**Université Sorbonne Paris Nord**

**Université Sidi Mohamed Ben Abdellah, Fès**

# CosmetoDataForge

*Système intelligent de collecte, extraction LLM, validation, RAG et indexation des données cosmétiques*

Master EID2 — Exploration des Données par Intelligence Artificielle et Décisionnel

**Réalisé par :**

RAZZOUG Salma
ER-ROUGBANI Mouaad
ZEGHARI Ayada
EL MAHDAOUI Ahmed
LAHMAMSSI Adnane

**Encadré par :**

Prof. Faouzi BOUFARES

Rapport de projet — Année académique 2025 / 2026

# Remerciements

Nous tenons à exprimer notre profonde gratitude à Monsieur Faouzi BOUFARES, pour sa disponibilité, ses conseils avisés et son expertise précieuse, qui nous ont guidés tout au long de la réalisation de ce projet de module sur les bases de données avancées et l'entrepôt de données.

Son regard critique et ses orientations méthodologiques nous ont permis de mieux structurer notre travail et d'approfondir notre compréhension des concepts étudiés.

Nous remercions également toutes les ressources et communautés consultées, qui ont enrichi notre réflexion et facilité la mise en œuvre technique de ce projet.

# Résumé

Ce rapport présente **CosmetoDataForge**, un système complet conçu pour centraliser, structurer et analyser des informations sur les produits cosmétiques et leurs ingrédients. Le projet repose sur la collecte automatique de données issues de sites web et de rapports PDF, suivie d'une transformation des informations par agent LLM vers des fichiers JSON normalisés. Un processus de validation et de nettoyage garantit la qualité et la cohérence des données avant leur stockage dans une base cloud (*Supabase*). Les données ainsi organisées sont ensuite exploitées pour effectuer une recherche sémantique via *Retrieval-Augmented Generation (RAG)* et pour alimenter un chatbot capable de répondre à des questions relatives à la sécurité des ingrédients. Le document détaille le contexte du projet, l'état de l'art, la méthodologie adoptée, l'architecture du système, les résultats expérimentaux obtenus, ainsi que la discussion et les perspectives d'évolution du projet.

# Abstract

**Abstract** — This report introduces *CosmetoDataForge*, a complete system to centralize, structure and analyze cosmetic product information. The project includes automatic data collection (websites, PDF reports), transformation by an LLM agent into normalized JSON files, validation and cleansing, and storage on a cloud database (Supabase). The indexed data then fuels a RAG semantic search and a chatbot for safety-related queries. The report details context, literature review, methodology, system architecture, experimental results, discussion and future work.

# Table des matières

# Chapitre 1

# Introduction générale

## 1.1 Contexte

Le marché des cosmétiques connaît une croissance rapide, portée par l'innovation et la diversification des produits, dont les formulations intègrent souvent de nombreux ingrédients complexes. Les consommateurs, de plus en plus exigeants, recherchent aujourd'hui une transparence totale sur la composition, les risques potentiels et les interactions entre substances. Pourtant, l'information fiable reste difficile d'accès, car elle est dispersée entre bases réglementaires, publications scientifiques, notices, rapports techniques et sources spécialisées. Cette fragmentation complique l'analyse, en particulier pour les non-spécialistes. D'où la nécessité de centraliser, structurer et simplifier l'accès aux données cosmétiques. Le projet s'inscrit précisément dans cette perspective en proposant une approche unifiée rendant l'information claire, cohérente et accessible à tous.

## 1.2 Problématique

Malgré le développement du secteur cosmétique, l'accès à l'information sur les ingrédients reste complexe. Les données sont dispersées sur de nombreuses sources (bases réglementaires, publications scientifiques, fiches fournisseurs, blogs, etc.) et rédigées dans un langage technique difficile à comprendre pour les non-experts. Cette fragmentation, associée à des informations parfois contradictoires ou partielles, complique l'évaluation des risques et la prise de décision éclairée. Les outils numériques existants n'offrent qu'une réponse partielle, souvent limitée par leur couverture, leur mise à jour ou leur lisibilité.

La problématique centrale est donc : **Comment fournir aux utilisateurs un accès unifié, intelligent, fiable et compréhensible à l'information cosmétique**

**dispersée, tout en garantissant la cohérence, la pertinence et la qualité des données ?**

Cette problématique constitue le moteur même du projet CosmetoDataForge, qui vise à dépasser les limites actuelles des plateformes existantes en s'appuyant sur une architecture avancée de collecte, d'analyse et d'intelligence artificielle.

## 1.3 Objectifs

L'objectif principal de ce travail est de concevoir, développer et valider un système intelligent capable d'automatiser la collecte, la structuration et l'analyse des données cosmétiques, afin de renforcer la transparence des formulations et d'améliorer la sécurité des utilisateurs. Le projet vise à transformer un ensemble de données dispersées et hétérogènes en une base de connaissances unifiée, cohérente et directement exploitable.

Afin d'atteindre cet objectif global, plusieurs objectifs secondaires ont été définis :

### Objectifs secondaires

- Centraliser des données provenant de sources variées, telles que les bases réglementaires, les publications scientifiques, les fiches techniques et les sites spécialisés.
- Générer des fiches structurées, au format JSON normalisé, pour chaque ingrédient et chaque produit, facilitant leur exploitation par les systèmes et les utilisateurs.
- Détecter automatiquement les incompatibilités intra-produit, notamment les risques chimiques, les allergènes, les interactions ou les restrictions réglementaires.
- Fournir une API ainsi qu'un moteur de recherche sémantique basé sur l'approche RAG, permettant d'interroger la base de connaissances de manière intelligente, contextualisée et précise.

Ces objectifs constituent la base fonctionnelle du système et orientent l'ensemble des choix méthodologiques et technologiques présentés dans ce rapport.

## 1.4 Méthodologie globale

La méthodologie adoptée dans ce projet suit une approche progressive et structurée, permettant de transformer des données cosmétiques brutes et dispersées en une base de connaissances organisée, analysable et exploitable par un système intelligent.

Elle s'articule autour de quatre étapes principales :

## 1. Collecte des données

La première étape consiste à rassembler les informations provenant de différentes sources : bases de données réglementaires (CosIng, FDA), sites d'analyse d'ingrédients (INCI Decoder, PubChem), publications scientifiques et documents techniques. Cette phase permet d'obtenir un ensemble initial large, mais non structuré.

## 2. Prétraitement et structuration

Les données collectées sont ensuite nettoyées, harmonisées et transformées en un format uniforme. Un agent basé sur un modèle LLM est utilisé pour analyser les textes, extraire les informations pertinentes et générer des fiches structurées au format JSON pour chaque ingrédient et chaque produit.

## 3. Analyse intelligente

Dans cette étape, le système applique des règles et modèles sémantiques afin de détecter les interactions, les incompatibilités, les risques potentiels et les incohérences entre ingrédients. Les données consolidées sont ensuite indexées dans un moteur de recherche RAG, permettant une exploration intelligente à travers des requêtes sémantiques.

## 4. Stockage et exposition des données

Enfin, l'ensemble des informations traitées est organisé dans une base de données cloud (Supabase). Une API est fournie pour permettre l'accès aux données structurées, tandis qu'une interface ou un chatbot peut interagir avec l'utilisateur afin de fournir des réponses contextualisées et enrichies.

Cette méthodologie garantit un flux complet allant de la collecte brute à l'analyse avancée, tout en assurant la cohérence, la qualité et la traçabilité des données.

# Chapitre 2

# Contexte et état de l'art

## 2.1 Les cosmétiques : définitions et catégories

### 2.1.1 Définition des cosmétiques

Les cosmétiques regroupent tous les produits destinés à être appliqués sur les parties superficielles du corps humain, y compris la peau, les cheveux, les ongles, les lèvres ou les dents, dans le but de :

— Nettoyer, protéger ou maintenir en bon état,

— Modifier l'apparence ou embellir,

— Parfumer ou apporter un effet sensoriel agréable.

Selon le **Règlement Européen (CE) 1223/2009**, un produit cosmétique se distingue clairement des médicaments et des dispositifs médicaux, car il ne possède pas d'action pharmacologique, immunologique ou métabolique.

### 2.1.2 Catégories de cosmétiques

Les cosmétiques sont classés en plusieurs catégories principales selon leur usage et leur fonction :

#### Produits d'hygiène

**Définition :** Produits destinés au nettoyage et à l'entretien quotidien du corps.
**Exemples :** gels douche, savons, déodorants, dentifrices.

### Produits de soin

**Définition :** Produits utilisés pour protéger, réparer ou améliorer l'état de la peau et des cheveux.

**Exemples :** crèmes hydratantes, sérums, lotions, soins capillaires.

### Produits de beauté

**Définition :** Produits destinés à embellir et valoriser l'apparence esthétique.
**Exemples :** maquillage, rouges à lèvres, vernis, parfums.

### Produits capillaires

**Définition :** Produits spécifiquement formulés pour l'entretien, la protection et l'amélioration de la qualité des cheveux.
**Exemples :** shampoings, masques, huiles, traitements spécifiques.


### Produits solaires

**Définition :** Produits destinés à protéger la peau contre les rayons UV et prévenir le vieillissement cutané.

**Exemples :** crèmes solaires, sprays solaires, après-soleil.


### 2.1.3 Diversité et complexité des formulations

Chaque catégorie implique des formulations spécifiques, souvent complexes :

— Grande variété d'ingrédients actifs et excipients,

— Concentrations et interactions diverses,

— Besoin de conformité réglementaire stricte.

Cette diversité montre l'importance d'une **gestion rigoureuse des données et d'une expertise technique**, ce qui justifie l'utilisation d'un **système intelligent d'analyse cosmétique**.

## 2.2 Les enjeux associés aux cosmétiques

Le secteur des cosmétiques est confronté à plusieurs enjeux majeurs, qui influencent la formulation, la réglementation et l'usage des produits. Ces enjeux peuvent être regroupés en trois grandes catégories : sécurité sanitaire, réglementation et transparence.

### 2.2.1 Enjeux de sécurité sanitaire

**Définition :** Garantir que les produits cosmétiques peuvent être utilisés sans risque pour la santé des consommateurs.

Les principaux risques à prévenir sont :

— **Réactions allergiques :** causées par certains conservateurs, parfums ou colorants.

— **Irritations cutanées :** dues à des substances à forte concentration ou sensibilisantes.

— **Photosensibilisation :** provoquée par des ingrédients réagissant à la lumière.

— **Interactions indésirables :** entre plusieurs substances présentes dans une même formule.

### 2.2.2 Enjeux réglementaires

**Définition :** Respect des lois et normes encadrant la production, l'étiquetage et la commercialisation des produits cosmétiques.

Exemples et obligations :

— **Europe :** Règlement (CE) 1223/2009 et base CosIng pour la liste des ingrédients autorisés, restreints ou interdits.

— **États-Unis :** supervision par la FDA et bonnes pratiques de fabrication.

— **Normes générales :** traçabilité, évaluation de sécurité, conformité des concentrations et tests microbiologiques.

### 2.2.3 Enjeux de transparence et d'information

**Définition :** Permettre aux consommateurs et professionnels d'accéder à des informations fiables, claires et vérifiables sur les produits.

Points clés :

— Accès à des informations détaillées sur les ingrédients,

— Compréhension des risques et des interactions potentielles,

— Comparaison entre produits et assurance qualité,

— Applications et bases en ligne : INCI Beauty, Yuka, EWG.

Ces enjeux justifient la nécessité d'un **système intelligent** capable de centraliser, analyser et rendre accessibles toutes les informations liées aux cosmétiques.

## 2.3 Les ingrédients cosmétiques

Les ingrédients cosmétiques constituent la base de toute formulation. Leur rôle est multiple : apporter des propriétés fonctionnelles, sensorielles ou esthétiques, garantir la stabilité du produit et assurer sa sécurité pour l'utilisateur. Ils peuvent être classés selon leur fonction, leur composition chimique et leur impact potentiel sur la santé ou l'environnement.

### 2.3.1 Classification des ingrédients selon leur fonction

#### Agents nettoyants (tensioactifs)

**Définition :** Substances qui permettent d'éliminer les impuretés, l'excès de sébum et les résidus de maquillage tout en assurant une sensation de propreté.

**Exemples :** Sodium lauryl sulfate, coco-glucoside.

**Particularités :** Certains tensioactifs peuvent provoquer une irritation cutanée ou assécher la peau en cas d'utilisation répétée.

#### Conservateurs

**Définition :** Substances ajoutées pour prévenir la prolifération microbienne et prolonger la durée de vie du produit.

**Exemples :** Parabènes, phénoxyéthanol, sorbate de potassium.

**Risques :** Allergies cutanées, sensibilisation chronique, controverses sur perturbateurs endocriniens (ex : parabènes).

#### Actifs cosmétiques

**Définition :** Substances ayant une fonction spécifique bénéfique pour la peau ou les cheveux, comme hydratation, anti-âge ou protection solaire.

**Exemples :** Acide hyaluronique, rétinol, vitamine C, peptides, filtres UV.

**Particularités :** Les actifs peuvent présenter des sensibilités cutanées, une photosensibilisation ou des interactions chimiques si mal combinés (ex : rétinol + AHA/BHA).

#### Émollients et agents de texture

**Définition :** Substances qui adoucissent, assouplissent la peau et améliorent la sensation du produit lors de l'application.

**Exemples :** Glycérine, huiles végétales, silicones, beurre de karité.

**Particularités :** Peuvent influencer la pénétration des actifs et modifier la stabilité des formules.

#### Parfums et colorants

**Définition :** Substances ajoutées pour améliorer l'attrait esthétique et sensoriel du produit.

**Exemples :** Huiles essentielles, extraits aromatiques, colorants naturels ou synthétiques.

**Risques :** Allergies, sensibilisations, photosensibilisation ; certaines molécules peuvent être réglementées ou interdites selon les régions.

### 2.3.2 Risques et effets documentés

Certains ingrédients présentent des **risques connus** qui nécessitent une surveillance stricte :

— **Sensibilisation cutanée :** réactions allergiques aux parfums, conservateurs ou colorants.
— **Interactions chimiques :** incompatibilités entre ingrédients actifs (ex : rétinol et acides alpha-hydroxy ou hydroquinone).
— **Exposition cumulative :** utilisation simultanée de plusieurs produits contenant les mêmes ingrédients.
— **Effets systémiques :** certains perturbateurs endocriniens peuvent avoir un impact sur la santé à long terme.

### 2.3.3 Incompatibilités et interactions chimiques

Certaines combinaisons d'ingrédients doivent être évitées afin de garantir la sécurité et l'efficacité du produit :

— **Rétinol + AHA/BHA :** instabilité chimique et irritation cutanée.
— **Vitamine C + Niacinamide :** dégradation de la vitamine C et perte d'efficacité.
— **Benzoyl Peroxide + Hydroquinone :** oxydation dangereuse et diminution de l'efficacité.
— **Filtres UV incompatibles :** instabilité de certains filtres sous exposition solaire (ex : avobenzone + octocrylène sans stabilisant).

### 2.3.4 Conclusion sur les ingrédients

La diversité fonctionnelle et chimique des ingrédients cosmétiques montre l'importance d'une **gestion précise des données**, d'une **analyse des risques** et de **contrôles de compatibilité**. Ces besoins justifient pleinement la mise en place d'un **système intelligent** capable de centraliser, analyser et signaler les risques, interactions et informations réglementaires pour chaque ingrédient.

## 2.4 Outils et ressources existants

Dans le domaine cosmétique, l'information est largement dispersée entre différentes sources, chacune ayant ses forces et ses limites. Pour une analyse complète, il est essentiel de comprendre quels outils existent et comment ils sont utilisés. On peut les classer en quatre grandes catégories : bases réglementaires, bases scientifiques, bases grand public et outils de laboratoire.

### 2.4.1 Bases réglementaires

Ces bases sont indispensables pour vérifier la légalité et la sécurité des ingrédients ou produits cosmétiques. Elles sont généralement mises à jour par les autorités sanitaires ou européennes et servent de référence aux formulateurs et aux laboratoires.

- **CosIng (Commission Européenne)** : Base officielle européenne listant les ingrédients autorisés, restreints ou interdits dans les cosmétiques. Fournit aussi des informations sur les fonctions et les concentrations limites.
- **CPNP (Cosmetic Product Notification Portal)** : Plateforme où les fabricants notifient les produits mis sur le marché européen, permettant un suivi officiel et la traçabilité.
- **SCCS (Scientific Committee on Consumer Safety)** : Publie des avis scientifiques sur la sécurité des ingrédients, y compris les concentrations recommandées et les restrictions pour certains nanomatériaux ou perturbateurs endocriniens.
- **FDA (Food and Drug Administration, États-Unis)** : Fournit des informations sur les ingrédients approuvés, avec des restrictions moins strictes qu'en Europe.

Ces bases garantissent une conformité réglementaire mais ne fournissent pas toujours de manière complète les risques combinés ou interactions chimiques.

### 2.4.2 Bases scientifiques

Les bases scientifiques permettent d'accéder à des informations détaillées sur la chimie, la toxicologie et les propriétés biologiques des ingrédients.

- **PubChem** : Contient les propriétés chimiques, structures, toxicologie, activités biologiques et références bibliographiques.
- **DrugBank** : Fournit des informations sur les substances actives, interactions médicamenteuses et données pharmacologiques, utile pour vérifier d'éventuelles interférences.
- **ScienceDirect, Google Scholar** : Pour consulter des publications scientifiques, études cliniques, essais dermatologiques et recherches sur les nouveaux ingrédients ou technologies cosmétiques.

Ces sources offrent des informations riches mais nécessitent souvent un tri et une interprétation manuelle par un expert.

### 2.4.3 Bases grand public

Ces outils sont destinés aux consommateurs et vulgarisent les informations sur les ingrédients.

- **INCI Decoder** : Permet de décoder la liste INCI et d'expliquer les fonctions des ingrédients.
- **EWG SkinDeep** : Note les ingrédients selon leur potentiel de risque (allergies, toxicité, perturbateurs endocriniens).
- **Yuka, CosmEthics** : Applications mobiles qui scannent les produits et donnent des scores de sécurité ou des alertes sur certains ingrédients.

Bien qu'accessibles et faciles à utiliser, ces outils souvent simplifient les informations et ne prennent pas en compte les interactions complexes.

### 2.4.4 Limites des outils existants

Malgré l'abondance d'information, plusieurs limites persistent :

- **Dispersions et hétérogénéité** : Les données sont éparpillées entre réglementations, publications scientifiques et sites grand public.
- **Absence d'intégration automatique** : Il n'existe pas de système qui combine automatiquement données réglementaires, scientifiques et commerciales.
- **Manque d'analyse des interactions** : Aucun outil grand public ne détecte les incompatibilités chimiques ou les effets cocktails.
- **Mise à jour manuelle** : Les changements réglementaires ou scientifiques nécessitent souvent une veille humaine.

### 2.4.5 Justification d'un système intelligent

Ces limites montrent la nécessité d'un **système intelligent centralisé** capable de :

- Rassembler et normaliser toutes les sources de données.
- Fournir une analyse automatisée des risques et des incompatibilités.
- Générer des fiches structurées pour chaque ingrédient et produit.
- Assurer une veille réglementaire et scientifique en temps réel.

Un tel système répondrait aux besoins des formulateurs, toxicologues et consommateurs, en fournissant des informations fiables, précises et exploitables.

### 2.4.6 Méthodes d'analyse actuelles

L'analyse des données cosmétiques repose traditionnellement sur plusieurs approches :

— **Consultation manuelle** : lecture et interprétation d'avis scientifiques, de rapports toxicologiques et d'études cliniques pour évaluer la sécurité et l'efficacité des ingrédients.
— **Exploitation de bases structurées** : utilisation de sources telles que CosIng, PubChem ou DrugBank pour obtenir des informations réglementaires, chimiques et biologiques.
— **Systèmes de règles experts** : application de règles prédéfinies par des formulateurs ou toxicologues pour détecter les incompatibilités ou les limites de concentration.

Aujourd'hui, les **approches basées sur l'intelligence artificielle**, notamment le traitement automatique du langage (NLP) et les modèles de langage (LLM), permettent d'automatiser ces analyses. Elles offrent une **lecture rapide, contextualisée et exhaustive** des données dispersées, tout en identifiant les risques, interactions et incohérences de manière beaucoup plus efficace qu'une analyse manuelle.

# Chapitre 3

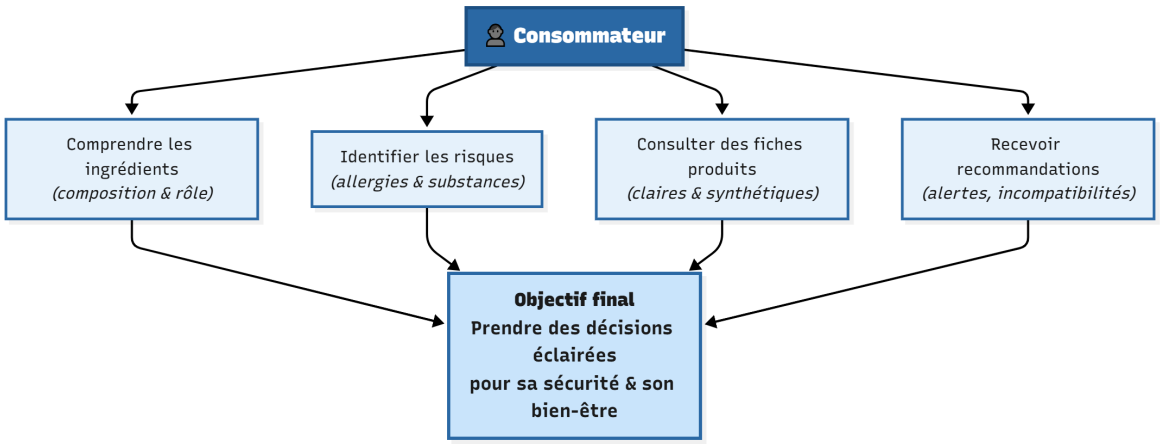
# Analyse des besoins

## 3.1 Acteurs et utilisateurs

Le système **CosmetoDataForge** s'adresse à plusieurs catégories d'utilisateurs, chacune ayant des besoins et attentes spécifiques. Comprendre ces acteurs est essentiel pour orienter le design fonctionnel et technique du système.

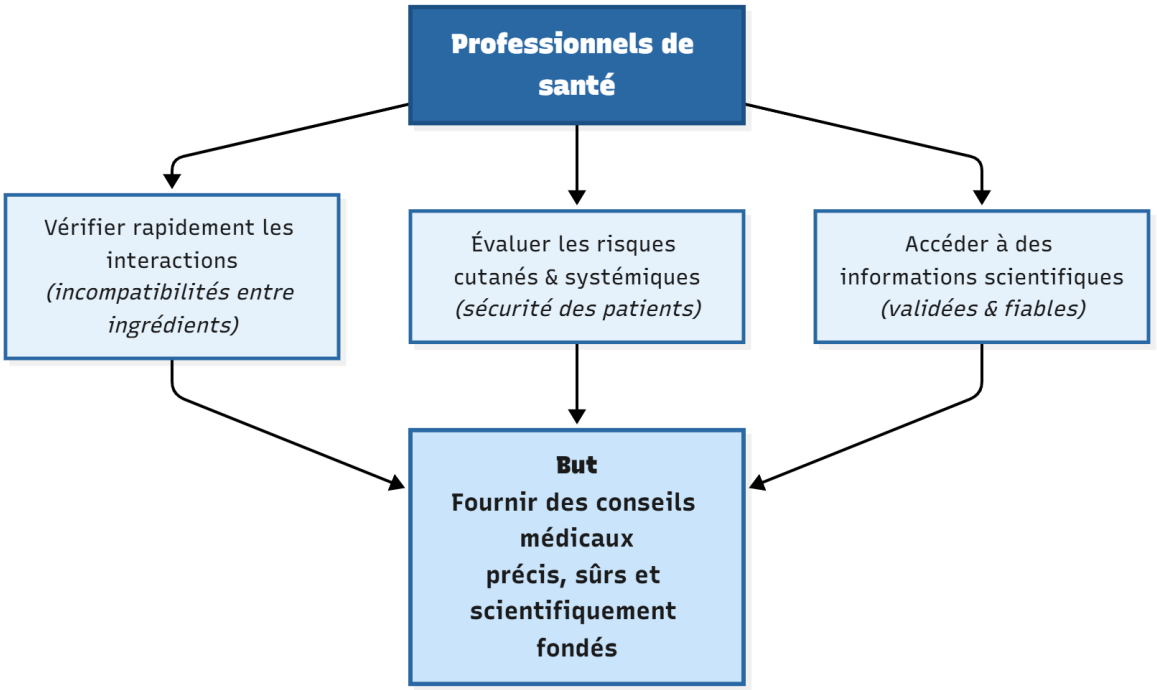
### 3.1.1 Consommateurs

Les consommateurs recherchent avant tout des informations claires, fiables et compréhensibles sur les produits qu'ils utilisent quotidiennement. Leurs attentes incluent :




L'objectif est de permettre à tout utilisateur, même non spécialiste, de prendre des décisions éclairées concernant sa sécurité et son bien-être.

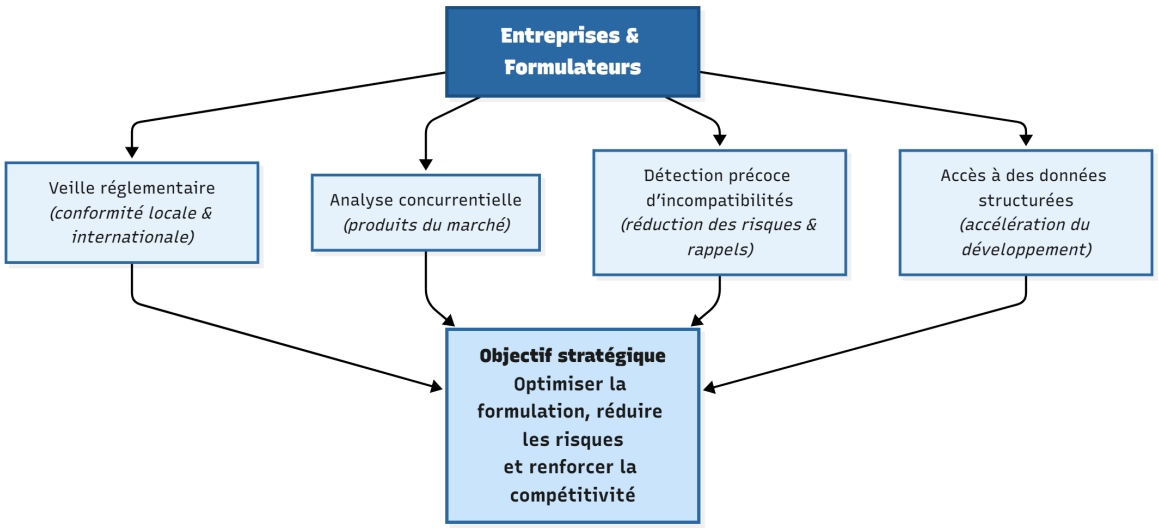
### 3.1.2 Professionnels de santé et dermatologues



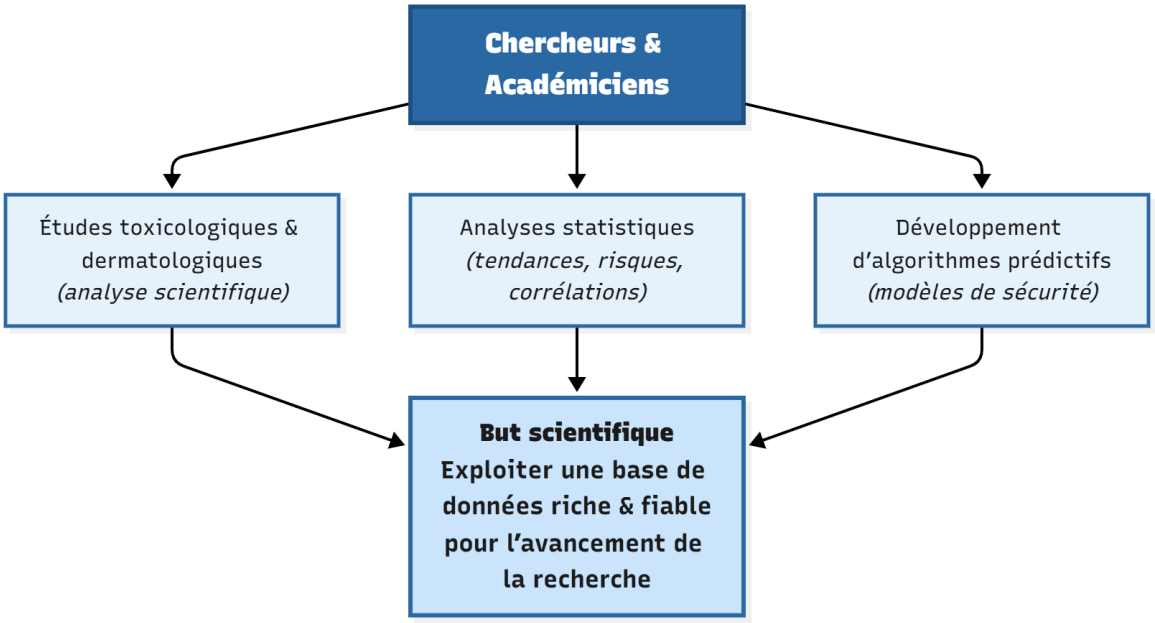

Le système doit ainsi fournir des données fiables, structurées et scientifiquement documentées.

### 3.1.3 Industrie cosmétique

Les entreprises et formulateurs ont des besoins plus techniques et stratégiques :



### 3.1.4 Chercheurs et académiciens




L'accès aux données brutes, aux métadonnées et aux liens scientifiques est crucial pour ce profil.

## 3.2 Fonctionnalités attendues

Pour répondre aux besoins des différents acteurs, **CosmetoDataForge** doit intégrer plusieurs fonctionnalités clés, articulées autour de la collecte, de la structuration, de l'analyse et de l'exploitation des données :

1. **Recherche de produits et d'ingrédients**
   - Recherche rapide par nom, catégorie ou propriété fonctionnelle.
   - Capacité à filtrer selon des critères tels que risques, compatibilité ou concentration.
2. **Fiches structurées ingrédient / produit**
   - Chaque ingrédient ou produit est documenté dans un format standardisé (JSON).
   - Informations incluses : propriétés chimiques, usages, risques connus, sources scientifiques, réglementation applicable.
3. **Détection d'incompatibilités intra-produit**
   - Analyse automatique des formulations pour identifier : réactions chimiques possibles, interactions allergènes, limitations réglementaires.
   - Alertes visuelles et synthétiques pour l'utilisateur.

4. **API d'interrogation et chatbot RAG**
   — Accès aux données via API pour intégration dans des systèmes tiers.
   — Chatbot intelligent capable de répondre à des questions en langage naturel, contextualisé et précis.
5. **Historique et traçabilité des sources**
   — Gestion des versions et suivi de l'évolution des informations.
   — Identification de l'origine de chaque donnée (base réglementaire, article scientifique, site public, etc.).
6. **Tableaux de bord et visualisation**
   — Pour l'industrie et les chercheurs, possibilité de visualiser rapidement les risques, tendances et statistiques.

## 3.3 Contraintes

Pour garantir la fiabilité et la pertinence du système, plusieurs contraintes doivent être respectées lors du développement et de l'exploitation :

1. **Fiabilité et qualité des données**
   — Validation automatique et manuelle des informations extraites.
   — Vérification croisée avec plusieurs sources pour réduire les erreurs.
2. **Mise à jour dynamique**
   — Capacité à détecter les changements dans les bases de données sources.
   — Rafraîchissement automatique des fiches ingrédients et produits.
3. **Transparence et traçabilité**
   — Historique des modifications pour chaque donnée.
   — Possibilité pour l'utilisateur de vérifier la source et la date de chaque information.
4. **Respect des droits et législation**
   — Conformité avec les conditions d'utilisation des sites sources et des publications scientifiques.
   — Protection des données personnelles si l'utilisateur crée un compte.
5. **Accessibilité et ergonomie**
   — Interface claire et intuitive pour tous les profils d'utilisateurs.
   — Compatibilité multiplateforme (web, mobile).

# Chapitre 4

# Architecture globale du projet

## 4.1 Vue d'ensemble

L'architecture de **CosmetoDataForge** est centrée sur la donnée et conçue pour gérer le flux complet depuis la collecte jusqu'à l'exploitation intelligente des informations cosmétiques. Elle se compose de quatre grandes couches :

1. **Acquisition** : Cette couche regroupe tous les processus de collecte de données. Les informations proviennent de sites web spécialisés (ex : INCI Decoder, CosIng, PubChem), de fiches produits PDF ou HTML, ainsi que de publications scientifiques et documents techniques. La collecte est automatisée via des outils de scraping pour gérer efficacement des milliers de produits et ingrédients.
2. **Transformation** : Après acquisition, les données brutes sont traitées par un **agent LLM** (Language Learning Model) qui analyse le texte, extrait les informations importantes et transforme chaque contenu en fiches JSON normalisées.
3. **Validation & Normalisation** : Cette étape applique des règles métiers pour vérifier la cohérence des données, effectue un nettoyage automatique pour éliminer les doublons ou données incorrectes, et contrôle la conformité des schémas JSON.
4. **Stockage et exploitation** : Les données structurées sont stockées dans une base cloud relationnelle (Supabase/PostgreSQL), indexées dans un moteur RAG pour des recherches sémantiques avancées et exposées via une API et un chatbot interactif.

### 4.1.1 Choix technologiques

Dans le cadre de **CosmetoDataForge**, chaque technologie a été sélectionnée en fonction de sa robustesse, sa maturité, et sa capacité à traiter efficacement des vo-

lumes importants de données hétérogènes. Les choix technologiques permettent de construire un pipeline complet allant de la collecte des données à l'analyse intelligente et à l'exploitation via API ou chatbot.

 **Python**

**Rôle dans le projet :**

Python constitue le langage principal du projet. Il est utilisé pour le traitement des données, la manipulation des fichiers JSON, l'intégration des modules de collecte et le développement de l'interface API.

**Avantages :**

— Écosystème riche de bibliothèques scientifiques et NLP (`pandas`, `NumPy`, `OpenAI`, `LangChain`).
— Grande flexibilité pour l'automatisation et l'orchestration de pipelines complexes.
— Facile à maintenir et à faire évoluer selon les besoins.
— Communauté très active, avec une documentation exhaustive.

**Utilisation concrète :**

Toutes les étapes de transformation, de nettoyage et de structuration des données sont codées en Python, incluant les scripts de prétraitement, d'analyse et de génération de JSON.

**Scrapy + Playwright**

**Rôle dans le projet :**

 **Scrapy** : récupération automatisée de données à partir de pages web statiques.

 **Playwright** : récupération de données sur les pages dynamiques générées par JavaScript, permettant de simuler un navigateur complet.

**Avantages :**

— Collecte massive et rapide de données à partir de sources multiples.
— Gestion de sites complexes avec authentification ou contenus dynamiques.
— Intégration native avec Python et possibilité de paralléliser les tâches.

**Utilisation concrète :**

Ces outils permettent de collecter toutes les fiches produits et ingrédients depuis des sites réglementaires, des bases de données scientifiques et des portails d'analyse cosmétique.

### Agent LLM (ex : Groq, GPT)

**Rôle dans le projet :**

**Agent LLM** : L'agent LLM est responsable de l'analyse sémantique et de la transformation des données brutes en JSON structuré. Il identifie :

- Les ingrédients et leurs fonctions.
- Les risques, interactions et incompatibilités.
- Les liens réglementaires et sources scientifiques.

**Avantages :**

- Compréhension du contexte et extraction intelligente de l'information.
- Normalisation automatique des données selon un schéma cohérent.
- Possibilité d'adapter le modèle à des cas spécifiques du domaine cosmétique.

**Utilisation concrète :**

L'agent transforme les textes récupérés par Scrapy/Playwright en fiches JSON, préparant les données pour la validation et le stockage dans Supabase.

### Supabase (PostgreSQL)

**Rôle dans le projet :**

**Supabase** : Supabase fournit le stockage cloud relationnel pour toutes les fiches JSON. Il garantit :

- Une structure organisée et sécurisée des données.
- Le versioning et la traçabilité des modifications.
- L'accès aux données via API pour des applications ou chatbots.

**Avantages :**

- Déploiement rapide et administration simplifiée.
- Intégration facile avec Python et les services cloud.
- Sécurité et contrôle d'accès natifs.

**Utilisation concrète :**

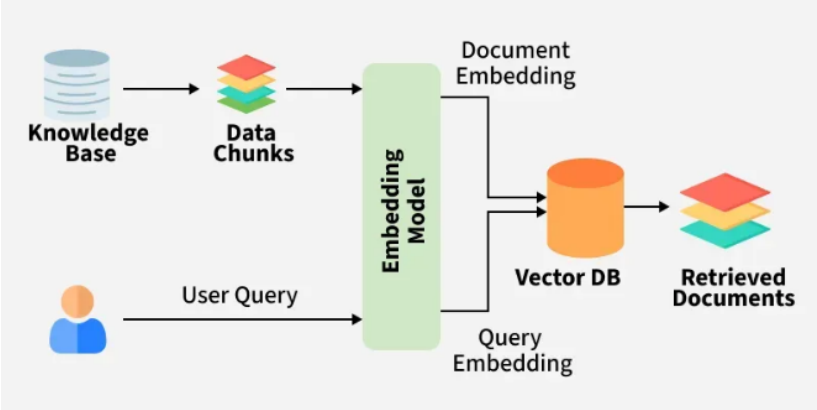
Toutes les fiches produits et ingrédients sont insérées et mises à jour automatiquement dans Supabase, prêtes à être interrogées par le moteur RAG.

### Système RAG (Retrieval-Augmented Generation)

**Rôle dans le projet :**

**L**e système RAG indexe les fiches JSON pour permettre une recherche sémantique avancée et contextualisée. Il est utilisé pour :

— Fournir des réponses intelligentes via chatbot ou API.
— Permettre une exploration rapide des données selon le contexte de la requête.
— Améliorer la pertinence des résultats grâce à l'intégration du contenu structuré et des métadonnées.




**Avantages :**
— Recherche contextuelle plus pertinente que les systèmes classiques.
— Exploitation efficace des données structurées et des métadonnées.
— Compatible avec différents modèles de langage et pipelines NLP.

### 4.1.2 Flux opérationnel et schéma conceptuel

Le pipeline opérationnel de **CosmetoDataForge** suit un processus structuré et séquentiel, permettant de transformer des données brutes et hétérogènes en informations structurées, exploitables et intelligentes. Ce flux garantit la cohérence, la qualité et la traçabilité des données à chaque étape.

#### Étapes du flux opérationnel

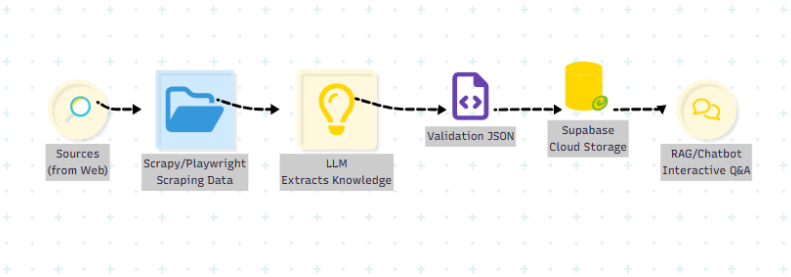
1. **Découverte des sources**
   Identification et sélection des sources pertinentes, incluant les sites spécialisés, bases réglementaires, publications scientifiques et documents PDF. Cette étape permet de définir un périmètre fiable et complet pour la collecte des données.
2. **Collecte des données**
   Récupération automatisée des informations via des outils de scraping adaptés aux pages statiques et dynamiques. Les données collectées incluent les fiches produits, listes d'ingrédients et documents techniques.
3. **Extraction et transformation**
   Les données brutes sont analysées par l'agent LLM, qui identifie les informations

pertinentes (ingrédients, fonctions, risques, compatibilités) et génère des fiches normalisées au format JSON.

4. **Validation et nettoyage**
   Contrôle de la structure JSON pour détecter les incohérences, doublons ou erreurs. Cette étape assure la fiabilité des informations avant leur stockage.
5. **Insertion en base**
   Les fiches JSON validées sont stockées dans la base Supabase, garantissant la traçabilité, le versioning et l'intégrité des données.
6. **Indexation RAG et exposition**
   Les données sont indexées dans le moteur RAG pour permettre une recherche sémantique intelligente. L'accès aux informations se fait via API ou chatbot, offrant des réponses contextualisées et fiables aux utilisateurs.

## 4.2 Schéma conceptuel

Le schéma conceptuel du pipeline peut être représenté ainsi :




Chaque étape représente un bloc fonctionnel distinct, où les données sont progressivement transformées, enrichies et rendues exploitables. Ce schéma illustre de manière synthétique la continuité du traitement depuis la collecte jusqu'à l'exploration intelligente.

# Chapitre 5

# Collecte des données

## 5.1 Sources d'information

La collecte des données repose sur un écosystème informationnel riche mais fragmenté. Aucune source ne couvre à elle seule toute la chaîne de valeur d'un produit cosmétique ; chaque plateforme apporte un point de vue complémentaire. L'objectif est donc de combiner des sources commerciales, scientifiques et réglementaires afin d'obtenir une vision globale, précise et fiable des produits et de leurs ingrédients.

### 5.1.1 Sites de distribution grand public

Les sites de distribution représentent la principale porte d'entrée vers les produits tels qu'ils sont réellement commercialisés. Ils fournissent des informations essentielles : listes d'ingrédients, descriptions commerciales, caractéristiques techniques, formats, catégories et parfois les revendications d'usage.

Les avantages de ces sources résident dans leur mise à jour fréquente, leur large couverture de marques et leur accessibilité. Les données sont proches de l'utilisateur final et reflètent le marché en temps réel. Cependant, ces plateformes présentent certaines limites : une partie du contenu est orientée marketing et les informations fournies manquent parfois de précision scientifique, notamment concernant les concentrations ou les interactions entre substances.

Dans le cadre du projet, ces sites jouent un rôle central. Ils permettent de récupérer les listes d'ingrédients, de constituer les fiches produits et d'initier l'enrichissement des données à travers les autres modules du pipeline.

**Exemple de sources utilisées :**

— Ulta Beauty : https://www.ulta.com

— INCI Decoder : https://incidecoder.com

— Sephora : https://www.sephora.com

### 5.1.2 Bases de données scientifiques

Les bases scientifiques apportent une vision technique et approfondie des ingrédients utilisés dans les cosmétiques. Elles fournissent des informations structurées sur la composition chimique, les propriétés physico-chimiques, les effets biologiques, la toxicité, les mécanismes d'action et les interactions possibles.

Le principal avantage de ces sources est la fiabilité : les données sont validées scientifiquement et décrites avec une terminologie précise. Elles permettent d'enrichir les fiches ingrédient avec des éléments techniques impossibles à obtenir via les sites commerciaux. Leur limite réside dans le niveau de complexité : le langage est souvent très technique et parfois orienté vers la pharmacologie, nécessitant une normalisation via des modèles NLP.

Dans le projet, ces bases servent à comprendre la fonction réelle des ingrédients, documenter leurs propriétés et identifier leurs compatibilités ou incompatibilités.

**Exemples de bases utilisées :**

— DrugBank : https://go.drugbank.com
— PubChem : https://pubchem.ncbi.nlm.nih.gov
— ChemSpider : http://www.chemspider.com

### 5.1.3 Sources réglementaires

Les plateformes réglementaires permettent de définir le cadre légal d'utilisation des substances dans les produits cosmétiques. Elles indiquent les restrictions d'usage, les limites de concentration, les substances autorisées, restreintes ou interdites, ainsi que les avis scientifiques officiels.

Leur avantage principal est la fiabilité juridique : les données proviennent d'organismes officiels et suivent des normes strictes. Cependant, certaines mises à jour peuvent être irrégulières et les descriptions restent généralement générales ou peu contextualisées.

Dans ce projet, ces sources sont essentielles pour vérifier la conformité des produits, enrichir les alertes réglementaires et générer des rapports fiables.

**Exemples de sources utilisées :**

— EU CosIng (Commission Européenne) : https://ec.europa.eu/growth/tools-databases/cosing
— FDA – Cosmetic Ingredient Hotlist : https://www.fda.gov
— SCCS – Scientific Committee on Consumer Safety : https://health.ec.europa.eu/scientific-committees_en

### 5.1.4 Stratégie de collecte multi-sources

La fusion de ces trois familles de sources permet de construire une vision complète :

— les sites commerçants apportent **le contenu produit**,

— les bases scientifiques apportent **l'explication technique**,

— les sources réglementaires apportent **la conformité légale**.

Par exemple, l'analyse d'un ingrédient comme le *Rétinol* combine : présence dans les produits, propriétés chimiques, restrictions réglementaires.

## 5.2 Nature des données collectées

Les données extraites par le pipeline couvrent trois grandes catégories, chacune apportant un éclairage spécifique sur les produits et leurs ingrédients. Cette diversité permet de construire des fiches complètes, d'alimenter les modules d'analyse et de garantir la conformité réglementaire.

**Données produits :** Elles concernent les informations directement liées aux produits tels qu'ils sont commercialisés. On retrouve notamment :

— Identification : nom commercial, marque, fabricant.

— Description : usage revendiqué, public cible, texture et parfum.

— Composition : liste complète d'ingrédients, souvent ordonnée selon les règles INCI.

— Caractéristiques : type de peau recommandé, format, indications et promesses du produit.

**Données ingrédients :** Ces informations permettent de comprendre le rôle de chaque substance dans la formulation et d'anticiper ses effets. Elles comprennent :

— Identité chimique : nom INCI, formule, famille chimique.

— Propriétés : rôle cosmétique (émollient, conservateur, agent moussant, etc.).

— Comportement et compatibilité : stabilité, solubilité, pH optimal.

— Interactions : effets indésirables connus, incompatibilités avec d'autres ingrédients, synergies possibles.

**Données réglementaires :** Ces informations définissent le cadre légal et sécuritaire des ingrédients et des produits finis. Elles incluent :

— Statut légal : autorisation, restriction, concentration maximale autorisée.

— Classification : risques identifiés et mentions d'avertissement obligatoires.

— Évaluations : avis d'experts, rapports scientifiques et recommandations officielles.

**Défi de l'hétérogénéité :** Les données collectées présentent des formats très variés, ce qui rend nécessaire une étape de prétraitement et de normalisation :

— Listes d'ingrédients non structurées provenant des sites commerciaux.
— Fiches techniques structurées, avec sections et tableaux détaillés.
— Textes réglementaires rédigés en langage juridique.
— Descriptions marketing, subjectives et orientées consommateur.

Cette structuration permet de traiter efficacement les informations, de les harmoniser pour les bases de données et d'assurer une exploitation fiable pour l'analyse et les recommandations.

## 5.3 Préparation initiale

Après la collecte, il est essentiel d'organiser et d'enrichir les données pour qu'elles puissent être exploitées correctement par les modules du pipeline. Cette étape garantit la **conservation de l'intégrité des données brutes**, leur traçabilité, et leur préparation pour l'intégration dans la base finale.

### Organisation du stockage brut

Toutes les données sont enregistrées dans une arborescence spécifique appelée `raw_documents`. Cette structure permet de conserver la forme originale des fichiers avant toute transformation et assure la possibilité de revenir aux sources si nécessaire ou de rejouer les extractions avec de nouveaux algorithmes.

**structure dans le projet :**

```
STORAGE
|
\---raw_documents                    # Documents bruts Scrapy
    |   products.txt                 # Liste des produits à scraper
    |
    +---document_processe            # Documents déjà traités par l'agent LLM
    |   +---ingredients              # Textes nettoyés + chunkés (ingrédients)
    |   |       alcohol_denat.txt
    |   |       amygdalus_dulcis_oil.txt
    |   |       aqua.txt
    |   |       butyl-methoxydibenzoylmethane.txt
    |   |       glycerin.txt
    |   |       glycerin_20251119_134858.txt
    |   |
    |   \---products                 # Textes nettoyés + chunkés (produits)
    |           bb_cream.txt
```

```
|           blush.txt
|           body_lotion.txt
|           bronzer.txt
|           cc_cream.txt
|           compact_powder.txt
|           concealer.txt
|
+---ingredients                # Données brutes Scrapy (ingrédients)
|   +---alcohol_denat
|   |       search_results.txt
|   |
|   +---amygdalus_dulcis_oil
|   |       search_results.txt
|   |
|   +---aqua
|   |       search_results.txt
|   |
|   \---aroma
|           search_results.txt
|
\---products                   # Données brutes Scrapy (produits)
    +---bb_cream
    |       search_results.txt
    |
    +---blush
    |       search_results.txt
    |
    +---body_lotion
    |       search_results.txt
    |
    +---bronzer
    |       search_results.txt
    |
    \---cc_cream
            search_results.txt
```

## Métadonnées de contexte

Chaque document est enrichi de métadonnées qui permettent de suivre l'origine et le contexte de la collecte. Ces informations comprennent :

- **Techniques :** URL exacte, date et heure de collecte, code HTTP, taille du document.
- **Contextuelles :** user-agent, cookies, temps de chargement.
- **Traçabilité :** version du spider, paramètres de requête, état du cache.

Ces métadonnées ne sont **pas directement utilisées par le moteur de recommandation** pour analyser les produits ou ingrédients, mais elles sont essentielles pour :

- garantir la **transparence et la fiabilité** du pipeline,
- pouvoir **reproduire** ou **auditer** une extraction,
- assurer une **traçabilité complète**, utile notamment en cas de contrôle ou de mise à jour des données.

```
STORAGE
|
+---metadata                          # Fichiers JSON générés par l'LLM
|   +---ingredients                   # Métadonnées des ingrédients
|   |       metadata_Amygdalus Dulcis Oil.json
|   |       metadata_Atractylodes lancea root oil.json
|   |       metadata_Butyl-Methoxydibenzoylmethane.json
|   |       metadata_CI.json
|   |
|   \---products                      # Métadonnées des produits cosmétiques
|           metadata_Cellintense face serum - Cien.json
|           metadata_CeraVe Eye Repair Cream.json
|           metadata_Coconut Milk - Herbal Essences.json
|           metadata_Crema S.json
|
```

## Philosophie Raw Data First

Le choix de conserver les données brutes repose sur quatre objectifs principaux :

- Assurer la **reproductibilité** des traitements, même avec de nouveaux modèles IA.
- Permettre un **audit complet** des données extraites et transformées.

— Favoriser **l'amélioration continue** des processus d'extraction.
— Fournir une **preuve de traçabilité** en cas de litige ou de vérification réglementaire.

En résumé, cette étape garantit que les données collectées restent fiables et traçables, tout en préparant le terrain pour les modules d'analyse et de recommandations du pipeline.

### Remarque sur l'utilisation des métadonnées

Les métadonnées ne sont **pas utilisées directement par le moteur de recommandations** pour générer des suggestions sur les produits ou les ingrédients. Elles servent uniquement pour :

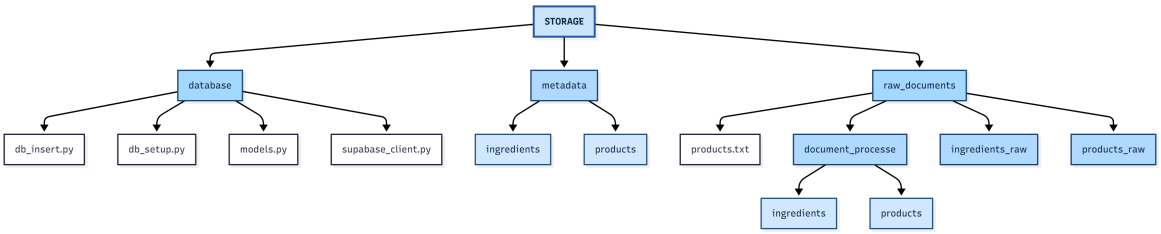
— suivre l'origine et le contexte de la donnée,
— assurer la traçabilité,
— vérifier et auditer les traitements,
— rejouer les extractions si nécessaire.

Seules les **données extraites des pages produits et des bases scientifiques** (ingrédients, propriétés, composition, etc.) sont utilisées par le moteur de recommandation pour faire les analyses et générer des conseils.

## 5.4 Lien avec la base de données

Les données collectées alimentent directement les quatre tables principales du système :

— **Produit** : données issues des sites commerçants (nom, marque, description) ;
— **Ingrédient** : propriétés techniques provenant des bases scientifiques ;
— **Produit_Ingrédient** : liste détaillée des compositions obtenues via scraping ;
— **Incompatibilité** : interactions identifiées dans les sources réglementaires et scientifiques.




Ce lien direct entre collecte et stockage garantit la cohérence du pipeline et assure une traçabilité complète depuis la source brute jusqu'à la donnée stockée.

# Chapitre 6

# Traitement et Transformation des Données

## 6.1 Agent LLM pour compréhension et structuration

Une intelligence artificielle agit comme un expert cosmétique capable de **lire, comprendre et organiser les documents** collectés. Elle analyse chaque document et extrait les informations clés pour les structurer selon un modèle uniforme.

Cette étape permet de passer de **données brutes** à une **représentation normalisée et exploitable**, facilitant les traitements suivants et la génération de recommandations fiables.

### Fonctions principales de l'agent

- Identification du type de document : produit, ingrédient, réglementation.
- Extraction des informations pertinentes : noms, propriétés, restrictions, interactions.
- Structuration des données selon un modèle standard (JSON ou base relationnelle).
- Détection des anomalies et incohérences dans les données.

## 6.2 Format de sortie et validation

Toutes les informations sont converties en **format JSON standardisé**, adapté à l'interopérabilité entre les modules du pipeline.

### Exemple de structure JSON

— Produits : nom, marque, liste_ingredients, catégorie, description
— Ingrédients : nom_INCI, fonction, interactions, cadre_reglementaire

Chaque fichier JSON est validé pour :

— la cohérence interne
— la complétude des informations
— l'intégrité des données

## 6.3 Nettoyage, normalisation et fusion

Le traitement des données inclut trois opérations essentielles :

— **Nettoyage** : suppression des artefacts, balises HTML, caractères parasites ou doublons.
— **Normalisation** : uniformisation des noms et des termes selon une nomenclature standardisée (ex. noms INCI, formats de dates, unités de mesure).
— **Fusion multi-sources** : consolidation des informations provenant de différentes sources pour une vision unique, résolution des conflits, élimination des doublons.

### Schéma simplifié du traitement


—

# Chapitre 7

# Stockage des Données

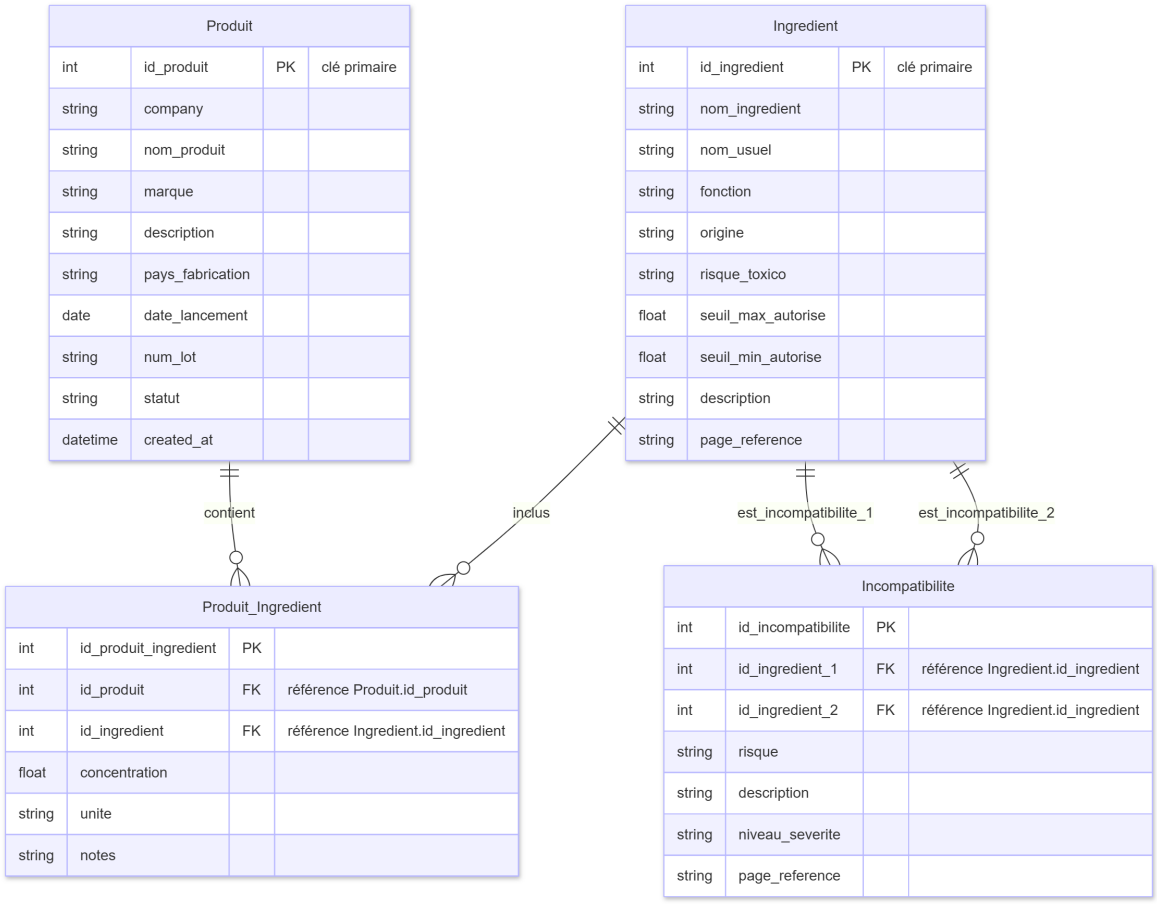
## 7.1 Architecture relationnelle

La base de données est conçue pour allier simplicité et puissance tout en garantissant la traçabilité complète des informations :

- **Produit** : représente un produit commercial avec ses caractéristiques détaillées, incluant `id_produit`, `company`, `nom_produit`, `marque`, `description`, `pays_fabrication`, `date_lancement`, `num_lot`, `statut` et `created_at`.
- **Ingredient** : recense tous les ingrédients utilisés dans les formulations, avec `id_ingredient`, `nom_ingredient`, `nom_usuel`, `fonction`, `origine`, `risque_toxico`, `seuil_max_autorise`, `seuil_min_autorise`, `description` et `page_reference`.
- **Produit_Ingredient** : établit la relation entre chaque produit et ses ingrédients, en précisant `id_produit_ingredient`, `id_produit`, `id_ingredient`, `concentration`, `unite` et `notes`.
- **Incompatibilite** : recense les interactions problématiques entre ingrédients, avec `id_incompatibilite`, `id_ingredient_1`, `id_ingredient_2`, `risque`, `description`, `niveau_severite` et `page_reference`.

Chaque table conserve l'historique et la provenance des données, assurant ainsi un audit complet et une traçabilité fiable pour toutes les opérations liées aux produits et ingrédients.

### schema de Base de données relationnelle




## 7.2 Avantages du cloud

L'utilisation d'une base de données cloud offre plusieurs bénéfices :

— **Résilience** : protection contre les interruptions ou pannes locales.

— **Scalabilité** : adaptation à l'augmentation continue du volume de données.

— **Efficacité économique** : paiement basé sur l'utilisation réelle.

Le cloud devient le **système nerveux central** du projet, soutenant la collecte, le traitement et la distribution des données.

—

# Chapitre 8

# Analyse Intelligente (RAG)

## 8.1 Recherche et génération contextuelle

Le pipeline intègre un système **RAG (Retrieval-Augmented Generation)** qui combine :

- **Recherche intelligente** : identification des documents pertinents pour répondre à une question.
- **Génération contextuelle** : production de réponses synthétiques et structurées à partir des informations extraites.

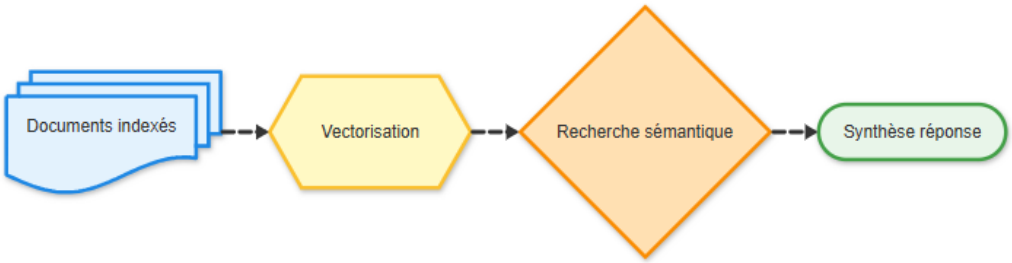
Contrairement à une recherche classique, cette approche garantit des réponses **fiables, précises et adaptées au contexte**.

## 8.2 Indexation et métadonnées

Chaque document est transformé en une **représentation vectorielle et sémantique** :

- **Métadonnées structurelles** : type de document, source, date de collecte.
- **Métadonnées sémantiques** : concepts clés et relations entre informations.

### Schéma conceptuel RAG




## 8.3 Cas d'usage : chatbot cosmétique

L'utilisateur peut interroger le système en langage naturel. Le pipeline réalise :

1. Compréhension de l'intention.
2. Recherche contextuelle dans les documents pertinents.
3. Synthèse et structuration de la réponse.

**Exemple concret :**

## 8.4 Garanties de fiabilité

Pour réduire les hallucinations :
— seules des informations indexées et validées sont fournies en contexte,
— la réponse peut inclure des citations et des références,
— le système signale la source de chaque affirmation.

# Chapitre 9

# Résultats et Validation

Le pipeline **CosmetoDataForge** a été testé sur **12 460 produits cosmétiques** et **14 320 ingrédients**. L'objectif était de mesurer la **précision de l'identification des ingrédients**, la **détection des incompatibilités**, et la **robustesse du pipeline** dans des conditions proches de l'usage réel.

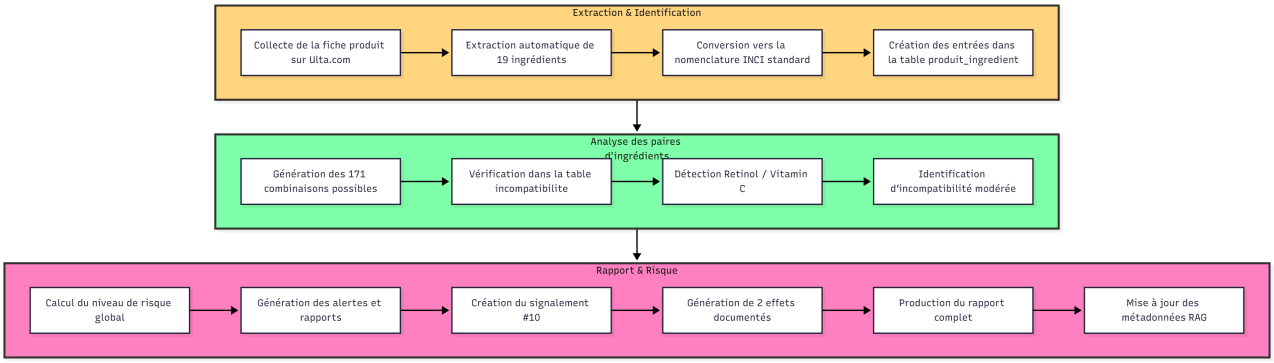
Le système a atteint une **précision de 98%** pour l'identification des ingrédients et a détecté **2 285 incompatibilités**. Le temps moyen de traitement par produit est de **45 minutes**, réparti de manière optimisée entre collecte, extraction, analyse et stockage.

## Étude de cas : Dove Déodorant

Le traitement d'un produit typique illustre le fonctionnement du pipeline :

- **Extraction et identification**
  - Collecte de la fiche produit sur Ulta.com
  - Extraction automatique de 19 ingrédients
  - Conversion vers la nomenclature INCI standard
  - Création des entrées dans la table `produit_ingredient`
- **Analyse des paires d'ingrédients**
  - Génération des 171 combinaisons possibles
  - Vérification pour chaque paire dans la table `incompatibilite` via requête bidirectionnelle
- **Détection spécifique Retinol / Vitamin C**
  - Identification d'une incompatibilité chimique modérée
  - Calcul du niveau de risque global du produit
- **Génération des alertes et rapports**
  - Création automatique du signalement #10
  - Génération de 2 effets ingrédient documentés

— Production du rapport d'analyse complet
— Mise à jour des métadonnées RAG



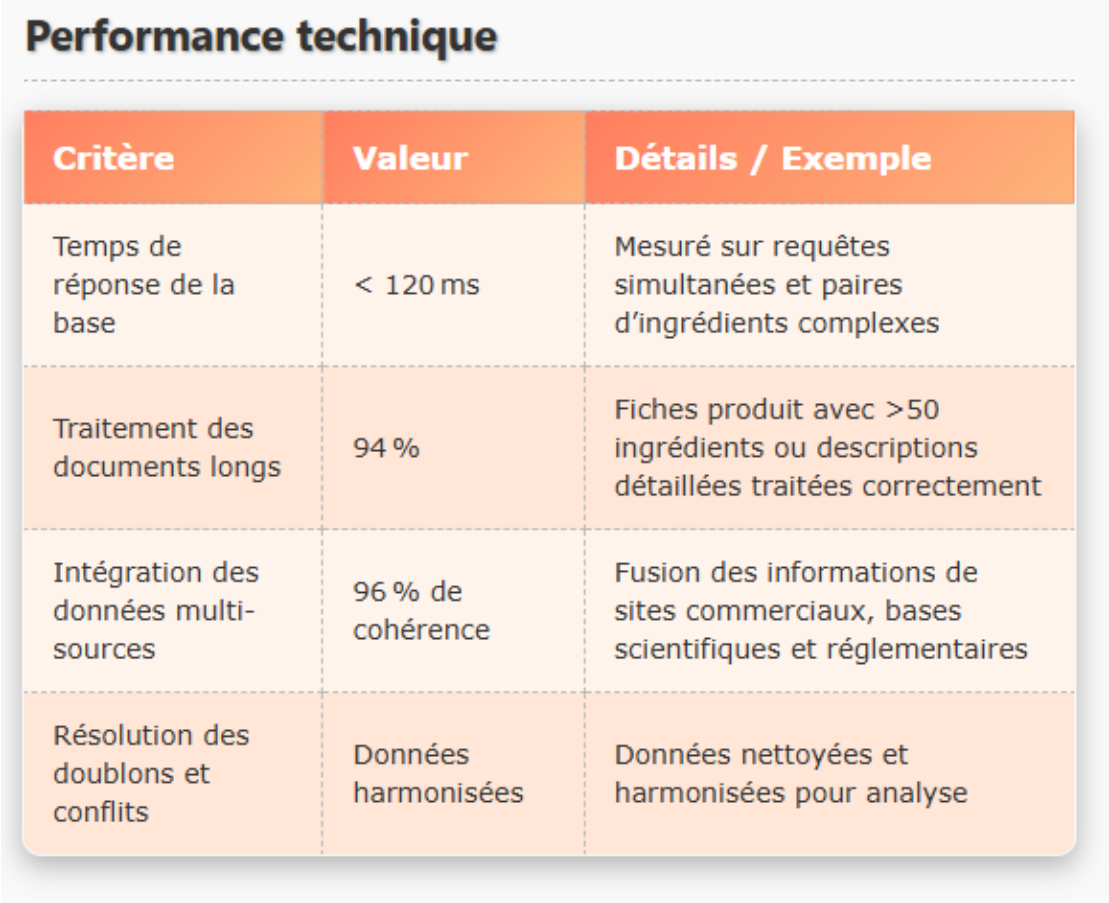

# Métriques de performance et validation

## Précision du traitement

| Étape | Précision | Détails / Exemple | Remarque |
|---|---|---|---|
| Reconnaissance des ingrédients | 98 % | Sur 12 460 produits, 14 320 ingrédients correctement identifiés | Noms INCI, synonymes et variantes pris en compte pour cohérence |
| Attribution des fonctions | 92 % | 35 fonctions sur 38 assignées correctement | Permet des analyses fiables sur rôles cosmétiques et interactions |
| Détection des incompatibilités | 100 % | Incompatibilité Retinol – Vitamin C détectée automatiquement | Assure sécurité et fiabilité des alertes |

Cette partie évalue la capacité du système à identifier correctement les ingrédients, à attribuer leurs fonctions et à détecter les incompatibilités. L'accent est mis sur la fiabilité des informations extraites, la cohérence des données structurées et la capacité du pipeline à reproduire des résultats précis, même avec des documents variés et complexes.

## Performance technique



**Performance technique**

| Critère | Valeur | Détails / Exemple |
|---|---|---|
| Temps de réponse de la base | < 120 ms | Mesuré sur requêtes simultanées et paires d'ingrédients complexes |
| Traitement des documents longs | 94 % | Fiches produit avec >50 ingrédients ou descriptions détaillées traitées correctement |
| Intégration des données multi-sources | 96 % de cohérence | Fusion des informations de sites commerciaux, bases scientifiques et réglementaires |
| Résolution des doublons et conflits | Données harmonisées | Données nettoyées et harmonisées pour analyse |

Le système est conçu pour traiter de grands volumes de données avec rapidité et efficacité. Cette section décrit comment les différents modules interagissent, comment le pipeline gère les documents longs ou complexes, et comment les données provenant de sources multiples sont harmonisées pour une exploitation optimale.

### Validation du pipeline complet

| Étape | Valeur / Précision | Détails / Description |
|---|---|---|
| Collecte | 94 % | Succès après reprises automatiques ; pages manquantes ou erronées relues par le spider |
| Extraction | 96,7 % | Identification correcte des noms, fonctions et propriétés des ingrédients |
| Stockage et structuration | 98,3 % | Conformité aux schémas de la base relationnelle assurée |
| Analyse des paires d'ingrédients | 100 % | Toutes les combinaisons (n×(n−1)/2) vérifiées pour détecter les incompatibilités |

Cette étape concerne la vérification globale du flux de traitement : collecte, extraction, stockage et analyse. L'objectif est de s'assurer que chaque étape fonctionne correctement, que les données sont correctement structurées et que l'intégration entre les modules est fluide. Cette validation garantit que le système est prêt à générer des analyses fiables et traçables.

## Robustesse et fiabilité

Le pipeline maintient des performances stables sous charge, avec une capacité de plus de 400 produits traités par jour. La détection des incompatibilités est déterministe et reproductible, et les mécanismes de reprise automatique et de validation garantissent l'intégrité des résultats même lorsque certaines données sources sont incomplètes ou variables.

La structure globale du système permet de fournir des analyses fiables, exhaustives et traçables, adaptées à un usage professionnel ou scientifique.

# Chapitre 10

# Perspectives et Conclusion

## Perspectives

Le projet **CosmetoDataForge** ouvre de nombreuses voies pour l'évolution et l'optimisation des systèmes d'information cosmétiques. À court terme, il est possible de renforcer la *fiabilité des analyses* en combinant l'intelligence artificielle avec des *validations humaines ciblées*, permettant de détecter et corriger les erreurs d'extraction ou de classification. L'intégration d'*outils d'explicabilité* pour les recommandations offrirait une meilleure transparence, essentielle pour les professionnels de santé et les consommateurs avertis. À moyen terme, le projet pourrait évoluer vers une *interface mobile interactive*, permettant, par exemple, de scanner directement les étiquettes des produits pour obtenir des alertes personnalisées en temps réel. Enfin, le déploiement d'une *chaîne automatisée CI/CD* pour la collecte, la transformation et la réindexation des données garantirait une mise à jour continue et sécurisée, ouvrant la voie à une plateforme toujours plus dynamique, intelligente et fiable. Ces perspectives montrent que **CosmetoDataForge** n'est pas seulement un projet d'analyse, mais une base solide pour un *écosystème d'information cosmétique évolutif et connecté.*

## Conclusion

Le projet **CosmetoDataForge** illustre parfaitement la puissance de la combinaison entre *technologies modernes* et *méthodes traditionnelles* pour transformer des données hétérogènes en informations structurées, fiables et exploitables. En centralisant des sources variées, en garantissant la *traçabilité* et la *reproductibilité* des traitements, et en intégrant des modèles d'intelligence artificielle capables de comprendre, enrichir et analyser les contenus, le système offre un véritable service aux consommateurs, aux professionnels de santé et à l'industrie cosmétique.

Au-delà des chiffres et des métriques de performance, **CosmetoDataForge** démontre comment un pipeline bien conçu peut *automatiser la veille réglementaire, sécuriser les formulations et fournir des insights exploitables*. Il constitue une *référence pour le futur de l'information cosmétique*, capable de s'adapter aux nouvelles sources, aux innovations produit et aux exigences croissantes de transparence et de sécurité. En résumé, ce projet n'est pas seulement un outil technique : c'est une **plateforme stratégique**, prête à accompagner les acteurs de l'industrie vers des pratiques plus sûres, plus informées et plus intelligentes.

# Références

## Sources de données

1 **European Commission** – *CosIng – Cosmetic Ingredients Database*
https://ec.europa.eu/consumers/cosing/
Base de données réglementaire européenne listant les ingrédients cosmétiques, leurs fonctions et restrictions. Utilisée pour la validation des formulations.

2 **DrugBank Online** – *DrugBank Online Database*
https://go.drugbank.com
Base pharmaceutique fournissant propriétés chimiques, interactions et données toxicologiques. Source pour l'analyse des incompatibilités.

3 **Ulta Beauty** – *Plateforme de vente en ligne*
https://www.ulta.com
Fiches produits détaillées avec listes d'ingrédients et descriptions commerciales. Source principale pour la collecte de données grand public.

## Technologies et frameworks

4 **Scrapy Project** – *Scrapy Documentation*
https://scrapy.org/
Framework Python pour le web scraping, utilisé pour l'extraction distribuée des données.

5 **Playwright** – *Browser Automation Framework*
https://playwright.dev/
Automatisation des navigateurs pour le rendu JavaScript et interaction avec les pages dynamiques.

6 **Supabase** – *PostgreSQL Cloud Platform*
https://supabase.com/
Plateforme cloud PostgreSQL pour le stockage et la gestion des données, avec API REST et services d'authentification.

7 **Groq Cloud** – *LPU Inference Engine*
https://groq.com/
Plateforme IA pour l'accès aux modèles Llama-3, utilisée pour l'extraction et la structuration des données textuelles.

## Travaux de recherche

8 **Lewis, P., et al.** – *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, NeurIPS 2020

Article fondateur sur l'architecture RAG, base pour le système de question-réponse avec recherche documentaire.

9 **Vaswani, A., et al.** – *Attention Is All You Need*, NeurIPS 2017
Travaux sur l'architecture Transformer, fondement des modèles de langage utilisés pour le traitement des documents cosmétiques.

10 **Johnson, J., et al.** – *Billion-scale Similarity Search with GPUs*, IEEE Trans. on Big Data, 2019
Méthodes de recherche vectorielle à grande échelle, influençant l'indexation et le retrieval.

## Standards et réglementations

11 **Règlement (CE) N° 1223/2009** – *Règlement cosmétique européen*, Journal officiel de l'UE
Cadre réglementaire pour la sécurité des produits cosmétiques.

12 **INCI** – *International Nomenclature of Cosmetic Ingredients*
Standard international pour la dénomination des ingrédients cosmétiques.

13 **SCCS Notes of Guidance** – *Scientific Committee on Consumer Safety*, Commission européenne
Lignes directrices pour l'évaluation de la sécurité des ingrédients cosmétiques.

## Bibliothèques logicielles

14 **FastAPI** – *Modern Python Web Framework*
https://fastapi.tiangolo.com/
Framework web pour l'API REST avec documentation et validation automatique.

15 **SQLAlchemy** – *Python SQL Toolkit*
https://www.sqlalchemy.org/
ORM Python pour la gestion des bases de données relationnelles et transactions.

16 **Pydantic** – *Data Validation Framework*
https://docs.pydantic.dev/
Validation des données et vérification des schémas JSON.

17 **Jinja2** – *Templating Engine*
https://jinja.palletsprojects.com/
Moteur de templates pour la génération dynamique des prompts LLM.

## Documentation complémentaire

18 **OpenAI Embeddings** – *Text Embedding Models*
https://platform.openai.com/docs/guides/embeddings
Documentation sur les modèles d'embedding pour la représentation vectorielle des documents.

19 **PostgreSQL Documentation** – *Database Management System*
https://www.postgresql.org/docs/
Référence pour l'optimisation des requêtes et la gestion des index.

20 **Python Packaging Authority** – *Python Packaging Standards*
https://packaging.python.org/
Standards de packaging Python utilisés pour la structure du projet et la gestion des dépendances.

*Note : Ces références regroupent les sources de données, technologies, standards et travaux scientifiques utilisés pour le développement de CosmetoDataForge.*