

Combining LLMs-based Conversational Agents and Ontologies for Open Data Research

Antoine Dupuy¹[0000–0003–4237–2462], Nathalie Aussenac-Gilles¹[0000–0003–3653–3223], Christophe Baehr²[0000–0002–1230–893X], and Cassia Trojahn³[0000–0003–2840–005X]

¹ IRIT, Université de Toulouse, UT2, CNRS, Toulouse, France
firstname.lastname@irit.fr

² CNRM UMR-3589, Université de Toulouse, Météo-France, CNRS, Toulouse, France
christophe.baehr@meteo.fr

³ Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, Grenoble, France
cassia.trojahn-dos-santos@univ-grenoble-alpes.fr

Abstract. Open Science has significantly increased the availability of heterogeneous scientific datasets. However, these datasets are often described with poor metadata, which makes it difficult to identify data relevant to a specific user’s needs. The pertinent data sets may be hard to find when described with poor metadata, or if users’ needs are expressed using a different vocabulary. This paper proposes an approach that combines semantically enriched metadata with LLM-based agents that interpret natural language queries to manage the gap between users’ needs and dataset descriptions, and to support the retrieval of relevant datasets. It enables the extraction and refinement of user needs, as well as the generation of justifications for the retrieved results. To assess the performance of the proposed system, an evaluation was conducted across multiple Earth Observation (EO) data request scenarios. Four LLM agents have been evaluated (LLaMA 3.3 70B, Mistral Saba 24B, Deepseek-R1, and Qwen 32B) using metrics such as answer relevancy, contextual precision, recall, and faithfulness. The results, conducted with the DeepEval library with the LLaMA 3 8B model, show relatively high scores for answer relevance and contextual precision, especially with the LLaMA and Deepseek-R1 models.

Keywords: Semantic Metadata · LLM · LLM-based agent · Dataset retrieval · Query interpretation · Knowledge graph.

1 Introduction

Over the past few years, public and research institutions have developed policies that encourage the sharing of scientific data through open-access infrastructures. Open data portals offer access to a variety of types of data (environmental, geospatial, sensor-based data, etc.). Yet, they remain challenging to navigate, and their content often lacks sufficient contextual information for users to assess their relevance [23]. An obstacle lies in the lack of fully described metadata and

the heterogeneity of metadata schemes [29, 22]. These schemes differ across institutions, projects, and domains, often reflecting distinct conceptualizations and terminologies. In the field of Earth Observation (EO), for instance, this challenge is compounded by the multidimensional nature of the data (spatial, temporal, thematic, and technical), which increases the need for structured, interoperable, and semantically rich metadata [27, 20, 13] and the mismatches between the formal descriptions used by data producers and the natural-language expressions of user needs. Research in dataset search and semantic technologies has explored the use of ontologies [15, 7], and, more recently, natural language interfaces to improve data access and query formulation. Ontologies provide formal vocabularies for describing a domain of interest. At the same time, advances in Large Language Models (LLMs) have enabled new ways to interpret user input and generate responses in flexible, human-readable formats.

This paper proposes K2K (Knowledge-To-Knowledge), a system that integrates LLM-based conversational agents with a knowledge graph, addressing the following challenges: (1) harmonization of heterogeneous metadata, datasets originating from a variety of domains, by structuring metadata around a shared ontological backbone. (2) semantic interpretation of user queries, where users express their needs in natural language, using domain-specific or informal terms that do not align with the metadata vocabulary – LLMs can serve as semantic bridges, mapping user input onto the ontology and identifying relevant search parameters; (3) iterative dialogue, where users can articulate their needs in the precise terms expected by the interface – a conversational paradigm, supported by an LLM capable of interpreting incomplete or vague input, can help users articulate their needs more clearly through interactive refinement, (4) providing explanations of the search results relevance, as users must understand why specific results were returned. The K2K architecture is structured using different agents that are in charge of different tasks (an interaction agent, a dataset extraction agent, and a response construction agent). The use case in this paper focuses on a specific type of dataset search, the Earth Observation (EO) data. To evaluate the approach, a comparative study across eight representative EO data search scenarios was conducted. Four language models (LLaMA 3.3 70B, Mistral Saba 24B, Deepseek-R1 70B, and Qwen 32B) were assessed on multiple criteria, including answer relevancy, contextual precision, recall, and faithfulness. We further explored the impact of ontological context on interaction quality through a complementary evaluation scenario.

The rest of this paper is organized as follows. Section 2 discusses the main related work. Section 3.1 presents the K2K architecture. Section 4 describes the evaluation protocol and results, which are discussed in Section 5. Finally, Section 6 summarizes the paper findings and outlines future research directions.

2 Related Work

The use of semantic technologies and natural language interfaces for finding datasets has grown in various scientific fields. Initiatives have aimed to orga-

nize metadata using ontologies [17, 24, 13]. This organization enables semantic search and linked data publishing. However, only a limited number of systems move beyond simply structuring metadata. They actively assist users in interacting, refining queries, and explaining results. Platforms like the European Data Portal and related initiatives use standards like DCAT-AP to share linked metadata across countries [19]. While technically strong, these portals usually target expert users and require knowledge of dataset structure and domain vocabularies. Google Dataset Search uses structured metadata from schema.org to index datasets on the web and include them in search results [8, 6]. However, its discovery method relies on keywords and lacks features to clarify meanings or improve queries. In a related effort to enhance dataset discoverability and semantic connection, Ahmad et al. suggest the ORKG-Dataset content type, an extension of the Open Research Knowledge Graph platform. This offers a FAIR-compliant semantic model for research datasets, directly linking them to their related scientific publications [1].

Several initiatives have addressed the formalization of semantics and compliance with the FAIR principles, in particular for EO and meteorological data. The works of [28] and [4] propose knowledge graph models for Météo-France data and FAIR-compliant meteorological resources that rely on domain ontologies and semantic web standards for spatio-temporal data representation. These models improve data interoperability but remain disconnected from interactive exploration interfaces.

In parallel, the rise of LLM-based systems has led to more flexible, user-friendly data interfaces. Recent work examines retrieval-augmented generation (RAG) architectures that leverage external knowledge bases to enhance LLM performance. Ronzano et al. [25] show that ontological knowledge can improve the connection between user input and structured concepts. Similarly, [3] finds that LLMs perform better on queries based on RDF graphs than on SQL-based systems. However, these systems primarily focus on increasing factual or matching accuracy, rather than enabling complete cycles of human-machine interaction. Within the context of dataset search, conversational or interactive approaches remain rare. Efforts such as PaperQA [21] and Serajah and al. [26], although using LLMs to target scientific question answering and data extraction, focus on literature mining. Crucially, most existing systems neglect the explanatory dimension of information retrieval. Very few explain why particular datasets are selected or how they relate to a user’s original query. This has been highlighted as a critical shortcoming in reviews of generative information retrieval systems and LLM interfaces [2, 12].

K2K builds upon these approaches in several key respects: (1) it uses knowledge graph for EO datasets representation and LLM-based agents to interpret user queries and retrieve suitable, (2) it emphasizes natural language interaction, including iterative query refinement, enabling non-expert users to articulate complex information needs progressively, (3) it incorporates a justification mechanism, summarizing the search criteria used, the selected datasets and articulates why they are relevant to the query.

3 K2K, A System Based on LLM-Based Agents and Knowledge Graph

The K2K system has been designed to address the barriers users encounter when exploring datasets. By integrating LLM-based agents and a knowledge graph, K2K enables users to express information needs in natural language and receive results with system justifications.

3.1 System Overview

The architecture of K2K adopts a modular pipeline to support user interaction, as presented in Figure 1. It is composed of (1) a web interface, through which users submit queries in natural language. These queries are transmitted to (2) a service module, which serves as the orchestrator for the entire interaction. It coordinates with (3) an endpoint module, which hosts the knowledge graph using the DATA-FW and SSN-extended ontologies. This graph provides a semantic context describing EO datasets, including their structure, observed properties, spatial-temporal scope, and quality. Finally, (4) the LLM-based agents module includes two specialized agents. The interaction agent interprets the user’s intent and connects it with the domain knowledge. The response construction agent generates a clear and structured answer, providing justifications on the process.

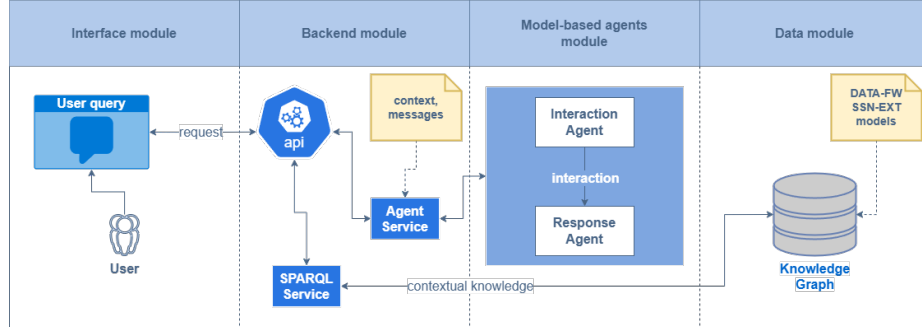


Fig. 1. K2K Architecture Overview

3.2 Ontology Population

The ontology of the K2K system is populated with catalogs from different open data sources (using a Python script). These catalogs, based mostly on DCAT properties, were retrieved in RDF format and matched with the semantic structure of DATA-FW [11] and SSN-extended [10, 18, 5]. When it receives a query, the service module retrieves the knowledge graph, structured according to the

DATA-FW and the SSN-extended models. It supplies it as context to the interaction agent. This agent is responsible for analyzing the user’s request. It is provided with the full semantic context of the EO datasets to interpret the user’s intent and align it with the datasets’ metadata. It can detect when a query lacks specific properties in the knowledge graph that would help refine the search, and it uses the properties already present in the query to select the datasets that best match it.

3.3 User query interpretation

At the core of K2K’s conversational mechanism is the interaction agent, which plays a central role in translating the user’s natural language query into a semantically structured representation suitable for information retrieval. This agent is designed to interpret not only the explicit terms used by the user, but also the underlying intent and contextual expectations conveyed through language.

When a user submits a query, such as “Which datasets could help me study sea surface temperature near the Canary Islands in 2023?”, the service module retrieves the knowledge graph. It transmits it, together with the original query, to the interaction agent.

The interaction agent leverages an LLM that uses the ontology provided in context to extract criteria from the user query. This design follows current research trends, which demonstrate the effectiveness of injecting ontological context into LLM agents to improve their ability to identify relevant concepts and structure [25, 3]. The model identifies key semantic dimensions of the request (the observed phenomenon, spatial coverage, and time frame) and maps them onto the ontology’s structure and vocabulary.

3.4 Response construction

After interpreting the user’s request and retrieving relevant datasets from the interaction agent, the response construction agent builds the final answer. It operates on the interaction agent’s output and the associated metadata from the selected datasets.

This agent is prompted with a detailed instruction set that guides the formatting of the response into five key components: (1) a summary of the user’s search criteria, (2) a list of selected datasets with their names and descriptions, (3) an association of each dataset with the relevant downloadable files, (4) a justification of why these datasets were selected based on the user’s request, and (5) an explanation. The objective is to organize results in a way that enhances user understanding and decision-making, and reduces ambiguity in interpreting them [5, 9].

4 Evaluation

The evaluation of the K2K system was designed to assess the response of its conversational agents, focusing on their ability to interpret user queries, retrieve

relevant datasets, and generate coherent, intelligible explanations. This section presents the evaluation protocol, the test scenarios, and the results for four different LLM-based agents, as well as results without the semantic context.

4.1 Experimental Setup

The K2K system follows a modular, containerized architecture designed to support the interpretation of natural-language queries over metadata. The system is composed of the following modules implementations, all deployed with Docker containers for portability and reproducibility: (1) a ReactJS web-based frontend allowing users to formulate natural language queries, (2) a NodeJS-based backend divided into two core services: the first communicates with LLM-based agents and the second handles the data extraction from the knowledge graph, (3) a Python Flask application that hosts two LLM-based agents, the interaction agent and the response agent, using the Groq inference API to access models and (4) a Virtuoso triple store, which houses structured metadata on EO datasets structured according to the DATA-FW and the SSN-extended models.

A set of eight user scenarios (S1 to S8) was created, each corresponding to a distinct thematic or analytical task commonly encountered in EO applications (Table 1), to benchmark four recent models: (1) LLaMA 3.3 70B, (2) Mistral Saba 24B, (3) Deepseek-R1 70B, and (4) Qwen 32B. These scenarios span climate monitoring, vegetation and soil analysis, hydrology, air quality, and long-term environmental change, thereby ensuring broad coverage of use cases, sensor types, and temporal scopes.

To evaluate the system, a benchmark of 200 natural language queries was built, using EO scenarios S1-S8 for each model. This rotation ensures a balanced distribution of query types and mitigates potential bias from over-reliance on a single scenario. To ensure the integrity of the evaluation, agents were reset between each test to prevent context contamination and eliminate noise propagation across queries.

The evaluation relies on four metrics derived from literature on LLM evaluation [14] [16], aiming to evaluate the model’s efficiency in retrieving information and generating answers. The Answer Relevancy captures how well the model generates answers that are directly pertinent to the query content and address the core inquiry. The Contextual Precision and Contextual Recall metrics measure the specificity, correctness, and completeness of the generated explanations in relation to the actual dataset metadata. Faithfulness assesses whether these explanations are consistent with the factual content of the retrieved resources. These metrics were computed automatically using the deepeval library⁴, which internally uses a local LLaMA 3 8B model (run via the Ollama⁵ framework) to score generations based on reference criteria.

Each model (LLaMA 3.3 70B, Mistral Saba 24B, Deepseek-R1, and Qwen-32B) was integrated into both the interaction agent (responsible for understand-

⁴ <https://github.com/confident-ai/deepeval>, 28/04/2025

⁵ <https://github.com/ollama/ollama>, 28/04/2025

ID	User Request	Objectives
S1	What Sentinel-3 datasets allow for analyzing sea surface temperature in 2023?	Verify that the system correctly selects Sentinel-3 products (SLSTR, OLCI) and explains their spatial and temporal resolution.
S2	I want to study the evolution of terrestrial water storage in France over the past five years.	Test the retrieval of GRACE/GRACE-FO gravimetric data, SMOS soil moisture, and Météo-France precipitation data, and check the explanation of the relationships between these datasets.
S3	What datasets are available to analyze CO2 concentration in France between 2015 and 2023?	Verify correct extraction of Sentinel-5P (TROPOMI), OCO-2, and CAMS, and ensure that their spatial and spectral resolutions are properly justified.
S4	I want to compare the evolution of vegetation biomass and soil moisture in France. What data can I use?	Test the combination of Sentinel-1 (biomass radar), MODIS NDVI, SMOS soil moisture, and evaluate the clarity of the justification.
S5	What Sentinel-2 and Landsat datasets can be used to calculate the evolution of NDVI in France since 2000?	Verify the retrieval of time series from Sentinel-2 (since 2015) and Landsat (since 1982), and test the explanation regarding the relevance of NDVI/EVI indices.
S6	How can we monitor temperature anomalies in France?	Test the retrieval of ERA5, MODIS LST, and Météo-France climate data, and ensure that the explanation includes appropriate scales of analysis.
S7	What datasets allow the study of the relationship between precipitation and flooding in France?	Verify the selection of satellite data (GPM/TRMM), hydrological data from the Copernicus Climate Data Store, in situ data from Vigicrues, and assess the relevance of the explanations.
S8	What Sentinel-5P datasets can be used to assess air pollution in France?	Verify the retrieval of NO2, SO2, O3, and PM2.5 products, and ensure the explanation addresses their limitations and strengths.

Table 1. Evaluation queries and corresponding experimental objectives

ing and reformulating the query) and the response agent (responsible for producing the explanation and justification). A total of 200 queries were distributed across four model configurations: LLaMA 3.3 70B was tested with 75 queries, Mistral Saba 24B and Deepseek-R1 70B with 50 queries each, and Qwen 32B with 25 queries.

In addition to these model-based evaluations, a complementary experiment was conducted to assess the behavior of the conversational agents when the

interaction agent operated without access to the ontological context, to isolate the impact of structured knowledge on system performance [25, 12, 3].

4.2 Results

The evaluation results for each LLM-based agent are summarized in Tables 2 to 5, which report the results for the four metrics used: Answer Relevancy, Contextual Precision, Contextual Recall, and Faithfulness. These results provide a comparative view of how each configuration performs across different aspects of response quality.

Metric	Mean	Median	Std Dev	Min	25%	75%	Max
Answer Relevancy	0.728	0.727	0.139	0.125	0.667	0.800	1.000
Contextual Precision	0.891	1.000	0.236	0.000	0.833	1.000	1.000
Contextual Recall	0.659	0.615	0.170	0.333	0.529	0.732	1.000
Faithfulness	0.722	0.750	0.273	0.000	0.625	1.000	1.000

Table 2. Results on 75 queries interpreted by agents based on the Llama 3.3 70B model.

Metric	Mean	Median	Std Dev	Min	25%	75%	Max
Answer Relevancy	0.685	0.710	0.179	0.300	0.556	0.794	1.000
Contextual Precision	0.825	1.000	0.302	0.000	0.833	1.000	1.000
Contextual Recall	0.608	0.556	0.136	0.444	0.533	0.662	1.000
Faithfulness	0.716	0.800	0.267	0.000	0.644	0.851	1.000

Table 3. Results on 50 queries interpreted by agents based on the Mistral Saba 24B model.

Metric	Mean	Median	Std Dev	Min	25%	75%	Max
Answer Relevancy	0.735	0.750	0.179	0.250	0.625	0.875	1.000
Contextual Precision	0.812	0.833	0.150	0.500	0.750	1.000	1.000
Contextual Recall	0.618	0.611	0.139	0.333	0.556	0.722	0.889
Faithfulness	0.838	0.875	0.172	0.400	0.750	1.000	1.000

Table 4. Results on 50 queries interpreted by agents based on the Deepseek-R1 70B model.

In the dimension of "Contextual Precision", which concerns the ability of the model only to reference entities that are in the knowledge graph in the given context, Qwen agents performed the best with the highest mean score (mean =

Metric	Mean	Median	Std Dev	Min	25%	75%	Max
Answer Relevancy	0.642	0.667	0.184	0.250	0.500	0.750	1.000
Contextual Precision	0.934	1.000	0.122	0.667	0.833	1.000	1.000
Contextual Recall	0.588	0.556	0.160	0.333	0.500	0.667	0.889
Faithfulness	0.783	0.800	0.166	0.500	0.667	0.889	1.000

Table 5. Results on 25 queries interpreted by agents based on the Qwen 32B model.

0.934), followed by LLaMA 3.3 70B (mean = 0.891), Mistral Saba 24B (mean = 0.825), and lastly DeepSeek-R1 70B (mean = 0.812).

In "Contextual Recall", the ability to cover all relevant concepts in the context, LLaMA agents outperformed the others (mean = 0.659), with DeepSeek and Mistral having nearly the same recall value (0.618 and 0.608, respectively), with Qwen having the least recall (mean = 0.588), which could imply that Qwen is omitting information from the context in its response.

For "Answer Relevancy", DeepSeek had the highest mean (0.735), followed by LLaMA (0.728) and Mistral (0.685). Qwen had the lowest relevancy score (0.642), perhaps due to its shorter and less elaborate answers.

Finally, with "Faithfulness" defined as the match between a generated explanation and the retrieved data in the knowledge graph, DeepSeek again had the best mean score (0.838), followed by Qwen (0.783), LLaMA (0.722), and Mistral (0.716). These scores being so close to each other shows that the models, despite their stylistic and depth differences, still retained a good ability to justify and accurately generate outputs.

Figure 2 visually compares these results across models, highlighting the trade-offs among precision, completeness, and explanatory quality.

Notably, LLaMA 3.3 70B and DeepSeek-R1 70B stand out as the best agents, providing strong performance across all areas. Qwen shows promise in generating high-precision results, while Mistral keeps a strong level of faithfulness. This supports the idea that using different strengths of various LLMs could be helpful.

Finally, Table 6 presents the results of an ablation study. In this study, the interaction agent worked without access to the ontological context for LLaMA 3.3 70B. While contextual precision stayed high (mean = 0.934), recall fell sharply (mean = 0.556), and the variability in faithfulness increased. This outcome suggests that the ontology is important for producing complete, grounded results, especially in complex query scenarios.

Metric	Mean	25%	Median	75%	Min	Max
Answer Relevancy	0.713	0.638	0.727	0.755	0.556	0.929
Contextual Precision	0.934	0.833	1.000	1.000	0.756	1.000
Contextual Recall	0.556	0.500	0.529	0.557	0.333	1.000
Faithfulness	0.790	0.750	0.817	0.892	0.000	1.000

Table 6. Results on 20 queries interpreted by agents based on the LLaMA 3.3 70B model without ontological context.

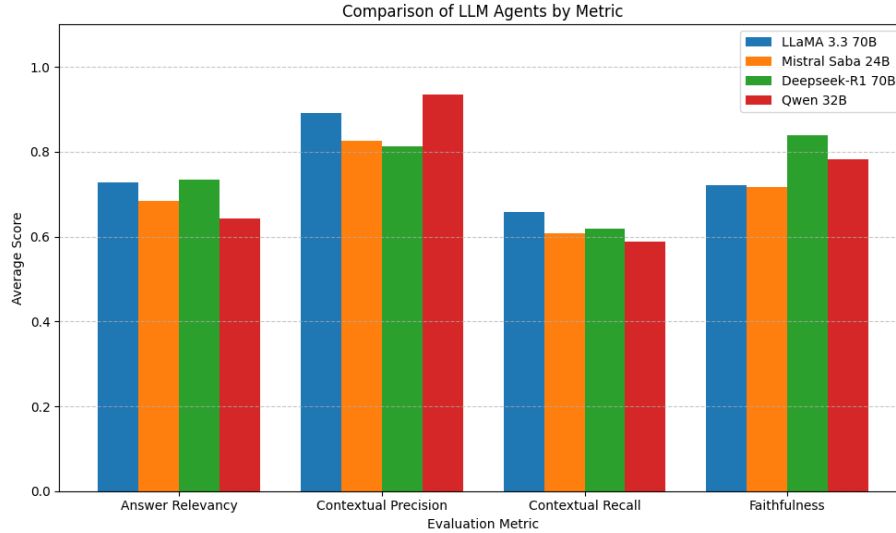


Fig. 2. Comparison of LLM Agents by Metric.

5 Discussion

The evaluation results presented in Section 4 confirm the relevance of the K2K architecture for enabling effective and interpretable access to EO datasets. Several observations can be drawn from the analysis of performance across models and configurations.

First, the overall performance of the conversational agents highlights the feasibility of relying on LLM-based agents for query interpretation and justification in structured data contexts. The best-performing configurations (notably LLaMA 3.3 70B and Deepseek-R1) demonstrate that these models can understand complex, multidimensional queries and produce semantically grounded explanations, mainly when supported by a knowledge graph. These results reinforce the central hypothesis of our work: combining domain ontologies with language models improves not only retrieval relevance but also the explainability of the results.

Second, differences observed across the four evaluation metrics reveal distinct behavioral tendencies. For instance, while Qwen achieved the highest precision, it suffered from lower contextual recall, suggesting a trade-off between conciseness and completeness. Deepseek provided a more balanced profile, especially in faithfulness. This makes it a solid choice for tasks centered on explanations. These results show the importance of selecting the right model. They also suggest that using hybrid architectures, which mix different models for interpreting and generating responses, could boost performance.

Third, the ablation study evaluating system performance without an ontological context shows an apparent degradation in contextual recall, despite high

precision. This result may confirm that the ontology serves as a critical backbone for aligning user queries with metadata semantics, particularly for tasks that require fine-grained reasoning over spatial, temporal, and thematic attributes.

Finally, the explicit justification mechanism implemented in the response construction agent appears to be effective in producing transparent outputs. The consistently high faithfulness scores across models suggest that the prompt structure and the system-imposed formatting constraints contribute positively to the clarity and trustworthiness of responses.

The study also reveals some limitations and challenges. The evaluation protocol, while systematic, was based on a limited set of scenarios, metrics, and test queries. Future work should also explore a broader range of user intents and profiles, including more ambiguous or exploratory queries. Moreover, user-centered evaluations (e.g., involving usability studies or qualitative feedback) are needed to complement the automated metrics and assess the actual impact of the explanations on decision-making.

6 Conclusion and Future Work

This paper introduced K2K, a modular conversational system for semantically grounded exploration of open datasets. By combining LLM-based agents with a domain-specific ontology and a structured justification mechanism, K2K supports natural language interaction and improves transparency in data retrieval.

The evaluation across four recent LLM architectures demonstrates that this approach achieves high levels of answer relevance, contextual precision, and faithfulness, particularly when supported by structured ontological knowledge. The results confirm the importance of integrating formal semantics into language-based interaction pipelines.

Future work will explore hybrid agent configurations and the integration of domain-specific vocabularies (e.g., Global Change Master Directory Keywords, GCMD) as prior context for LLM-based agents. It will expand the evaluation protocol and integrate user-centered evaluations. We also aim to enhance adaptability based on user expertise and to improve grounding for partial or vague queries.

References

1. Ahmad, R.A., D’Souza, J., Zloch, M., Otto, W., Rehm, G., Oelen, A., Dietze, S., Auer, S.: Toward FAIR semantic publishing of research dataset metadata in the open research knowledge graph, <http://arxiv.org/abs/2404.08443>
2. Alaofi, M., Arabzadeh, N., Clarke, C.L.A., Sanderson, M.: Generative information retrieval evaluation. <https://doi.org/10.48550/ARXIV.2404.08137>, <https://arxiv.org/abs/2404.08137>, version Number: 2
3. Allemang, D., Sequeda, J.: Increasing the LLM accuracy for question answering: Ontologies to the rescue! (2024), <http://arxiv.org/abs/2405.11706>

4. Annane, A., Kamel, M., Trojahn, C., Gilles, N.A., Comparot, C., Baehr, C.: Improving FAIRness of the SYNOP meteorological data set with semantic metadata. *International Journal of Metadata, Semantics and Ontologies* **16**(2), 118–137 (2023). <https://doi.org/10.1504/IJMSO.2023.135332>, <http://www.inderscience.com/link.php?id=135332>
5. Armant, V., Vargas-Rojas, F., Agazzi, V., Desconnets, J.C., Mougnot, I., Beretta, V., Debard, S., Symeonidou, D., Mouakher, A., Guérin, J., Catry, T., Roux, E.: Leveraging Knowledge Graphs for Earth System Dataset Discovery. In: Demartini, G., Hose, K., Acosta, M., Palmonari, M., Cheng, G., Skaf-Molli, H., Ferranti, N., Hernandez, D., Hogan, A. (eds.) *Lecture Notes in Computer Science. Lecture Notes in Computer Science*, vol. Volume 15233 LNCS, pp. 271 – 288. Springer, Baltimore (Maryland), United States (Nov 2024). https://doi.org/10.1007/978-3-031-77847-6_15, <https://hal.science/hal-04823866>
6. Benjelloun, O., Chen, S., Noy, N.: Google dataset search by the numbers. <https://doi.org/10.48550/ARXIV.2006.06894>, <https://arxiv.org/abs/2006.06894>, version Number: 1
7. Brewster, C., Nouwt, B., Raaijmakers, S., Verhoosel, J.: Ontology-based access control for FAIR data **2**(1), 66–77 (2020). https://doi.org/10.1162/dint_a_00029, <https://direct.mit.edu/dint/article/2/1-2/66-77/9993>
8. Brickley, D., Burgess, M., Noy, N.: Google dataset search: Building a search engine for datasets in an open web ecosystem. In: *The World Wide Web Conference*. pp. 1365–1375. ACM. <https://doi.org/10.1145/3308558.3313685>, <https://dl.acm.org/doi/10.1145/3308558.3313685>
9. Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., Zhao, J., He, X.: Exploring large language model based intelligent agents: Definitions, methods, and prospects (2024). <https://doi.org/10.48550/ARXIV.2401.03428>, <https://arxiv.org/abs/2401.03428>, version Number: 1
10. Compton, M., Barnaghi, P., Bermudez, L., García-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., Huang, V., Janowicz, K., Kelsey, W.D., Le Phuoc, D., Lefort, L., Leggieri, M., Neuhaus, H., Nikolov, A., Page, K., Passant, A., Sheth, A., Taylor, K.: The ssn ontology of the w3c semantic sensor network incubator group. *Journal of Web Semantics* **17**, 25–32 (2012). <https://doi.org/https://doi.org/10.1016/j.websem.2012.05.003>, <https://www.sciencedirect.com/science/article/pii/S1570826812000571>
11. Dupuy, A., Trojahn, C., Aussenac-Gilles, N., Baehr, C.: Data-fw: An ontology network for annotating open datasets. *ACM* (2024)
12. Edwards, C.: Hybrid context retrieval augmented generation pipeline: LLM-augmented knowledge graphs and vector database for accreditation reporting assistance. <https://doi.org/10.48550/ARXIV.2405.15436>, <https://arxiv.org/abs/2405.15436>, version Number: 1
13. Furner, J.: Definitions of “metadata”: A brief survey of international standards. *Journal of the Association for Information Science and Technology* **71**(6), E33–E42 (2020). <https://doi.org/https://doi.org/10.1002/asi.24295>, <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24295>
14. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H.: Retrieval-augmented generation for large language models: A survey (2023). <https://doi.org/10.48550/ARXIV.2312.10997>, <https://arxiv.org/abs/2312.10997>, version Number: 5

15. Guizzardi, G.: Ontology, ontologies and the “i” of FAIR **2**(1), 181–191 (2020). https://doi.org/10.1162/dint_a_00040, <https://direct.mit.edu/dint/article/2/1-2/181-191/10008>
16. Hu, Y., Lu, Y.: RAG and RAU: A survey on retrieval-augmented language model in natural language processing (2024). <https://doi.org/10.48550/ARXIV.2404.19543>, <https://arxiv.org/abs/2404.19543>, version Number: 1
17. Jacobsen, A., Kaliyaperumal, R., Da Silva Santos, L.O.B., Mons, B., Schultes, E., Roos, M., Thompson, M.: A generic workflow for the data FAIRification process **2**(1), 56–65. https://doi.org/10.1162/dint_a_00028, <https://direct.mit.edu/dint/article/2/1-2/56-65/9988>
18. Janowicz, K., Haller, A., Cox, S.J., Le Phuoc, D., Lefrançois, M.: SOSA: A lightweight ontology for sensors, observations, samples, and actuators **56**, 1–10 (2019). <https://doi.org/10.1016/j.websem.2018.06.003>, <https://linkinghub.elsevier.com/retrieve/pii/S1570826818300295>
19. Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., Hauswirth, M.: Linked data in the european data portal: A comprehensive platform for applying dcat-ap. In: Lindgren, I., Janssen, M., Lee, H., Polini, A., Rodríguez Bolívar, M.P., Scholl, H.J., Tambouris, E. (eds.) *Electronic Government*, vol. 11685, pp. 192–204. Springer International Publishing. https://doi.org/10.1007/978-3-030-27325-5_15
20. Křemen, P., Nečaský, M.: Improving discoverability of open government data with rich metadata descriptions using semantic government vocabulary **55**, 1–20. <https://doi.org/10.1016/j.websem.2018.12.009>, <https://linkinghub.elsevier.com/retrieve/pii/S1570826818300714>
21. Lála, J., O’Donoghue, O., Shtedritski, A., Cox, S., Rodrigues, S.G., White, A.D.: PaperQA: Retrieval-augmented generative agent for scientific research. <https://doi.org/10.48550/ARXIV.2312.07559>, <https://arxiv.org/abs/2312.07559>, version Number: 2
22. Mi, J.: Making open resources discoverable: Collaborative approaches for enhanced access **8**(4), 17–29 (2024). <https://doi.org/10.23974/ijol.2024.vol8.4.350>, <https://journal.calaijol.org/index.php/ijol/article/view/350>
23. Quarati, A.: Open government data: Usage trends and metadata quality **49**(4), 887–910 (2023). <https://doi.org/10.1177/01655515211027775>, <http://journals.sagepub.com/doi/10.1177/01655515211027775>
24. RDA FAIR Data Maturity Model Working Group: FAIR data maturity model: specification and guidelines. Publisher: Research Data Alliance Version Number: 1 (2020). <https://doi.org/10.15497/RDA00050>, <https://zenodo.org/record/3909563#.YGRNnq8za70>
25. Ronzano, F., Nanavati, J.: Towards ontology-enhanced representation learning for large language models (2024). <https://doi.org/10.48550/ARXIV.2405.20527>, <https://arxiv.org/abs/2405.20527>, version Number: 1
26. Serajeh, N.T., Mohammadi, I., Fuccella, V., De Rosa, M.: LLMs in HCI data work: Bridging the gap between information retrieval and responsible research practices. <https://doi.org/10.48550/ARXIV.2403.18173>, <https://arxiv.org/abs/2403.18173>, version Number: 1
27. Specka, X., Gärtner, P., Hoffmann, C., Svoboda, N., Stecker, M., Einspanier, U., Senkler, K., Zoader, M.M., Heinrich, U.: The bonares metadata schema for geospatial soil-agricultural research data - merging inspire and datcite metadata schemes **132**, 33–41. <https://doi.org/10.1016/j.cageo.2019.07.005>, <https://linkinghub.elsevier.com/retrieve/pii/S009830041930086X>

28. Yacoubi Ayadi, N., Faron, C., Michel, F., Gandon, F., Corby, O.: A model for meteorological knowledge graphs: Application to météo-france data. In: Di Noia, T., Ko, I.Y., Schedl, M., Ardito, C. (eds.) *Web Engineering*, vol. 13362, pp. 283–299. Springer International Publishing. https://doi.org/10.1007/978-3-031-09917-5_19, series Title: *Lecture Notes in Computer Science*
29. Zhu, J.: Unlocking potential: Harnessing the power of metadata for discoverability and accessibility **43**(3), 249–256 (2023). <https://doi.org/10.3233/ISU-230202>, <https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/ISU-230202>