# Interpreting User Needs with LLMs-based Conversational Agents and Knowledge Graphs: An Earth Observation Use Case

Antoine Dupuy[1], Nathalie Aussenac-Gilles[1], Christophe Baehr[2] and Cassia Trojahn[3]

[1]IRIT, Université de Toulouse, UT2, CNRS, Toulouse, France
[2]CNRM UMR-3589, Université de Toulouse, Météo-France, CNRS, Toulouse, France
[3]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, Grenoble, France

**Abstract**
Open Science has broadened access to scientific datasets. However, identifying relevant ones to specific user needs remains challenging due to the volume, diversity and poor metadata. This paper proposes to integrate semantically enriched metadata with LLM agents to interpret user natural language queries, to extract user intent and to generate justifications for retrieved results. Experiments with different LLMs highlight the potential of such approach for scientific dataset retrieval.

## 1. Introduction

Public and research institutions have increasingly promoted open access to scientific data, leading to a proliferation of repositories offering environmental, geospatial and sensor-based datasets. However, openness alone does not guarantee usability. Many portals remain difficult to navigate and often lack sufficient context for users—especially non-experts—to assess data relevance [1]. A key barrier is the limited availability and heterogeneity of metadata [2, 3]. In Earth Observation (EO), this issue is amplified by the data's multifaceted spatial, temporal, thematic and technical dimensions, emphasizing the necessity for structured and semantically enriched metadata.

Research in information retrieval, semantic web and natural language interfaces has improved metadata access and query formulation [4, 5, 6, 7], with more recently LLMs enabling intuitive user input interpretation and readable responses. Yet, many systems rely on rigid and heterogeneous templates, assume domain knowledge expertise and lack explanations for results. Recent work calls for better alignment of user intent with metadata and transparent, explainable retrieval [8, 9]. Close to ours, in [10], ontologies are injected into language models — a strategy we adopt with the DATA-FW [11] ontology – to structure notions such as datasets, users and data quality. In [12], combining a LLM with an ontology yields more relevant results than relying solely on a traditional relational data source.

This paper proposes to integrate LLM-based conversational agents with a domain-specific knowledge graph to retrieve datasets. The approach builds upon previous work in several key respects: (1) a knowledge graph to represent EO datasets metadata; (2) LLM-based agents to interpret user queries and retrieve suitable datasets; (3) natural language interaction, including iterative query refinement, enabling non-expert users to progressively articulate complex information needs; (4) a justification mechanism, summarizing the search criteria used, the selected datasets and explaining why they are relevant to the query. The performance of four LLMs (LLaMA 3.3 70B, Mistral Saba 24B, Deepseek-R1 and Qwen 32B) is evaluated across scenarios in the EO domain, using the Deepeval framework [13, 14] with metrics such as relevance, precision, recall and faithfulness.

*ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan*
antoine.dupuy@irit.fr (A. Dupuy); nathalie.aussenac-gilles@irit.fr (N. Aussenac-Gilles); christophe.baehr@meteo.fr (C. Baehr); cassia.trojahn-dos-santos@univ-grenoble-alpes.fr (C. Trojahn)
0000-0003-4237-2462 (A. Dupuy); 0000-0003-3653-3223 (N. Aussenac-Gilles); 0000-0002-1230-893X (C. Baehr); 0000-0003-2840-005X (C. Trojahn)

The rest of the paper is organized as follows. Section 2 introduces the proposed architecture. Section 3 discusses evaluation and results. Section 4 concludes with a synthesis of our findings and outlines perspectives for future work.

# 2. K2K: An approach Based on LLM Agents and KG

## 2.1. Main steps of the approach

K2K (Knowledge-To-Knowledge) addresses the barriers that (non-expert) users encounter when exploring datasets from different domains. By integrating LLM-based agents, relevant contextual knowledge and an ontology representing datasets metadata, it enables users to express their information needs in natural language and receive results with justifications (Figure 1). The code and the evaluation datasets[1], as well as the prototype[2] are available online.

As depicted in Figure 1, K2K processes user queries through a modular architecture that integrates domain-specific data, semantic metadata and LLM-based agents modules. Each of these modules is responsible for a well-delimited functional role. When a user submits a query (1, 2), the selected domain (e.g., Meteorology, Earth Observation) and the current conversation determine the contextual data scope and the historical dialogue memory to be retrieved. The domain context (3) is filtered out using a TF.IDF-based context selector that identifies the most relevant textual chunks to be used as grounding information. Meanwhile, the SPARQL library selectively retrieves semantic triples (4) from the ontology graph hosted on a SPARQL endpoint. This ontology layer is built on top of established vocabularies such as DCAT3, DUV, RDF Data Cube, DOV and FOAF. It is structured as a network in the DATA-FW ontology that define the concept of a dataset and its links to platforms, producers, structure, etc. and its properties.

The retrieval step includes filtering ontology entities to reduce redundancy (e.g., multilingual labels, duplicate predicates) and reduce the number of tokens while preserving semantic richness. All selected data, including the user query, context snippets, relevant triples (from the ontology) and dialogue history are passed to the LLM-based agents (5) module and integrated into the agents prompts. This module contains specialized agents: (i) the Query Analyst Agent analyzes the user request in relation to the ontology and identifies missing or ambiguous criteria (6), (ii) the Data Identification Agent identifies appropriate datasets by triggering web search through a tool-integrated LLM and parses the results to extract download links and metadata (7) and (iii) the Response Agent synthesizes a fluent, structured response from the output of the previous agents, ensuring readability and traceability (8). Agents operate in isolation but communicate through JSON structures or structured outputs, enabling easier parsing, traceability and logging. The final answer, together with links to any matched datasets, is rendered in the user interface, completing the interaction cycle (9, 10). Moreover, this architecture enables flexible domain adaptation by changing the corpus of specific-domain data while keeping the agent logic reusable.

## 2.2. Example of User interaction

An illustrative interaction is presented in Figure 2. The user initiates the dialogue with the query: "Which datasets are available to analyze $CO_2$ concentration in France between 2015 and 2023?". This query triggers (1) the complete K2K pipeline: the system extracts ontology entities (classes, object properties and data properties) associated with datasets,retrieves relevant contextual knowledge from domain-specific resources (here Earth Observation) using tf.idf scores and generates a natural-language response. The response comprises several elements. The system provides (2) the criteria extracted from the user query, (3) the dataset identified as relevant and a justification explaining why this dataset was selected. It also issues a disclaimer highlighting possible limitations of the dataset and requests further details such as preferred file type (CSV, JSON, etc.), specific producers, or direct dataset sources. In the

[1] https://github.com/DupuyAntoine/K2K
[2] http://ec2-13-60-15-33.eu-north-1.compute.amazonaws.com:3000

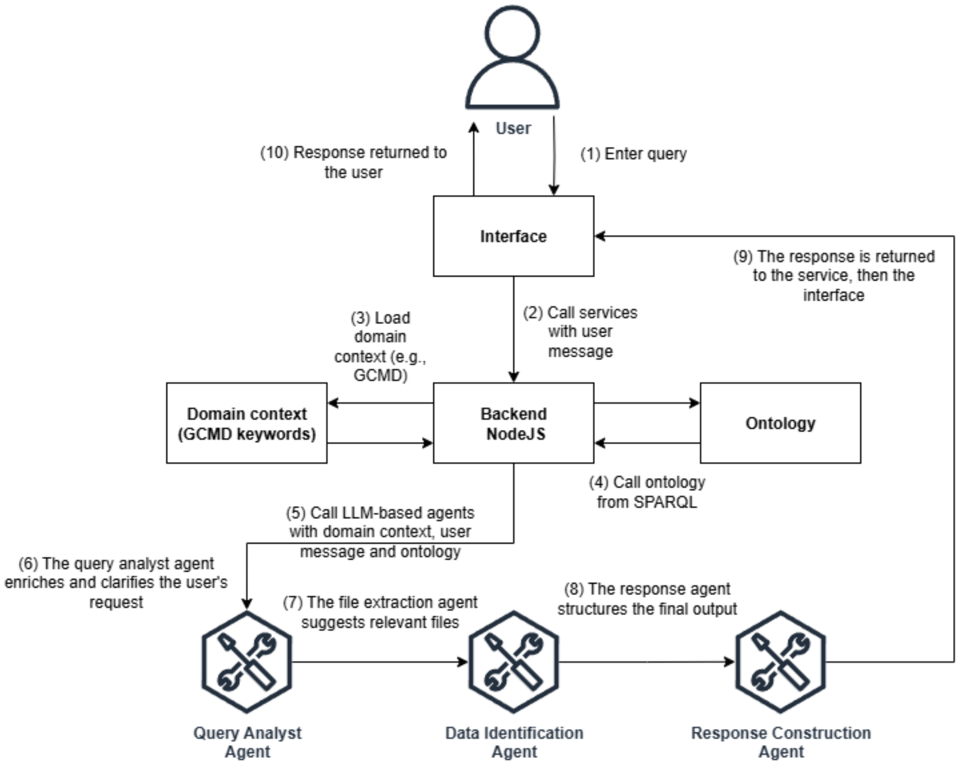


**Figure 1:** K2K System Overview

example, the user responds by noting that (4) the dataset proposed by the system is not appropriate, as it contains predictive rather than observational data for the requested period. The user further specifies a preference for CSV format and for data provided by ADEME. In turn, the system (5) reformulates the updated search criteria, acknowledges that the previous dataset did not meet the user's needs and apologises. It then continues by asking for clarification: whether the user is seeking $CO_2$ concentration or $CO_2$ emissions (6) and which temporal resolution (daily, weekly, monthly, yearly) is desired. At this stage, the system reports (7) that no dataset matching all the refined constraints has been identified yet. However, it emphasizes that it is keeping track of the user's evolving requirements and (8) encourages the user to provide further details that could guide the search more effectively.

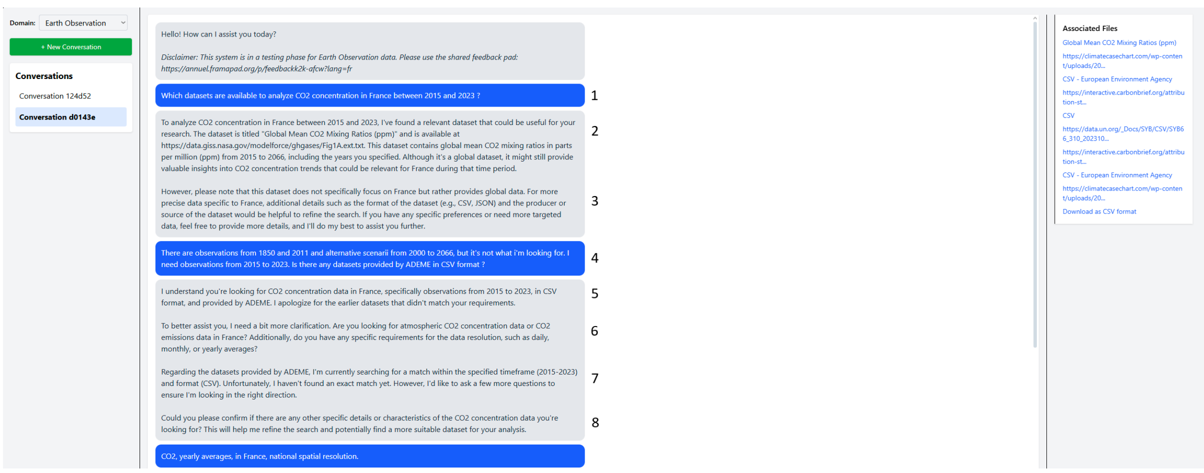


**Figure 2:** Query "Which datasets are available to analyze CO2 concentration in France between 2015 and 2023 ?" (Part 1)

As illustrated in Figure 3, the user provides additional details: (9) he specifies that he is interested in $CO_2$ data (without clarifying whether this refers to emissions or concentrations), with yearly averages, at the national resolution and focused on France. The system responds by (10) reformulating and confirming the updated search criteria, following its established practice of maintaining explicit traceability of user constraints. Since the distinction between emission and concentration remains unresolved, the system (11) reiterates its request for clarification and encourages the user to provide further details if possible. At this point, the system presents (12) four candidate datasets related to $CO_2$ concentration, each accompanied by an explanation of its relevance with respect to the specified query. The response is concluded with (13) a renewed request for clarification on whether the user's interest lies in concentration or emission data, along with an invitation to refine the query to achieve more accurate results. Throughout the interaction, (14) the datasets retrieved by the system are directly accessible to the user via the file panel displayed on the right-hand side of the interface, ensuring transparency and immediate usability.


**Figure 3:** Query "Which datasets are available to analyze CO2 concentration in France between 2015 and 2023 ?" (Part 2)

## 3. Evaluation

A benchmark of 8 natural language queries was created, each corresponding to a common theme encountered in EO scenarios. All material is available online[3]. The evaluation adopts 4 metrics [15, 16]: Answer Relevance (how well answers address the query), Contextual Precision and Recall (accuracy and completeness of explanations against dataset metadata) and Faithfulness (consistency with factual content). These were computed automatically using the Deepeval library[4], which leverages a local LLaMA 3 8B model running via Ollama to evaluate the generated responses. The 8 queries were distributed across the models 200 times: 75 were allocated to LLaMA, 50 to Mistral and to Deepseek-R1 and 25 to Qwen. Figure4 compares these results.

LLaMA 3.3 70B and DeepSeek-R1 70B are the most balanced models, performing well across all metrics. Qwen excels in precision, while Mistral demonstrates strong faithfulness. It demonstrates that LLM-based agents, particularly LLaMA 3.3 70B and DeepSeek-R1, are effective at interpreting complex queries and generating grounded explanations in structured data contexts. An ablation study confirms the ontology's role in preserving contextual recall. The system justification mechanism improves response transparency, as shown by consistently high faithfulness scores (as in Figure 2, step 3 and in Figure 3, step 12).

## 4. Conclusion and Future work

This paper presented K2K, a system that combines LLM-based agents, ontologies, domain-specific metadata context and a language interaction to improve transparency in data retrieval. Results corrob-

[3]https://github.com/DupuyAntoine/K2K/tree/main/ai-agent/src/agents/evaluation
[4]https://github.com/confident-ai/deepeval, 28/04/2025

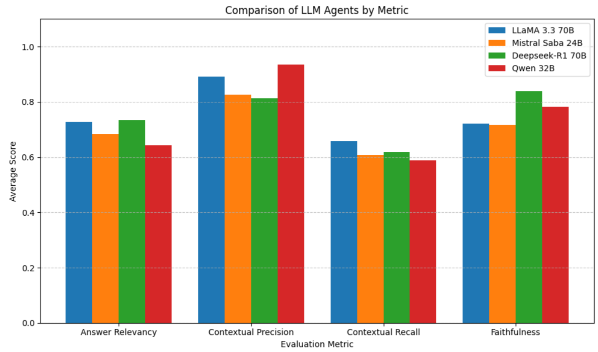


**Figure 4:** Comparison of LLM Agents by Metric.

orate the positive impact of integrating ontologies into language-based interaction pipelines. Future work will explore hybrid agent configurations, usage of embedding models to select relevant chunks of domain-specific data context (instead of TF.IDF), improvement of the evaluation protocol and integration of user-centered evaluations.

## Declaration on Generative AI

In accordance with the CEUR-WS Policy on AI-Assisting Tools [5], the authors disclose the following:

**Tools and services used.** During the preparation of this manuscript, we used the following generative AI tools: ChatGPT (OpenAI).

**Contributions of each tool.** ChatGPT was employed solely for *language polishing, paraphrasing and stylistic refinement*. In no instance was it used to replace original scientific contributions, develop arguments, draw conclusions or generate novel technical content.

**Human oversight and responsibility.** All AI-generated suggestions were meticulously reviewed, edited and validated by the authors. We take full responsibility for the final content, correctness and integrity of the manuscript. We affirm that the core scientific insights, results and reasoning remain entirely the work of the human authors.

The contributions made by generative AI are acknowledged here in this dedicated section, in compliance with the transparency and accountability requirements mandated by CEUR-WS.

## References


[1] A. Quarati, Open government data: Usage trends and metadata quality 49 (2023) 887–910. URL: http://journals.sagepub.com/doi/10.1177/01655515211027775. doi:10.1177/01655515211027775.

[2] J. Zhu, Unlocking potential: Harnessing the power of metadata for discoverability and accessibility 43 (2023) 249–256. URL: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/ISU-230202. doi:10.3233/ISU-230202.

[3] J. Mi, Making open resources discoverable: Collaborative approaches for enhanced access 8 (2024) 17–29. URL: https://journal.calaijol.org/index.php/ijol/article/view/350. doi:10.23974/ijol.2024.vol8.4.350.

[4] G. Guizzardi, Ontology, ontologies and the "i" of FAIR 2 (2020) 181–191. URL: https://direct.mit.edu/dint/article/2/1-2/181-191/10008. doi:10.1162/dint_a_00040.

[5] C. Brewster, B. Nouwt, S. Raaijmakers, J. Verhoosel, Ontology-based access control for FAIR data 2 (2020) 66–77. URL: https://direct.mit.edu/dint/article/2/1-2/66-77/9993. doi:10.1162/dint_a_00029.


[5] https://ceur-ws.org/GenAI/Policy.html

[6] K. Janowicz, A. Haller, S. J. Cox, D. Le Phuoc, M. Lefrançois, SOSA: A lightweight ontology for sensors, observations, samples, and actuators 56 (2019) 1–10. URL: https://linkinghub.elsevier.com/retrieve/pii/S1570826818300295. doi:10.1016/j.websem.2018.06.003.

[7] M. Compton, P. Barnaghi, L. Bermudez, R. García-Castro, O. Corcho, S. Cox, J. Graybeal, M. Hauswirth, C. Henson, A. Herzog, V. Huang, K. Janowicz, W. D. Kelsey, D. Le Phuoc, L. Lefort, M. Leggieri, H. Neuhaus, A. Nikolov, K. Page, A. Passant, A. Sheth, K. Taylor, The ssn ontology of the w3c semantic sensor network incubator group, Journal of Web Semantics 17 (2012) 25–32. URL: https://www.sciencedirect.com/science/article/pii/S1570826812000571. doi:https://doi.org/10.1016/j.websem.2012.05.003.

[8] V. Armant, F. Vargas-Rojas, V. Agazzi, J.-C. Desconnets, I. Mougenot, V. Beretta, S. Debard, D. Symeonidou, A. Mouakher, J. Guérin, T. Catry, E. Roux, Leveraging Knowledge Graphs for Earth System Dataset Discovery, in: G. Demartini, K. Hose, M. Acosta, M. Palmonari, G. Cheng, H. Skaf-Molli, N. Ferranti, D. Hernandez, A. Hogan (Eds.), Lecture Notes in Computer Science, volume Volume 15233 LNCS of *Lecture Notes in Computer Science*, Springer, Baltimore (Maryland), United States, 2024, pp. 271 – 288. URL: https://hal.science/hal-04823866. doi:10.1007/978-3-031-77847-6_15.

[9] Y. Cheng, C. Zhang, Z. Zhang, X. Meng, S. Hong, W. Li, Z. Wang, Z. Wang, F. Yin, J. Zhao, X. He, Exploring large language model based intelligent agents: Definitions, methods, and prospects, 2024. URL: https://arxiv.org/abs/2401.03428. doi:10.48550/ARXIV.2401.03428, version Number: 1.

[10] F. Ronzano, J. Nanavati, Towards ontology-enhanced representation learning for large language models, 2024. URL: https://arxiv.org/abs/2405.20527. doi:10.48550/ARXIV.2405.20527, version Number: 1.

[11] A. Dupuy, C. Trojahn, N. Aussenac-Gilles, C. Baehr, Data-fw: An ontology network for annotating open datasets, ACM, 2024.

[12] D. Allemang, J. Sequeda, Increasing the LLM accuracy for question answering: Ontologies to the rescue!, 2024. URL: http://arxiv.org/abs/2405.11706. arXiv:2405.11706 [cs].

[13] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: NeurIPS Datasets and Benchmarks Track (Round 2), 2021. URL: https://arxiv.org/abs/2104.08663.

[14] N. Ferro, M. Maistro, Evaluation of IR Systems, Technical Report, ACM / SIGIR / Tutorial / survey, 2024. URL: https://doi.org/10.1145/nnnnnnn.nnnnnnn.

[15] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2023. URL: https://arxiv.org/abs/2312.10997. doi:10.48550/ARXIV.2312.10997, version Number: 5.

[16] Y. Hu, Y. Lu, RAG and RAU: A survey on retrieval-augmented language model in natural language processing, 2024. URL: https://arxiv.org/abs/2404.19543. doi:10.48550/ARXIV.2404.19543, version Number: 1.