**Hyper Open X**

**Documentation**

**ProspectIA**

**By Quinten Finance**

**Octobre 2025**

**CONFIDENTIEL**

SOMMAIRE

1. ***Objectifs du projet***
2. ***Présentation des travaux réalisés***
   1. ***Architecture globale***
   2. ***Infrastructure et déploiement***
   3. ***Zoom sur les briques fonctionnelles***
   4. ***Zoom sur l'interface et son utilisation***
   5. ***Usage des briques HOX***
   6. ***Fonctionnalité d'émergence***
   7. ***Exemple de résultats***
   8. ***Organisation des codes***
3. ***Statut du produit final livré et tâches restantes***
4. ***Annexes & Références***
5. **Objectifs du projet**

Le projet s'inscrit dans un contexte où de nombreux secteurs ont besoin d'une connaissance précise du métier des entreprises françaises afin de :

* + - Les financer.
    - Acquérir de nouveaux clients.
    - Identifier des fournisseurs.

Acquérir cette connaissance nécessite en premier lieu de collecter des données qualitatives et à jour.

Cependant, en France, malgré un vivier en open source à fort potentiel, les données sont peu exploitables et les acteurs du marché ne répondent que partiellement au besoin.

Cela se traduit par des processus manuels chronophages pour comprendre leur métier et disposer des principaux indicateurs d'une entreprise.

Aujourd'hui, les spécialistes français des données d'entreprises sont en retard par rapport à des acteurs d'autres pays, notamment aux États-Unis (DowJones, Hoovers, Bloomberg...), mais dont les outils ne tiennent pas compte des spécificités françaises et européennes.

Chacun des acteurs français de la donnée d'entreprise ne répond que partiellement au besoin et présente plusieurs limites. Ils s'appuient notamment sur des données : ​

* + - "Froides" et non représentatives de l'état en temps réel des entreprises.
    - Déclaratives et donc peu précises, parfois éloignées de la situation réelle des entreprises.
    - Incomplètes et ne permettent pas une bonne segmentation.

**Ce projet ambitionne de renforcer nos développements et de passer à l'échelle notre collecte de données, afin de développer la meilleure plateforme de prospection autour de l'écosystème des entreprises françaises.**

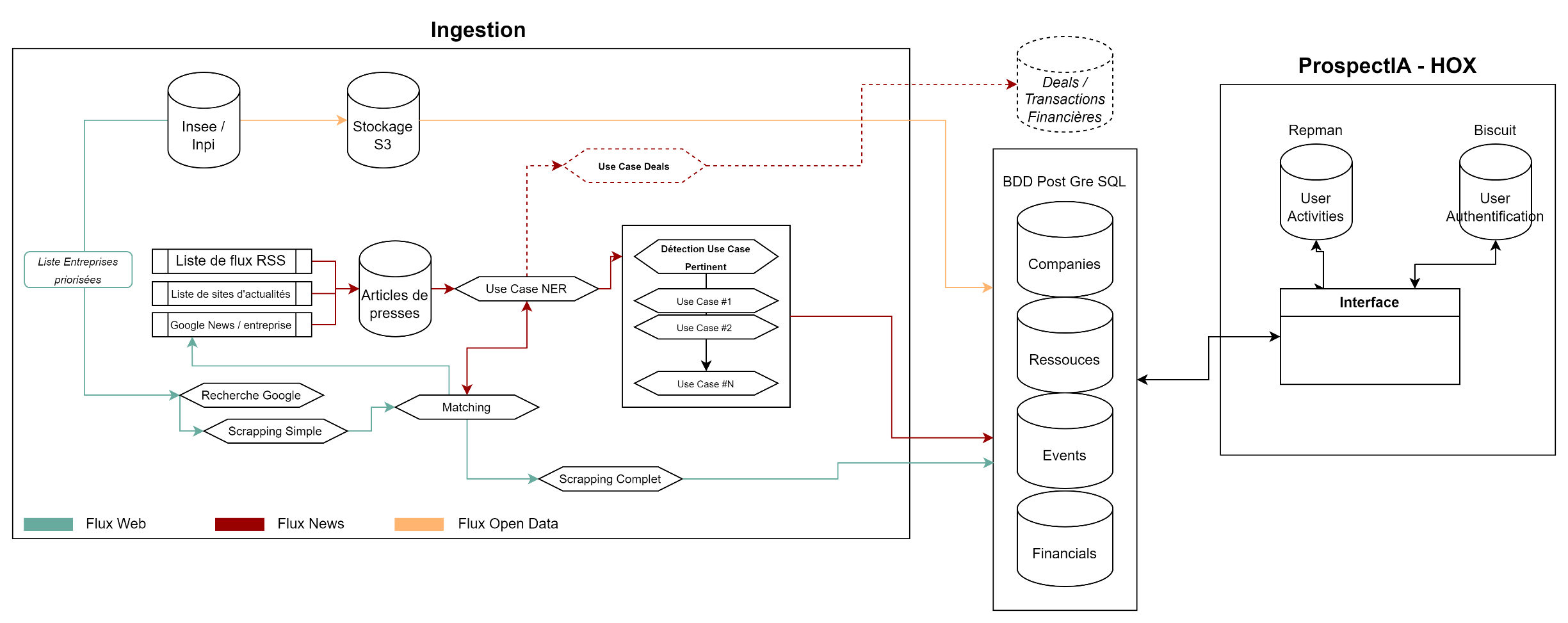
Notre plateforme permettra notamment de :

* + - Accéder à des données chaudes, au plus proche des actualités "temps réel" des entreprises.
    - Mettre à disposition de nouveaux critères de segmentation des entreprises, non disponibles aujourd'hui sur le marché de la prospection.
    - Établir un lien multi-sources entre open data, site internet propriétaire et actualités.

**Ainsi, le principal objectif de ce projet est de livrer une interface complète et fonctionnelle (équivalent à un TRL 7), permettant de réaliser de la prospection sur un secteur sélectionné, alliant segmentation innovante et données d’actualités. Cette interface sera un démonstrateur pour ensuite développer un outil sur mesure à coûts réduits, centré autour de l’usage des données presse et** d'**un usage M&A.**

1. **Présentation des travaux réalisés.**
   1. Architecture globale

La solution se décompose en trois parties :

* Une brique d'ingestion
* Une base de données
* Et une brique d’interface

La plus complexe et complète est la brique d'ingestion, qui collecte, récupère et met en forme des informations provenant de trois types de sources différents : des données open-source, des actualités de presse et des données issues de sites d'entreprises.

* 1. Infrastructure et déploiement

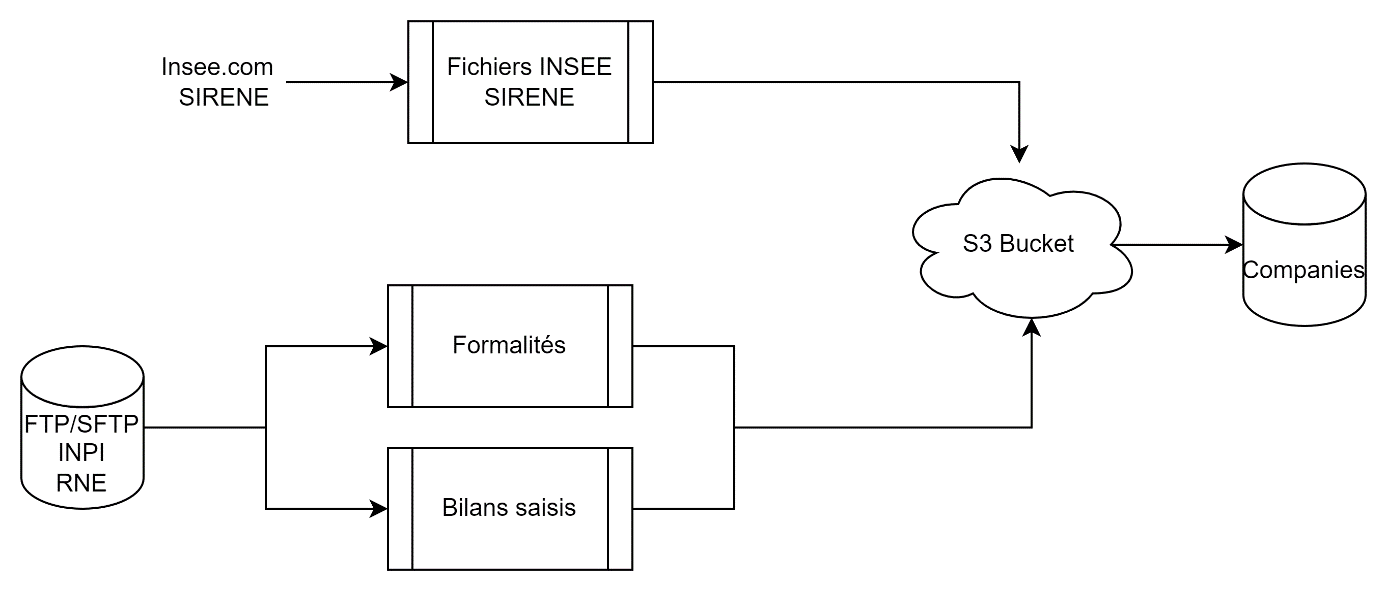
L’ensemble de la solution est déployé sur un environnement souverain OVH, hébergé en France en région Parisienne. 

Le projet principal est composé de 4 machines :

* BDD : contient et stocke l'ensemble des bases de données (PostgreSQL, Repman - MariaDB)
* Extraction : applique les pipelines web (scrapping, application de LLM, structuration et stockage)
* Ingestion : héberge des services partagés : API de matching, parseur d'adresse postale, etc.
* WebAPP : héberge l'interface et ses outils, dont Biscuit.

Toutes ces instances communiquent avec un bucket S3 pour partager des fichiers plats/lourds.

* 1. Zoom sur les briques fonctionnelles.
     1. **Flux d'ingestion open source.**



L’ingestion des données issues de l’INPI (Institut National de la Propriété Industrielle) et de l’INSEE représente une étape essentielle dans la constitution et la mise à jour de notre base de données sur les entreprises. Ces données, accessibles publiquement via un serveur FTP, se présentent sous la forme d’archives ZIP contenant des fichiers JSON. Deux grandes familles de jeux de données sont exploitées :

* Les formalités : il s’agit des dossiers d’identité complets des entreprises. Chaque fichier JSON regroupe des informations structurées telles que le SIREN, la raison sociale, les adresses, les représentants légaux, les dates d’ouverture et de clôture, ainsi que la date de dernière mise à jour.
* Les bilans saisis : ces fichiers contiennent les comptes annuels déclarés par les entreprises, également indexés par le SIREN. Six formats de bilans différents coexistent, chacun présentant une structure propre.

Les archives sont publiées selon une nomenclature standardisée, par exemple :  
stock\_formalites\_XX\_XX\_XXXX.zip ou stock\_bilans\_saisis\_XX\_XX\_XXXX.zip,  
où la date intégrée correspond à la dernière mise à jour du jeu de données.

Un pipeline d’ingestion automatique a été mis en place pour assurer la synchronisation continue avec le FTP de l’INPI. Le processus vérifie systématiquement la disponibilité de nouvelles versions des fichiers et déclenche le téléchargement uniquement lorsqu’une mise à jour est détectée. Les fichiers récupérés sont ensuite stockés sur un bucket S3, garantissant la traçabilité et la persistance des données sources.

Concernant les informations sur les entreprises (formalités), celles-ci sont principalement récupérées depuis l’INSEE, qui procède à un retraitement préalable des données issues de l’INPI afin d’en garantir la cohérence et la fiabilité.

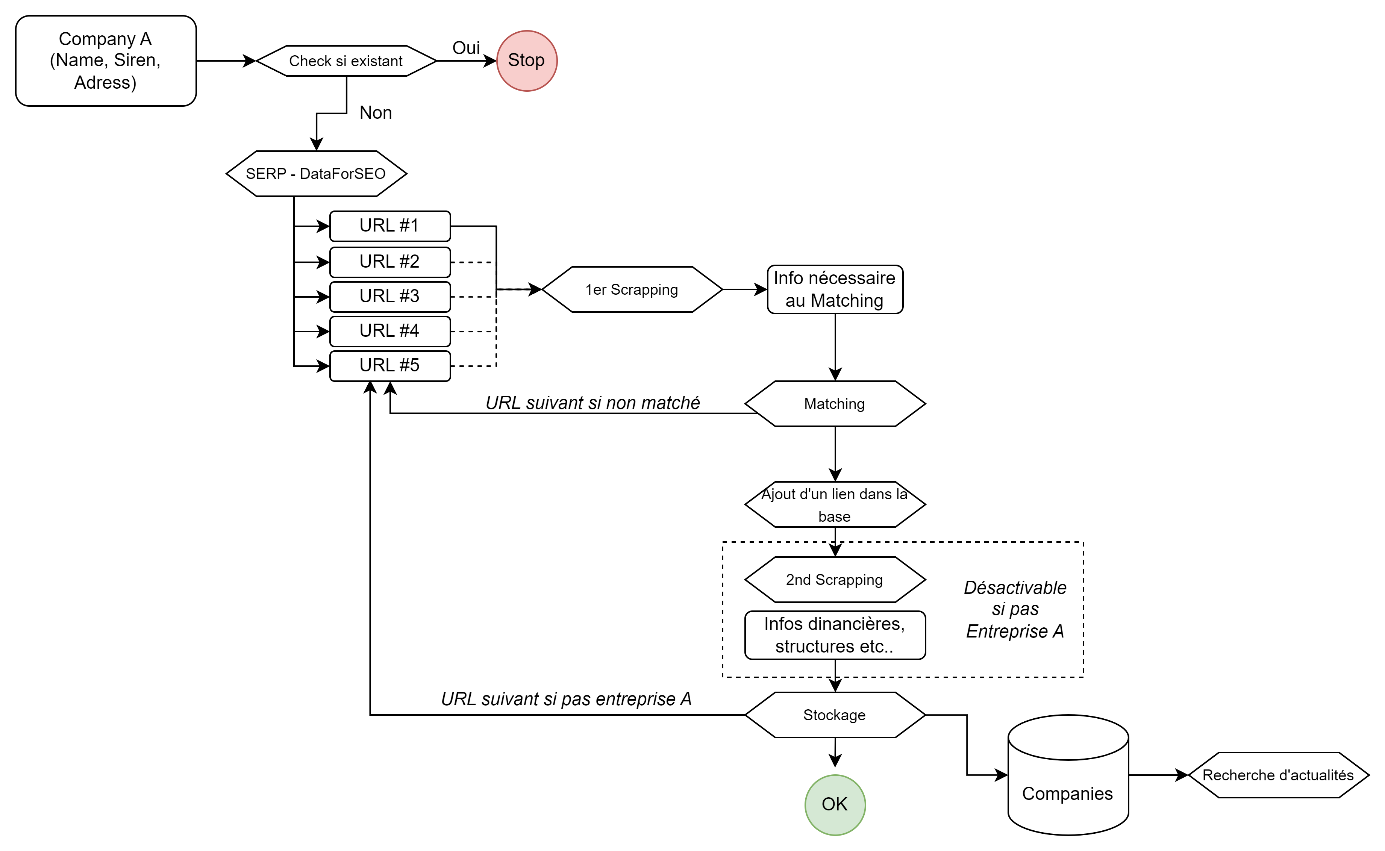
Pour les bilans saisis, un travail d'uniformisation a été nécessaire afin d'extraire les bonnes informations selon le type de bilan rencontré. Un fichier de mapping (au format CSV) a été élaboré : il définit, pour chaque structure de bilan, les champs pertinents à extraire (actif, passif, résultat, etc.). Ces extractions sont ensuite intégrées dans la base de données relationnelle, dans les tables companies et financial\_records, avec un lien direct vers l'entreprise concernée via son SIREN.

L’ensemble de cette architecture garantit un processus d’ingestion robuste, automatisé et traçable, permettant d’alimenter en continu notre entrepôt de données à partir des sources officielles ouvertes de l’INPI et de l’INSEE.

* + 1. **Flux d'ingestion Web**

L’objectif du flux Web est d'extraire et de structurer automatiquement les informations publiées par les entreprises sur leurs sites web officiels (compréhension du métier, organisation, données financières, recrutement, actualités…), puis de les relier à d’autres sources de données via un identifiant commun.  
Ce processus permet d'obtenir une vision "chaude" de l'entreprise à partir de leur site officiel complétant ainsi les données déclaratives "froides" issues des flux open source (INPI, INSEE). La sirénisation, c'est-à-dire l'association fiable entre un site web et le numéro SIREN d'une entreprise, constitue l'enjeu central de ce flux : elle garantit que les informations extraites du web sont correctement rattachées à la bonne entité dans la base de données.  
Le processus s’articule autour de plusieurs étapes successives : identification du site web, extraction et structuration des contenus, vérification et sirénisation, puis enrichissement et stockage final des données.

Finalement, l'étape finale pour parfaire notre outil est la recherche des actualités de cette entreprise. Aujourd'hui centrée sur les informations financières, cette étape permet à notre outil de prendre tout son sens, en ayant une connaissance à 360°, via ce que communique l'entreprise mais aussi ce qui est communiqué sur elle.



***Recherche sur le site web***

Le flux démarre avec les informations d'identification d'une entreprise cible : son nom commercial, ses identifiants légaux et son adresse. Ces trois éléments constituent le point de départ pour l'ensemble des recherches et vérifications qui suivent. Avant tout traitement, le système vérifie si l'entreprise est déjà associée à un site web dans la base de données. Cette vérification évite les traitements redondants et les duplications. Si une liaison existe déjà, le processus s'arrête immédiatement. Dans le cas contraire, le flux continue. Le système effectue alors une recherche Google via l'API SERP en combinant le nom de l'entreprise et sa localisation géographique. Cette requête permet de récupérer les premiers résultats de recherche organique, typiquement entre 10 et 20 URLs candidates. Les résultats sont automatiquement filtrés pour exclure les sites non pertinents tels que LinkedIn, Pappers, Societe.com ou autres annuaires d'entreprises. Les URLs restantes sont ensuite classées par score de pertinence, calculé en fonction de la correspondance entre le nom de l'entreprise et le domaine du site, ainsi que d'autres critères de proximité.

***Extraction, liaison et SIRENisation***

Le contenu HTML de chaque page candidate est récupéré et nettoyé. Avant l'extraction proprement dite, un modèle de langage analyse l'ensemble des URLs disponibles sur le site web pour identifier les pages les plus pertinentes. Ce mécanisme, qui s’adapte parfaitement aux différentes structures de sites internet, permet de limiter les coûts en scrapant un nombre limité de pages.  
Une fois les pages pertinentes sélectionnées, le texte est extrait et structuré automatiquement par un modèle de langage. Ce premier scraping collecte les informations générales de l'entreprise nécessaires à son identification : nom, adresse, pays, identifiants légaux et participants clés.  
Les informations extraites sont ensuite transmises à l'algorithme de liaison (voir ii.c.iv - Algorithme de liaison), qui recherche dans la base de données l'entreprise correspondante. Si un match est trouvé, une liaison est créée entre le site web et le SIREN de l'entreprise, puis enregistrée dans PostgreSQL. S'il n’y a pas de match ou si l’entreprise mappée n’est pas l’entreprise cible, le système continue d’explorer les URLs restants.

***Enrichissement et stockage final***

Une seconde étape de scraping est ensuite réservée par défaut à l'entreprise cible. Un modèle de langage sélectionne intelligemment les pages web les plus pertinentes parmi l'ensemble du site. Le volume de pages explorées est adapté à la taille de l'entreprise. Les informations collectées incluent la structure organisationnelle, les données financières, la compréhension du secteur d’activité et, pour les fonds d'investissement, des détails additionnels spécifiques (secteurs, montants, portefeuille).

L’extraction des informations liées au métier s’effectue en deux temps. Dans un premier temps, le système génère automatiquement des questions métier adaptées au contexte de l’entreprise (mots-clés, produits et services, clientèle cible, etc.). Dans un second temps, à partir des réponses obtenues et de l’analyse du contenu, il identifie de nouvelles pages web pertinentes, affinant ainsi progressivement la compréhension de l’activité de l’entreprise.

En complément, le système collecte des articles de presse récents via Google News, ciblant spécifiquement les actualités liées aux événements financiers et stratégiques (levées de fonds, fusions-acquisitions, etc.) (voir ii.c.iii)).

À l’issue du flux d’ingestion, l’ensemble des informations extraites est stocké en base de données, contribuant à l’enrichissement de l’interface utilisateur. Les données textuelles, notamment d’activités, sont quant à elles converties en embeddings, permettant de faire émerger les entreprises via la recherche de proximité sémantique (voir ii.f – Fonctionnalité d’émergence).

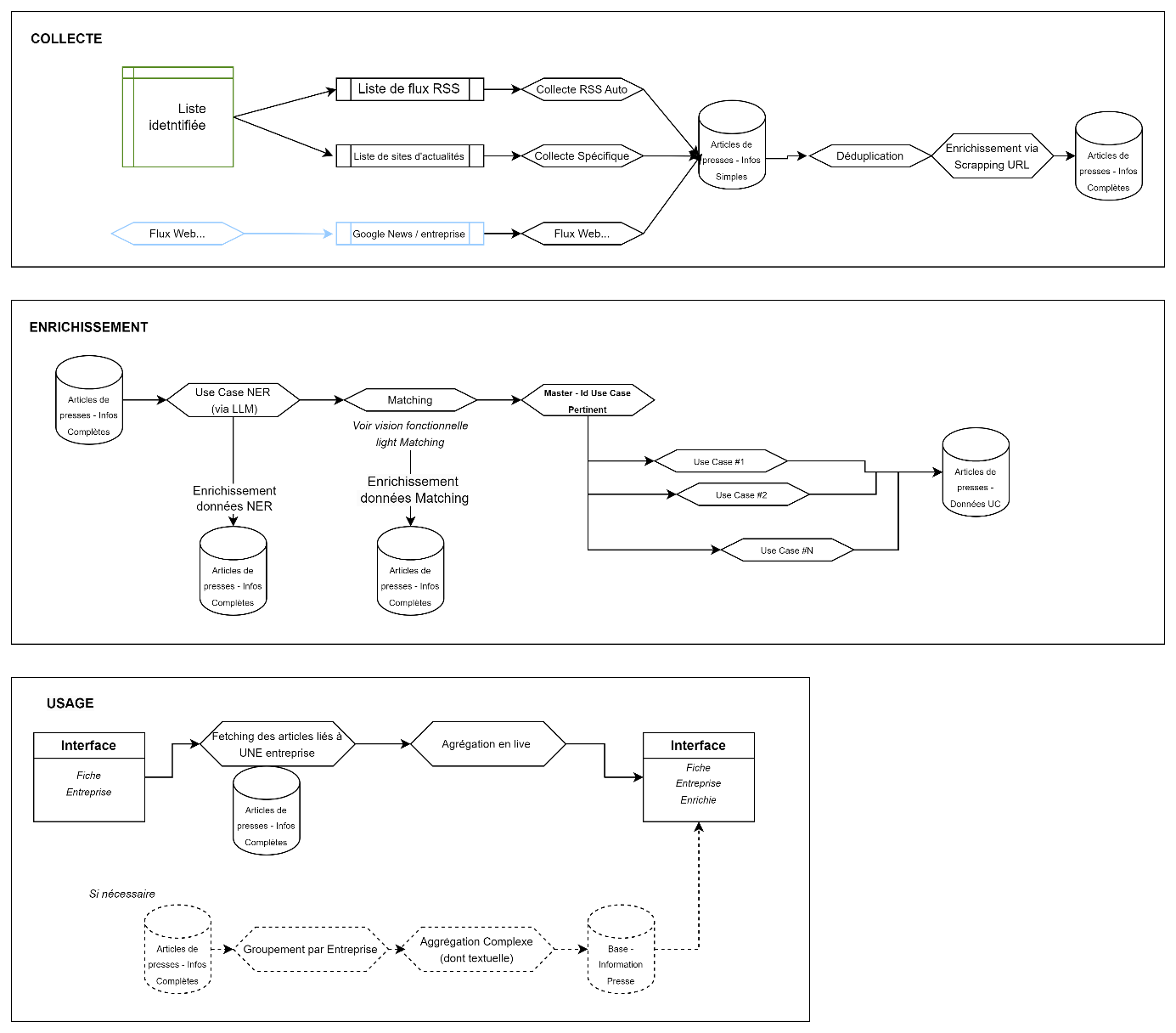
* + 1. **Le texte ne contient pas de fautes d'orthographe ou de grammaire.**

Le flux Presse constitue un élément central du projet, représentant la véritable valeur ajoutée par rapport aux solutions concurrentes. Il a pour objectif de collecter, filtrer et structurer les informations issues des médias en ligne, afin d’identifier et d'analyser les événements récents susceptibles d'affecter les entreprises référencées dans la base.

Ce flux complète ainsi la vision "froide" fournie par les bases de données institutionnelles (INSEE, INPI, etc.) par une vision "chaude" issue de l'actualité, en capturant à la fois ce que l'entreprise communique elle-même et ce qui est communiqué à son sujet dans la presse.

Le processus s’articule autour de plusieurs étapes successives : l’identification des sources de presse à traiter, l’extraction et la structuration du contenu des articles, le matching des informations avec les entreprises présentes dans la base, puis la catégorisation thématique et l’application de cas d'usage.

Ce cheminement garantit une collecte d’informations complète, cohérente et directement exploitable pour nos besoins.



***Identification des sources à suivre de façon régulière.***

Pour obtenir une couverture complète des sites web traitant de transactions financières, nous avons d’abord sélectionné un ensemble de sites français reconnus comme des sources fiables d’information économique et financière. À partir de cette liste initiale, des recherches Google ont été effectuées de manière régulière sur plusieurs années, afin d’identifier et de collecter les articles pertinents.

Les résultats obtenus ont ensuite été filtrés à l’aide d’un LLM chargé d’évaluer la pertinence de chaque contenu. Une seconde phase de recherche Google a ensuite été mise en œuvre, cette fois sur un ensemble élargi de sites identifiés au cours de la première phase. L’ensemble des résultats a ensuite fait l’objet d’un processus de nettoyage, de déduplication et de structuration.

Enfin, une analyse complémentaire a été menée pour qualifier la nature des domaines collectés. Un second passage par un LLM a permis de distinguer les véritables sites d’actualité des pages institutionnelles, blogs ou plateformes promotionnelles. Cette étape de classification a conduit à la constitution d’un corpus de sources médiatiques pertinentes.

***Extraction et structuration des articles.***

Une fois le corpus de sources défini, le flux procède au traitement *— via une ingestion régulière —* des articles collectés via ces sources. Pour chaque article, le contenu est scrapé afin d’en extraire les éléments clés nécessaires à l’analyse, à savoir : le titre, le corps du texte et la date de publication, lorsque ces informations sont disponibles.

Les données brutes ainsi collectées sont ensuite nettoyées et normalisées par un LLM, qui supprime les balises inutiles et reformate le texte afin de garantir une structure homogène.

Le LLM génère ensuite une sortie standardisée contenant les champs suivants :

* Titre : titre de l'article,
* content : contenu de l'article,
* published\_date : date de publication,
* company\_mentioned : indicateur booléen précisant si une entreprise est mentionnée (True ou False),
* entities : liste structurée des entreprises identifiées, incluant le company\_name (le nom de l’entreprise), l’activité et les adresses associées, lorsqu’elles sont mentionnées.

À l'issue de cette étape :

* Si aucune entreprise n’est mentionnée, l’article et les informations associées sont stockés dans une table intitulée Article, et le processus s’arrête à ce stade ;
* Dans le cas contraire, s'il y a mention d'au moins une entreprise, le traitement se poursuit vers l'étape de liaison.

***Liaison entre les articles et les entreprises***

L'étape de liaison vise à associer les entreprises mentionnées dans les articles aux entreprises déjà référencées dans la base de données d'entreprises. Elle se déroule en deux phases successives : l'extraction d'informations détaillées, puis le rapprochement avec la base d'entreprises existante.

La première phase consiste à extraire, à partir du titre, du contenu et de la date de publication de l’article, les informations nécessaires à la mise en relation des entreprises mentionnées avec les entités déjà enregistrées. Un LLM identifie et extrait, pour chaque entreprise citée, les éléments suivants :

* company\_name : nom officiel de l’entreprise, incluant sa forme légale le cas échéant,
* activity : description synthétique de l’activité principale,
* company\_size : taille estimée (*petite*, *moyenne*, *grande, non spécifiée*),
* company\_country : pays d’origine,
* addresses : liste des adresses mentionnées (siège, bureaux, sites industriels, etc.),
* participants : personnes clés associées à l’entreprise (dirigeants, fondateurs, administrateurs, etc.).

L’ensemble de ces informations est ensuite stocké dans une table intitulée Analyse, qui établit la correspondance entre les informations extraites et l’article d’origine.

Lors de la deuxième phase, les informations extraites sont ensuite transmises à l’algorithme de liaison (voir ii.c.iv - Algorithme de liaison). Celui-ci vérifie si les entreprises identifiées dans l’article correspondent à des entreprises déjà référencées dans la base d’entreprises.

* Si aucune correspondance n'est trouvée, le processus s'arrête à ce stade et l'article est enregistré dans la base avec les champs supplémentaires is\_matched = False et matching\_results = None.
* En revanche, si au moins une correspondance est identifiée, les résultats correspondants sont enregistrés et le processus se poursuit vers les étapes suivantes du flux d’ingestion.

Le texte "***Déduplication" ne comporte pas de fautes d'orthographe ou de grammaire.***

Une fois le matching réalisé, une étape de déduplication est effectuée afin d’identifier les articles traitant du même sujet, et ainsi éviter la redondance d’informations. Cette étape permet de ne pas traiter plusieurs fois des articles relatant le même événement et de limiter la récupération d’informations déjà identifiées. Seuls les articles portant la mention is\_duplicated = False sont pris en compte pour les étapes suivantes, les plus anciens faisant foi comme source principale d’information.

Pour cela, chaque article est vectorisé à partir de son titre et de son contenu, puis comparé à l’ensemble des articles publiés dans une fenêtre temporelle de 14 jours (avant et après la date de publication). Cette comparaison repose sur une mesure de similarité vectorielle, avec un seuil fixé de manière arbitraire à 0,2 pour détecter les fortes proximités de contenu.

En cas de similarité élevée, un LLM intervient pour déterminer s'il s'agit réellement d'un doublon ou d'un article distinct.

* Si le LLM conclut à un doublon, l’article est enregistré dans la base avec les champs is\_duplicated = True et duplicated\_with renseigné avec l’identifiant de ou des articles d’origine, et le processus s’arrête là pour ces articles.
* Dans le cas contraire, les champs sont définis comme is\_duplicated = False et duplicated\_with = None, et le processus se poursuit vers la dernière étape du flux.

***Catégorisation***

Enfin, les articles non dupliqués sont classés selon leur thématique principale. La classification est assurée par un LLM, qui s’appuie sur un ensemble de catégories prédéfinies couvrant l’ensemble du spectre médiatique : *Financial*, *Investment*, *Growth*, *ESG*, *R&D*, *Market*, *Politics*, *Sports*, *Arts*, *Technology*, *Health*, *Education*, *Environment*, *Social*, *Media*, *Legal*, *Deals* et *Others*. Un même article peut être associé à plusieurs catégories.

À l’issue de cette dernière étape, les articles sont stockés dans la base de données avec le champ supplémentaire "catégories", qui enregistre les résultats de la classification générée par le LLM.

**Application des cas d'usage (en cours d'implémentation) :**

Une fois l’ingestion des articles de presse effectuée, l’étape suivante consistera à appliquer différents cas d'usage aux articles enregistrés dans la base Article, en fonction de leur catégorie.

Ainsi, selon la thématique associée à chaque article, des LLM spécialisés seront exécutés. Ces modèles prendront en entrée le contenu de l’article ainsi que le titre et auront pour rôle d’extraire les informations spécifiques aux cas d'usage définis.

Par exemple, si un article est associé à la catégorie *Deals*, un LLM dédié à l’extraction d’informations transactionnelles (type d’opération, parties impliquées, montants, etc.) sera lancé. Les résultats extraits seront ensuite stockés dans la table Analyse (avec le champ use\_case =“Deals”) pour être exploités dans les étapes d’analyse ou de reporting ultérieures.

* + 1. **Algorithme de liaison**

Une fois l'ensemble des sources collectées, il est crucial de réussir à faire le lien entre chacune d'entre elles ; et en l'occurrence de réussir à lier chaque document à une ou des entreprises identifiables (c'est-à-dire ayant un identifiant unique).

Cette étape est particulièrement complexe car elle nécessite rapidité mais aussi précision, parmi un univers des possibles très large et contenant de multiples homonymes. Ainsi, lier un article qui mentionne l'entreprise ABC à une entreprise juste parce qu’elle s’appelle aussi ABC serait une erreur importante : il existe plusieurs dizaines d’entreprises avec ce nom.  
  
Il faut donc mettre en place un algorithme poussé capable de prendre en compte un grand nombre d’éléments de contexte supplémentaires. Parmi ces éléments, on retrouve :

* Les adresses mentionnées dans les textes (précises ou larges - au minimum un pays)
* Les personnes mentionnées
* Le contexte métier de l’entreprise (ABC est-elle un opérateur téléphonique ou un salon de coiffure dans le texte ?)
* Les identifiants mentionnés (SIREN/SIRET, numéro de TVA...) notamment sur les sites internet

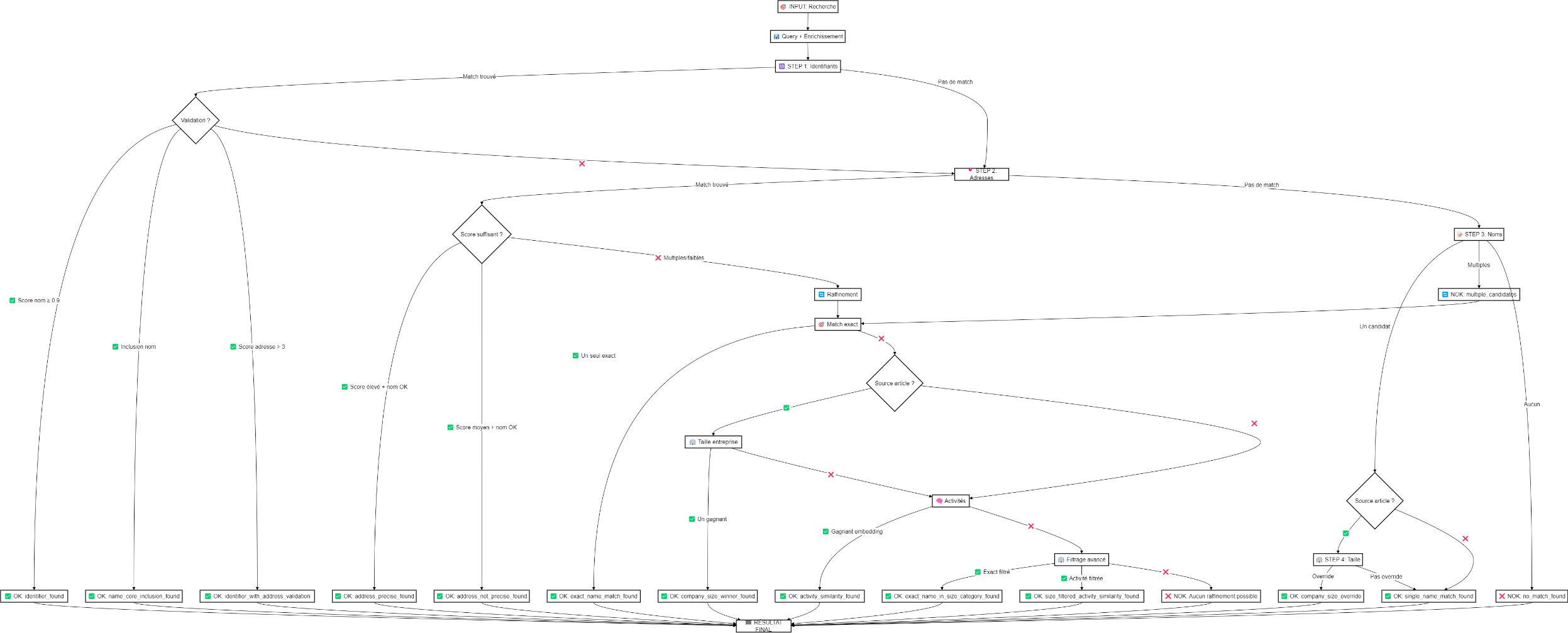
Il faut également faire attention à bien utiliser les données disponibles en open source. L’INSEE et l’INPI fournissent notamment un ensemble d’adresses possibles pour une entreprise (notamment en regroupant par Headquarters).

Il est très important de retraiter au mieux les noms des entreprises pour les comparer : On mentionnera rarement le mot « Groupe » dans un article de presse pour parler de « ABC Groupe » par exemple. Donc la distance sémantique doit être réalisée extrêmement finement pour être la plus pertinente possible.

Finalement, le nombre de candidats est également important à prendre en compte : si on ne remonte qu'un seul candidat en recherchant mon entreprise "ABC", on peut être plus confiant sur la liaison que si l'on remonte 10 candidats avec un nom similaire ou proche.

En sortie de cet algorithme, on retrouve un statut (OK/NOK), un label (explicabilité de la décision) et une liste de candidats.

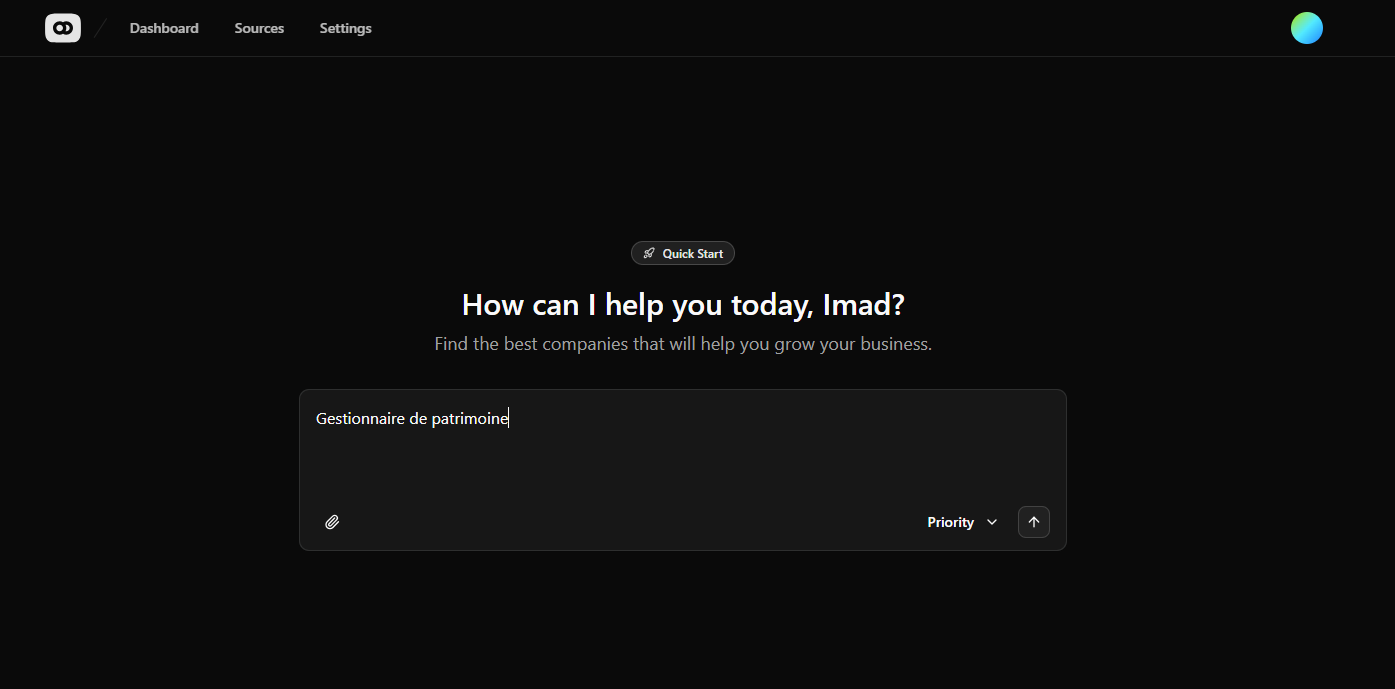
Cet algorithme se base également sur une librairie C++ "Libpostal" permettant de retravailler facilement les adresses postales en les décortiquant proprement.



* 1. Zoom sur l'interface et son utilisation

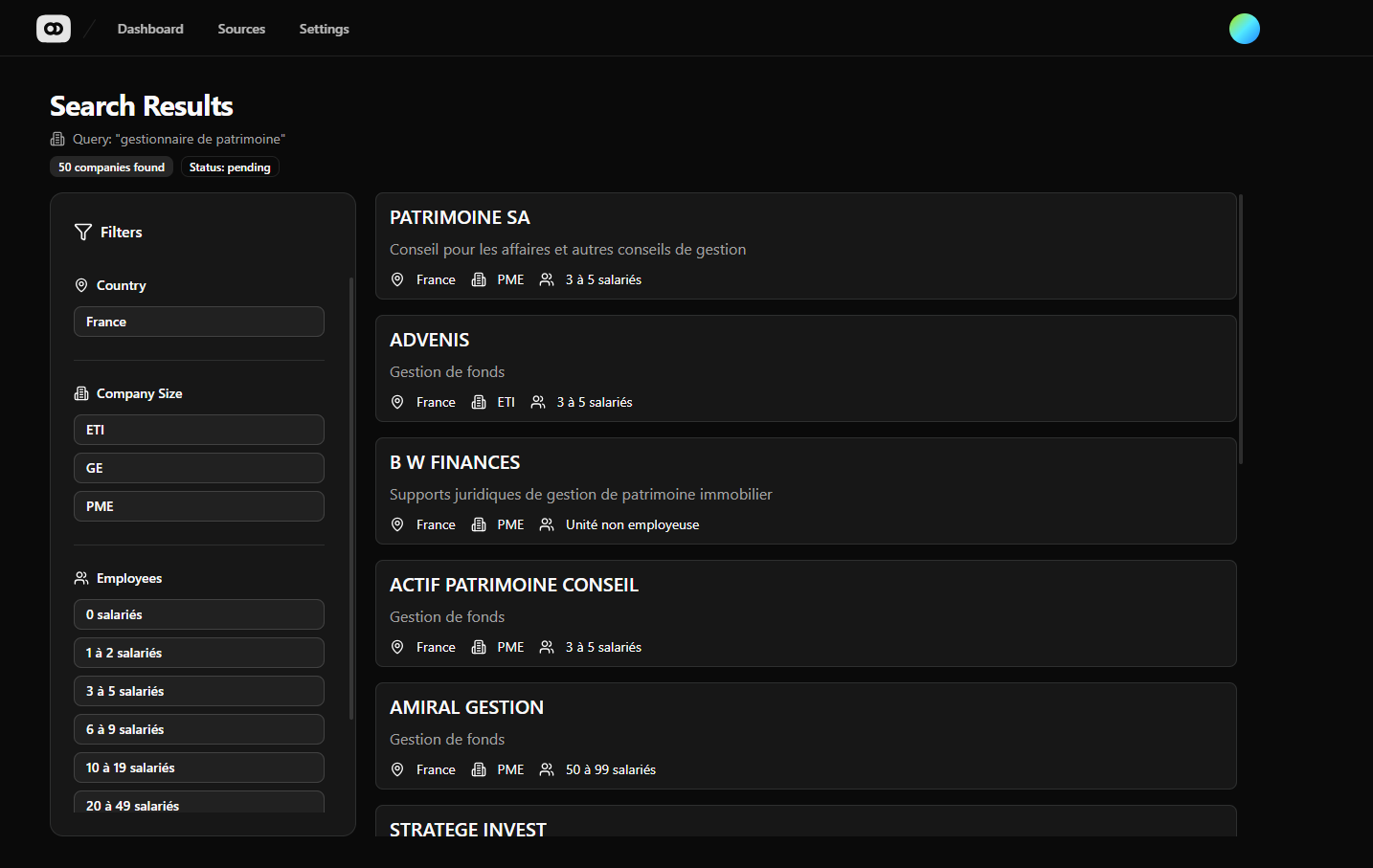
L'interface est décomposée en deux pages principales :

* Une page d'accueil, qui contient le moteur de recherche de l'outil. C'est la porte d'entrée principale.



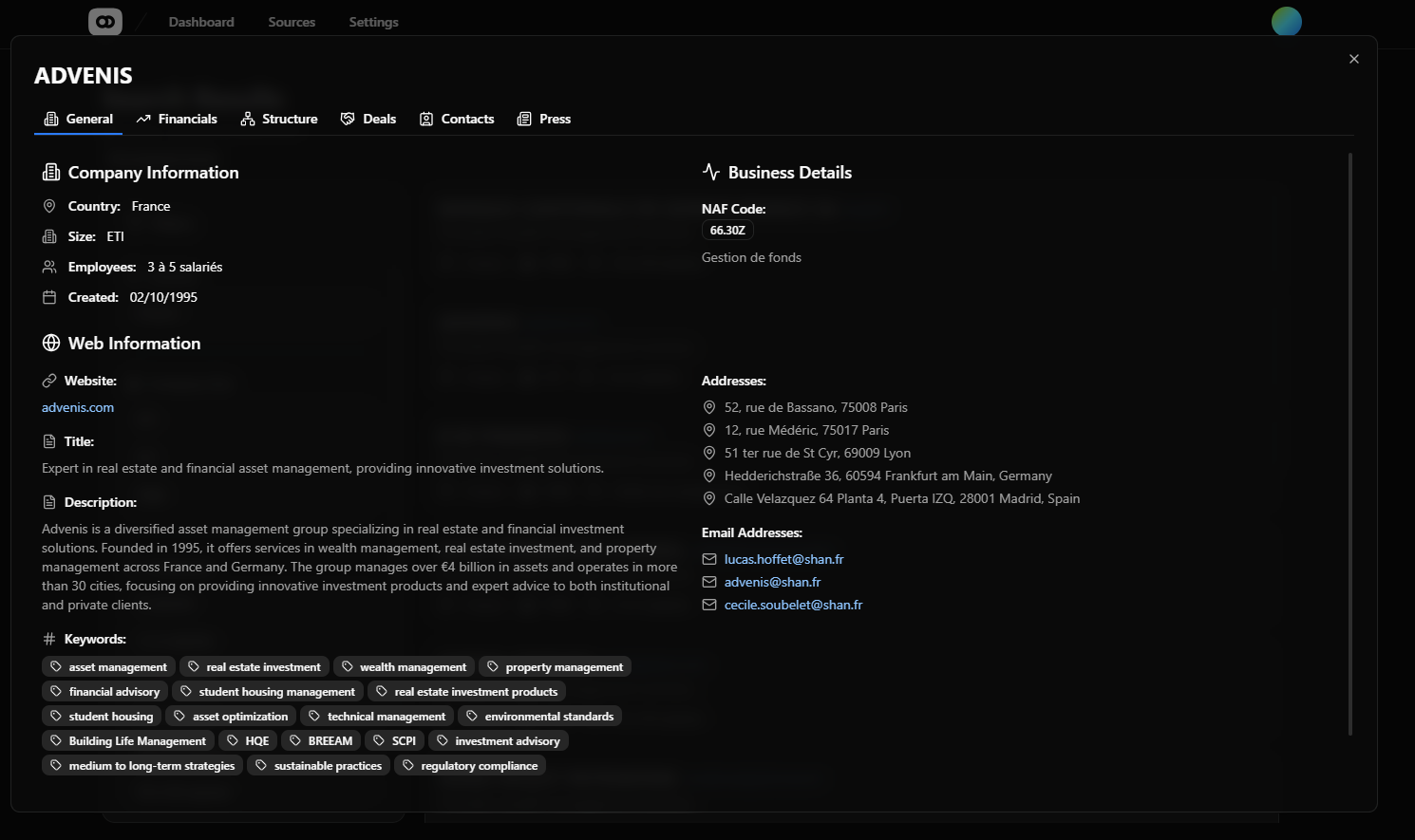
*La page d'accueil, contenant le moteur de recherche, est ici utilisée avec la requête "Gestionnaire de patrimoine".*

* Une page de résultats, permettant de se balader à travers les entreprises remontées :

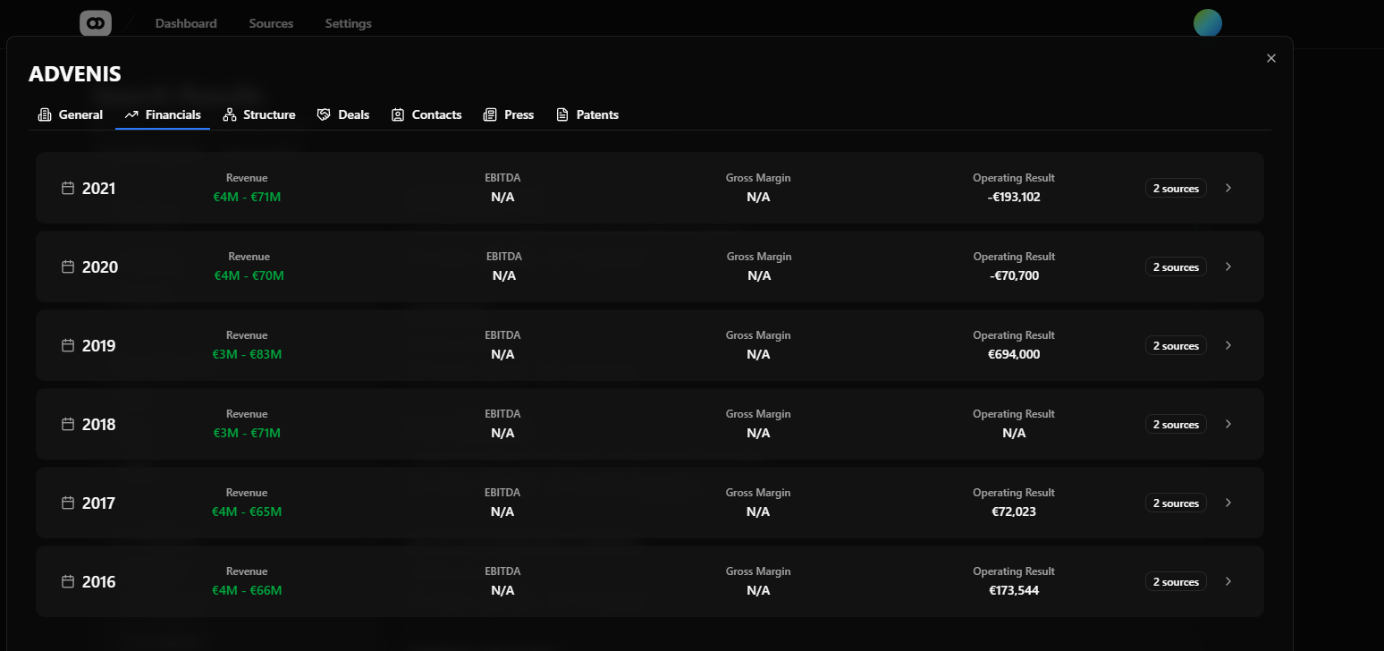
**

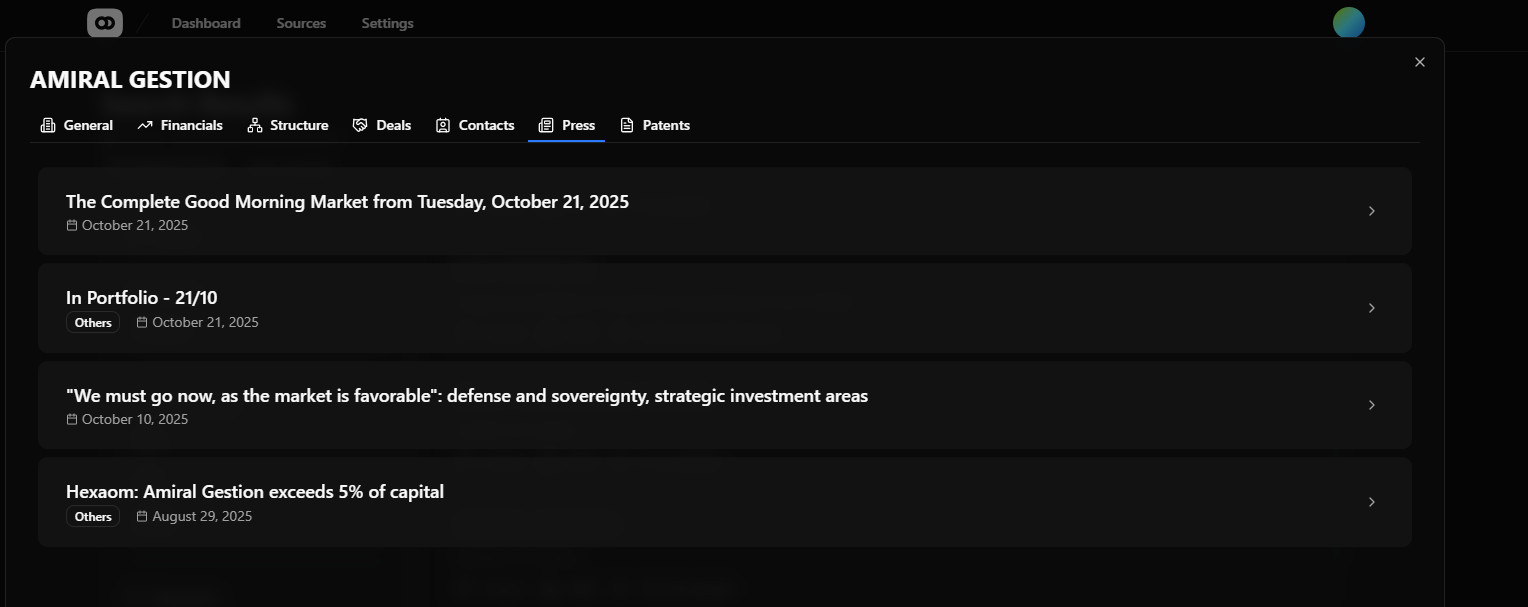
*La page de résultats, listant les meilleurs candidats et permettant de les parcourir via des critères de segmentation présents sur la gauche.*

* Un ensemble d'informations sur chaque entreprise, notamment issues des sites internet :

**

*Un premier onglet contenant des informations générales sur l'entreprise.*

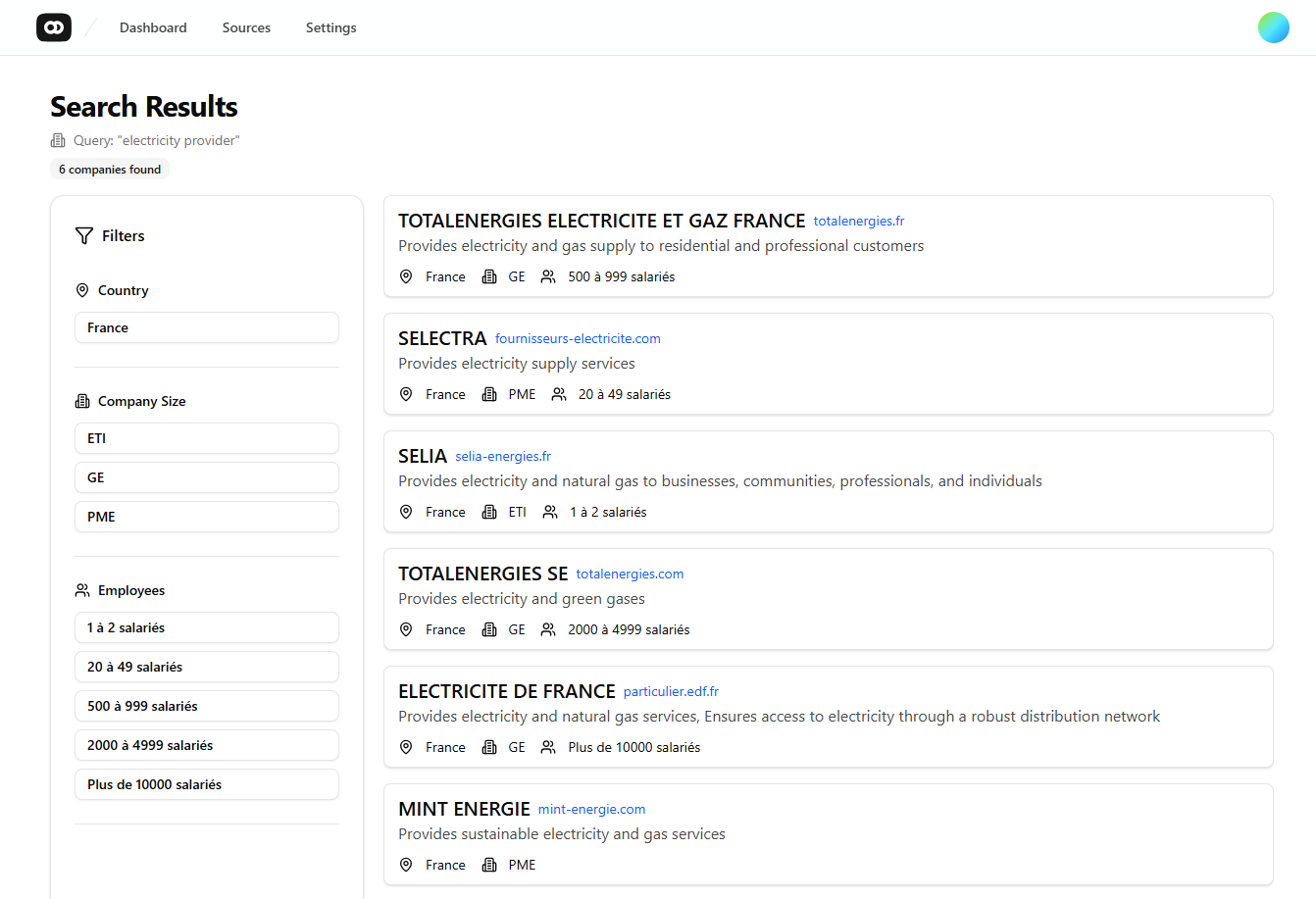
*Un second contenant des informations financières sur l’entreprise*

**

*Un troisième onglet contenant des articles de presse liés à l’entreprise.*

Ces onglets ont vocation à être complétés. Plus d’informations dans la partie *III - Statut du produit final livré et reste à faire.*

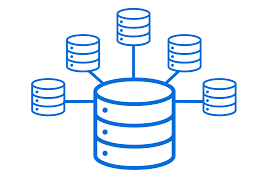
À noter que l'interface est également accessible en « mode léger » :



Cette interface repose sur l'usage des briques HOX Repman et Biscuit, détaillées ci-dessous.

* 1. Usage des briques HOX :

Un des objectifs du projet était d'intégrer dans notre outil les briques Repman et Biscuit.

Nous avons choisi de les intégrer là où leur apport était le plus important : au service de l’interface.

**Repman**, déployé avec des bases MariaDB, contient l'ensemble des données utilisateurs de notre outil. Repman nous garantit ainsi une sauvegarde multi-base de ces données cruciales ; mais aussi une scalabilité certaine lorsque le nombre d'utilisateurs augmentera.

**Biscuit** nous permet, quant à lui, de gérer les utilisateurs et les droits d'accès. Il s'agit donc d'un outil JWT, permettant de gérer l'authentification des utilisateurs et leur donnant ainsi accès à l'interface via un mécanisme de jeton géré par le navigateur web.

* 1. Fonctionnalité d'émergence

Le système d'émergence permet d'identifier automatiquement des entreprises à partir d'une description d'activité. L'utilisateur saisit une phrase exprimant l'activité recherchée, et le service fait remonter les entreprises dont les activités correspondent le mieux à cette description.

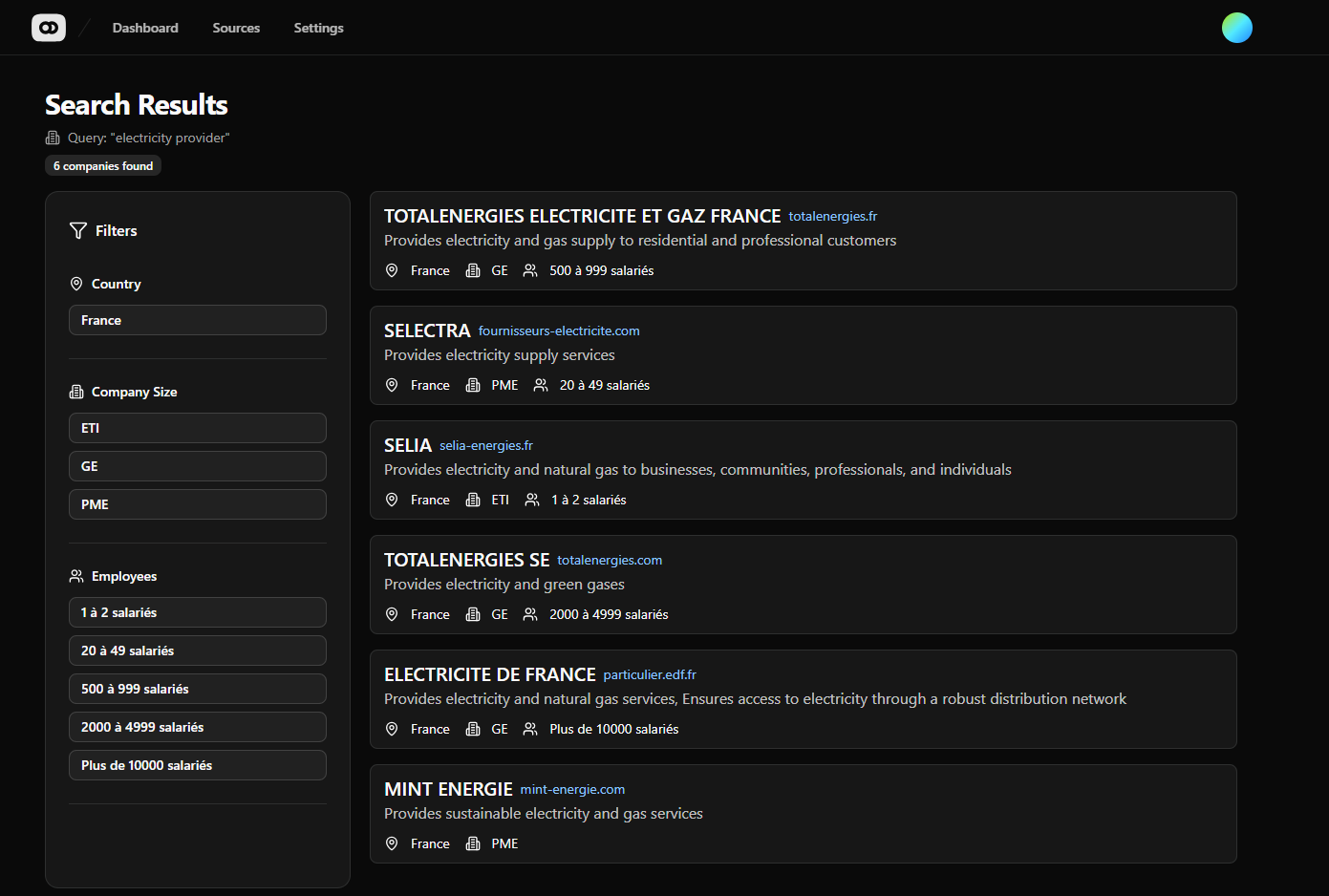
Le processus débute par une analyse de la requête utilisateur afin d'en extraire les éléments essentiels. Le système identifie l'activité principale, reformule les termes clés et distingue les informations contextuelles comme la localisation ou le type d'entreprise. Cette étape garantit une compréhension claire de la demande, même lorsqu'elle est formulée de manière libre ou imprécise.

L’activité identifiée est ensuite comparée à une base d’activités d’entreprises. Grâce à une approche sémantique, le système évalue la proximité de sens plutôt que la simple similarité des mots. Cette recherche permet de détecter des correspondances pertinentes, y compris lorsque les termes diffèrent, et d’attribuer un score de similarité à chaque entreprise, traduisant la cohérence entre sa description et la requête initiale.

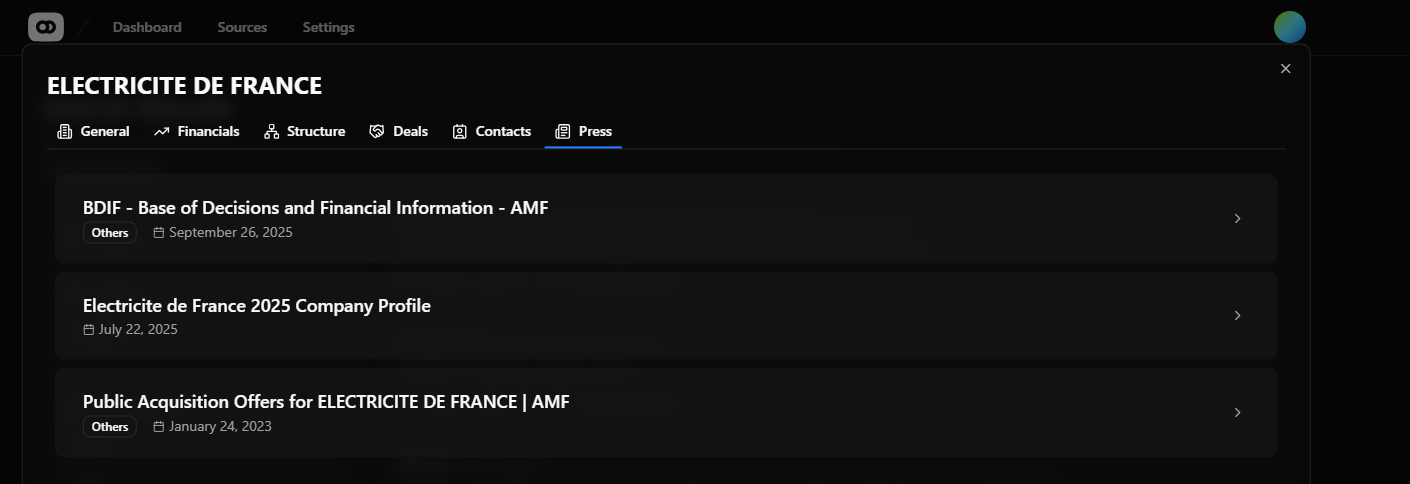
Enfin, les entreprises retenues sont enrichies et reclassées grâce à l’analyse des mots-clés. Le système croise les mots-clés générés à partir de la requête avec ceux stockés en base, issus de la structuration des informations provenant des sites web des entreprises. Ce croisement affine la pertinence et aboutit à un score hybride, combinant la similarité sémantique et la correspondance des mots-clés. Ce score global garantit que les entreprises les plus pertinentes remontent en tête des résultats.

* 1. Exemple de résultats.

En cherchant des fournisseurs d'électricité (rappel : notre base est pour l'instant limitée en volumétrie, les résultats ne peuvent donc pas être exhaustifs), on tombe sur les principaux fournisseurs facilement et simplement. On peut alors les filtrer pour ne garder que les grandes entreprises, par exemple.



Également, on retrouve certains articles de presse parlant d’éléments financiers de ces entreprises :

****

* 1. Organisation des codes

L’outil s’appuie sur un ensemble de 15 dépôts Git, dont certains ne sont que des POC ayant pour but de dérisquer des sujets divers.

Ms-app : contient le front-end et le back-end de l'application.

Ms-company : micro-service principal permettant les appels à d'autres micro-services spécifiques.

Ms-news : micro-service responsable de l'orchestration complète du processus d'ingestion de la presse.

Ms-textlab : micro-service assurant les appels aux LLM pour des tâches dédiées au traitement de texte telles que l'analyse, le résumé, la classification ou la reformulation des articles.

Qmd-compass : Contient le système principal d'analyse et de structuration des informations d'entreprises à partir de sites web. Héberge également le service d'émergence des entreprises à partir de leurs activités.

Is\_seo\_projet : Contient l'API de recherche SERP pour récupérer des URLs d'entreprises et actualités financières.

Address\_parser : héberge libpostal pour faciliter le parsing des adresses

Deals\_monitoring : consolide des articles en une base de deals

Ms\_ingestion : microservice d'ingestion des données (récupération et traitement mensuel des informations récentes de l'INSEE + INPI, récupération des fichiers ZIP et CSV -> nettoyage -> stockage intermédiaire -> envoi vers la base principale d'interface pour stockage final)

Ms\_postgres : Lance les dockers Postgres et PostgresVector

Insee\_API : collecte les données de l'INSEE

INPI\_API : collecte les données de l'INPI

HOX\_master\_api : POC de mécanisme agentique pour la recherche d'informations en ligne

OCR\_PaddlePaddle : POC sur les techniques d'OCR

API\_secured\_with\_biscuit : POC sur l'intégration de Biscuit dans l'outil

Ces services sont déployés sur les machines détaillées en partie II-b. Infrastructure et déploiement.

1. **Statut du produit final livré et tâches restantes**

Ainsi, après une année de travail d'une équipe aux profils multiples, nous avons à notre disposition un outil avancé, qui combine des données dont l'accès est compliqué, et qui cherche à les faire parler pour les mettre en valeur.

Cet outil est aujourd'hui déployé et fonctionnel – ce qui était l'objectif de cette subvention - mais doit encore être amélioré légèrement pour représenter un démonstrateur convaincant pour ses futurs clients.

Ainsi, les onglets de l’interface ont vocation à être complétés dans les prochains mois, notamment en les complétant par :

* Leurs transactions financières.
* L'identification des contacts au sein de chaque entreprise.
* Potentiellement leurs marques, des informations sur les structures voire leurs brevets.

**De plus, l'intégralité de ces informations est amenée à être transformée en données de segmentation, et donc être utilisable comme filtres, ce qui représentera un point d'atterrissage suffisant pour être considéré comme un démonstrateur efficace.**

À noter que les principaux défis et difficultés ont été de réussir à faire en sorte que les différents flux arrivent à communiquer entre eux proprement, sans perte d’information. L’architecture de stockage a notamment été cruciale pour que l’interface réussisse à lier toutes ces données, et soit de plus évolutive pour le futur. De plus, le passage à l’échelle (c'est-à-dire la capacité à collecter de l’information sur plusieurs milliers d’entreprises par jour) a été difficile à atteindre, tout en cherchant à garder les coûts d’infrastructure raisonnables.

En complément de ces avancées, il pourrait être intéressant d'envisager :

* Une extension du périmètre : augmenter le nombre d’entreprises présentes dans notre base (45 000 aujourd’hui), idéalement autour des secteurs d’intérêt pour des potentiels clients. L’ambition ultime est de couvrir les 3,5 millions d’entreprises commerciales françaises en activité, et de les mettre à jour régulièrement.
* Finaliser la mise en place d'une base de transactions financières, également utilisable comme critère de segmentation.
* Avancer sur le Business Plan et les premiers contacts commerciaux, notamment pour anticiper les étapes manquantes pour une personnalisation et un déploiement d'une telle solution chez un client.

1. **Annexes & Références**

### ***Études de marché & tendances globales***

Le marché de l’Open Source Intelligence (OSINT) connaît une croissance rapide. Plusieurs sources estiment qu’il pourrait atteindre entre **50 et 90 milliards USD** au cours de la décennie à venir, avec un taux de croissance annuel compris entre 15 % et 25 % (Future Market Insights, 2025; IMARC Group, 2024; Verified Market Research, 2024). Cette croissance est alimentée par la montée des menaces cybernétiques, l’explosion des sources de données accessibles et le besoin croissant d’analyses automatisées et scalables.

Parallèlement, l’adoption des plateformes d’intelligence commerciale (sales intelligence) a un impact significatif sur l’efficacité des équipes commerciales. Selon Markets & Markets (2025), les organisations utilisant ces solutions enregistrent des **taux de gain supérieurs de 32 %** et des cycles de vente plus courts de **28 %**, démontrant l’intérêt stratégique de l’intégration d’outils OSINT et d’intelligence commerciale dans les processus de prospection et de veille.

### ***Analyse concurrentielle***

Plusieurs acteurs se démarquent sur les segments OSINT/IA, veille média et prospection data-driven :

| **Segment** | **Acteurs** | **Positionnement clé** |
| --- | --- | --- |
| OSINT / IA | Energent.ai, Signal AI | Extraction et analyse automatisée de données ouvertes (médias, forums, documents publics). |
| Veille presse / médias | Aleria | Surveillance médiatique et détection de signaux faibles pour la stratégie marketing et communication. |
| Prospection data-driven | Zefram, ZoomInfo | Fourniture de données firmographiques et signaux d’intention pour optimiser les processus de vente outbound ou ABM. |

Ce panorama concurrentiel met en évidence les opportunités de **différenciation** : qualité et couverture des données, intelligence prédictive ou encore conformité réglementaire.

### ***Synthèse et justification***

Les perspectives du marché confirment un essor rapide des technologies d'intelligence open source et commerciale. Dans ce contexte concurrentiel, le projet présenté devra s'appuyer sur des éléments différenciateurs clairs : **qualité des données**, **intelligence prédictive**, **intégration fluide avec les systèmes existants** et **garantie de conformité et de confidentialité des données**.

***Liste des références***

1. Future Market Insights. (2025). Open Source Intelligence Market. <https://www.futuremarketinsights.com/reports/open-source-intelligence-market> [↩](https://chatgpt.com/c/68fba0f3-d62c-832d-959f-7da7c06f4d5e#user-content-fnref-1)
2. IMARC Group. (2024). Open Source Intelligence Market Report. <https://www.imarcgroup.com/open-source-intelligence-market> [↩](https://chatgpt.com/c/68fba0f3-d62c-832d-959f-7da7c06f4d5e#user-content-fnref-2)
3. Verified Market Research. (2024). Open Source Intelligence (OSINT) Market Analysis. <https://www.verifiedmarketresearch.com/product/open-source-intelligence-osint-market> [↩](https://chatgpt.com/c/68fba0f3-d62c-832d-959f-7da7c06f4d5e#user-content-fnref-3)
4. Markets & Markets. (2025). Choosing the Right Sales Intelligence Platform – Buyer’s Guide 2025. <https://www.marketsandmarkets.com/AI-sales/choosing-the-right-sales-intelligence-platform-buyers-guide-2025> [↩](https://chatgpt.com/c/68fba0f3-d62c-832d-959f-7da7c06f4d5e#user-content-fnref-4)