



HydroGraphRAG: exploiter les graphes de connaissances pour une interrogation fiable des données de production hydroélectrique en langage naturel

Nassim Boudjenah ⁽¹⁾, **Fatma-Zohra Hannou** ⁽²⁾, **Leila Hassani** ⁽²⁾, **Yann Rebours** ⁽³⁾,
Marie Jubault ⁽²⁾

(1) INRIA, CNRS, Centrale Lille

(2) EDF R&D, Saclay

(3) EDF Hydro- Centre D'ingénierie Hydraulique (CIH)



Sommaire

1. Introduction & Contexte

2. Cas d'usage

3. Approche

4. Implémentation & résultats

5. Conclusion et Perspectives



N°1

Contexte & Motivation

Besoins et périmètre

► Grand Modèles de langage LLM - RAG

Les LLM sont largement populaires pour leurs capacités de génération et de compréhension.

→ Leur déploiement dans des contextes industriels reste limité à cause du risque d'hallucination et leur limite d'adaptation aux contextes métier spécifiques.

Ces limites peuvent freiner leur adoption et déploiement.

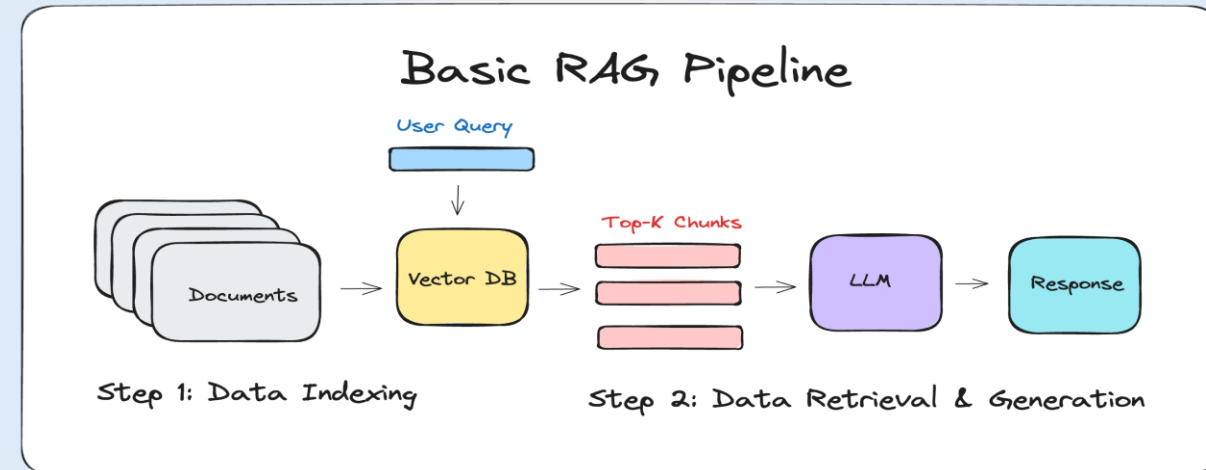
Des solutions existent pour fiabiliser leurs réponses :

- Fine-tuning
- **RAG** (Retrieval-Augmented Generation).

Les systèmes RAG 'classiques' restent limités.

Ils peinent à capturer les **relations** entre différents documents ou à exploiter pleinement des informations de nature structurelle.

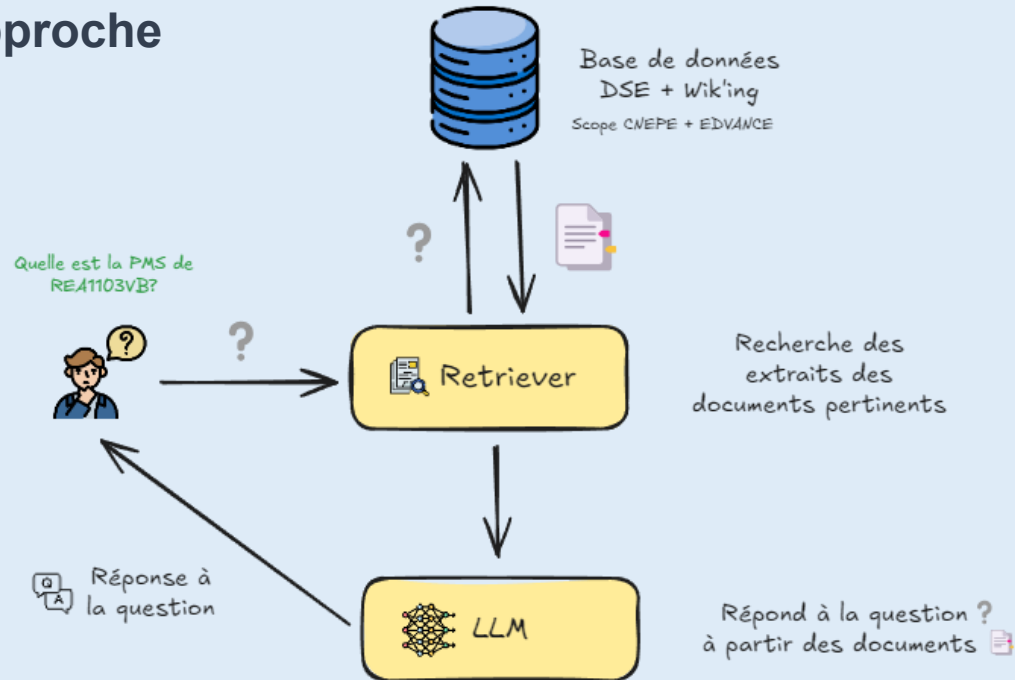
Ils montrent aussi leurs limites dans les **contextes longs**, où la surcharge d'informations empêche de tout restituer de manière pertinente.



► Chatbot Nucléaire

Développement d'un chatbot pour l'interrogation des bases documentaire métier (production nucléaire)

Approche



Système de question-réponse sur DSE (Dossiers de Systèmes Elémentaires) de l'ECM (+ base Wik'ing) avec RAG naïf.

Modèle : Mixtral 8x7B non fine-tuné.

Taux d'hallucination global : 15%
Taux d'hallucination sur les extractions d'information : 11%
Taux d'hallucination sur les synthèses : 25%

Résultats

The screenshot shows the chatbot's user interface. At the top, it says "Recherche sur les Systèmes Elémentaires (Wik'ing, ECM et EASYDOC)". Below this, there is a prompt "Posez votre question." followed by a text input field containing "Quel est le rôle du système SAP ?". There is also a "Champ d'application" (Application field) dropdown menu currently set to "Tous les documents". Below that, there is a section "Filtrer sur les systèmes élémentaires :" with a text input field containing "Appuyez sur 'Entrée' pour rajouter un système". At the bottom, there is an "Envoyer" (Send) button.

Taux de bonnes réponses global : 40%
Taux de bonnes réponses sur les extractions d'information : 33%
Taux de bonnes réponses sur les synthèses : 50%

Motivation

► Evaluation chatbot hydro (RAG classique)

Prompt: Quels sont les équipements des GU ?

Peux tu lister les équipements des GU ?



Portail Multi IA Gen - C2-Interne-Mixtral-8x7b — 10:12

Les Gardes de l'Union (GU) sont équipés d'une armure légère ou intermédiaire, d'un casque, de gants et de genouillères. Ils portent également une ceinture utilitaire avec des outils et des munitions.

En termes d'armes, ils sont équipés d'un fusil d'assaut standard, d'un pistolet de secours et d'une baïonnette. Ils peuvent également être équipés de grenades à fragmentation ou fumigènes.

Les officiers supérieurs des GU peuvent être équipés de pistolets mitrailleurs ou de fusils de sniper

LLM Sans RAG

Peux tu lister les équipements des GU ? dans le contexte production hydraulique EDF ?



Portail Multi IA Gen - C2-Interne-Mixtral-8x7b — 10:14

Dans le contexte de la production hydraulique EDF, les Gardes de l'Union (GU) sont équipés de manière similaire à leurs homologues dans d'autres contextes, mais avec un équipement supplémentaire adapté à leur mission de protection des installations hydroélectriques EDF.

Les GU affectés à la sécurité des centrales hydroélectriques EDF sont équipés d'une armure légère

LLM Sans RAG prompt précisant le contexte



Chat'Hy Méca CUTMH

C2-Interne-Mistral-Large ▼

peux tu lister les équipements par GU ?



Chat'Hy Méca CUTMH - C2-Interne-Mistral-Large — 09:59

Les documents fournis ne spécifient pas explicitement les équipements par Groupe Usine (GU) individuel. Cependant, ils mentionnent les équipements pour des groupes réversibles à axe vertical et des groupes Pelton verticaux. Voici une liste des équipements mentionnés pour ces groupes :

Groupes Réversibles à Axe Vertical

- **Robinet sphérique de garde à l'amont**
- **Pompe turbine :**
 - Tournant à 428 tr/min
 - Fournissant 230 MW sous la chute nette maxi en turbine
 - Refoulant environ 52 m³/s sous la hauteur nette mini de refoulement en pompe
- **Convertisseur statique de fréquence :**
 - Pour le démarrage en pompe
 - En secours, en dos à dos
- **Vanne de garde aval :**
 - Avec deux tabliers
 - Chacun avec deux clapets de sécurité

LLM avec RAG



Motivation

► Evaluation chatbot hydro (RAG classique)

Prompt: Que signifie CUTMH ?

Chat'Hy Méca CUTMH C2-Interne-Mistral-Large ▼

(GU) mentionnés dans les documents ? techniques des équipements des GEH ? mentionnent les équipements des GU ?

que signifie CUTMH ?

Chat'Hy Méca CUTMH - C2-Interne-Mistral-Large — 10:36

Le terme "CUTMH" n'est pas explicitement défini dans les extraits de documents fournis. Cependant, en se basant sur le contexte des documents, il semble que "CUTMH" soit une abréviation ou un acronyme utilisé dans le cadre des communications et des publications techniques relatives à l'hydroélectricité et à la maintenance des équipements hydrauliques. Voici quelques indices contextuels qui peuvent aider à comprendre ce que pourrait signifier "CUTMH" :

- **Communication et Partage de Connaissances :**
 - Les documents mentionnent que les articles CUTMH permettent de partager des expériences vécues et d'accumuler du retour d'expérience (REX) sur des thématiques précises.
 - [Document:CUTMH-2019-Com.-4.13.txt](#)

Message

Cet agent peut faire des erreurs. Vérifiez les informations importantes.

LLM avec RAG

► Les graphes de connaissances

Les graphes de connaissances sont des structures de données qui permettent de représenter la connaissance métier et sa sémantique à travers des nœuds et relations explicites.



L'approche GraphRAG est une approche émergente permettant d'exploiter les graphes de connaissances pour fiabiliser et améliorer les performances des LLMs. Les graphes apportent la structuration des connaissances permettant de mieux cibler l'information pertinente (retrieving).

La **disponibilité** des graphes de connaissances reste un prérequis important dans ce type d'approche.

Questions de recherche



Comment peut-on exploiter les graphes de connaissances dans une approche RAG pour améliorer la précision des réponses générées par les grands modèles de langage (LLM) ?

Verrous techniques

- La transformation naïve des graphes en vecteurs peut **dégrader les liens sémantiques**
- Les graphes de connaissances peuvent être extrêmement large : identification d'un **sous-graphe pertinent** nécessaire
- Déterminer la meilleure manière de représenter le sous-graphe trouvé pour le donner au modèle

N°2

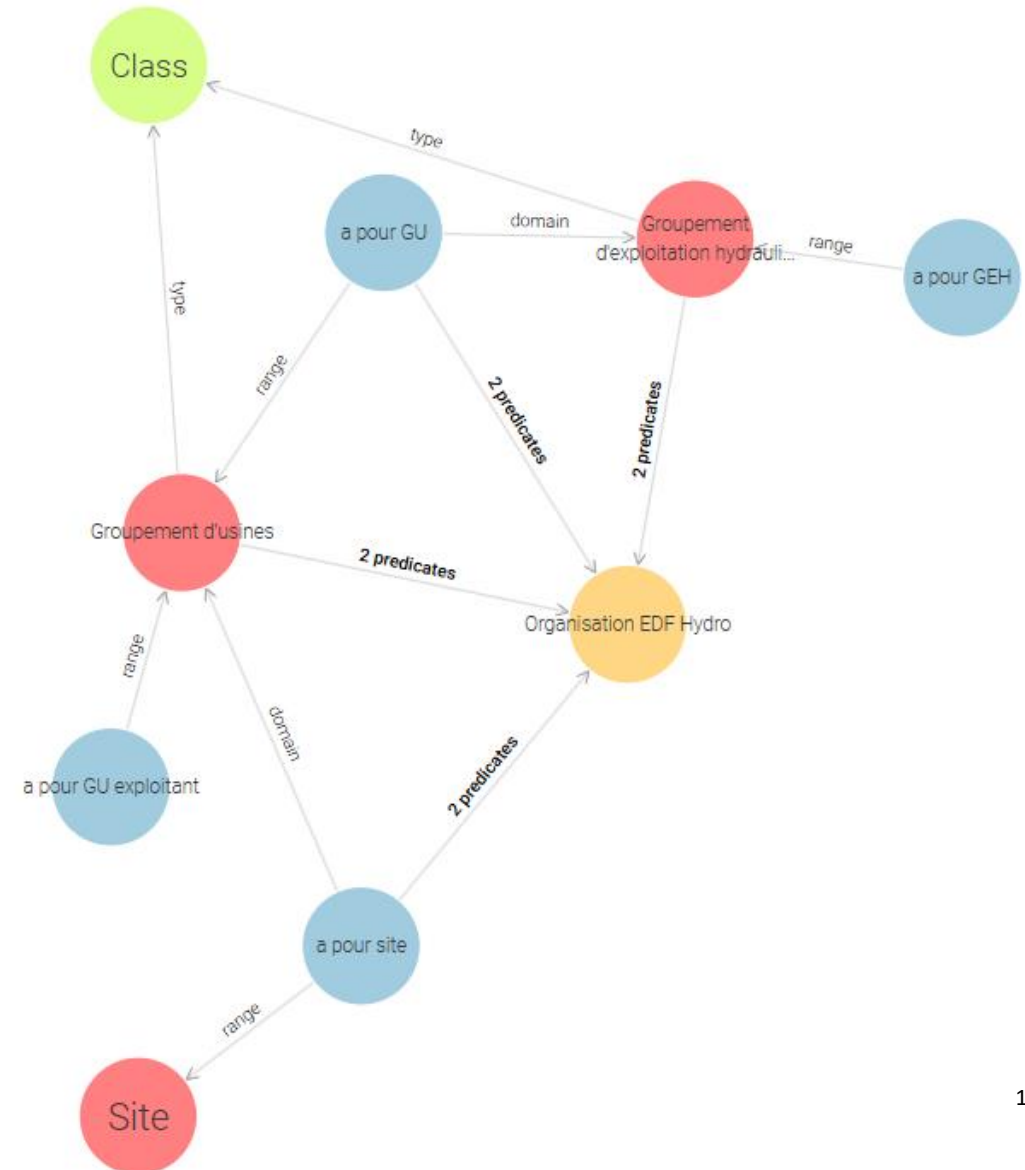
Cas d'usage

Production Hydraulique

Modèle et Méthodologie

Données - Ontologies -

- Données EDF Hydro
- 19 ontologies couvrant les périmètres techniques, RH, production, communications, GED, capteurs
- Il repose sur plusieurs ontologies, chacune définissant un périmètre sémantique propre (ex. : Organisation, BMI). Ces ontologies sont interconnectées, certaines importent et réutilisent des concepts d'autres ontologies du graphe.
- Des ontologies externes sont également exploitées pour enrichir la modélisation

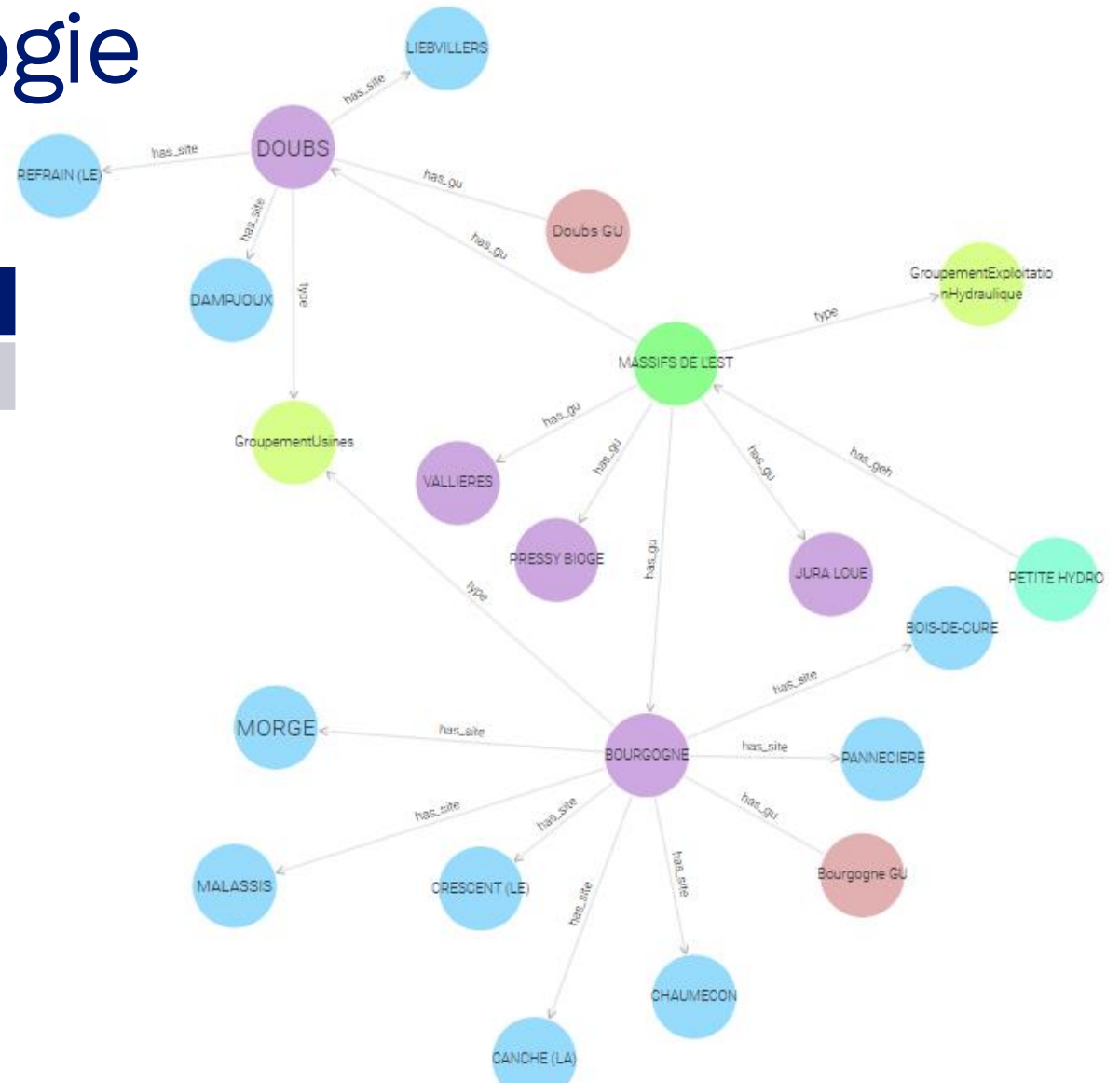


Modèle et Méthodologie

Données - Instances du graphe -

Nombre de Nœuds	Nombre de Relations
11481435	21429347

- ~300 mo d'instances de données sur différents périmètres: infrastructures des sites, usines, leurs caractéristiques techniques.
- Des documents textes sur les communications et les interventions des équipes régulièrement générées.
- Il s'agit également d'un graphe dense, présentant beaucoup plus de relations que de nœuds, ce qui reflète une forte connectivité entre les entités.
- Ces caractéristiques ont un impact direct sur les solutions techniques, notamment en termes de performance et de scalabilité



N°3

Approche

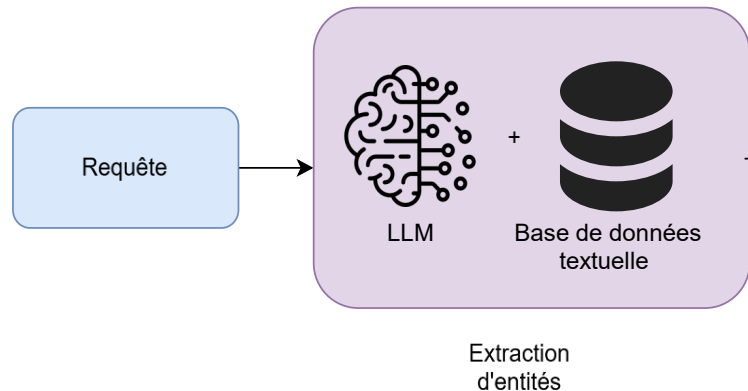
Chaîne de traitements

HydroGraphRAG

► De la question au contexte pertinent

L'approche **HydroGraphRAG** transforme une question utilisateur en langage naturel en un contexte graphe exploitable par le LLM. Elle repose sur trois tâches: **Extraction d'entités**, **extraction de sous graphes** et **prompting du LLM**.

L'objectif de l'approche est d'utiliser la question utilisateur pour extraire un sous graphe pertinent du graphe de connaissance, afin de l'utiliser comme contexte enrichi durant la phase du prompting LLM.



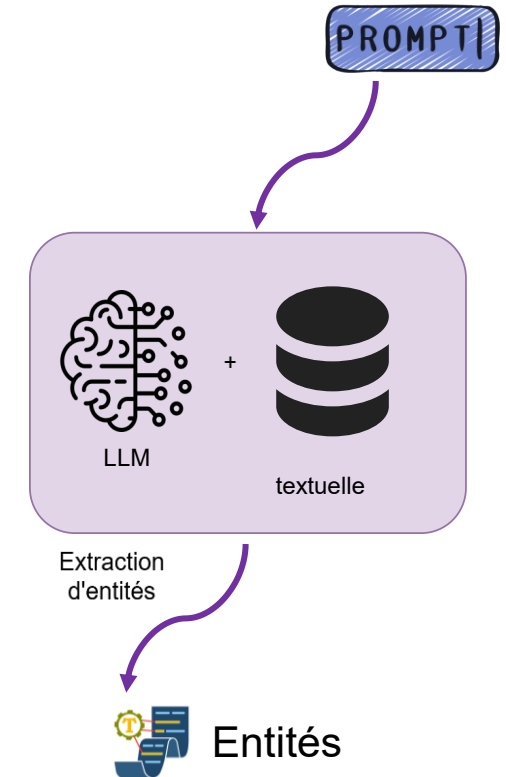
Exemple: Citez tous les sites associés aux groupements d'exploitation hydraulique GEH.

► L'extraction d'entités

Objectif : identifier les **entités pertinentes** présentes dans la requête textuelle pour les relier au graphe par la suite.

Cette requête est traitée par un LLM en **few-shot learning** afin d'extraire des entités susceptibles de correspondre à des entités présentes dans le graphe.

Cette approche est à la fois plus généralisable et moins coûteuse que des alternatives reposant sur des techniques classiques de reconnaissance d'entités nommées (NER).



Entités de l'exemple: Citez tous les **sites** associés aux **groupements d'exploitation hydraulique**.

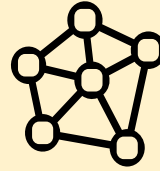
HydroGraphRAG

► L'extraction des sous graphes



Tâche réalisée en deux étapes:
L'appariement et l'identification des sous graphes

Entités



Graphe de connaissance

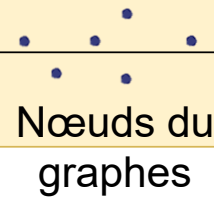
Sous-graphe

Extraction des Sous graphes



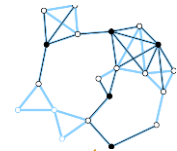
Entités

Appariement



Nœuds du graphes

Identification des sous graphes



Sous-graphe

► L'appariement

Objectif : relier chaque entité extraite au bon **nœud** du graphe pour exploiter le contexte associé.

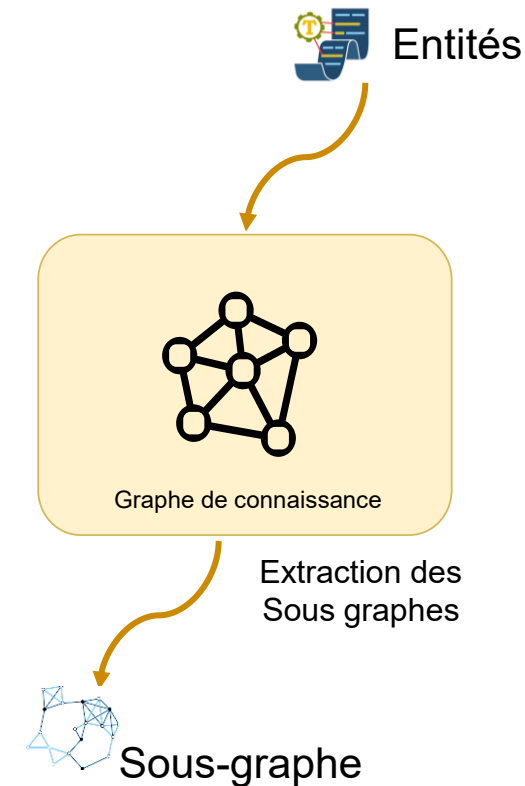
Les approches par embeddings (**Sentence-BERT**) fonctionnent mal car les noms de nœuds sont très courts.

BM25 donne de meilleurs résultats, mais reste limité par les variations linguistiques et les fautes.

Pour améliorer la robustesse, nous combinons **BM25** et **distance d'édition** (en weighted **Reciprocal Rank Fusion**).

$$\text{wRRF}(d) = \sum_{i=1}^n w_i \cdot \frac{1}{k + r_i(d)}$$

Plutôt que de garder un seul nœud, nous retenons les k plus proches afin de mieux gérer les ambiguïtés.



Noeuds de l'exemple:

<https://hydro.data.edf.fr/ontologie/organisation/Site>
<https://hydro.data.edf.fr/ontologie/organisation/GroupementExploitationHydraulique>

► L'identification des sous graphes

Objectif: Capturer les sous graphes pertinents pour créer un contexte enrichi

Les approches naïves (chemin le plus court, graphe minimal) sont insuffisantes car elles ne reflètent pas toujours la pertinence sémantique.

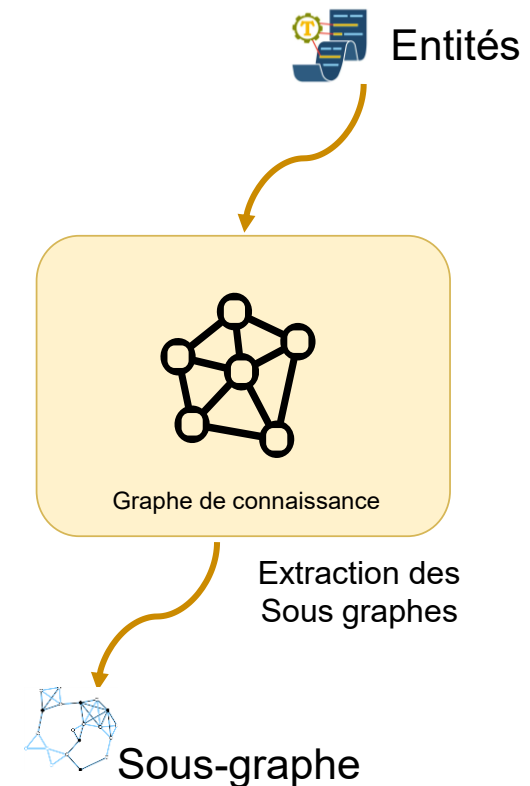
Cette tâche s'effectue en 2 étapes:

- Une **propagation des scores** est effectuée dans le graphe à partir des nœuds candidats retenus
- Exploration **DFS élaguée par ces scores de pertinence** (score transmis = $\alpha \times \text{score} / \text{degré sortant}$)

Les chemins sont conservés uniquement s'ils relient deux nœuds d'intérêt, ou écartés s'ils deviennent non pertinents.

Cette méthode réduit le bruit et améliore la pertinence du sous-graphe extrait.

Le graphe exploré est augmenté par l'**ontologie**



Sous-graphes de l'exemple:
(MASSIFS DE L'EST,has_gu, JURA LOUE)
(JURA LOUE,has_site, CRISSEY)
.....

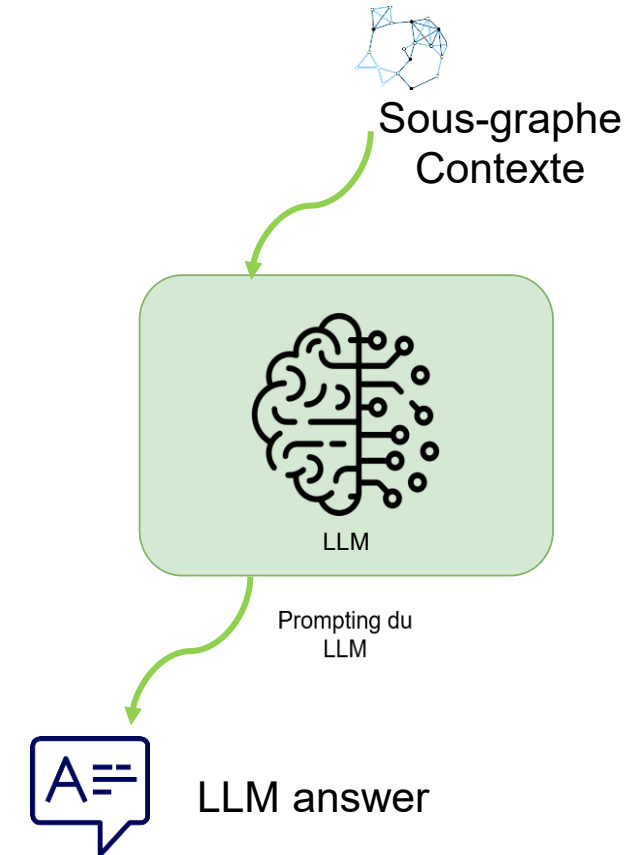
HydroGraphRAG

► Le prompting du LLM

Objectif:

Le graphe est représenté sous forme de triplets enrichis (type, label) pour faciliter la compréhension.

Le LLM génère ensuite une réponse en langage naturel à la requête de l'utilisateur.



La réponse finale (extrait):

Le groupement d'exploitation hydraulique Massifs de l'Est est associée aux sites suivants : SERRE (LA) - RAVILLOLES- ... - CRISSEY - ...

N°4

Implémentation & Résultats

Implémentation

► Pipeline implémentée

HydroGraphRAG est implémentée en trois modules:

1. Extractions d'entités basée sur un LLM (**Qwen2.5-14B-Instruct-1M**)
2. Retrieval graphe, utilisant un appariement lexical et un parcours de graphe avec élagage sémantique
3. Génération finale du LLM (**Qwen2.5-32B-Instruct**) contrainte par le contexte graphe.



► Quelques chiffres

Infrastructure	2-4 GPUs sur DGX (A100/H100)
Hyperparamètres appariement	N=30, k=4, weights= (0.15, 0.85)
Hyperparamètres propagation	Alpha=0.8, min_score=0.05, theta=0.005
Baseline DFS	N_hop=2

► Protocole expérimental

Deux évaluations sont considérées :

Retrieval (évaluation quantitative) :

comparaison directe du graphe extrait avec le **graphe de référence** obtenu via une requête **SPARQL** .

```
3 PREFIX ghs: <https://hydro.data.edf.fr/ontologie/ecsh/Egf/>
4 PREFIX ecsh: <https://hydro.data.edf.fr/ontologie/ecsh/>
5 PREFIX ght: <https://hydro.data.edf.fr/ontologie/ecsh/EquipementTurbine/>
6 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
7
8 SELECT DISTINCT ?codeECSH ?riverLabel
9 WHERE {
10   ?eqt a ecsh:Equipement ;
11       gh:code_ecsh ?codeECSH .
12   ?sef gh:has_eqt ?eqt .
13   ?gef gh:has_sef ?sef .
14   ?gef ghs:has_chute ?chute .
15   ?chute ghc:on_river ?river .
```

Pipeline entière (évaluation qualitative):

comparaison avec le résultat du chatbot RAG **documentaire** d'EDF Hydo.

HYDRO-Chat'Hy Méca CUTMH

Fournit un REX mécanique pour l'hydraulique. Se base sur les données CUTMH

[Documents \(2228\)](#) [Réglages IA](#) [Utilisateurs](#) [Paramètres](#)

► Dataset d'évaluation

50 requêtes métier issues des utilisateurs EDF Hydraulique

Requêtes SPARQL définies et exécutées par des experts sur le même dépôt GraphDB

Résultats

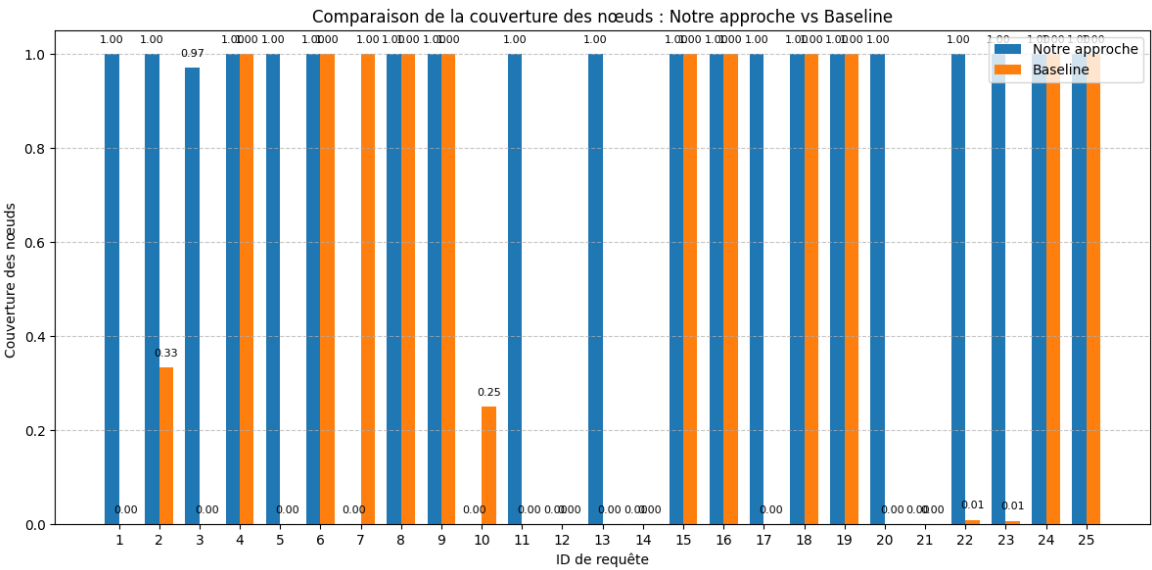
► Approche d'extraction de sous graphes vs. Voisinage direct

La qualité des sous-graphes extraits est mesurée en termes de couverture (infos pertinentes retrouvées) et de bruit (infos non pertinentes), pour les nœuds et les relations.

Comparaison avec une baseline simple : **extraction directe du voisinage des nœuds**, une méthode classique encore largement utilisée

Métrique	Couverture		Bruit	
	Nœud	Relation	Noeud	Relation
Notre approche	0.80	0.78	0.46	0.50
Voisinage directe	0.46	0.42	0.97	0.99

$$\text{Bruit}_{\text{nœuds}} = \frac{|\text{nœuds}_{\text{extrait}} \setminus \text{nœuds}_{\text{ref}}|}{|\text{nœuds}_{\text{extrait}}|}, \quad \text{Couverture}_{\text{nœuds}} = \frac{|\text{nœuds}_{\text{extrait}} \cap \text{nœuds}_{\text{ref}}|}{|\text{nœuds}_{\text{ref}}|}$$



Comparaison Couverture et bruit (notre approche vs. Voisinage direct)

► Analyse quantitative

► Couverture

- Pour **76% des requêtes**, notre approche récupère le contexte de façon précise et fournit une réponse qui correspond exactement à celle attendue.
- Pour **80 % des requêtes**, elle parvient à couvrir la réponse, au moins partiellement avec une couverture supérieure à **50 %**.
- Seules **20 % des requêtes complexes** n'ont pas pu être traitées, en raison d'entités ambiguës difficiles à apparier, entraînant l'échec de l'extraction d'un sous-graphe pertinent, ou l'explosion de la taille du contexte récupéré.

► Bruit

- Pour **32 % des requêtes**, notre approche récupère le contexte exact **sans aucun bruit**.
- Pour **52% des requêtes** le bruit est faible. Les autres requêtes avec un bruit important comportent notamment des cas d'échec d'appariement ou de graphes très dense et volumineux.



Notons toutefois que le bruit **n'est pas aussi critique** que la couverture. Même avec un contexte bruité, le LLM reste capable de fournir une réponse correcte, exacte, tant que les **éléments pertinents** sont présents.

Résultats

► Analyse quantitative

Requête 3:

- Est-ce que le site de Couesque contient des instruments du fabricant Endress+Hauser ?

C'est une requête où le contexte concerne deux instances

Méthodes	Nombre de Nœuds	Nombre de Relations
Requête SPARQL	6	7
Notre approche	53	52
En commun	6	7

Réponse de notre Graph RAG

Oui, le site de Couesque contient plusieurs instruments fabriqués par Endress+Hauser.
Voici quelques exemples : PROMAG 50W ou 53W, CERABAR T PMCPMP131, et PROSONIC FLOW 91.
Il y a donc bien des instruments du fabricant Endress+Hauser sur ce site.

Malgré un contexte bruité, le chatbot parvient à renvoyer une réponse finale exacte sans superflus.

► Limitations

On distingue deux types de requêtes complexes où l'extraction échoue :

Ambiguïtés d'entités : par exemple, les entités *managers* et *auteurs* sont toutes deux des nœuds de type *personne* dans le graphe, ce qui rend l'appariement difficile.

Propagation sur nœuds très connectés : lorsque le sous-graphe traverse des nœuds à forte connectivité (ex. *sites*), les scores de pertinence chutent trop vite, même pour des nœuds pourtant pertinents

Requête 5: Recupère les communications CUTMH ou M. VOLLE est mentionné.

C'est une requête sur le contenu des communication CUTMH

Méthodes	Nombre de Nœuds	Nombre de Relations
Notre approche	0	0

Réponse de notre Graph RAG

Je ne dispose pas d'assez d'informations pour répondre à cette question.

► Analyse qualitative

Résultats préliminaires de l'évaluation effectuée sur quelques requêtes utilisateur (10).
L'évaluation a été effectuée manuellement.



HYDRO-Chat'Hy Méca CUTMH Mistral Small 3.1

Voici la liste des Unités de Production (UP) et les Communications CUTMH associées, extraites du contexte fourni :

- **UP Alpes**
 - **CUTMH-2006-Com.-4.62.txt** : Les réunions se tiennent avec une périodicité de 24 mois, en alternance COCCHE et CUTMH, puis COGECH et COMEP. Les UP Alpes ont organisé des réunions en 2004, 2007, 2011, 2014, et 2017.
- **UP Centre**
 - **CUTMH-2006-Com.-4.62.txt** : Les réunions se tiennent avec une périodicité de 24 mois, en alternance COCCHE et CUTMH, puis COGECH et COMEP. Les UP Centre ont organisé des réunions en 2004, 2008, 2011, 2015, et 2018.
- **UP Est**
 - **CUTMH-2006-Com.-4.62.txt** : Les réunions se tiennent avec une périodicité de 24 mois, en alternance COCCHE et CUTMH, puis COGECH et COMEP. Les UP Est ont organisé des réunions en 2005, 2008, 2012, 2015, et 2018.
- **UP Sud Ouest**
 - **CUTMH-2006-Com.-4.62.txt** : Les réunions se tiennent avec une périodicité de 24 mois, en alternance COCCHE et CUTMH, puis COGECH et COMEP. Les UP Sud Ouest ont organisé des réunions en 2005, 2009, 2012, 2016, et 2019.

Voici la liste des Unités de Production et des Communications CUTMH associées :

****ALPES UniteProduction****

- Communication [CUTMH-2002-Com-1-54]
- Communication [CUTMH-2002-Com-2-57]
- Communication [CUTMH-2002-Com-2-42]
- Communication [CUTMH-2004-Com-4-07]
- Communication [CUTMH-2009-Com-1-02]
- Communication [CUTMH-2009-Com-4-56]
- Communication [CUTMH-2011-Com-1-10]
- Communication [CUTMH-2011-Com-1-14]
- Communication [CUTMH-2011-Com-2-07]
- Communication [CUTMH-2011-Com-2-12]
- Communication [CUTMH-2011-Com-4-16]
- Communication [CUTMH-2013-Com-1-05]

Dans **80% des cas** où les deux systèmes répondent, la réponse de HydroGraphRAG est plus exhaustive comparé au RAG documentaire (échec du RAG documentaire sur certains cas).
Pour les requêtes où GraphHydro ne répond pas, le RAG documentaire fournit une réponse partielle, voire aucune réponse.

N°5

Conclusion & Perspectives

Conclusion

HydroGraphRAG est un chatbot combinant les LLMs et les graphes de connaissances, avec une pipeline en 3 modules: **extraction d'entités, extraction des sous graphes, et prompting du LLM.**

Approche GraphRAG : premiers résultats prometteurs avec **76%** de taux de réussite dans la phase de retrieving. **Evaluation quantitative effectuée par rapport aux requêtes SPARQL incluant les requêtes courantes même avec acronymes** : cela permet aux utilisateurs non experts d'interroger le graphe de connaissances et d'obtenir des réponses pertinentes et exhaustives.

Analyse qualitative démarrée par rapport à un RAG documentaire sur le même périmètre, avec des premiers résultats positifs. Pour toutes les requêtes où le chatbot fournit une réponse, celle-ci est plus exhaustive et précise que celle du chatbot RAG classique qui traite les documents séparément.

Limites identifiées: cas de requêtes avec vocabulaire ambigu, où contexte récupéré très volumineux.

La suite ?

Amélioration sur les entités ambiguës ou nécessitant l'extraction de chemins longs et complexes

Evaluation plus exhaustive de la pipeline GraphRAG end-to-end / évaluation par le métier

Système de routage: choisir dynamiquement le meilleur moteur selon la nature de la question et les garanties requises (frugalité, exhaustivité, traçabilité,...)

Moteurs possibles: RAG vectoriel, Graph RAG, moteur de requête,...



Merci

