

UNIVERSITÉ SULTAN MOULAY SLIMANE ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES KHOUREBGA

Master IMSD

Développement d'un Outil d'Exploration et Visualisation des Articles Scientifiques

Réalisé par :

M. TAFahi Soufiane

M .KHEMMI Omar

M. HAFIDI Yassine

M .JARMOUNI Abdellah

Encadré par :

M. HAFIDI Imad

03 Juin 2025

Plan

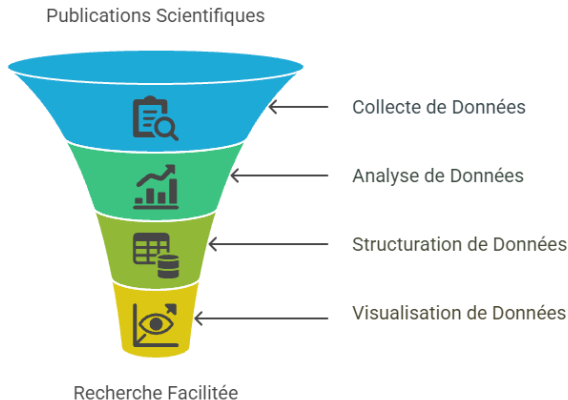
- 1 Introduction
- 2 Étude de l'existant
- 3 Analyse et spécification du besoin
- 4 Conception
- 5 Réalisation
- 6 Démonstartion
- 7 Conclusion

Introduction

Dans un contexte où les publications scientifiques connaissent une croissance exponentielle, il devient crucial de développer des outils permettant :

- L'extraction automatisée des connaissances,
- Leur structuration,
- Et leur visualisation intuitive.

Objectif général :



Contexte :

- Les bases comme arXiv, PubMed ou Semantic Scholar offrent un accès massif aux articles scientifiques.
- Problèmes actuels :
 - Difficulté à naviguer dans le volume croissant de données.
 - Manque d'outils visuels intelligents pour identifier rapidement les concepts clés et les collaborations.

Limitations des solutions existantes

- Recherche basée uniquement sur les mots-clés.
- Peu d'analyses relationnelles ou temporelles.
- Visualisation souvent absente ou rigide.

Analyse et spécification du besoin

Besoins fonctionnels



Recherche avancée

Recherche par auteur, mot-clé, institution



Extraction d'entités

Extraction automatique de concepts, résultats, institutions



Visualisation

Visualisation des relations scientifiques, des tendances temporelles

Exigences non fonctionnelles :

- Interface intuitive et responsive.
- Système modulaire et performant.
- Portabilité (exécutable sur différentes plateformes).

Conception

Architecture générale :

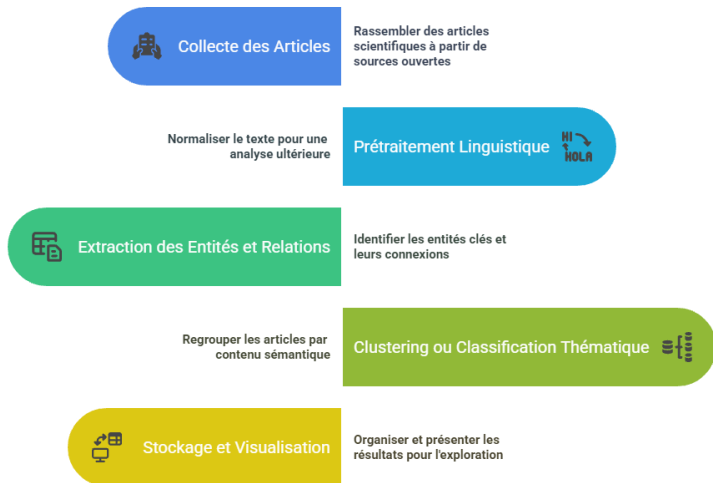
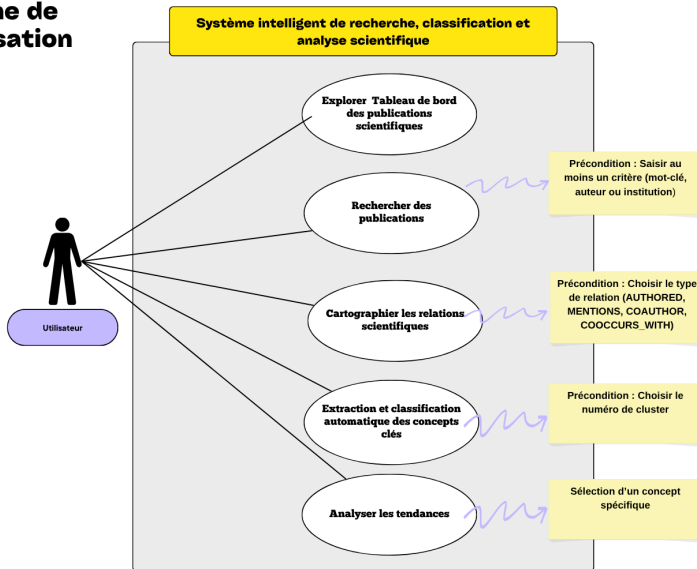
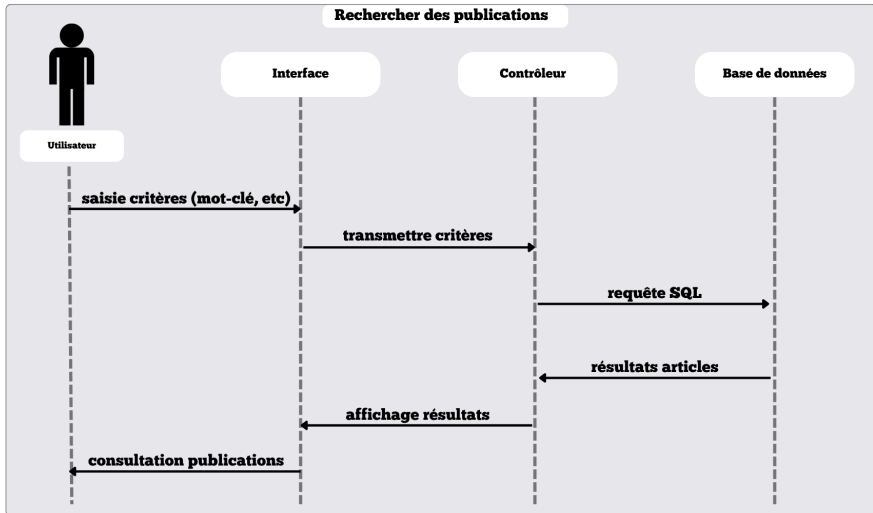
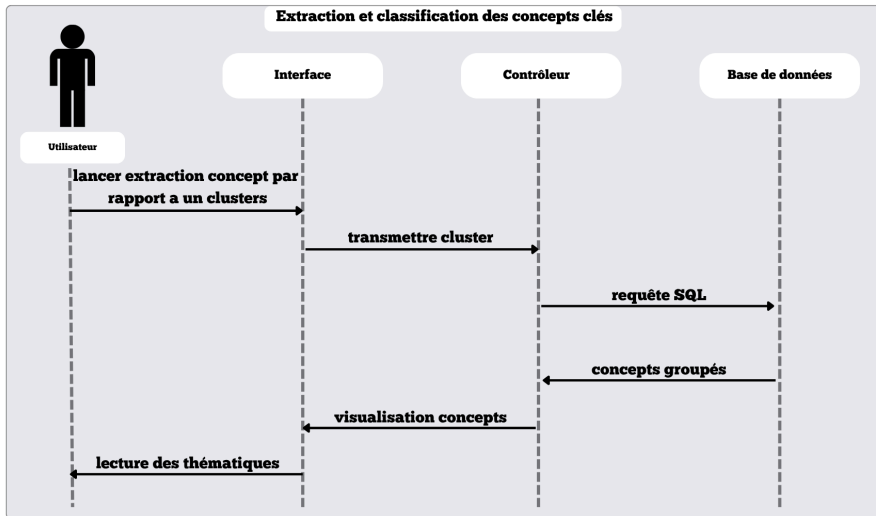


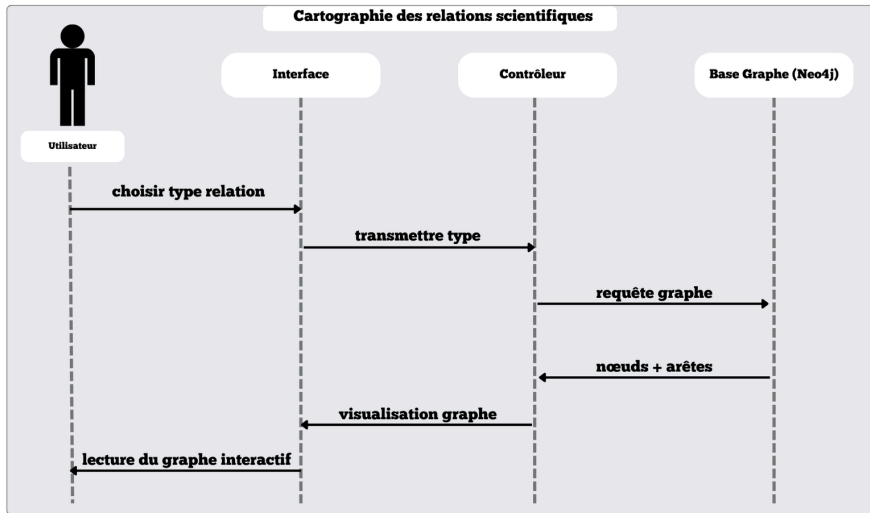
Diagramme de cas d'utilisation

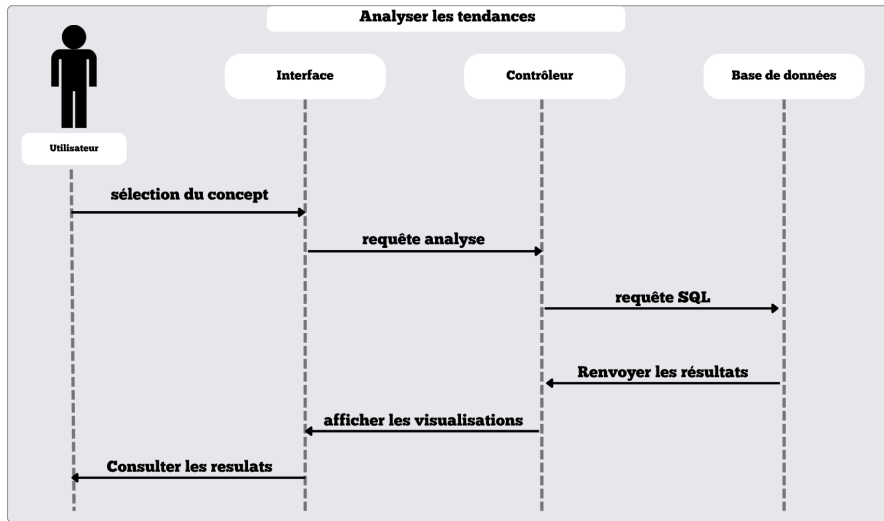




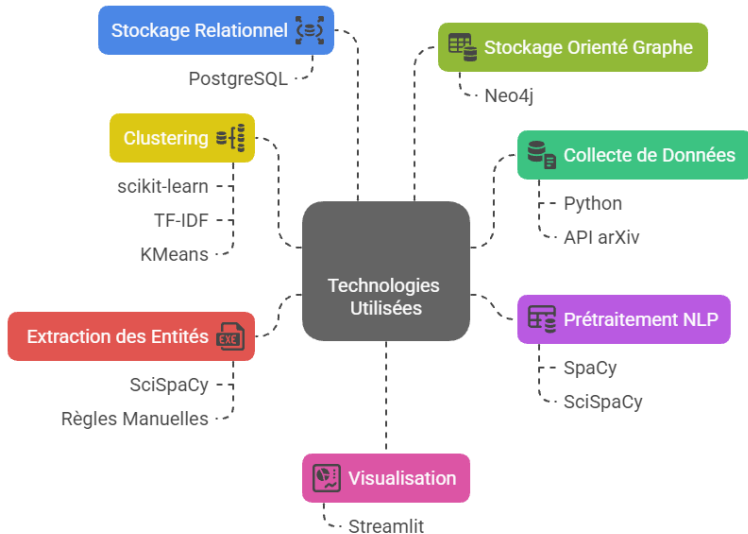


Conception





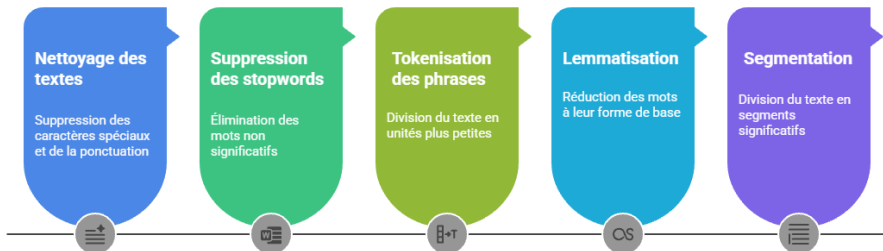
Conception



1 Collecte des données :

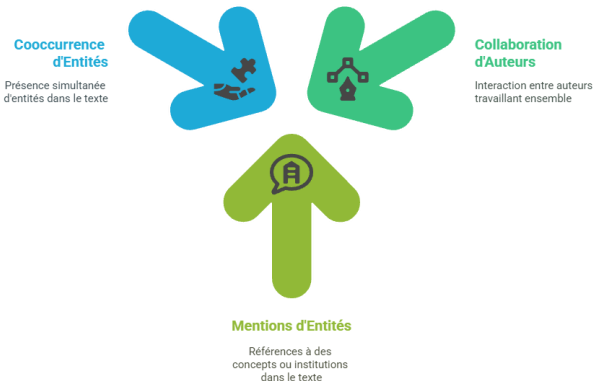
- Extraction de **20 000** articles scientifiques à partir de l'API arXiv, en ciblant plusieurs catégories du domaine informatique (ex : cs.AI, cs.LG).
- Les métadonnées extraites incluent : **le titre, le résumé, les auteurs, la date de publication, et les liens ...**

2 Prétraitement linguistique :



3 Extraction d'entités et de relations

- Utilisation de techniques NER (reconnaissance d'entités nommées) pour identifier automatiquement :
Auteurs, Institutions, Concepts, Résultats expérimentaux.
- Extraction des relations telles que :



4 Clustering thématique automatique

- Vectorisation des textes par TF-IDF sur les concepts extraits.
- Application de l'algorithme KMeans pour regrouper les publications selon leurs thématiques.
- Création de 6 clusters permettant d'identifier les grands axes de recherche représentés dans le corpus.

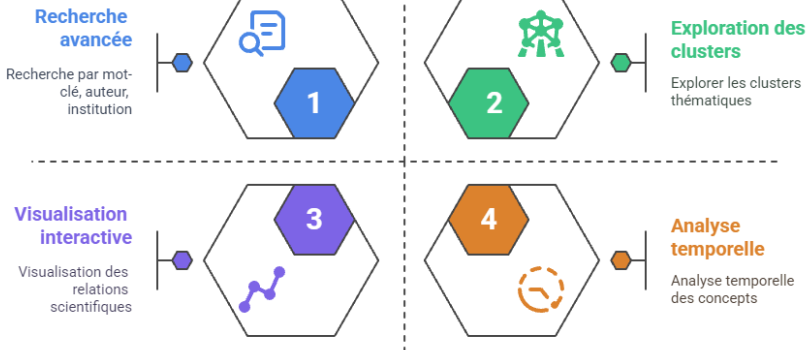
5 Structuration et stockage des données

- Utilisation d'une base relationnelle (PostgreSQL) pour stocker les métadonnées des articles.
- Utilisation d'une base orientée graphe (Neo4j) pour modéliser les entités et leurs relations sous forme de graphe exploratoire.

6 Visualisation interactive :

Développement d'une interface utilisateur simple et fonctionnelle via un tableau de bord web.

Fonctionnalités logicielles



Conclusion

Le projet a permis de concrétiser une solution complète d'analyse scientifique, avec des résultats tangibles en matière d'extraction de connaissances, de structuration sémantique et de visualisation interactive. Nous avons mis en œuvre des compétences transversales en traitement du langage, manipulation de bases de données et conception d'interfaces web.

Perspectives d'amélioration

- Intégration d'autres sources : PubMed, Semantic Scholar
- Amélioration NER (ex : BioBERT)
- Personnalisation de l'interface et analyses
- Déploiement web accessible publiquement

Merci Pour votre attention