Pipeline d'Aide à la Découverte et l'Utilisation de Données Ouvertes basées sur les LLM

Antoine Dupuy, Nathalie Aussenac-Gilles, Christophe Baehr, Cassia Trojahn

▶ **To cite this version:**

Antoine Dupuy, Nathalie Aussenac-Gilles, Christophe Baehr, Cassia Trojahn. Pipeline d'Aide à la Découverte et l'Utilisation de Données Ouvertes basées sur les LLM. Atelier IACD @EGC 2025 : Intelligence Artificielle Centrée sur les Données, Sana Sellami; Frédéric Flouvat, Jan 2025, Strasbourg, France. 10 p. ⟨hal-05300087⟩

HAL Id: hal-05300087

https://hal.science/hal-05300087v1

Submitted on 6 Oct 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Pipeline d'Aide à la Découverte et l'Utilisation de Données Ouvertes basées sur les LLM

Antoine Dupuy*, Nathalie Aussenac-Gilles*
Christophe Baehr**,** Cassia Trojahn***

*IRIT, Université de Toulouse, UT2, CNRS, Toulouse, France
firstname.lastname@irit.fr,
https://www.irit.fr/
**CNRM UMR-3589, Université de Toulouse, Météo-France, CNRS, Toulouse, France
christophe.baehr@meteo.fr
https://www.umr-cnrm.fr/
***Université Grenoble Alpes
cassia.trojahn-dos-santos@univ-grenoble-alpes.fr

Résumé. L'utilisation des grands modèles de langues (LLM) et d'agents conversationnels pour la réalisation de tâches spécifiques telles que la programmation, la recherche d'information dans des documents, les systèmes de questions/réponses et les systèmes de recommandations, permettent des améliorations significatives dans la réponse aux besoins des utilisateurs. Dans le domaine de la découverte de données ouvertes, en particulier, la compréhension des besoins des utilisateurs finaux est un défi majeur. Des travaux proposent notamment des systèmes de génération augmentée par récupération (RAG) impliquant des ontologies associées à des agents conversationnels pour améliorer la qualité des réponses aux requêtes utilisateur, données par un LLM. Dans cette optique, cet article propose un pipeline d'aide à la découverte de jeux de données basés sur leurs métadonnées, décrites par les ontologies, dont des métadonnées sur utilisation par d'autres utilisateurs. Ce pipeline s'appuie sur des agents conversationnels basés sur le large modèle de langage Llama 3.1 70B et s'appuyant sur une base de connaissance ontologique, DATA-FW. Les résultats sont prometteurs mais de nouveaux travaux doivent être réalisés pour améliorer le système, notamment sur l'extraction de données depuis les plateformes publiques.

1 Introduction

Combler l'écart entre les besoins en données des utilisateurs finaux et les données partagées par les producteurs, qui les génèrent principalement pour leurs propres usages, constitue un défi majeur dans de nombreux domaines. Prenons l'exemple des données météorologiques : le producteur Météo France utilise ses propres données pour prévoir le temps, étudier le changement climatique, analyser l'environnement et créer des produits pour divers secteurs tels que l'agriculture, l'aviation, le ferroviaire et la santé. Cependant, les utilisateurs finaux peuvent avoir des besoins spécifiques et une compréhension particulière des données, comme une entreprise

utilisant des grues à Toulouse, souhaitant connaître le nombre de jours dans l'année où la vitesse du vent est trop élevée pour manœuvrer les grues. Cette divergence entre les besoins des producteurs et ceux des utilisateurs soulève une question clé : comment aligner efficacement les besoins en données des utilisateurs finaux avec les informations issues du vocabulaire des producteurs ?

Les données produites par ces producteurs sont généralement mises à disposition sous forme de volumes importants de données ouvertes sur le web. Elles peuvent être consultées sous des licences ouvertes via différents portails, tels que les portails gouvernementaux pour les données publiques (par exemple, data.gouv en France [1] ou data.gov [2] aux États-Unis, des portails européens comme le Portail Européen des Données [3]), des portails de services publics (par exemple, la Bibliothèque nationale de France [4]), ou encore des portails de données scientifiques comme Copernicus [5] pour les sciences de la Terre. Cependant, ces données sont souvent publiées avec des métadonnées insuffisantes, ce qui complique la tâche d'identifier les jeux de données adaptés aux besoins des utilisateurs, comme indiqué par Ahmad et al. (2024).

Pour répondre à cet objectif, nous proposons un système composé de quatre agents conversationnels basés sur le LLM Llama 3.1 70B et sur l'ontologie de métadonnées DATA-FW [6] qui sert de base de connaissances pour représenter les métadonnées des jeux de données ainsi que sur les informations des portails de données ouvertes (Dupuy et al. (2024)).

La suite de l'article est organisée comme suit. La Section 2 introduit un exemple illustratif de recherche de données, en prenant le cas spécifique d'une entreprise de construction utilisant des grues pour ses activités. La Section 3 spécifie l'implémentation technique du système. La Section 4 indique les différentes étapes de création du système et le pipeline de données. La Section 5 présente les premiers résultats du système. La Section 6 présente les outils et travaux réalisés utilisant le framework RAG pour répondre aux demandes d'utilisateurs dans divers domaines. Enfin, la Section 7 résume les contributions de l'article et discute des perspectives pour les travaux futurs.

2 Exemple d'illustration

Prenons l'exemple d'une entreprise de construction exerçant son activité dans la zone de Toulouse Métropole et recherchant des données sur la force du vent pour calculer le nombre de jours où les grues de l'entreprise ne seront pas utilisables. Dans ce cas, l'utilisateur peut formuler une requête en langage naturel, telle que : "Je recherche des données d'observation fiables de force du vent pour savoir combien de jours dans l'année l'activité de mes grues sera interrompue."

Notre système commence par analyser cette requête à l'aide du LLM pour en extraire les critères essentiels : le domaine (météorologie), le paramètre recherché (la force ou la vitesse du vent), la qualité (les certifications), et l'usage (projets scientifiques, usages similaires). Il peut demander des informations complémentaires à l'utilisateur, comme par exemple la loca-

1. https://www.data.gouv.fr/fr/
2. https://www.data.gov/
3. https://ec.europa.eu/info/statistics/eu-open-data-portal_en
4. https://data.bnf.fr/
5. https://www.copernicus.eu/en/access-data
6. DATA-FW est disponible ici : https://w3id.org/data-fw

lisation des données recherchées ou la période qu'il souhaite analyser. Ensuite, il explore les métadonnées descriptives pour trouver des jeux de données pertinents en cherchant les critères sélectionnés dans la requête utilisateur et filtre les résultats en se basant sur les métadonnées de qualité et d'usage. Cet exemple montre comment l'intégration d'une ontologie et d'agents conversationnels peut transformer la recherche de données.

Dans la Section 3, nous proposons une implémentation technique combinant l'utilisation de l'ontologie DATA-FW et un agent conversationnel à base de grand modèle de langue (LLM).

3 Implementation Technique

La génération augmentée par récupération (RAG) est un framework d'intelligence artificielle qui améliore les capacités des grands modèles de langue en intégrant des sources de connaissances externes (Wiratunga et al. (2024), Gao et al. (2023), Jeong et al. (2024), Hu et Lu (2024), Zhao et al. (2024), Alaofi et al. (2024), Edwards (2024)). Le fonctionnement de RAG repose sur la récupération d'informations pertinentes à partir d'une base de connaissances, que l'on utilise ensuite pour enrichir l'entrée du modèle de langage, afin que le modèle génère des réponses plus précises, actualisées et contextuellement pertinentes. Cette approche permet de dépasser des limites des LLM telles que le risque d'hallucinations dans les résultats. Notre implémentation repose sur (i) l'ontologie DATA-FW en tant que base de connaissances qui modélise les métadonnées des jeux de données à l'aide de l'ontologie DATA-FW, permettant de modéliser les métadonnées des jeux de données et (ii) quatre agents conversationnels basés sur un LLM qui captent et affinent la requête utilisateur (Hoseini et al. (2024)), la compare aux métadonnées représentées dans l'ontologie, récupèrent les données sur les plateformes publiques et construisent une réponse composée des données et d'explications sur le choix des données effectuées par les agents.

3.1 Architecture Proposée

Quatre agents interagissent pour répondre à la requête utilisateur et fournir des données pertinentes dans son cas d'utilisation. Chaque agent a un rôle spécifique dans le traitement de la requête (Cheng et al. (2024)).

Un agent d'interaction utilisateur analyse la requête initiale de l'utilisateur et identifie les informations manquantes, comme des précisions sur la localisation, le format des données (CSV, JSON, GeoJSON, etc.), les critères de qualité ou l'objectif d'utilisation des données. Si ces éléments ne sont pas présents dans la requête, l'agent engage une conversation pour les demander explicitement à l'utilisateur, assurant ainsi une recherche plus ciblée et pertinente.

Un agent de génération de requêtes SPARQL traduit les besoins exprimés par l'utilisateur en requêtes SPARQL en testant les valeurs des critères sélectionnés dans la requête utilisateur sur des triplets présents dans la base de connaissance basée sur l'ontologie DATA-FW. Ces requêtes permettent d'interroger les métadonnées des jeux de données pour identifier celles qui correspondent aux critères de l'utilisateur.

Un agent d'extraction des données se charge d'interagir avec les services associés, tels que des API ou des connecteurs externes, référencés dans l'ontologie. Son rôle est de récupérer uniquement les données nécessaires pour répondre à la demande de l'utilisateur, réduisant ainsi les volumes inutiles et optimisant les ressources.

Enfin, un agent de construction de réponse assemble les résultats sous une forme compréhensible et informative pour l'utilisateur. Il explique les choix effectués par le système, en justifiant pourquoi certains jeux de données ou données spécifiques ont été recommandés. Cette transparence renforce la confiance de l'utilisateur dans les réponses fournies.

Ces interactions entre l'utilisateur, les agents et l'ontologie DATA-FW sont illustrés en Figure 1.

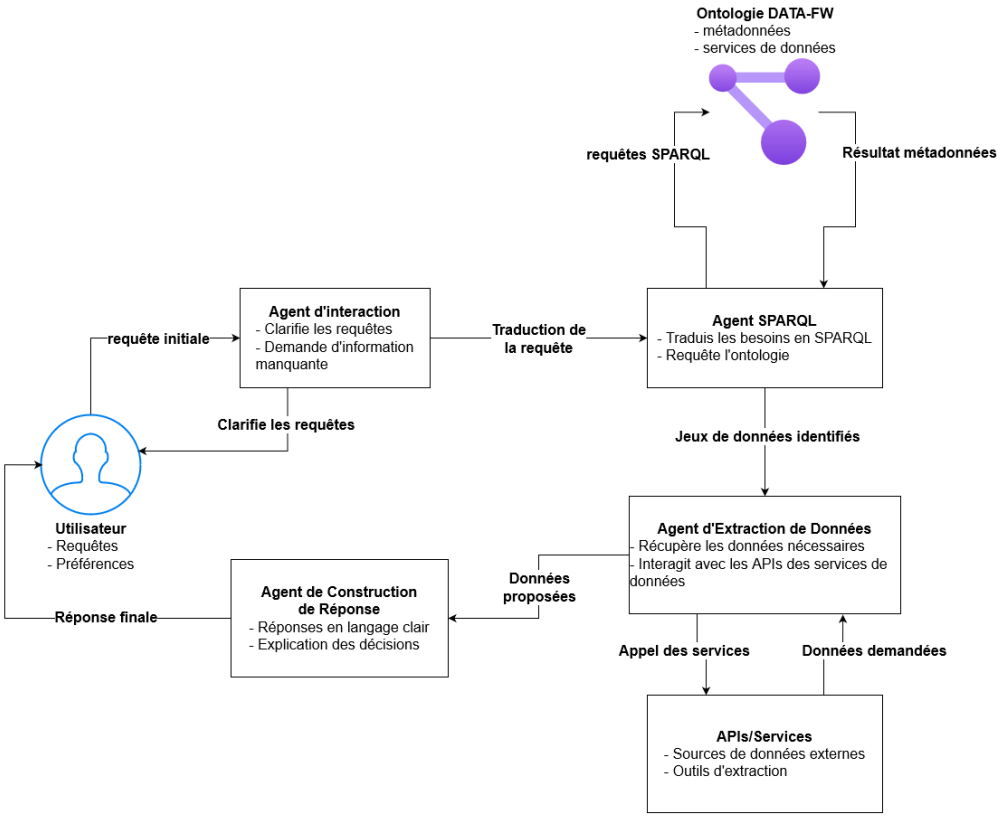

FIG. 1 – Interactions entre l'utilisateur, les agents conversationnels et l'ontologie DATA-FW.

3.2 Ontologie DATA-FW

L'ontologie DATA-FW, proposée par Dupuy et al. (2024), a été modélisée en utilisant Web Ontology Language (OWL). Elle inclut des classes et des propriétés représentant quatre dimensions de métadonnées : les métadonnées de base (descriptive, de provenance, de droits d'accès et d'historique de version), structurelles, qualitatives et d'usage, ainsi que des relations entre ces classes (par exemple, un jeu de données est lié à son producteur ou à des outils compatibles) (Annane et al. (2021), Barbosa et al. (2014)). Pour représenter ces dimensions, elle réutilise plusieurs vocabulaires et ontologies pour la description des métadonnées des jeux de données et leur réutilisation : Data Catalog Vocabulary (DCAT 3), RDF Data Cube, Data

Quality Vocabulary (DQV), Dataset Usage Vocabulary (DUV) et Friend Of A Friend (FOAF). Les métadonnées représentées par l'ontologie DATA-FW ne sont pas spécifiques à un domaine d'étude. Elle peut être étendue par des ontologies de domaine, comme par exemple l'ontologie Semantic Sensor Network (SSN[7]) dans le cadre de données d'observations de la Terre via la représentation des capteurs et des observations (Haller et al. (2018)).

Dans la Section 4, nous expliquons le pipeline de recherche de données, en détaillant les étapes qui permettent la réalisation de l'outil de recherche de données.

4 Pipeline

Le pipeline du système proposé se déroule en plusieurs étapes clés : l'intégration des jeux de données dans l'ontologie DATA-FW, l'interaction de l'utilisateur avec des agents basés sur le grand modèle de langage Llama 3.1 70B, la recherche et le filtrage des jeux de données à partir de la requête de l'utilisateur et la génération de réponses, comme illustré par la Figure 2.

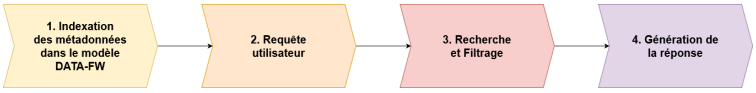

FIG. 2 – Etapes Clés du Pipeline.

4.1 Intégration des Jeux de Données

L'intégration des métadonnées des jeux de données est la première étape du pipeline d'aide à la découverte de jeux de données. Lors de cette étape, les informations relatives aux jeux de données seront extraites de plateformes publiques via les interfaces de programmation d'application (APIs). Le catalogue de chaque plateforme est extrait dans un fichier JSON, comme illustré sur la Figure 3.

Les fichiers JSON sont ensuite traités par un script Python pour structurer et insérer les métadonnées selon le modèle de l'ontologie DATA-FW, que nous présentons dans la Section 3.2.

4.2 Interaction Utilisateur

L'utilisateur interagit avec l'outil en formulant ses besoins en langage naturel. Cette requête peut être raffinée pour obtenir des informations complémentaires. Cette interaction entre l'utilisateur et l'outil permet de construire une requête regroupant les différents besoins de l'utilisateur pour les comparer aux métadonnées structurés dans l'ontologie DATA-FW.

4.3 Recherche et Filtrage

Les agents conversationnels du prototype, basés sur le modèle Llama 3.1, interprètent cette requête pour en identifier les critères principaux, recherchent dans l'ontologie les métadonnées des jeux de données qui correspondent au besoin de l'utilisateur, interroge les services

7. https://w3c.github.io/sdw-sosa-ssn/ssn/

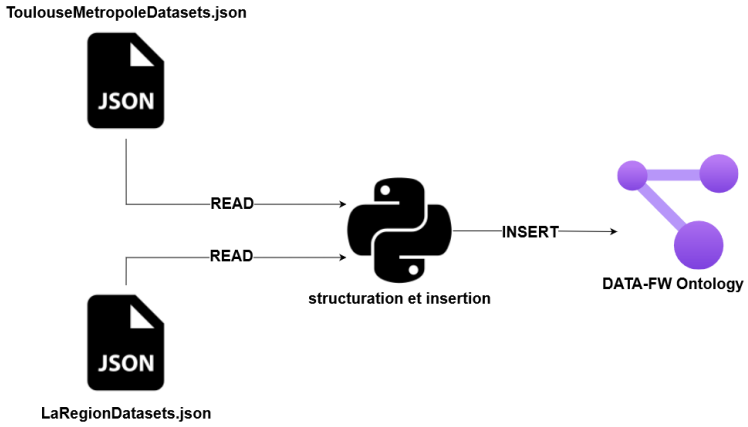

FIG. 3 – Extraction des Métadonnées des Jeux de Données des Plateformes de Toulouse Métropole et de la Région Occitanie

de données pour trouver les jeux de données correspondant. Les résultats sont filtrés et classés en fonction de paramètres spécifiques, par exemple si il y a un besoin exprimé sur des certifications de qualité.

4.4 Génération de Réponses

L'agent de construction de réponses génère une réponse en langage clair, expliquant les choix effectués et proposant des recommandations personnalisées. Ce processus a pour but que chaque utilisateur reçoit des suggestions adaptées à ses attentes, tout en offrant des options pour ajuster les critères ou explorer de nouvelles pistes.

5 Résultats Initiaux

Les résultats initiaux obtenus sont prometteurs. L'utilisation de l'ontologie s'avère efficace pour restreindre la recherche aux données pertinentes contenues dans celle-ci, réduisant ainsi les risques d'hallucinations du système. Les données utilisées pour le prototype sont celles des plateformes de Toulouse Métropole [8] et de La Région Occitanie [9]. Elles ont été récupérées via les APIs de ces plateformes [10] [11]. L'agent d'interaction utilisateur contribue à une meilleure spécification des requêtes en guidant l'utilisateur pour compléter les informations manquantes (Chapman et al. (2020), Gwizdka (2023), Lafia et al. (2024)), ce qui permet de filtrer les résultats et de produire des requêtes SPARQL plus ciblées vers l'ontologie. Par ailleurs, l'agent de construction de réponse fournit des réponses détaillées tout en justifiant les choix effectués par le système. Ces explications permettent à l'utilisateur de mieux comprendre le fonctionne-

8. https://data.toulouse-metropole.fr
9. https://data.laregion.fr/pages/accueil/
10. https://data.toulouse-metropole.fr/api/v1/console/datasets/1.0/search/
11. https://data.laregion.fr/api/explore/v2.1/console

ment de l'outil et obtenir une meilleure visualisation des données proposées, facilitant ainsi ses recherches futures (Peña et al. (2014), Wu et al. (2019)). Cependant, des limites ont été identifiées, notamment avec l'agent de construction de requête SPARQL dont les requêtes SPARQL renvoient des résultats erronés et l'agent d'extraction des données, qui tend à proposer des résultats non pertinents et nécessite une révision de ses instructions. Par exemple, lors d'un test, un utilisateur demandant des données sur la force du vent a reçu des suggestions portant sur des données relatives aux musées parisiens, illustrant un problème de cohérence entre la requête et les données extraites dû à un exemple de requête SPARQL sur la base de connaissance Wikidata de l'agent de construction de requêtes SPARQL.

6 Travaux Liés

Plusieurs systèmes de découverte de données ont été développés ces dernières années, comme Google Dataset SearchBrickley et al. (2019)Benjelloun et al. (2020)Sostek et al. (2024). Ces solutions se concentrent principalement sur les métadonnées descriptives, ce qui limite leur capacité à fournir des recommandations adaptées aux contextes d'usage spécifiques. Notre approche enrichit cette base en intégrant des dimensions comme la qualité des données, leur structure, et les usages antérieurs.

En parallèle, des recherches récentes sur les modèles de langage, notamment dans le cadre des architectures RAG, ont montré leur potentiel pour la génération de réponses complexes et adaptées au contexte de l'utilisateur (Li et al. (2023)). Il existe également des travaux dans divers domaines, comme la recherche d'information dans des documents et des bases de données, notamment dans le domaine du droit (Lála et al. (2023), Wiratunga et al. (2024), Yang et al. (2024b)), dans la recherche de littérature scientifique (Zhang et Kotanko (2024)), la programmation informatique (Yang et al. (2024a)) ou dans l'éducation et le développement des activités commerciales (Posedaru et al. (2024), Alqahtani et al. (2023)), où des systèmes basés sur des agents conversationnels et des modèles de langage sont utilisés pour extraire des informations pertinentes et répondre à des requêtes complexes.

7 Conclusion et Futurs Travaux

Ce système basé sur une architecture RAG offre une solution pour la découverte de jeux de données. En s'appuyant sur l'ontologie DATA-FW et les agents conversationnels alimentés par Llama 3.1, il permet une recherche simplifiée par rapport aux plateformes de données traditionnelles, par le biais d'une recherche en langage naturel, une personnalisation accrue des recommandations par l'intégration des paramètres de qualité et d'usage, et une meilleure compréhension des besoins des utilisateurs par l'intégration d'agents d'intelligence artificielle pour préciser la demande de l'utilisateur.

Cependant, plusieurs axes d'amélioration et de développement sont envisagés. Un benchmark entre différents modèles de langage open source permettra de comparer leurs performances pour optimiser les recommandations. Une collaboration avec la Région Occitanie permettra de définir des cas d'utilisation concrets pour adapter le système à des besoins territoriaux et tester son efficacité en situation réelle. L'intégration de nouvelles plateformes sectorielles ou internationales, comme Eurostat ou des catalogues spécialisés, viendra enrichir le catalogue de

données disponibles. Par ailleurs, les limites actuelles, telles que la dépendance à la qualité des métadonnées initiales et les biais potentiels des modèles de langage, nécessitent une attention particulière pour améliorer la fiabilité des recommandations. Enfin, l'ajout d'agents spécialisés, comme ceux dédiés à la génération de code ou à la recherche d'informations contextuelles sur les données, renforcera les fonctionnalités et l'utilité du système dans divers scénarios.

Références

Ahmad, R. A., J. D'Souza, M. Zloch, W. Otto, G. Rehm, A. Oelen, S. Dietze, et S. Auer (2024). Toward FAIR semantic publishing of research dataset metadata in the open research knowledge graph. http ://arxiv.org/abs/2404.08443.

Alaofi, M., N. Arabzadeh, C. L. A. Clarke, et M. Sanderson (2024). Generative information retrieval evaluation. https ://arxiv.org/abs/2404.08137.

Alqahtani, T., H. A. Badreldin, M. Alrashed, A. I. Alshaya, S. S. Alghamdi, K. Bin Saleh, S. A. Alowais, O. A. Alshaya, I. Rahman, M. S. Al Yami, et A. M. Albekairy (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy 19*(8), 1236–1242.

Annane, A., M. Kamel, C. Trojahn, N. Aussenac-Gilles, C. Comparot, et C. Baehr (2021). SYNOP data evaluation using FAIR maturity model. Research Report IRIT/RR–2021–03–FR, IRIT - Institut de Recherche en Informatique de Toulouse.

Barbosa, L., K. Pham, C. Silva, M. R. Vieira, et J. Freire (2014). Structured open urban data : Understanding the landscape. *Big Data 2*(3), 144–154.

Benjelloun, O., S. Chen, et N. Noy (2020). Google dataset search by the numbers. https ://arxiv.org/abs/2006.06894.

Brickley, D., M. Burgess, et N. Noy (2019). Google dataset search : Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference*, pp. 1365–1375. ACM.

Chapman, A., E. Simperl, L. Koesten, G. Konstantinidis, L.-D. Ibáñez, E. Kacprzak, et P. Groth (2020). Dataset search : a survey. *The VLDB Journal 29*(1), 251–272.

Cheng, Y., C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao, et X. He (2024). Exploring large language model based intelligent agents : Definitions, methods, and prospects. https ://arxiv.org/abs/2401.03428.

Dupuy, A., C. Trojahn, N. Aussenac-Gilles, et C. Baehr (2024). Data-fw : An ontology network for annotating open datasets. In *Proceedings of MTSR '24 : MTSR 2024 18th International Conference on Metadata and Semantics Research*. ACM.

Edwards, C. (2024). Hybrid context retrieval augmented generation pipeline : LLM-augmented knowledge graphs and vector database for accreditation reporting assistance. https ://arxiv.org/abs/2405.15436.

Gao, Y., Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, et H. Wang (2023). Retrieval-augmented generation for large language models : A survey. Version Number : 5.

Gwizdka, J. (2023). *Direct, Orienting, and Scenic Paths : How Users Navigate Search in a Research Data Archive*. ACM Conferences. Association for Computing Machinery.

Haller, A., K. Janowicz, S. J. Cox, M. Lefrançois, K. Taylor, D. Le Phuoc, J. Lieberman, R. García-Castro, R. Atkinson, et C. Stadler (2018). The Modular SSN Ontology : A Joint W3C and OGC Standard Specifying the Semantics of Sensors, Observations, Sampling, and Actuation. *Semantic Web – Interoperability, Usability, Applicability 10*(1), 9–32.

Hoseini, S., A. Burgdorf, A. Paulus, T. Meisen, C. Quix, et A. Pomp (2024). Towards llm-augmented creation of semantic models for dataspaces. In *The Second International Workshop on Semantics in Dataspaces, co-located with the Extended Semantic Web Conference*.

Hu, Y. et Y. Lu (2024). RAG and RAU : A survey on retrieval-augmented language model in natural language processing. https ://arxiv.org/abs/2404.19543.

Jeong, S., J. Baek, S. Cho, S. J. Hwang, et J. C. Park (2024). Adaptive-RAG : Learning to adapt retrieval-augmented large language models through question complexity. https ://arxiv.org/abs/2403.14403.

Lafia, S., A. Million, et L. Hemphill (2024). Exploratory and directed search strategies at a social science data archive. *IASSIST Quarterly 48*(1), online.

Li, X., K. Lv, H. Yan, T. Lin, W. Zhu, Y. Ni, G. Xie, X. Wang, et X. Qiu (2023). Unified demonstration retriever for in-context learning. https ://arxiv.org/abs/2305.04320.

Lála, J., O. O'Donoghue, A. Shtedritski, S. Cox, S. G. Rodriques, et A. D. White (2023). PaperQA : Retrieval-augmented generative agent for scientific research. https ://doi.org/10.48550/arXiv.2312.07559.

Peña, O., U. Aguilera, et D. López-de Ipiña (2014). Linked open data visualization revisited : A survey. submited to the Semantic Web Journal, https ://www.semantic-web-journal.net/system/files/swj937.pdf.

Posedaru, B.-S., F.-V. Pantelimon, M.-N. Dulgheru, et T.-M. Georgescu (2024). Artificial intelligence text processing using retrieval-augmented generation : Applications in business and education fields. *Proceedings of the International Conference on Business Excellence 18*(1), 209–222.

Sostek, K., D. M. Russell, N. Goyal, T. Alrashed, S. Dugall, et N. Noy (2024). Discovering Datasets on the Web Scale : Challenges and Recommendations for Google Dataset Search. *Harvard Data Science Review* .(Special Issue 4), online.

Wiratunga, N., R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. Nkisi-Orji, R. Weerasinghe, A. Liret, et B. Fleisch (2024). CBR-RAG : Case-based reasoning for retrieval augmented generation in LLMs for legal question answering. https ://arxiv.org/abs/2404.04302.

Wu, M., F. Psomopoulos, S. J. Khalsa, et A. De Waard (2019). Data discovery paradigms : User requirements and recommendations for data repositories. *Data Science Journal 18*, 3.

Yang, K., J. Liu, J. Wu, C. Yang, Y. R. Fung, S. Li, Z. Huang, X. Cao, X. Wang, Y. Wang, H. Ji, et C. Zhai (2024a). If LLM is the wizard, then code is the wand : A survey on how code empowers large language models to serve as intelligent agents. https ://arxiv.org/abs/2401.00812.

Yang, X., Z. Wang, Q. Wang, K. Wei, K. Zhang, et J. Shi (2024b). Large language models for automated q&a involving legal documents : a survey on algorithms, frameworks and

applications. *International Journal of Web Information Systems 20*(4), 413–435.

Zhang, H. et P. Kotanko (2024). #1506 uremic toxicity : gaining novel insights through AI-driven literature review. *Nephrology Dialysis Transplantation 39*, gfae069–0657–1506.

Zhao, P., H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, J. Jiang, et B. Cui (2024). Retrieval-augmented generation for AI-generated content : A survey. https ://arxiv.org/abs/2402.19473.

Summary

The use of Large Language Models (LLMs) and conversational agents for specific tasks such as programming, information retrieval from documents, question-and-answer systems, and recommendation systems enable significant improvements in meeting user needs. In the field of data search, understanding end-user needs is a major challenge. Some works propose Retrieval-Augmented Generation (RAG) systems that incorporate knowledge graphs, such as ontologies, combined with conversational agents powered by LLMs to enhance the quality of responses to user queries. In this context, our approach introduces a pipeline to facilitate dataset discovery based on dataset metadata, described by the ontologies, and their usage by other users associated with conversational agents from the large language model Llama 3.1 70B supported by an ontological knowledge base, DATA-FW. The results are promising, but further work is needed to improve the system, particularly in data extraction from public platforms.