bases d'un Groupement d'Intérêt Scientifique (GIS) consacré à la gouvernance environnementale. Dans cette perspective, la poursuite des recherches apparaît essentielle.

1. Loi n° 76-629 du 10 juillet 1976 relative à la protection de la nature : https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006068553
2. Décret n° 2016-519 du 28 avril 2016 portant réforme de l'autorité environnementale : https://www.legifrance.gouv.fr/loda/id/JORFTEXT000032465348
3. Rapport PEGASE : https://ittecop.fr/fr/ressources/telechargements/rapport-final/rapport-ittecop-apr-2017/1391-apr-2017-pegase-rf/file

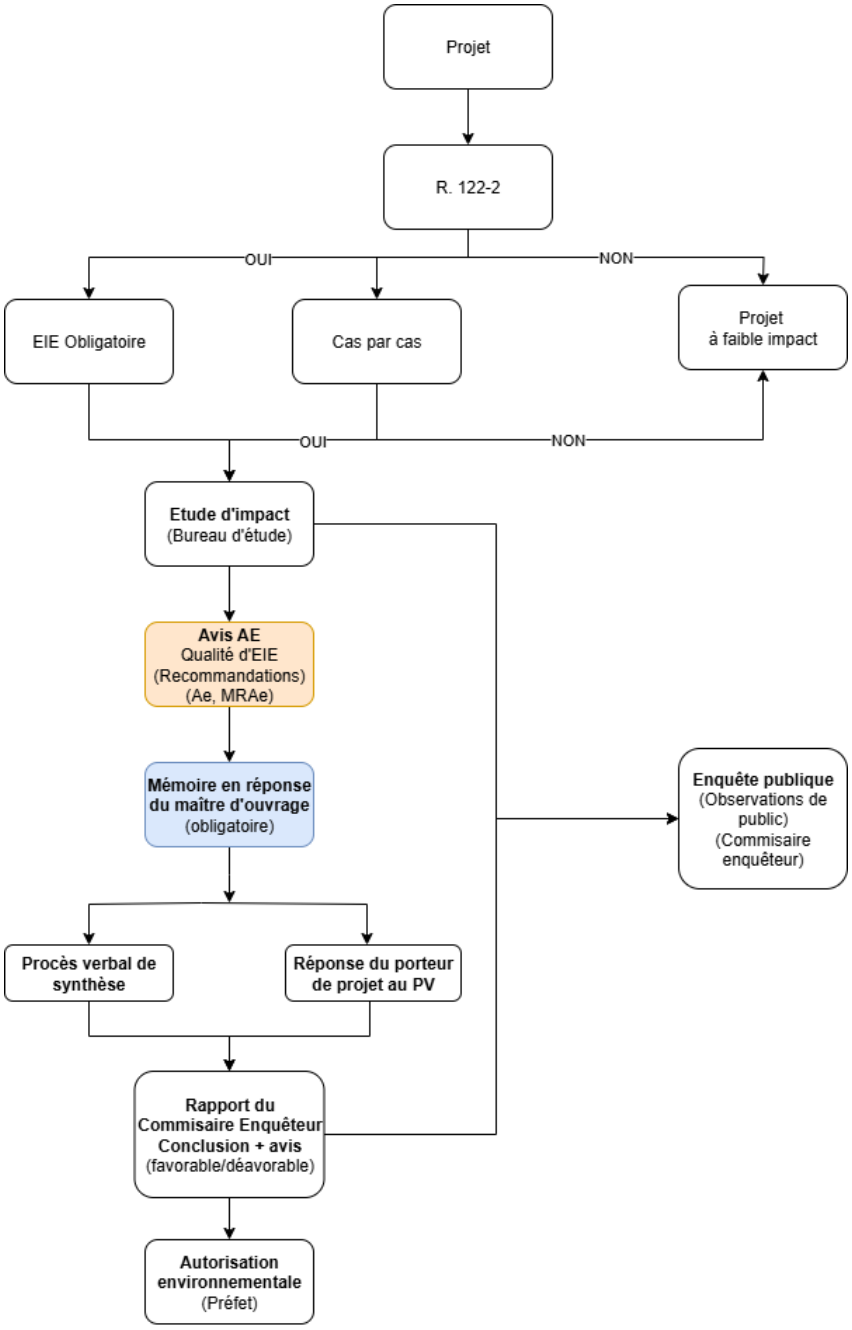

FIGURE 1 – Processus d'évaluation environnementale d'un projet

De manière générale, le processus d'évaluation environnementale d'un projet mobilise de nombreux acteurs, se déploie en plusieurs étapes successives et génère une volumétrie documentaire très importante. Cette abondance de documents rend l'analyse manuelle longue, coûteuse et, à grande échelle, difficilement réalisable.

C'est dans ce contexte que s'inscrit le projet de recherche OBSERVANCE, dont l'objectif est de développer un système automatisé d'analyse des documents générés tout au long du processus d'évaluation environnementale. Au cours de ce stage, un cadre méthodologique a été conçu et expérimenté afin d'extraire, d'apparier et d'évaluer les recommandations formulées dans les avis de l'Ae, ainsi que, dans un premier temps, les réponses apportées par les maîtres d'ouvrage dans leur mémoire en réponse.

Les résultats obtenus démontrent la faisabilité d'une telle approche et ouvrent la voie à des analyses élargies. Le pipeline mis en place a été intégré dans une application web, proposant une interface de consultation et d'exploration, accessible à l'adresse suivante : `http://34.38.26.53:8501/`.

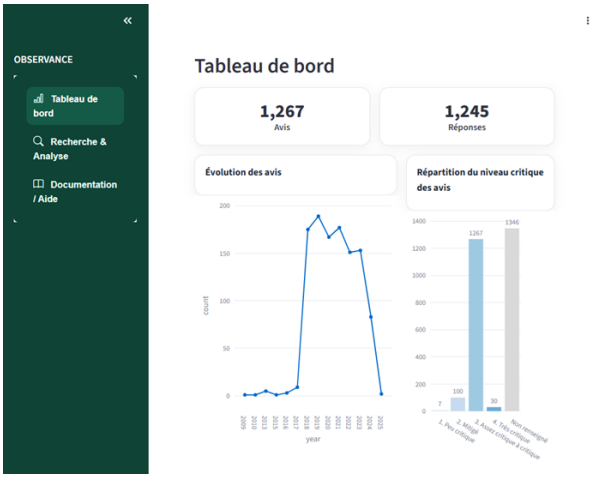

FIGURE 2 – (a) Interface de l'application OBSERVANCE

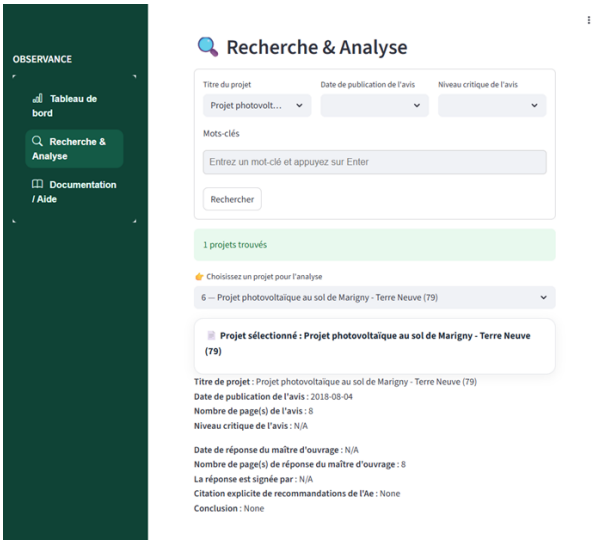

FIGURE 3 – (b) Interface de l'application OBSERVANCE

II. État de l'art

1. Contexte scientifique

L'objectif du projet OBSERVANCE étant d'analyser, à grande échelle et de manière automatisée, des documents administratifs, il s'inscrit naturellement dans le champ de l'*extraction d'information* (Information Extraction – IE), une branche du traitement automatique du langage naturel (TALN). L'enjeu de cette discipline est d'identifier, à partir de textes bruts, les éléments pertinents et de les structurer de façon exploitable, afin de faciliter l'accès aux informations utiles pour l'analyse et la prise de décision, tant pour les chercheurs que pour les décideurs et, plus largement, l'ensemble des parties prenantes.

Selon WIMALASURIYA et DOU 2010, l'IE consiste à « automatically retrieving certain types of information from natural language text ». Ces auteurs rappellent, en citant Russell et Norvig, que cette tâche vise plus précisément à «*process natural language text and to retrieve occurrences of a particular class of objects or events and occurrences of relationships among them*». Cette définition correspond parfaitement aux objectifs du projet OBSERVANCE, qui cherche à extraire et relier automatiquement des informations clés issues des documents de nature différente.

Afin de situer ce projet dans le paysage scientifique actuel, il est nécessaire d'examiner plusieurs initiatives récentes qui mobilisent des approches similaires d'extraction d'information appliquées à des corpus réglementaires ou administratifs. Deux projets illustrent particulièrement cette dynamique : **AI4PublicPolicy** et **Beyond Modeling**.

2. Travaux existants

2.1 AI4PublicPolicy

Il s'agit d'un projet européen (Horizon 2030) qui vise à doter les administrations publiques d'une plateforme cloud explicable (XAI) pour l'analyse de documents réglementaires, de données ouvertes et de retours citoyens multilingues. Basé sur une méthodologie inspirée de **CRISP-DM**, il combine collecte, préparation, analyse textuelle, extraction de règles interprétables (module QARMA) et visualisation interactive des résultats .

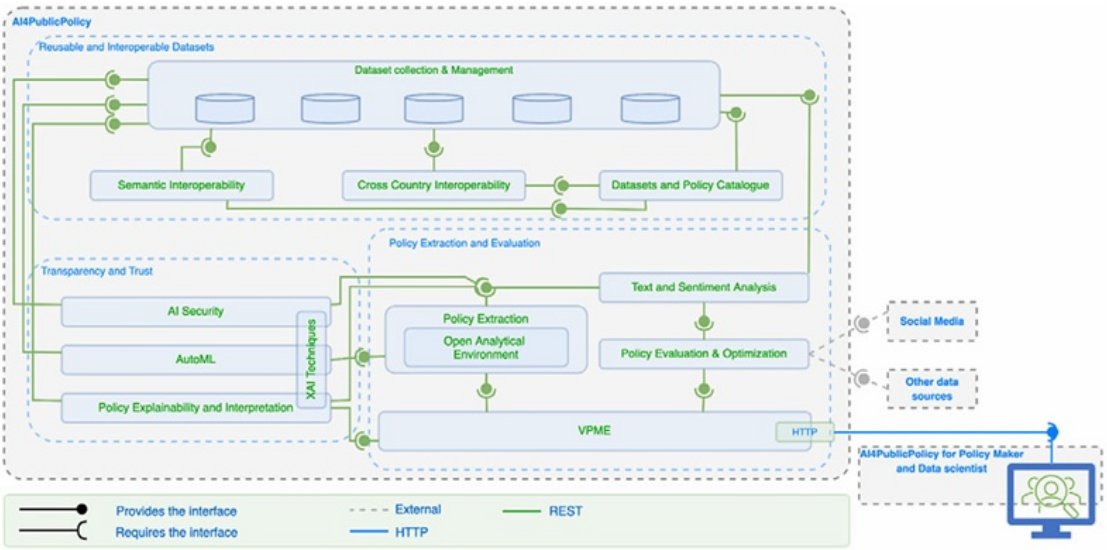

FIGURE 4 – Composants de l'architecture d'AI4PublicPolicy

Ce projet constitue une avancée notable en matière de conformité et de transparence, et ses expérimentations menées dans plusieurs villes européennes (Athènes, Lisbonne, Nicosie, Gênes, Bourgas) ont montré une réduction significative du temps d'évaluation des politiques publiques (PAPADAKIS et al. 2024).

Toutefois, l'approche d'**AI4PublicPolicy** reste difficilement transférable au projet OBSERVANCE. Ce dernier vise un appariement précis entre recommandations et réponses administratives, dans un cadre réglementaire spécifique. De plus, l'adaptation complète de cette infrastructure nécessiterait des moyens techniques et humains importants, peu compatibles avec un projet de stage exploratoire.

2.2 Beyond Modeling

Le projet **Beyond Modeling** (PLANAS et al. 2022) se focalise sur l'analyse automatisée de politiques environnementales. Il propose un pipeline modulaire incluant collecte automatique, prétraitement multiformat, labellisation assistée (Sentence-BERT), classification thématique et des incitations, extraction d'entités et de relations, puis construction d'un graphe de connaissances.

Les résultats indiquent une réduction du temps d'analyse de plusieurs semaines à quelques minutes, avec des performances robustes en classification (F1-score supérieur à 80% sur certaines tâches) et en extraction d'entités. Cette approche montre la faisabilité d'un traitement

NLP à grande échelle dans le domaine environnemental.

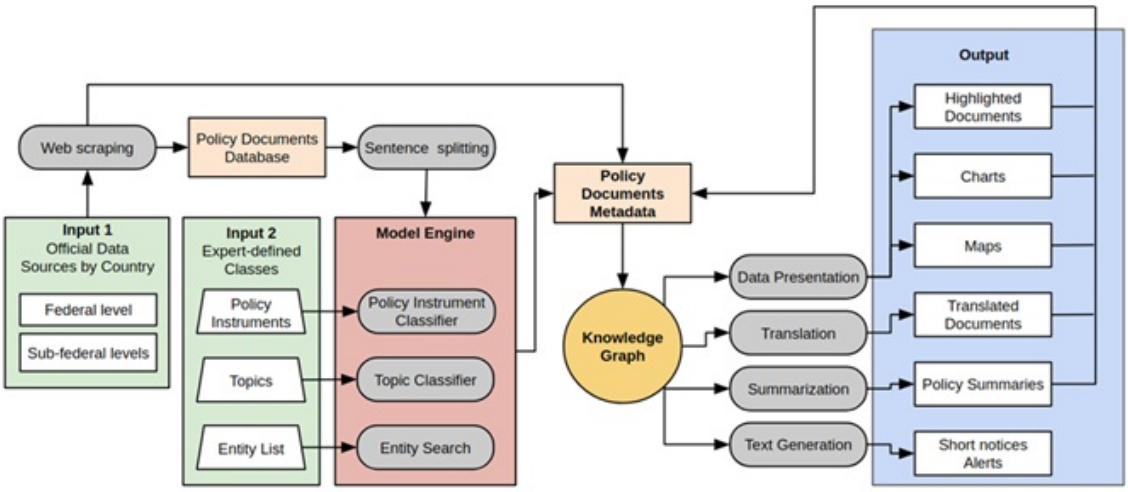

FIGURE 5 – Système de gestion des connaissances combinant différentes méthodologies de traitement du langage naturel (NLP) pour construire un pipeline de bout en bout

Ce projet partage plusieurs caractéristiques avec OBSERVANCE, notamment la gestion de documents administratifs non structurés, l'identification d'unités d'information pertinentes, et la structuration des résultats sous forme exploitable. Sa rigueur méthodologique et son architecture modulaire en font une source d'inspiration directe pour la conception du pipeline d'OBSERVANCE.

2.3 Positionnement

Le projet OBSERVANCE se distingue par son approche qui articule différents types de documents administratifs. Dans un premier temps, l'analyse porte principalement sur les avis émis par les autorités environnementales ainsi que sur les mémoires en réponse correspondants des maîtres d'ouvrage. L'objectif est d'évaluer dans quelle mesure les recommandations formulées par l'Ae sont effectivement appréciées et prises en compte par les maîtres d'ouvrage.

Une telle problématique dépasse les tâches classiques d'extraction d'information. Elle suppose également des opérations de **mise en relation et d'interprétation** entre ces deux sources textuelles qui sont longues, hétérogènes et, surtout pour les mémoires en réponse, peu structurées. Répondre à ces défis nécessite la conception d'une architecture spécifique, dont la section suivante présente les choix méthodologiques, le protocole et la démarche d'évaluation.

III. Matériel et Méthode

Le stage a consisté à concevoir et mettre en œuvre un pipeline d'extraction, de traitement et d'analyse automatisée des documents produits dans le cadre de l'évaluation environnementale. L'objectif méthodologique est double : (i) transformer un corpus hétérogène de documents PDF en données structurées exploitables, et (ii) extraire des indicateurs permettant d'évaluer la prise en compte effective des recommandations de l'Autorité environnementale (Ae) par les maîtres d'ouvrage. Les résultats attendus du pipeline ont été définis pour chaque type de document :

Sorties attendues :

Avis	1. Date de l'avis 2. Nombre de pages 3. Nombre de recommandations 4. Niveau critique (*peu critique, mitigé, assez critique à critique, très critique*)
Mémoire en réponse	1. Date de la réponse 2. Signataire(s) ou maître d'ouvrage identifié(s) 3. Nombre de pages 4. Citations explicites des recommandations de l'Ae (*totale, partielle, aucune*) 5. Conclusion synthétique (*volonté faible, réelle ou modérée du maître d'ouvrage*)

Deux phases successives ont été conduites :

— une **phase exploratoire**, sur 10 couples avis–réponse (issus du corpus du projet PEGASE) annotés manuellement, servant de jeu d'entraînement et d'évaluation ;

— une **phase d'extension**, appliquée à plus de 1200 couples issus du site `https://projets-environnement.gouv.fr`, permettant de tester la généralisabilité du pipeline.

L'approche adoptée dans le cadre du projet OBSERVANCE repose sur un pipeline modulaire en huit étapes (Figure 6), conçu pour transformer des documents PDF (avis de l'Ae et mémoire en réponse des maîtres d'ouvrage) en données structurées et interrogeables. Chaque étape est guidée par un objectif méthodologique précis, justifiée par des choix techniques, et évaluée à l'aide de métriques appropriées.

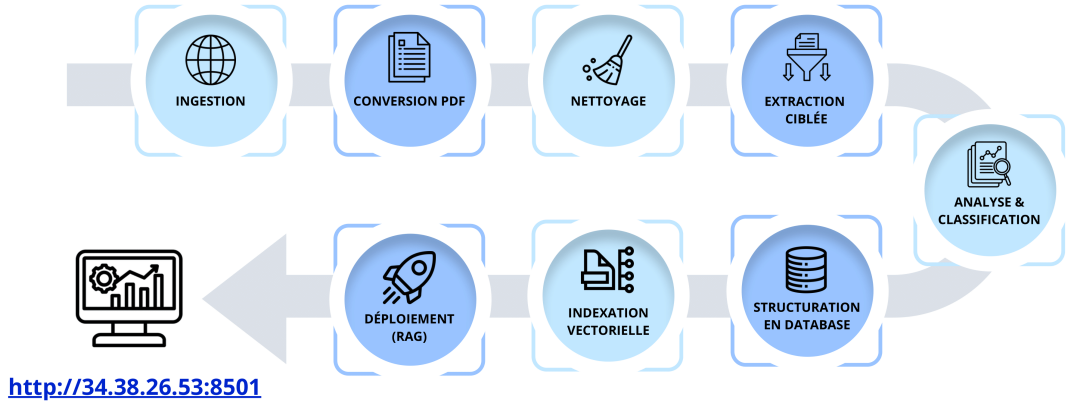

FIGURE 6 – Schéma du pipeline développé dans le cadre du projet OBSERVANCE

1. Ingestion

1.1. Collection des données

Le corpus de travail combine trois sources principales :

— **10 couples avis–réponse**, issus du programme PEGASE et utilisés pour la phase exploratoire ;

— **7651 avis** de l'Ae collectés via la plateforme `data.disclosure.ngo`, dépourvus de mémoires en réponse ;

— **2100 couples avis–réponse** identifiés sur le site `projets-environnement.gouv.fr`.

Les documents au format PDF ont été téléchargés via les liens répertoriés dans un fichier **.csv** de métadonnées mis à disposition par les sites. Toutefois, les avis issus de `data.disclosure.ngo` n'ont pas encore été exploités dans la suite des travaux, faute de mémoires en réponse associés. Pour le corpus venus du site `projets-environnement.gouv.fr`, après filtrage et exclusion des fichiers corrompus, le corpus final retenu pour l'expérimentation comprend environ **1200 couples avis–réponse**.

1.2. Résultats

Le corpus obtenu se caractérise par une hétérogénéité importante des documents collectés : certains sont textuels, d'autres hybrides (texte enrichi d'images scannées), et une fraction non négligeable correspond à des PDF entièrement scannés. Cette variabilité structurelle justifie la mise en place d'approches différenciées pour la conversion en texte brut. Une typologie détaillée de cette diversité est présentée en Annexes (Nature des documents).

2. Conversion des PDF

L'objectif de cette étape est de transformer des documents hétérogènes (PDF textuels, scannés ou hybrides) en texte brut exploitable pour l'analyse automatique. Trois approches ont été mises en œuvre et comparées :

— **Extraction directe** avec `pdfplumber`, adaptée aux documents nativement textuels ;
— **Reconnaissance optique de caractères (OCR)** via `pdf2image` et `pytesseract`, nécessaire pour les fichiers scannés ou mal structurés ;
— **Approche hybride** combinant extraction directe et OCR : les deux versions sont comparées, et la plus complète (longueur textuelle la plus élevée après nettoyage) est conservée.

2.1. Protocole d'évaluation

L'évaluation de la qualité des textes extraits repose sur des métriques standards issues de la littérature scientifique. Plusieurs travaux insistent sur la nécessité de comparer ces sorties avec un *ground truth* de référence, en mesurant à la fois la fidélité au niveau des caractères et des mots, ainsi que la qualité de la structure du document. Parmi les indicateurs les plus couramment utilisés, on retrouve :

— **Character Error Rate (CER) et Word Error Rate (WER)**, définis comme le rapport entre le nombre de substitutions, suppressions et insertions et la longueur du texte de référence. Ces métriques constituent la base de l'évaluation OCR (Echavarría Peláez et Le Meur 2025).
— **Longest Common Subsequence (LCS)** est une métrique de similarité qui identifie la plus longue séquence commune apparaissant dans deux textes. Elle est largement utilisée pour l'évaluation de la qualité des transcriptions ou des textes produits automatiquement (Reshma et Mathew 2015).
— **Des indicateurs plus récents tels que l'Exact Match et le F1-score**, utilisés dans le contexte des systèmes de *Retrieval-Augmented Generation* (RAG), afin de mesurer l'impact direct des erreurs OCR sur les tâches de recherche et de génération de contenu (J. Zhang et al. 2024 ; Yu et al. 2024).

Dans le cadre de ce projet, la qualité des textes extraits a été donc évaluée en comparant chaque sortie à un jeu de référence (*ground truth* ou vérité de terrain) constitué par transcription manuelle des **10 couples avis–réponse**. Les textes produits par les trois méthodes ont été alignés avec les références grâce à la **distance de Levenshtein** (LEVENSHTEIN 1966; HIRSCHBERG 1977) et à l'algorithme de **Longest Common Subsequence (LCS)** (RESHMA et MATHEW 2015), permettant de calculer plusieurs indicateurs complémentaires.

Character Error Rate (CER) : proportion d'erreurs au niveau des caractères, définie comme le rapport entre les substitutions (S), suppressions (D) et insertions (I), et le nombre total de caractères de référence N.

$$CER = \frac{S + D + I}{N}$$

Word Error Rate (WER) : analogue au niveau des mots, défini à partir des alignements entre texte prédit et texte de référence, avec S_w, D_w, I_w les substitutions, suppressions et insertions de mots, et N_w le nombre de mots de référence.

$$WER = \frac{S_w + D_w + I_w}{N_w}$$

F1-score lexical : mesure harmonique de la précision (P) et du rappel (R) calculés au niveau des mots, permettant d'évaluer la couverture lexicale tout en pénalisant les faux positifs.

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F1 = \frac{2 \cdot P \cdot R}{P + R}$$

où TP, FP, FN désignent respectivement les vrais positifs, faux positifs et faux négatifs au niveau des mots.

LCS-ratio : proportion de la plus longue sous-séquence commune (*Longest Common Subsequence*, LCS) entre la sortie et la référence, normalisée par la longueur du texte de référence N.

$$LCS\text{-ratio} = \frac{\mathrm{LCS}(y^{\text{réf}}, y^{\text{prédit}})}{N}$$

où LCS désigne la longueur de la plus longue sous-séquence commune.

2.2. Résultats

TABLE 1 – Performances comparées des trois méthodes d'extraction PDF

mode	CER	Word_precision	Word_recall	Word_f1	WER	lcs_ratio
avis_hybrid	0.121	0.86	0.876	0.868	0.143	0.858
avis_ocr	0.027	0.949	0.969	0.959	0.055	0.946
avis_plumber	0.125	0.856	0.852	0.854	0.156	0.844
reponse_hybrid	0.101	0.949	0.894	0.902	0.13	0.871
reponse_ocr	0.103	0.951	0.893	0.902	0.128	0.872
reponse_plumber	0.412	0.739	0.563	0.581	0.452	0.587

Les résultats montrent que :

— Pour les **avis**, l'approche OCR surpasse nettement les deux autres, avec un **CER moyen de 2.7%** et un **Word F1 de 0.96**.

— Pour les **mémoires en réponse**, les approches OCR et hybride donnent des performances proches, avec un **CER autour de 10%** et un **Word F1 ≈ 0.90**. L'extraction par `pdfplumber` seule montre des performances insuffisantes (CER > 40

— La distribution des scores (Figure 7) met en évidence une forte variabilité pour les mémoires extraits via `pdfplumber`.

— Enfin, la comparaison des métriques globales (Figure 8) confirme la robustesse de l'OCR sur l'ensemble du corpus, justifiant son intégration comme méthode par défaut dans le pipeline.

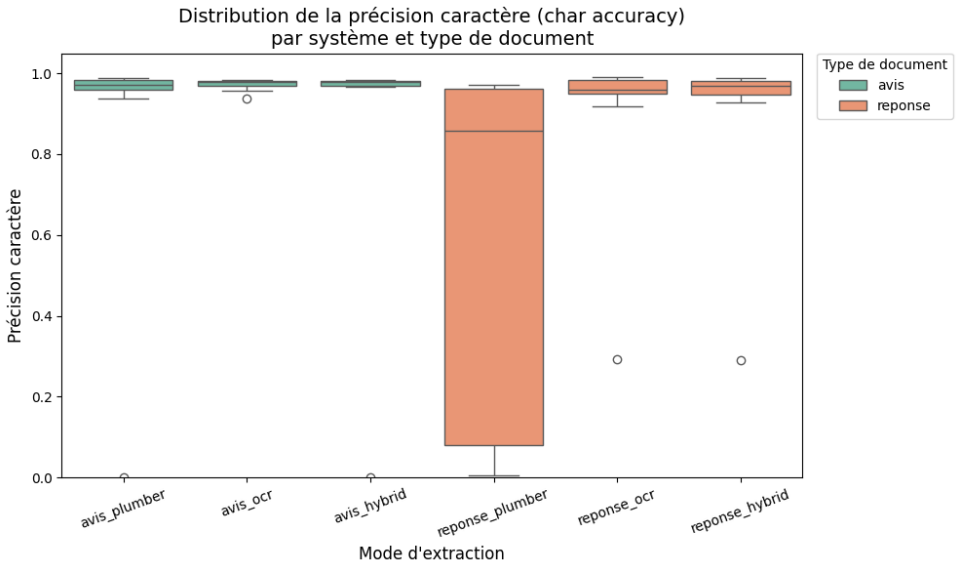

FIGURE 7 – Distribution de la précision caractère (char accuracy) par système et type de document.

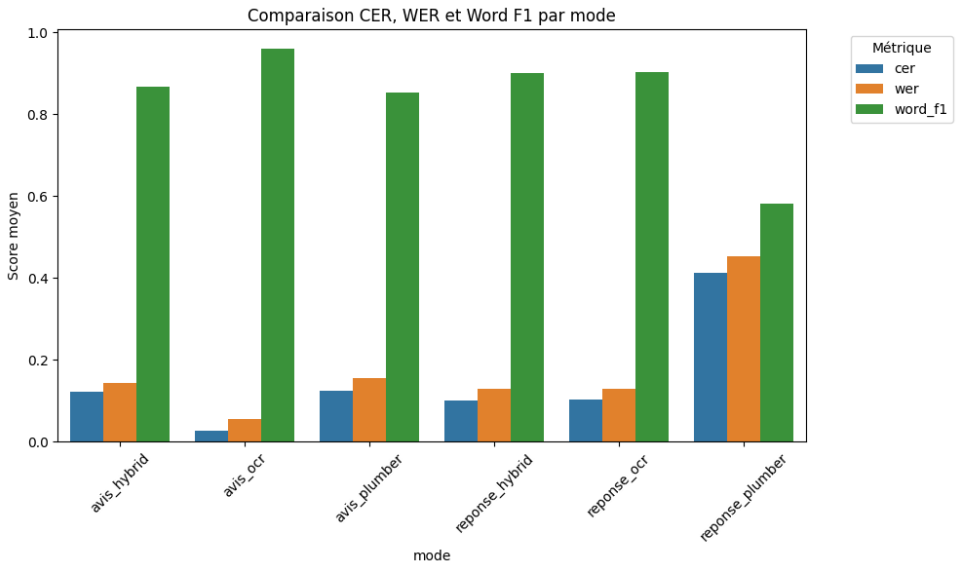

FIGURE 8 – Comparaison des scores moyens CER, WER et Word F1 par mode d'extraction.

3. Nettoyage

L'objectif de cette étape est de corriger les erreurs récurrentes générées lors de l'OCR, en particulier dans les **mémoires en réponse**, souvent issus de documents scannés et donc plus bruités que les avis. La stratégie retenue repose sur une approche empirique : l'ensemble de textes extraits a été comparé manuellement au *ground truth* afin d'identifier les erreurs les plus fréquentes (confusion de caractères, ponctuation défectueuse, substitutions récurrentes). Ces motifs ont ensuite été intégrés sous forme de règles de remplacement dans une fonction de normalisation appliquée automatiquement à l'ensemble du corpus.

3.1. Protocole d'évaluation.

Comme pour la conversion des PDF, l'impact du nettoyage a été évalué par comparaison avec le *ground truth*. Les textes avant et après nettoyage ont été alignés au niveau caractère et mot, et analysés selon les métriques standards : **CER**, **WER**, **F1-score lexical** et **LCS-ratio**.

3.2. Résultats.

(Voir Annexes pour les résultats détaillésRésultats des mesures

Les résultats montrent que :

— **Pour les avis :** la version OCR brute atteint déjà une qualité élevée (**CER = 2.7%**, **Word F1 = 0.96**). L'application du nettoyage conduit paradoxalement à une légère

TABLE 2 – Performances avant et après nettoyage (comparaison avec le *ground truth*).

mode	CER	word_precision	word_recall	word_f1	WER	LCS-ratio
avis_ocr	0.027	0.949	0.969	0.959	0.055	0.946
avis_cleaned	0.035	0.914	0.933	0.923	0.213	0.968
reponse_ocr	0.103	0.951	0.893	0.902	0.128	0.872
reponse_cleaned	0.109	0.957	0.9	0.909	0.643	0.897

dégradation des performances (Word F1 passant de 0.96 à 0.92, WER multiplié par quatre), ce qui suggère que certaines règles ont introduit des erreurs dans un texte déjà relativement propre. Le nettoyage ne présente donc pas de valeur ajoutée pour ce type de document.

— **Pour les mémoires en réponse :** Après nettoyage, la char_accuracy baisse légèrement (0.89 contre 0.90), mais les métriques lexicales s'améliorent : **Word Precision = 0.96** (contre 0.95 pour OCR brut) et **Word F1 = 0.91** (contre 0.90). Ces résultats indiquent que les règles de correction ciblées (apostrophes, unités, symboles) améliorent la lisibilité et la fidélité lexicale, même si le taux d'erreurs caractère reste élevé.

À ce stade, le nettoyage n'apparaît pas comme un levier d'amélioration significatif de la qualité des fichiers texte extraits. Même si une légère progression est observée sur certains mémoires en réponse, la forte hétérogénéité de ces documents ne permet pas de garantir l'efficacité durable des règles actuellement définies. Néanmoins, nous avons choisi de conserver cette étape dans le pipeline, car elle constitue un maillon important du processus et pourra être optimisée ultérieurement.

4. Extraction ciblée

Une fois les textes normalisés, plusieurs modules d'extraction ont été développés afin d'identifier les informations clés nécessaires à l'analyse. Cette étape est cruciale car elle conditionne la qualité des analyses ultérieures et permet de structurer les données brutes en unités exploitables.

— **Nombre de pages** : extrait directement à partir des métadonnées des fichiers PDF grâce à la librairie `PyPDF`.

— **Date de publication** : détectée par expressions régulières (`Regex`), puis normalisées. La stratégie consiste à sélectionner la date la plus récente mais antérieure à la date courante.

— **Sections clés** : les avis ont été segmentés en deux parties principales — *Synthèse* et *Avis détaillé* — à partir de la détection automatique des titres et motifs typographiques.

— **Recommandations** : compte tenu de leur structure (formulations débutant par « recommande » et se terminant par un point), elles ont été récupérées via des règles `Regex`.
— **Signataires** : l'identification des auteurs des mémoires en réponse (nom, prénom, fonction, organisation) a d'abord été tentée via un module de reconnaissance d'entités nommées (NER), sans succès. Une seconde approche, mobilisant un LLM local (`llama3.2`), a permis d'extraire correctement les signataires dans la majorité des cas.

4.1. Protocole d'évaluation.

Comme pour l'étape de conversion des PDF, l'évaluation de l'extraction ciblée a porté sur un jeu de validation constitué manuellement. Les critères mesurés concernaient à la fois la complétude et la précision des extractions.

4.2. Résultats

TABLE 3 – Résumé de l'évaluation de l'extraction des informations clés

Information	Méthode utilisée	Résultat	Complétude	Précision
Date de publication	Regex	Toutes les dates détectées	100%	100%
Nombre de pages	Métadonnées PDF	Nombre exact	100%	100%
Signataires	NER (échec)	Détection incorrecte	0%	0%
Signataires	LLM (`llama3.2`)	4/5 signataires trouvés	100%	80%
Sections clés et recommandations	Regex	Extraction complète	100%	100%

L'évaluation met en évidence des résultats contrastés selon la nature des informations extraites. Les métadonnées simples (nombre de pages) et les éléments textuels structurés (sections, recommandations) sont captés avec une précision élevée à l'aide de règles déterministes (regex, métadonnées PDF). Ces observations confirment que, contrairement à certaines idées reçues, les approches basées sur règles demeurent robustes et efficaces dans des contextes bien définis (CHITICARIU, LI et REISS 2013).

En revanche, l'identification des signataires illustre clairement les limites des approches classiques. Les méthodes de reconnaissance d'entités nommées (NER) échouent fréquemment lorsqu'elles sont appliquées à des documents longs et hétérogènes, souvent issus de l'OCR - un constat déjà rapporté dans la littérature sur les journaux historiques numérisés, où

les erreurs OCR et la variation orthographique entraînent une baisse significative du rappel (NEUDECKER 2016). De plus, même lorsqu'une liste d'entités est correctement détectée, déterminer avec précision lesquelles correspondent effectivement aux signataires du mémoire en réponse demeure une tâche complexe. Le recours à un modèle de langage tel que `llama3.2` a permis d'obtenir des résultats nettement supérieurs (80% de précision sur l'échantillon testé).

Toutefois, cette dépendance aux LLM soulève des enjeux de reproductibilité et de coût computationnel, déjà signalés par la littérature récente (STAUDINGER et al. 2024; YUAN et al. 2025). Dans ce cadre, un compromis doit être recherché entre la légèreté et la transparence des méthodes par règles, et la puissance mais aussi l'opacité et le coût des modèles LLM.

5. Analyse et classification

Cette étape s'est concentrée sur l'évaluation du **niveau critique** des avis émis par l'Ae, selon une échelle à quatre niveaux définie par le projet PEGASE :

1. *Peu critique* : avis globalement positif, peu ou pas de remarques.
2. *Mitigé* : avis contenant à la fois des points positifs et des critiques modérées.
3. *Assez critique à critique* : plusieurs critiques importantes, étude jugée incomplète.
4. *Très critique* : avis très négatif, étude jugée très insuffisante ou absente.

Plusieurs modèles de langage ont été testés afin d'automatiser cette tâche : **GPT-4** (OpenAI) et différents modèles open-source déployés via **Ollama** (*LLaMA 3.2*, *Qwen*, *Gemma*, *Mistral*, *DeepSeek*). Chaque avis, constitué de la *synthèse* et de l'*avis détaillé*, a été soumis à un prompt décrivant les critères de classification (voir Annexes Prompts utilisés). Le modèle devait produire une étiquette (1–4) correspondante au niveau critique du texte analysé.

5.1. Protocole d'évaluation

L'évaluation de la classification des niveaux de criticité repose sur la comparaison entre :
— les jugements de référence établis par les chercheur·es PEGASE,
— et les prédictions issues des LLMs, appliquées à la fois sur la **synthèse de l'avis** et sur l'**avis détaillé**.

Afin d'assurer la robustesse des résultats, chaque modèle a été exécuté sur **3 itérations**. Cette approche permet d'évaluer la **stabilité inter-runs** (YUAN et al. 2025).

Les performances de classification du **niveau critique des avis** ont été évaluées à l'aide de plusieurs métriques adaptées :

— **Accuracy** : proportion d'avis correctement classés par rapport au niveau de référence.

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{y_i^{\text{prédit}} = y_i^{\text{référence}}\}$$

où N est le nombre total d'avis testés et $\mathbf{1}$ la fonction indicatrice.

— **F1-score (micro)** : mesure globale de la qualité de classification multi-classes, combinant précision (P) et rappel (R) sur l'ensemble des niveaux de criticité.

$$F1 = \frac{2 \cdot P \cdot R}{P+R}, \quad P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}$$

où TP, FP, FN désignent respectivement les vrais positifs, faux positifs et faux négatifs.

— **F1-score macro** : afin de compenser le déséquilibre entre classes (certains niveaux de criticité étant plus fréquents), le F1 est également calculé séparément pour chaque niveau $c \in C$, puis moyenné.

$$F1_{\text{macro}} = \frac{1}{|C|}\sum_{c \in C} F1_c$$

— **Écart absolu moyen (MAE)** : étant donné que la tâche est une classification **ordinale** (les niveaux 1 à 4 sont ordonnés), il est pertinent de mesurer l'écart moyen entre le niveau attendu et le niveau prédit. Une erreur d'un niveau (ex. 2 vs 3) est ainsi moins pénalisée qu'une erreur de deux ou trois niveaux (ex. 1 vs 4).

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \left| y_i^{\text{prédit}} - y_i^{\text{référence}} \right|$$

— **Variance inter-runs** : pour analyser la stabilité des modèles (en particulier des LLMs locaux sensibles aux variations d'exécution), on calcule la variance des métriques (Accuracy, MAE, etc.) sur plusieurs itérations indépendantes K.

$$\text{Var} = \frac{1}{K}\sum_{k=1}^{K} \left(m_k - \bar{m}\right)^2$$

où m_k est la valeur de la métrique lors du run k et $\bar{m}$ la moyenne sur les K exécutions.

En complément, l'**effet de la longueur du contexte** a été testé en comparant la classification sur la *synthèse de l'avis* (texte court et condensé) et sur l'*avis détaillé* (plusieurs dizaines de pages). Ce facteur est critique, car il influence la capacité des modèles de langage

à conserver l'information pertinente dans des contextes longs, phénomène déjà documenté par LIU, SALAKHUTDINOV et LIANG 2023.

5.2. Résultats

Parmi les modèles testés, certains se sont révélés inadéquats pour la tâche. En particulier, **Gemma** s'est montré très limité, en classant systématiquement tous les avis dans une seule catégorie (niveau 1), ce qui le rend inutilisable dans le cadre de notre évaluation. De plus, **Mistral** et **DeepSeek** n'ont pas pu être évalués : le premier en raison de contraintes computationnelles, et le second à cause de restrictions d'accès (`sudo`) sur le serveur expérimental.

En conséquence, nous ne présentons ici que les résultats pour les modèles **GPT-4** et **LLaMA 3.2** (Voir Annexes pour les résultats détaillés Résultats des mesures).

TABLE 4 – Comparaison des performances de GPT-4 et LLaMA selon la longueur du texte (synthèse vs avis détaillé)

Modèle	Type de texte	Accuracy (moy.)	MAE (moy.)	F1-micro	F1-macro
GPT-4	Synthèse (3 runs)	0,60	0,467	0,60	≈ 0,505
GPT-4	Avis détaillé (3 runs)	0,50	0,60	0,50	≈ 0,167
LLaMA 3.2	Avis synthèse (3 runs)	0,50	0,60	0,50	≈ 0,167
LLaMA 3.2	Avis détaillé (3 runs)	0,3333	0,8667	0,3333	0,176

TABLE 5 – Variances inter-runs des métriques

Modèle	Type de texte	Variance Accuracy	Variance MAE	Variance F1-micro	Variance F1-macro
GPT-4	Synthèse (3 runs)	0	0.002	0	0.007
GPT-4	Avis détaillé (3 runs)	0	0	0	0
LLaMA 3.2	Avis synthèse (3 runs)	0	0	0	0
LLaMA 3.2	Avis détaillé (3 runs)	0.0089	0.0422	0.0089	0.00438

— **GPT-4** se montre robuste et cohérent sur les textes courts (*synthèses*), avec une précision moyenne de 60% et un MAE de 0.467. Ses erreurs demeurent généralement proches de la vérité (±1 niveau), ce qui explique un F1-macro autour de 0.50. En revanche, sur les avis longs, ses prédictions s'effondrent en raison d'une normalisation quasi systématique vers la classe « 3 » : l'accuracy tombe à 50% et le F1-macro à 0.167. Ce comportement illustre une limite bien documentée des LLMs sur les contextes étendus, comme l'a montré l'étude *Lost in the Middle : How Language Models Use Long Contexts* (LIU, SALAKHUTDINOV et LIANG 2023).

— **LLaMA 3.2** présente une dynamique inverse. Sur les synthèses, il tend lui aussi à prédire presque exclusivement la classe « 3 », obtenant des scores proches de ceux

de GPT-4 sur avis détaillés (accuracy 50%, MAE 0.60). Toutefois, sur les avis longs, ses sorties sont plus variées : si l'accuracy moyenne est inférieure (33%), la variance inter-runs est importante (acc=0.0089, MAE=0.0422, F1-macro=0.00438). Cela traduit une instabilité plus marquée, mais également une sensibilité accrue au contenu long par rapport à GPT-4.

Un essai complémentaire a également été mené avec **GPT-4 sur synthèse**, en comparant les configurations *few-shot* et *zero-shot* (voir Annexes Prompts utilisés). Les résultats confirment que, pour cette tâche, le **few-shot est plus performant que le zéro-shot**.

TABLE 6 – Résultats de GPT-4 en mode zero-shot et few-shot (synthèse)

Configuration	**Accuracy**	**MAE**	**F1-micro**	**F1-macro**
Zero-shot	**0,60**	0,50	0,60	0,429
Few-shot (moy. 3 runs)	**0,60**	0,467	0,60	≈ 0,505

En résumé, l'expérimentation met en évidence plusieurs aspects :

— **Validité :** GPT-4 fournit des résultats proches de la *préférence*, confirmant son potentiel pour l'automatisation partielle de l'évaluation des avis.

— **Mérite :** les modèles open-source locaux (LLaMA) peuvent produire des résultats acceptables, mais au prix d'une instabilité importante.

— **Limites :** la dépendance aux ressources matérielles et logicielles (cas Mistral et DeepSeek), ainsi que la sensibilité aux textes longs, constituent des freins importants.

5.3. Classification du mémoire en réponse

Afin d'évaluer la **volonté de prise en compte des recommandations**, les mémoires en réponse ont été classés selon un arbre de décision simple, basé sur deux critères principaux : la longueur du document et la présence (partielle ou totale) de citations explicites des recommandations de l'autorité environnementale.

- Si le mémoire en réponse contient moins de **4 pages**, il est directement classé comme une **volonté faible**.
- Si le mémoire dépasse **4 pages**, alors :
  - l'absence de citation des recommandations entraîne une classification en **volonté faible**,
  - une citation partielle conduit à une **volonté moyenne**,
  - une citation systématique et exhaustive conduit à une **forte volonté**.

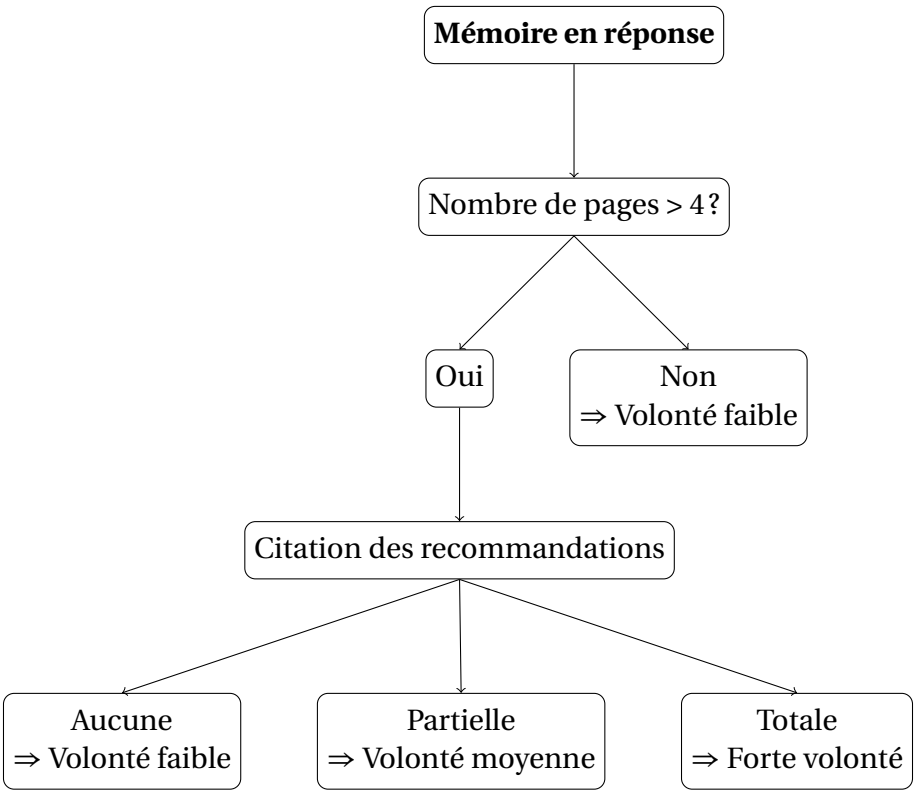

FIGURE 9 – Arbre de décision pour la classification des mémoires en réponse

6. Structuration en base de données

La structuration en base de données constitue une étape clé pour assurer la traçabilité, la reproductibilité et l'extensibilité des analyses menées dans ce projet. Après l'extraction et le prétraitement des informations issues des documents PDF, celles-ci ont été organisées dans une base relationnelle construite sous `SQLite`, enrichie d'un moteur de recherche plein texte (FTS5). Le choix d'un modèle relationnel répond aux principes fondateurs de la modélisation de données (CODD 1970).

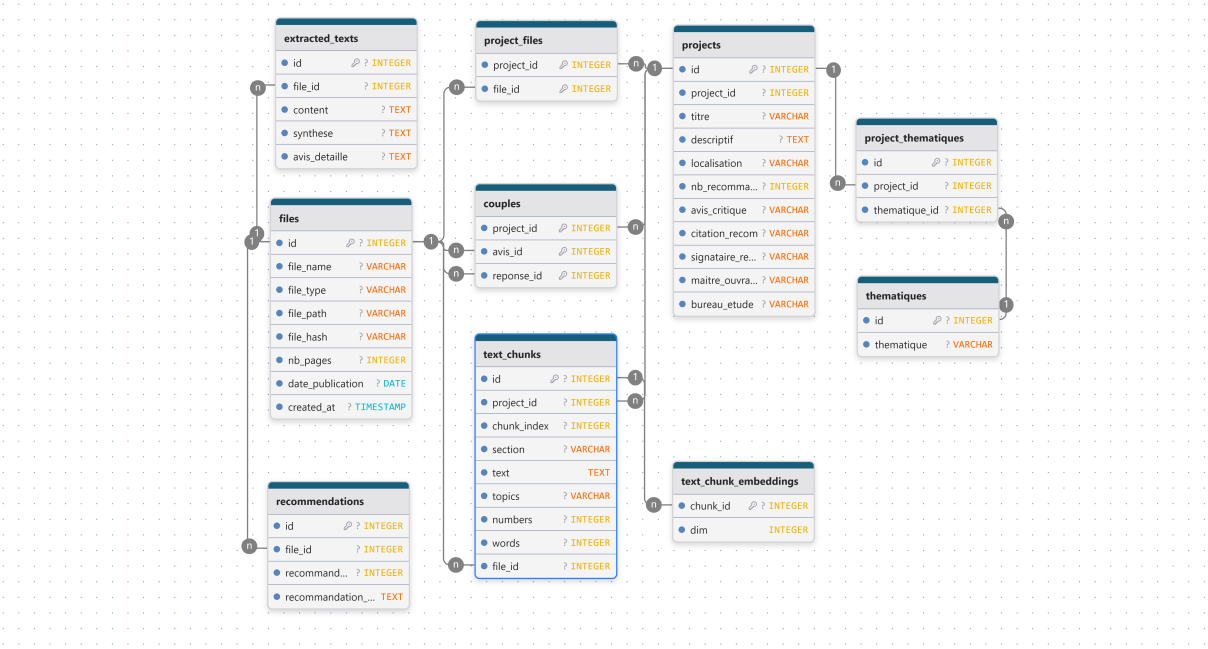

FIGURE 10 – Schéma relationnel de la base de données construite pour le pipeline OBSERVANCE

Le schéma conceptuel de la base repose sur plusieurs entités principales :

— **files** : centralise les métadonnées documentaires (nom, type, lien de téléchargement, date de publication, empreinte de hachage) et assure la traçabilité des sources ;

— **projects** : contient les informations contextuelles (titre, localisation, maître d'ouvrage, nombre de recommandations, niveau de criticité de l'avis) ;

— **extracted_texts** : stocke les contenus textuels normalisés (synthèse de l'avis, avis détaillé, extraits pertinents du mémoire en réponse du maître d'ouvrage) ;

— **recommandations** : conserve, pour chaque projet, l'index de recommandations ainsi que leurs contenus textuels normalisés ;

— **thematiques** et **project_thematiques** : dispositif relationnel permettant la gestion des associations plusieurs-à-plusieurs et la recherche thématique efficace ;

— **text_chunks** et **text_chunk_embeddings** : structures spécifiques destinées à la segmentation textuelle et à l'indexation vectorielle, facilitant l'intégration avec des modèles de type RAG (Retrieval-Augmented Generation).

Chaque table dispose d'une clé primaire unique et de clés étrangères explicites, assurant ainsi la cohérence des relations. Ce design respecte les normes N1 et N2 de conception : les attributs sont atomiques (1NF) et les dépendances fonctionnelles sont correctement séparées (2NF), conformément aux bonnes pratiques de conception relationnelle (ELMASRI et NAVATHE 2015).

L'intégration du module `FTS5` étend les capacités de `SQLite` en permettant la recherche textuelle plein texte dans les documents volumineux (SQLITE n.d.). Contrairement à une recherche naïve par `LIKE`, le moteur FTS construit un index inversé optimisé qui autorise des requêtes rapides, avec support de la proximité (`NEAR`), de la pondération par pertinence et de la combinaison logique (`AND`, `OR`). Ce mécanisme constitue un socle essentiel pour l'intégration du système RAG dans la prochaine étape.

Dans ce projet, les données extraites ont d'abord été stockées temporairement dans des fichiers `CSV` avant leur importation dans la base relationnelle. Le choix de `SQLite`, léger et portable, est justifié pour un prototype exploratoire. Néanmoins, ses limites apparaissent dans des environnements multi-utilisateurs ou lors du traitement de volumes plus importants. Une migration vers un SGBD plus robuste (PostgreSQL, MySQL) est alors envisageable.

7. Indexation vectorielle - RAG

Avec la volonté de dépasser le stade exploratoire et d'aboutir à un résultat concret, nous avons choisi de mettre en œuvre une approche de type **Retrieval-Augmented Generation (RAG)**. L'idée initiale était de permettre aux chercheurs de manipuler les données de manière la plus interactive possible. Toutefois, au fur et à mesure de l'avancement du projet, il est apparu que le RAG pouvait constituer une stratégie particulièrement prometteuse pour exploiter efficacement le corpus disponible.

Plutôt que de soumettre l'intégralité d'un document au modèle de langage, l'approche consiste à le segmenter en passages courts, puis à sélectionner uniquement les plus pertinents grâce à une recherche vectorielle (par ex. ChromaDB). Le modèle ne traite alors qu'un contexte ciblé, ce qui améliore sa robustesse et la qualité des justifications produites. De récents travaux confirment l'intérêt de cette stratégie : LEWIS et al. 2020 ont introduit le concept de RAG pour les tâches de *knowledge-intensive NLP*, et GAO et al. 2024 ont montré que cette approche facilite la gestion des documents longs en réduisant la surcharge cognitive

des modèles.

Afin de mettre en œuvre ce principe, les documents nettoyés ont été segmentés en **chunks** d'environ 180 mots (400–600 tokens), avec un chevauchement de 30 mots entre segments pour préserver le contexte (MICROSOFT AZURE ARCHITECTURE CENTER 2025). Les segments trop courts (<20 mots) ont été éliminés afin de réduire le bruit, puis chaque chunk a été enrichi de métadonnées structurées (nom de fichier, section, index, thématiques, occurrences numériques, longueur).

Ces unités textuelles ont ensuite été vectorisées à l'aide du modèle multilingue (multilingual-e5-base), entraîné pour la recherche sémantique (WANG et al. 2024). Les embeddings ont été normalisés (L_2) afin de stabiliser et d'accélérer le calcul de similarité (cosinus ou dot-product), tout en s'appuyant sur l'usage de la similarité cosinus pour les embeddings (REIMERS et GUREVYCH 2019). Les vecteurs ont ensuite été insérés par lots dans une base persistante `ChromaDB` au sein d'une collection dédiée. Cette indexation vectorielle constitue le socle d'une recherche dense par similarité permettant d'identifier rapidement les passages pertinents et de préparer l'intégration d'un système RAG.

7.1. Stratégie d'évaluation du RAG

L'évaluation d'un système de *Retrieval-Augmented Generation* (RAG) nécessite de prendre en compte à la fois la **qualité de la récupération des documents** et la **qualité des réponses générées**. Plusieurs métriques issues de la littérature ont été retenues pour définir notre stratégie d'évaluation :

— **Évaluation de la récupération (retrieval)** :
  — **Recall@k** : proportion de fois où la réponse de référence se trouve dans les k documents récupérés.

$$\text{Recall@k} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\text{doc\_réf} \in \text{Top-k}(q_i)\}$$

  (GUU et al. 2020; KARPUKHIN et al. 2020).
  — **Mean Reciprocal Rank (MRR)** : mesure de la position moyenne du document pertinent dans la liste des résultats.

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\text{rang}_i}$$

  (VOORHEES et TICE 2000).

- **Évaluation de la génération (generation)** :
  - **Exact Match (EM)** et **F1-score lexical** : comparaison mot à mot entre la réponse générée et la réponse de référence, fréquemment utilisés dans les benchmarks QA (J. ZHANG et al. 2024; YU et al. 2024).
  - **BLEU**, **ROUGE-L** et **METEOR** : métriques classiques de similarité textuelle, adaptées pour mesurer le recouvrement lexical entre génération et référence (LIN 2004; PAPINENI et al. 2002).
  - **BERTScore** : mesure sémantique basée sur des représentations contextualisées, plus robuste aux paraphrases (T. ZHANG et al. 2019).
- **Évaluation humaine** : Enfin, une analyse qualitative par des chercheurs peut compléter les métriques automatiques, selon trois dimensions principales :
  - *Fidélité* : la réponse respecte-t-elle fidèlement les documents récupérés ?
  - *Pertinence* : la réponse correspond-elle réellement à la question posée ?
  - *Lisibilité* : la réponse est-elle compréhensible et bien formulée ?

Dans le cadre de ce stage, **cette évaluation n'a pas pu être réalisée faute de temps et de ressources**, mais la stratégie a été formalisée de manière à pouvoir être appliquée dans un travail futur.

8. Déploiement

Enfin, le pipeline a été encapsulé dans une API et intégré au sein d'une application permettant de consulter et d'interroger le corpus de manière interactive. Pour la génération des réponses, un modèle local `llama3.2:1b` a été mobilisé, garantissant la production de formulations naturelles et adaptées aux requêtes des utilisateurs, tout en assurant une exécution maîtrisée sur des ressources limitées.

L'application a été développée sous `Streamlit`, avec l'ajout de composants personnalisés en `CSS/HTML` afin de proposer une interface ergonomique et visuellement attractive. Ce déploiement illustre la capacité du projet à dépasser le stade expérimental pour fournir un outil opérationnel, combinant rigueur scientifique (pipeline reproductible, base vectorielle) et accessibilité pratique (interaction temps réel, visualisation des résultats).

V. Conclusion

Ce stage s'inscrivait dans le cadre du projet OBSERVANCE, dont l'objectif est d'analyser la prise en compte des recommandations de l'Autorité environnementale (Ae) dans les projets soumis à évaluation environnementale. Plus précisément, il s'agissait de concevoir et d'expérimenter un pipeline automatisé capable d'extraire, de traiter et d'analyser un corpus hétérogène de documents administratifs (avis de l'Ae et mémoires en réponse des maîtres d'ouvrage), afin d'apporter aux chercheurs et décideurs un outil facilitant l'accès à l'information.

Les travaux menés ont permis de développer une architecture modulaire en huit étapes, intégrant des modules d'ingestion, de conversion PDF, de nettoyage, d'extraction ciblée, de classification, de structuration en base de données, d'indexation vectorielle et de déploiement via une application RAG. Les expérimentations ont mis en évidence plusieurs résultats encourageants :

— des performances quasi-parfaites pour l'extraction de métadonnées simples (dates, pagination) et pour la détection de sections clés et de recommandations grâce à des règles légères (regex) ;
— une faisabilité démontrée pour la classification du niveau critique des avis, notamment avec GPT-4, même si les modèles locaux testés se sont révélés instables ou peu performants ;
— une première application web interactive, construite avec Streamlit et intégrant un modèle LLM local (`llama3.2:1b`), offrant un prototype fonctionnel pour la consultation et l'interrogation des données.

Ces résultats valident la faisabilité technique du pipeline, tout en soulignant certaines limites, notamment l'identification automatique et fiable des signataires ainsi que la dépendance aux ressources computationnelles pour les modèles locaux.

Si les expérimentations menées sur un échantillon réduit de 10 projets ont donné des résultats satisfaisants, le passage à une mise en œuvre à grande échelle (plus de 1200 projets) a révélé de nouveaux défis. En particulier, le contrôle de la qualité des données d'entrée s'avère problématique : la vérification manuelle de l'ensemble des documents est difficile, voire impossible, et les défauts de qualité ne sont souvent détectés qu'au moment où les modèles échouent à traiter certains fichiers. Cette expérience souligne l'importance de définir des formats normalisés pour les documents administratifs (avis, mémoires en réponse, etc.) et de veiller à la complétude des métadonnées publiées en ligne. De telles améliorations sont indispensables pour permettre une exploitation et une analyse automatiques plus fluides et

plus fiables.

Au-delà du cadre du stage, ce travail illustre la capacité des approches de traitement automatique du langage (TAL) à traiter des corpus réglementaires complexes et ouvre la voie à des analyses systématiques à grande échelle de la gouvernance environnementale.

Des perspectives de développement sont à envisager. À court terme, l'amélioration des modules d'extraction complexes (notamment l'identification des signataires) et la stabilisation de la classification via des modèles open-source plus performants constituent des priorités. À moyen terme, le déploiement et l'évaluation du système RAG par des utilisateurs finaux permettront de valider son utilité et sa pertinence. Enfin, à plus long terme, une extension du pipeline à d'autres corpus réglementaires ou à des phases ultérieures de la procédure (enquêtes publiques, décisions préfectorales) pourrait enrichir l'analyse et offrir un outil encore plus complet pour la recherche et l'action publique.

En somme, ce stage a permis de poser les bases d'une analyse automatisée robuste et reproductible de la prise en compte des enjeux environnementaux dans les processus décisionnels, tout en ouvrant des perspectives prometteuses pour des recherches futures et des applications concrètes dans le champ de la gouvernance environnementale.

Bibliographie

CHITICARIU, Laura, Yunyao LI et Frederick R. REISS (oct. 2013). “Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!” In : *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Seattle, Washington, USA : Association for Computational Linguistics, p. 827-832. URL : https://aclanthology.org/D13-1079/.

CODD, Edgar F. (1970). “A Relational Model of Data for Large Shared Data Banks”. In : *Communications of the ACM* 13.6, p. 377-387. DOI : 10.1145/362384.362685. URL : https://dl.acm.org/doi/10.1145/362384.362685.

ECHAVARRÍA PELÁEZ, Andrés Felipe et Mathilde LE MEUR (2025). *CER et WER : métriques d'évaluation de modèles de reconnaissance automatique de manuscrits et d'imprimés anciens*. Rapp. tech. Rapport technique, GT2 « Acquisition de données et transcription assistée par ordinateur », UAR 3565 - CNRS Université Grenoble Alpes. Consortium-HN ARIANE - Analyses, Recherches, Intelligence Artificielle et Nouvelles Éditions numériques. URL : https://hal.science/hal-05267873v1.

ELMASRI, Ramez et Shamkant B. NAVATHE (2015). *Fundamentals of Database Systems*. 7e éd. Pearson. ISBN : 9780133970777. URL : https://www.pearson.com/en-us/subject-catalog/p/fundamentals-of-database-systems/P200000003546/9780137502523.

GAO, Yifan et al. (2024). “Retrieval-Augmented Generation for Large Language Models : A Survey”. In : *arXiv preprint arXiv :2312.10997*. URL : https://arxiv.org/abs/2312.10997.

GUU, Kelvin et al. (2020). “REALM : Retrieval-Augmented Language Model Pre-Training”. In : *Proceedings of the 37th International Conference on Machine Learning (ICML)*. T. 119. Proceedings of Machine Learning Research. PMLR, p. 3929-3938. URL : https://proceedings.mlr.press/v119/guu20a/guu20a.pdf.

HIRSCHBERG, Daniel S. (1977). “Algorithms for the Longest Common Subsequence Problem”. In : *Journal of the ACM* 24.4. Presents a linear-space algorithm for computing the Longest

Common Subsequence (LCS), p. 664-675. DOI : 10.1145/322033.322044. URL : https://dl.acm.org/doi/10.1145/322033.322044.

KARPUKHIN, Vladimir et al. (2020). “Dense Passage Retrieval for Open-Domain Question Answering”. In : *Proceedings of EMNLP 2020*. Online : Association for Computational Linguistics, p. 6769-6781. URL : https://aclanthology.org/2020.emnlp-main.550/.

LEVENSHTEIN, Vladimir I. (1966). “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals”. In : *Soviet Physics Doklady* 10.8. Introduces the edit distance, now commonly called Levenshtein distance, p. 707-710. URL : https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.

LEWIS, Patrick et al. (2020). “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In : *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, p. 9459-9474. arXiv : 2005.11401 [cs.CL]. URL : https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

LIN, Chin-Yew (2004). “ROUGE : A Package for Automatic Evaluation of Summaries”. In : *Text Summarization Branches Out (ACL Workshop)*. Barcelona, Spain : Association for Computational Linguistics, p. 74-81. URL : https://aclanthology.org/W04-1013/.

LIU, Nelson F., Ruslan SALAKHUTDINOV et Percy LIANG (2023). “Lost in the Middle : How Language Models Use Long Contexts”. In : *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. DOI : 10.48550/arXiv.2307.03172.

MICROSOFT AZURE ARCHITECTURE CENTER (2025). *RAG Chunking Phase*. “Fixed-size parsing, with overlap … allows for some overlap …”. URL : https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/rag/rag-chunking-phase (visité le 03/10/2025).

NEUDECKER, Clemens (mai 2016). “An Open Corpus for Named Entity Recognition in Historic Newspapers”. In : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia : European Language Resources Association (ELRA), p. 4348-4352. URL : https://aclanthology.org/L16-1689.

PAPADAKIS, Thanasis et al. (2024). “Explainable and transparent artificial intelligence for public policymaking”. In : *Data & Policy* 6, e10. DOI : 10.1017/dap.2024.3. URL : https://doi.org/10.1017/dap.2024.3.

PAPINENI, Kishore et al. (2002). “BLEU : a Method for Automatic Evaluation of Machine Translation”. In : *Proceedings of ACL 2002*. Association for Computational Linguistics, p. 311-318. DOI : 10.3115/1073083.1073135. URL : https://dl.acm.org/doi/10.3115/1073083.1073135.

PLANAS, Jordi et al. (2022). “Beyond modeling : NLP Pipeline for efficient environmental policy analysis”. In : *arXiv preprint arXiv :2201.07105*. Accepted at Fragile Earth workshop proceedings at KDD 2021. DOI : 10.48550/arXiv.2201.07105. URL : https://doi.org/10.48550/arXiv.2201.07105.

REIMERS, Nils et Iryna GUREVYCH (2019). “Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks”. In : *arXiv preprint arXiv :1908.10084*. URL : https://arxiv.org/abs/1908.10084.

RESHMA, V. M. et Linda Sara MATHEW (2015). “Longest Common Subsequence : A Method for Automatic Evaluation of Handwritten Essays”. In : *IOSR Journal of Computer Engineering (IOSR-JCE)* 17.6, Ver. IV. Application of LCS for automatic essay evaluation, p. 01-07. DOI : 10.9790/0661-17640107. URL : https://www.iosrjournals.org/iosr-jce/papers/Vol17-issue6/Version-4/A017640107.pdf.

SQLITE, Development Team (n.d.). *FTS5 — Full-Text Search Extension for SQLite*. https://www.sqlite.org/fts5.html. Consulté le 3 octobre 2025.

STAUDINGER, Moritz et al. (2024). “A Reproducibility and Generalizability Study of Large Language Models for Query Generation”. In : *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP '24)*. Tokyo, Japan : ACM. ISBN : 979-8-4007-0724-7/24/12. DOI : 10.1145/3673791.3698432. URL : https://doi.org/10.1145/3673791.3698432.

VOORHEES, Ellen M. et Dawn M. TICE (2000). “The TREC-8 Question Answering Track Report”. In : *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. NIST. URL : https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=151495.

WANG, Liang et al. (2024). “Multilingual E5 Text Embeddings : A Technical Report”. In : *arXiv preprint arXiv :2402.05672*. URL : https://arxiv.org/abs/2402.05672.

WIMALASURIYA, Daya C. et Dejing DOU (2010). “Ontology-based information extraction : An introduction and a survey of current approaches”. In : *Journal of Information Science* 36.3, p. 306-323. DOI : 10.1177/0165551509360123. URL : https://journals.sagepub.com/doi/10.1177/0165551509360123.

YU, Hao et al. (2024). *Evaluation of Retrieval-Augmented Generation : A Survey*. arXiv : 2405.07437 [cs.CL]. URL : https://arxiv.org/abs/2405.07437.

YUAN, Jiayi et al. (2025). *Give Me FP32 or Give Me Death? Challenges and Solutions for Reproducible Reasoning*. arXiv : 2506.09501 [cs.CL]. URL : https://arxiv.org/abs/2506.09501.

ZHANG, Junyuan et al. (2024). *OCR Hinders RAG : Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation (OHRBench)*. Utilise LCS pour la phase récupération

et EM/F1 pour la génération. arXiv : 2412.02592 [cs.IR]. URL : https://arxiv.org/abs/2412.02592.

ZHANG, Tianyi et al. (2019). “BERTScore : Evaluating Text Generation with BERT”. In : *arXiv preprint arXiv :1904.09675*. URL : https://arxiv.org/abs/1904.09675.

Annexes

Nature des documents

Le corpus étudié dans le cadre du projet est donc constitué principalement de fichiers PDF qui incluent :

— **Les avis de l'Autorité environnementale (Ae)**, souvent générés de manière standardisée.


MRAe
Mission régionale d'autorité environnementale
Région Hauts-de-France
Avis délibéré de la mission régionale
d'autorité environnementale
Hauts-de-France
sur le projet de création d'une plate-forme logistique par la
société
PRD sur les communes de ILLIES et SALOMÉ (59)
n°MRAe 2018-2404

FIGURE 11 – Avis MRAe

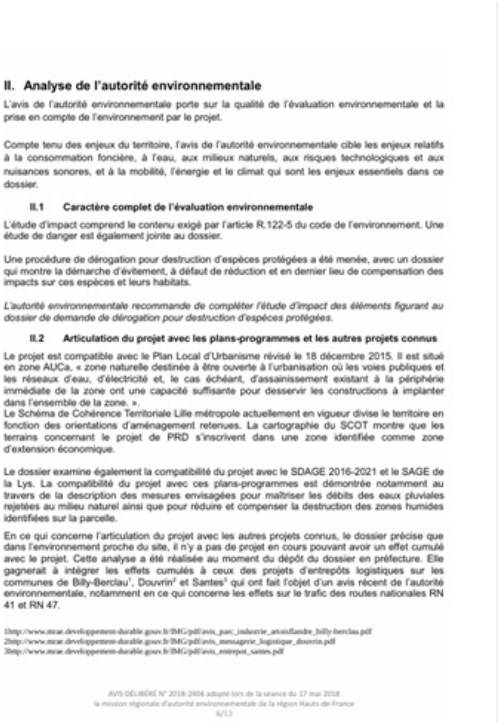
II. Analyse de l'autorité environnementale
L'avis de l'autorité environnementale porte sur la qualité de l'évaluation environnementale et la prise en compte de l'environnement par le projet.
Compte tenu des enjeux du territoire, l'avis de l'autorité environnementale cible les enjeux relatifs à la consommation foncière, à l'eau, aux milieux naturels, aux risques technologiques et aux nuisances sonores, et à la mobilité, l'énergie et le climat qui sont les enjeux essentiels dans ce dossier.
II.1 Caractère complet de l'évaluation environnementale
L'étude d'impact comprend le contenu exigé par l'article R.122-5 du code de l'environnement. Une étude de danger est également jointe au dossier.
Une procédure de dérogation pour destruction d'espèces protégées a été menée, avec un dossier qui montre la démarche d'évitement, à défaut de réduction et en dernier lieu de compensation des impacts sur ces espèces et leurs habitats.
L'autorité environnementale recommande de compléter l'étude d'impact des éléments figurant au dossier de demande de dérogation pour destruction d'espèces protégées.
II.2 Articulation du projet avec les plans-programmes et les autres projets connus
Le projet est compatible avec le Plan Local d'Urbanisme révisé le 18 décembre 2015. Il est situé en zone AUCa, « zone naturelle destinée à être ouverte à l'urbanisation où les voies publiques et les réseaux d'eau, d'électricité et, le cas échéant, d'assainissement existant à la périphérie immédiate de la zone ont une capacité suffisante pour desservir les constructions à implanter dans l'ensemble de la zone. ».
Le Schéma de Cohérence Territoriale Lille métropole actuellement en vigueur divise le territoire en fonction des orientations d'aménagement retenues. La cartographie du SCOT montre que les terrains concernant le projet de PRD s'inscrivent dans une zone identifiée comme zone d'extension économique.
Le dossier examine également la compatibilité du projet avec le SDAGE 2016-2021 et le SAGE de la Lys. La compatibilité du projet avec ces plans-programmes est démontrée notamment au travers de la description des mesures envisagées pour maîtriser les débits des eaux pluviales rejetées au milieu naturel ainsi que pour réduire et compenser la destruction des zones humides identifiées sur la parcelle.
En ce qui concerne l'articulation du projet avec les autres projets connus, le dossier précise que dans l'environnement proche du site, il n'y a pas de projet en cours pouvant avoir un effet cumulé avec le projet. Cette analyse a été réalisée au moment du dépôt du dossier en préfecture. Elle gagnerait à intégrer les effets cumulés à ceux des projets d'entrepôts logistiques sur les communes de Billy-Berclau[1], Douvrin[2] et Santes[3] qui ont fait l'objet d'un avis récent de l'autorité environnementale, notamment en ce qui concerne les effets sur le trafic des routes nationales RN 41 et RN 47.
1http://www.mrae.developpement-durable.gouv.fr/IMG/pdf/avis_parc_industrie_artois flandre_billy-berclau.pdf
2http://www.mrae.developpement-durable.gouv.fr/IMG/pdf/avis_messagerie_logistique_douvrin.pdf
3http://www.mrae.developpement-durable.gouv.fr/IMG/pdf/avis_entrepot_santes.pdf
AVIS DÉLIBÉRÉ N° 2018-2404 adopté lors de la séance du 17 mai 2018
la mission régionale d'autorité environnementale de la région Hauts-de-France
6/13

FIGURE 12 – Exemple d'une recommendation

— **Les mémoires en réponse** des maîtres d'ouvrage, beaucoup plus variables en format et en structure.

FIGURE 13 – Exemple de réponse du maître d'ouvrage

Ces documents présentent une grande hétérogénéité de nature et de format, pouvant être classés en trois grandes catégories :

PDF textuels Certains documents sont entièrement encodés en texte, permettant une extraction directe via des bibliothèques classiques (PyMuPDF, etc.). Ces fichiers sont les plus faciles à exploiter automatiquement.

PDF hybrides D'autres documents, plus fréquents, sont partiellement encodés : une partie du contenu est bien encodée en texte, tandis qu'une autre partie est composée d'éléments scannés ou d'images contenant du texte (par exemple, figures, tableaux, extraits de recommandation).

PDF image-only (scannés) Enfin, certains documents sont intégralement scannés, donc ne contiennent aucun texte encodé. L'accès à leur contenu repose uniquement sur des techniques de reconnaissance optique de caractères (OCR).

Résultats des mesures

TABLE 7 – Résultats des mesures "avis"

file	mode	levenshtein_char	char_accuracy	cer	word_precision	word_recall	word_f1	wer	lcs_ratio	exact_match
36_avis.txt	avis_plumber	1123	0.965334	0.034666	0.896256	0.895543	0.895900	0.105850	0.894150	0
10_avis.txt	avis_plumber	257760	0.000000	1.000000	0.005219	0.000852	0.001465	0.998467	0.001533	0
39_avis.txt	avis_plumber	890	0.959003	0.040997	0.943710	0.933942	0.938800	0.075190	0.924810	0
11_avis.txt	avis_plumber	316	0.986148	0.013852	0.950886	0.959249	0.955050	0.044562	0.955827	0
7_avis.txt	avis_plumber	380	0.983142	0.016858	0.962772	0.957289	0.960023	0.053246	0.946754	0
21_avis.txt	avis_plumber	8143	0.938767	0.061233	0.956214	0.913535	0.934388	0.093670	0.906330	0
8_avis.txt	avis_plumber	1298	0.972920	0.027080	0.958595	0.956113	0.957352	0.057789	0.942211	0
2_avis.txt	avis_plumber	814	0.969253	0.030747	0.957063	0.954269	0.955664	0.058380	0.941620	0
1_avis.txt	avis_plumber	376	0.988752	0.011248	0.972335	0.984525	0.978392	0.023310	0.976978	0
9_avis.txt	avis_plumber	380	0.984317	0.015683	0.959641	0.969100	0.964347	0.046617	0.953838	0
36_avis.txt	avis_ocr	828	0.974441	0.025559	0.913460	0.930362	0.921833	0.085953	0.915609	0
10_avis.txt	avis_ocr	2308	0.938812	0.061188	0.898319	0.937649	0.917563	0.114991	0.889832	0
39_avis.txt	avis_ocr	405	0.981344	0.018656	0.968563	0.984779	0.976604	0.034094	0.966467	0
11_avis.txt	avis_ocr	395	0.982685	0.017315	0.969897	0.991791	0.980722	0.031369	0.969323	0
7_avis.txt	avis_ocr	363	0.983896	0.016104	0.971823	0.982062	0.976915	0.032460	0.967878	0
21_avis.txt	avis_ocr	5813	0.956288	0.043712	0.951072	0.959019	0.955029	0.054791	0.945663	0
8_avis.txt	avis_ocr	1041	0.978282	0.021718	0.964180	0.979556	0.971807	0.039935	0.960692	0
2_avis.txt	avis_ocr	547	0.979338	0.020662	0.970885	0.981513	0.976170	0.034541	0.965833	0
1_avis.txt	avis_ocr	560	0.983248	0.016752	0.944701	0.960431	0.952501	0.056024	0.944894	0
9_avis.txt	avis_ocr	756	0.968799	0.031201	0.939640	0.986947	0.962713	0.065530	0.937611	0

TABLE 8 – Résultats des mesures "réponse"

file	mode	levenshtein_char	char_accuracy	cer	word_precision	word_recall	word_f1	wer	lcs_ratio	exact_match
9_reponse.txt	reponse_plumber	570	0.948248	0.051752	0.895967	0.901235	0.898593	0.108760	0.891876	0
36_reponse.txt	reponse_plumber	23302	0.813462	0.186538	0.889971	0.777293	0.829825	0.232254	0.767746	0
2_reponse.txt	reponse_plumber	6176	0.019838	0.980162	0.629630	0.017120	0.033333	0.985901	0.014099	0
1_reponse.txt	reponse_plumber	15539	0.005568	0.994432	0.125000	0.001233	0.002441	0.998767	0.001233	0
39_reponse.txt	reponse_plumber	172	0.972249	0.027751	0.929051	0.961310	0.944905	0.080357	0.922339	0
8_reponse.txt	reponse_plumber	117708	0.265147	0.734853	0.897071	0.240318	0.379083	0.762374	0.237626	0
21_reponse.txt	reponse_plumber	4328	0.970694	0.029306	0.914088	0.912431	0.913259	0.092632	0.907368	0
7_reponse.txt	reponse_plumber	6586	0.903923	0.096077	0.882281	0.892596	0.887408	0.164148	0.837749	0
11_reponse.txt	reponse_plumber	9508	0.007930	0.992070	0.333333	0.005453	0.010731	0.994547	0.005453	0
10_reponse.txt	reponse_plumber	197	0.967967	0.032033	0.899809	0.917315	0.908478	0.102140	0.899809	0
9_reponse.txt	reponse_ocr	170	0.984565	0.015435	0.927452	0.939447	0.933411	0.072310	0.928613	0
36_reponse.txt	reponse_ocr	5269	0.957820	0.042180	0.941442	0.951017	0.946205	0.074927	0.925827	0
2_reponse.txt	reponse_ocr	90	0.985717	0.014283	0.976167	0.989930	0.983000	0.028197	0.972195	0
1_reponse.txt	reponse_ocr	337	0.978433	0.021567	0.937853	0.954807	0.946254	0.066557	0.934625	0
39_reponse.txt	reponse_ocr	63	0.989835	0.010165	0.985337	1.000000	0.992614	0.014881	0.985337	0
8_reponse.txt	reponse_ocr	113413	0.291961	0.708039	1.000000	0.284621	0.443121	0.715379	0.284621	0
21_reponse.txt	reponse_ocr	6035	0.959135	0.040865	0.941287	0.950339	0.945792	0.061275	0.939309	0
7_reponse.txt	reponse_ocr	5638	0.917752	0.082248	0.912832	0.944587	0.928438	0.101987	0.901442	0
11_reponse.txt	reponse_ocr	505	0.947308	0.052692	0.953520	0.950920	0.952218	0.072938	0.927062	0
10_reponse.txt	reponse_ocr	287	0.953333	0.046667	0.937853	0.968872	0.953110	0.075875	0.926554	0

Prompts utilisés

```
Lis le texte suivant et classe le niveau de critique:
    1 - Peu critique (avis globalement positif, peu ou pas de
       remarques)
    2 - Mitigé (avis contenant à la fois des points positifs et
       des critiques modérées)
    3 - Assez critique à critique (plusieurs critiques importantes
       , étude jugée incomplète)
    4 - Très critique (avis très négatif, étude jugée très
       insuffisante ou absente)

Exemples:
    1 - "L'étude aborde de manière satisfaisante les thématiques.
       Pas de remarques majeures."
    2 - "L'avis note des aspects positifs mais aussi des réserves
       ou critiques modérées."
    3 - "L'étude est incomplète sur la biodiversité. Plusieurs
       impacts importants non analysés."
    4 - "Etude d'impact très insuffisante, thèmes majeurs non
       traités, impacts non étudiés."

Texte à évaluer :
{text}

Réponds uniquement par un chiffre (1, 2, 3 ou 4), sans explication
   .
```

```
Lis le texte suivant et classe le niveau de critique:
    1 - Peu critique (avis globalement positif, peu ou pas de
       remarques)
    2 - Mitigé (avis contenant à la fois des points positifs et
       des critiques modérées)
    3 - Assez critique à critique (plusieurs critiques importantes
       , étude jugée incomplète)
    4 - Très critique (avis très négatif, étude jugée très
       insuffisante ou absente)

Exemples:
    1 - "L'étude aborde de manière satisfaisante les thématiques.
       Pas de remarques majeures."
    2 - "L'avis note des aspects positifs mais aussi des réserves
       ou critiques modérées."
    3 - "L'étude est incomplète sur la biodiversité. Plusieurs
       impacts importants non analysés."
    4 - "Etude d'impact très insuffisante, thèmes majeurs non
       traités, impacts non étudiés."

Texte à évaluer :
{text}
Réponds uniquement par un chiffre (1, 2, 3 ou 4), sans explication
   .
```

```
Lis le texte suivant et classe le niveau de critique :
    1 - Peu critique (avis globalement positif, peu ou pas de
       remarques)
    2 - Mitigé (avis contenant à la fois des points positifs et
       des critiques modérées)
    3 - Assez critique à critique (plusieurs critiques importantes
       , étude jugée incomplète)
    4 - Très critique (avis très négatif, étude jugée très
       insuffisante ou absente)

Texte à évaluer :
{text}
Réponds uniquement par un chiffre (1, 2, 3 ou 4).
```

Résultats des mesures

TABLE 9 – Résultats de GPT-4 (3 runs) sur la synthèse et l'avis détaillé

Fichier	Réf.	Synthèse (3 runs)			Avis détaillé (3 runs)		
1_avis	2	3	3	3	3	3	3
2_avis	1	2	2	2	3	3	3
7_avis	4	4	4	4	3	3	3
8_avis	2	3	3	3	3	3	3
9_avis	2	3	3	3	3	3	3
10_avis	3	3	4	3	3	3	3
11_avis	3	3	3	3	3	3	3
21_avis	3	3	3	3	3	3	3
36_avis	3	3	3	3	3	3	3
39_avis	3	4	3	3	3	3	3

TABLE 10 – Résultats de LLaMA 3.2 (3 runs) sur la synthèse et l'avis détaillé

Fichier	Réf.	Synthèse (3 runs)			Avis détaillé (3 runs)		
1_avis	2	3	3	3	3	3	3
2_avis	1	3	3	3	4	4	2
7_avis	4	3	3	3	2	3	4
8_avis	2	3	3	3	3	4	3
9_avis	2	3	3	3	3	3	3
10_avis	3	3	3	3	3	2	3
11_avis	3	3	3	3	3	4	4
21_avis	3	3	3	3	3	3	4
36_avis	3	3	3	3	3	3	3
39_avis	3	3	3	3	2	4	3