Usages de l'IA avec le Thésaurus INRAE (et vice versa)

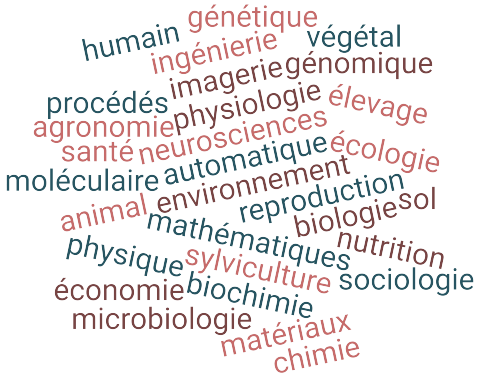
Thésaurus INRAE

Rencontres Hortidoc
18/11/2025 - Sophie Aubin, Magalie Weber, Maël Lesavourey

Notre contexte

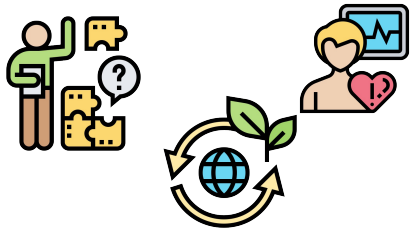
Riche, mais dispersé





INRAE

- **périmètre large**
- **communautés scientifiques diverses**
- **produits de recherche multiples**
  - article scientifique
  - logiciel
  - jeu de données
  - ouvrage
  - rapport
  - mémoire de thèse
- **objets d'étude complexes**



- **cadres de recherche variés**

* Notre problématique

Pour les chercheurs comme pour les décideurs

Exploiter des sources d'**informations variées**, hétérogènes, silotées… pour

- identifier rapidement les **informations pertinentes**,
- repérer les **experts** d'une thématique,
- comprendre les **liens entre les acteurs**

- ✓ *quelles personnes collaborent sur la séquestration du carbone dans le sols?*
- ✓ *quelles sont les contributions de l'Institut sur les transitions agroécologiques ?*
- ✓ *quels sont les domaines qui ont été financièrement les plus soutenus au sein de l'Institut?*
- ✓ *qui travaille sur les zoonoses dans la métropole Rennaise?*

* Nous expérimentons l'IA

- Au sein des communautés scientifiques (depuis longtemps)
- Au sein des directions d'appui

Pour trouver de nouvelles réponses, se décharger des tâches ingrates ou chronophages, pallier la réduction des ressources humaines…

Notre approche

✓ être prudent

✓ respecter la législation et la déontologie

✓ bénéficier de l'effort d'ouverture de la science

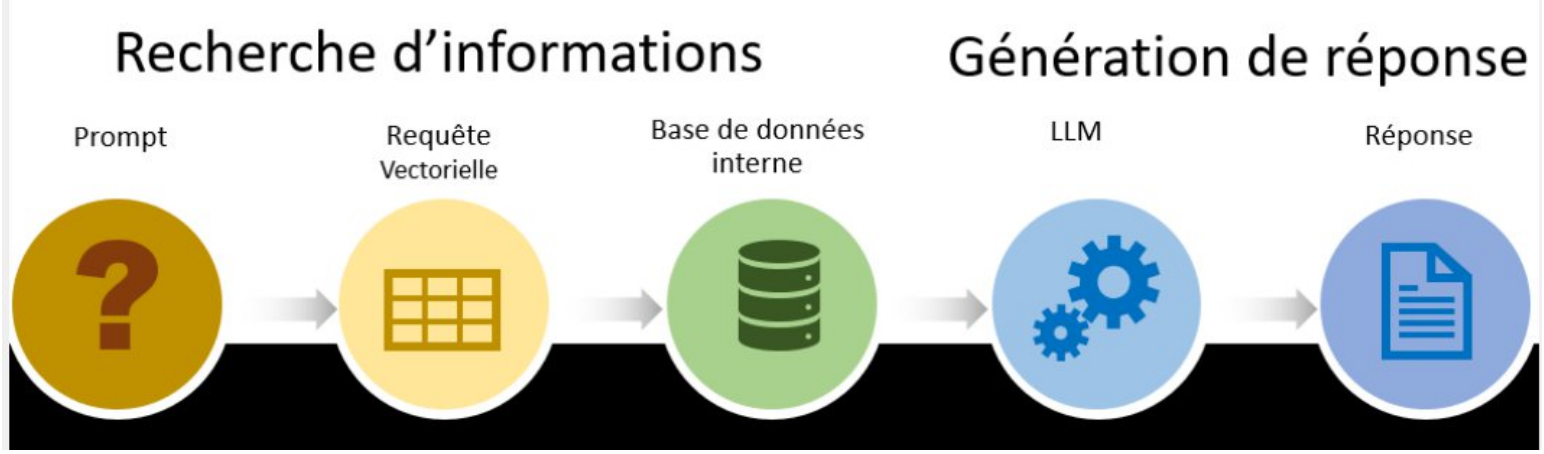
✓ être sobre

✓ envisager des approches hybrides pour tirer le meilleur parti des solutions

Retrieval Augmented Generation (RAG)

L'IA générative appliquée à un corpus maîtrisé

Technique qui **combine les capacités de génération de texte** des modèles de langage de grande taille (LLM) avec des techniques de **récupération d'information structurée (contexte)**, afin de mieux maîtriser le contenu des réponses produites par un LLM.



[Comment fonctionne le Retrieval Augmented Generation (RAG) ?]

Exploration des productions scientifiques à INRAE

à l'aide d'un système RAG basé sur un agent conversationnel

Données (open access)

- Multi-base : HAL et OpenAire (2019-2024), ScanR, BBI…
- Multi-objets: publications (extraits), jeux de données, projets de recherche
- Vocabulaire spécialisé : Thésaurus INRAE

Ce que fait le prototype

✓ Comprend une requête complexe

✓ Identifie les publications scientifiques pertinentes

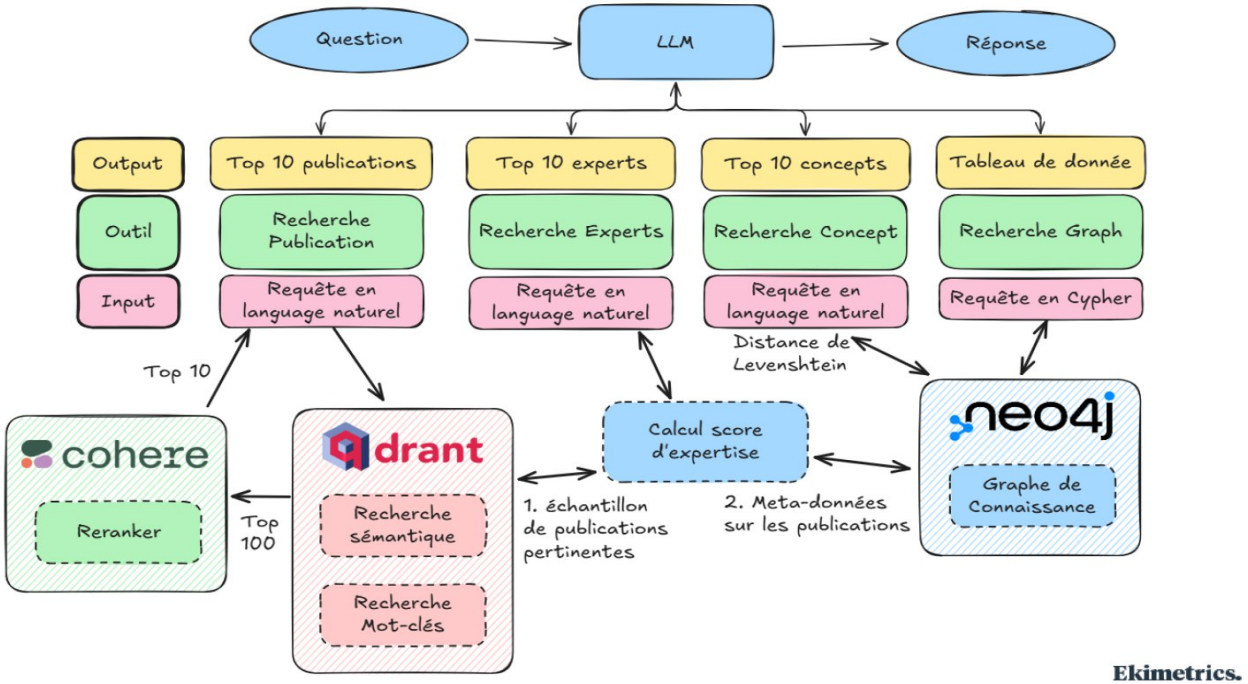
✓ Génère une synthèse sur le sujet

✓ Liste les experts

✓ Montre les collaborations

L'approche agentique

Des agents spécialisés dirigés par un LLM pour plus d'efficacité



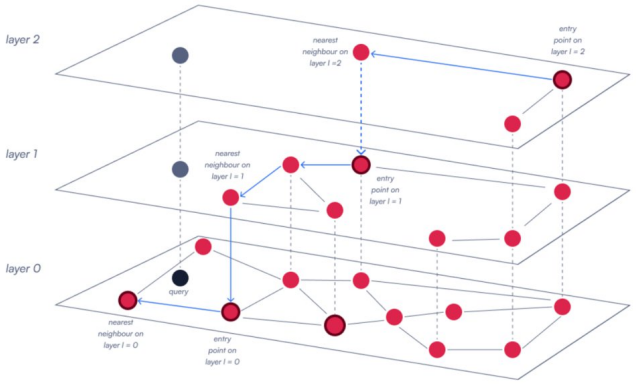
*L'«**orchestrateur**» (basé sur un LLM) détermine par rapport à une question utilisateur :*

- *la stratégie de recherche (NLP classique, requête Bdd, recherche complexe, …)*
- *le(s) agent(s) à mobiliser*

*Les résultats (d'une recherche, d'un agent) sont redirigés vers le **LLM** pour produire une réponse en langage naturelle*

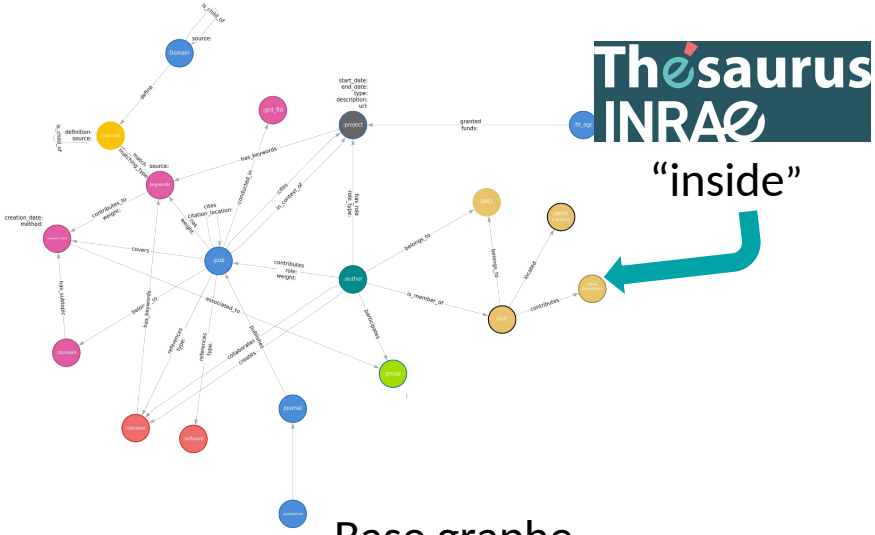
Expérimentation de l'approche Graph RAG

Augmenter la génération LLM avec des connaissances structurées



Base vectorielle

+



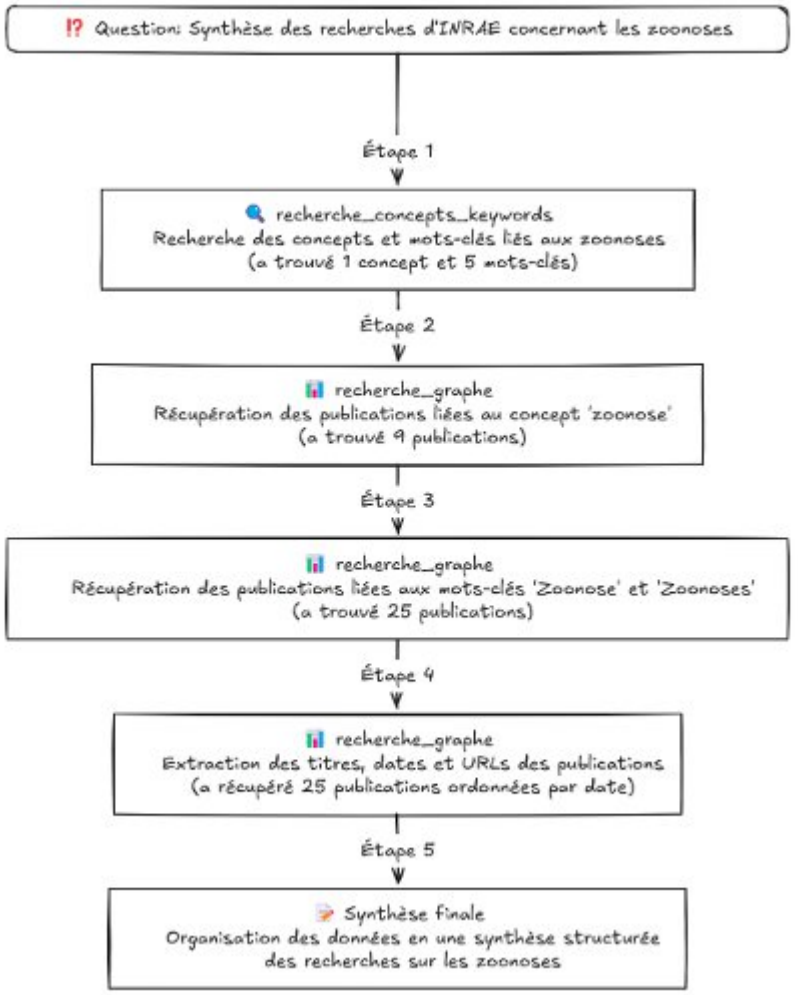
Base graphe

Forces : Simplicité relative de mise en œuvre, rapidité, efficacité prouvée pour améliorer la factualité et l'actualité des réponses pour des questions directes.

Forces : Capacité à modéliser et exploiter les relations, compréhension contextuelle approfondie, amélioration de l'interprétabilité via la visualisation du graphe

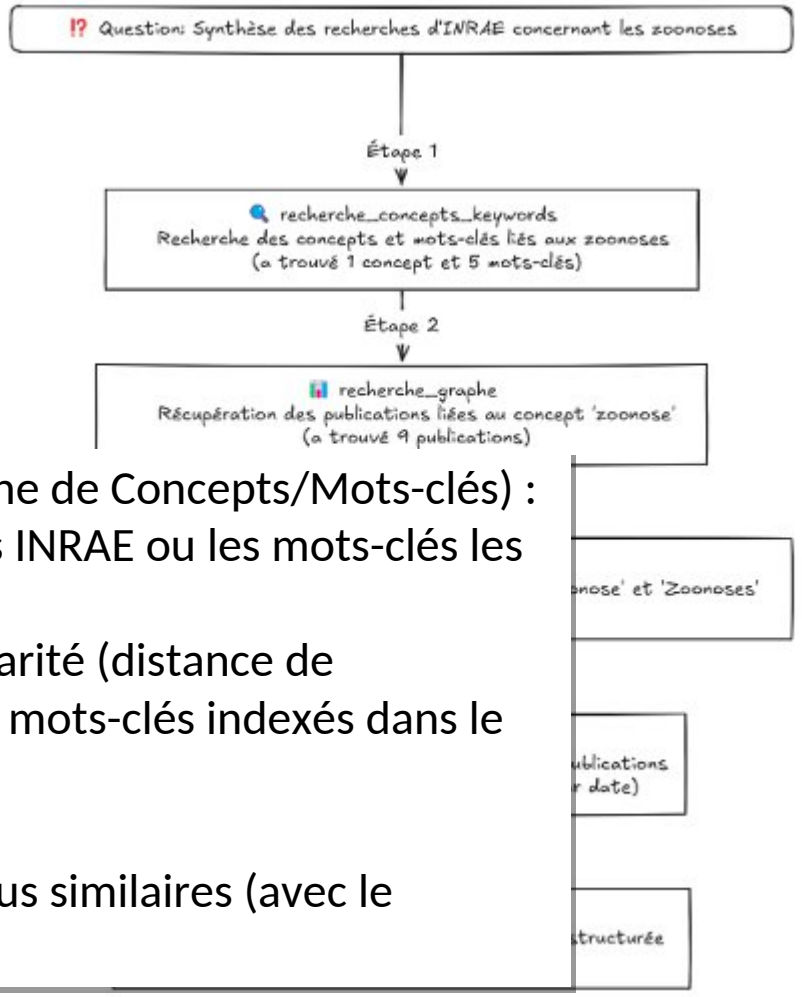
Utilisation : un exemple

- L'agent a commencé par identifier le concept 'zoonose' et les mots-clés associés.
- L'agent a ensuite récupéré toutes les publications pertinentes et les a organisées chronologiquement.
- Ce parcours a révélé que les recherches à INRAE sur les zoonoses, portent sur l'épidémiologie, la surveillance environnementale et la modélisation.



Utilisation : un exemple

- L'agent a commencé par identifier le concept 'zoonose' et les mots-clés associés.
- L'agent a ensuite récupéré toutes les publications pertinentes et les a



Outil recherche_concepts_keywords (Recherche de Concepts/Mots-clés) :

- Objectif : Identifier les concepts du thésaurus INRAE ou les mots-clés les plus proches d'un terme donné.
- Mécanisme : Effectue une recherche de similarité (distance de Levenshtein ou embedding) sur les concepts et mots-clés indexés dans le graphe
- Entrée : Terme ou sujet en langage naturel.
- Sortie : Liste des concepts ou mots-clés les plus similaires (avec le nombre de publications associées).

Limites et perspectives

Liées au Thésaurus INRAE

Améliorer la pertinence de la récupération : combler le fossé sémantique entre la formulation de la requête utilisateur et la manière dont l'information est exprimée dans les documents

Améliorer la qualité des sources : peu de jeux de données, publications… indexés avec le Thésaurus INRAE

- 12 900 / 500 000 notices HAL (<3%)
- 1 024 / 4 800 notices data INRAE (21%)
- 0 projet

—> utiliser l'IA pour assister ou automatiser l'ajout de mots-clés contrôlés

* Rendre le Thésaurus INRAE plus utilisable par les outils basés sur IA

Besoin d'enrichir le Thésaurus avec

- des termes anglais (réalisé en 2024-2025)
- des définitions (projet 2026)

Prochain objectif

Enrichissement semi-automatique du Thésaurus INRAE avec des définitions - Exploiter les techniques de RAG (Retrieval-Augmented Generation) pour collecter/générer des définitions de concepts à partir de publications scientifiques, en assurant la traçabilité des sources.

* Et beaucoup d'autres idées pour continuer à faire vivre le Thésaurus INRAE

- Trouver des équivalents en anglais
- Vérifier la cohérence des hiérarchies
- Vérifier la qualité des descriptions de concepts
- Proposer des emplacements pour les concepts
- Calculer des alignements (correspondances) avec d'autres vocabulaires
- …

Merci de votre attention

Contact : vocabulaires-ouverts@inrae.fr

thesaurusINRAE@inrae.fr