# Can Knowledge Graphs and Retrieval-Augmented Generation be combined to Explain Query/Answer Relationships Truthfully?

Vincent Armant[1,*], Amira Mouakher[1,*], Felipe Vargas-Rojas[1], Danai Symeonidou[3], Joris Guérin[1], Isabelle Mougenot[1] and Jean-Christophe Desconnets[2]

[1]ESPACE-DEV, IRD, Université de Montpellier, Université de Perpignan, Université Antilles, Université de Guyane, Université de la Réunion, Montpellier, France

[2]Mission Science Ouverte, IRD, Montpellier, France

[3]UMR MISTEA, INRAE, Université de Montpellier, Montpellier, France

## Abstract

In recent years, there has been a significant increase in the adoption of Large Language Models (LLMs) by users in both academic and industrial fields. These powerful tools are progressively challenging the dominance of traditional keyword-based search engines in various fields. While advancements like Retrieval-Augmented Generation (RAG) are enabling LLMs to provide provenance and explanations, their widespread adoption remains hindered by some well-known limitations, including hallucinations (factual inconsistencies), outdated knowledge, and answer precision. In contrast, classical search engines do not suffer from these issues, thanks to recent progress that has made them both efficient and accurate. However, their output may lack interpretability compared to LLMs. This vision paper proposes a novel explanation system that bridges this gap. By integrating Knowledge Graphs with RAG, we aim to elucidate the semantic relationships between retrieved resources and user queries. Addressing this research question has the potential to enhance user trust and confidence in the utilization of explainable search engines.

## Keywords

XAI, Explanation, Resource Discovery, Knowledge Graph, Retrieval-Augmented Generation, Large Language Model



## 1. Introduction

In the context of high-volume, multi-source data, sophisticated solutions like Elasticsearch [1, 2] offer scalable and near-real-time search capabilities for efficient document retrieval. These approaches prioritize efficiency, leveraging indexed data structures to match user queries with relevant documents. However, less focus has been placed on explaining the semantic relationships between the retrieved documents and the user's intent. Conversely, entity-centric approaches that implement Applied Ontologies and KGs [3, 4] have demonstrated effectiveness in facilitating resource discovery within heterogeneous and multi-source data environments. Structured Knowledge Graphs which consist of resource descriptions and their interrelationships following a data model, are well-suited for explaining the semantic relevance between retrieved resources and the initial user request. Their straightforward knowledge representation makes them inherently understandable to many users. However, the applicability of KGs for explainability becomes less clear when dealing with large Knowledge Bases (KBs) containing blank nodes, intricate relationships (transitive, complex, or nested), and alignment across different KGs. In such scenarios, KGs may compromise their self-explanatory nature in favor of data integration

*DAO-XAI 2024: 4th International Workshop on Data meets Applied Ontologies in Explainable AI, October 19, 2024, Santiago de Compostela, Spain*

*Corresponding author.
vincent.armant@ird.fr (V. Armant); amira.mouakher@univ-perp.fr (A. Mouakher); felipe.vargas-rojas@ird.fr (F. Vargas-Rojas); danai.symeonidou@inrae.fr (D. Symeonidou); joris.guerin@ird.fr (J. Guérin); isabelle.mougenot@umontpellier.fr (I. Mougenot); jean-christophe.desconnets@ird.fr (J. Desconnets)
0000-0001-5797-7734 (V. Armant); 0000-0002-1346-3851 (A. Mouakher); 0000-0002-3530-7964 (F. Vargas-Rojas); 0000-0003-1152-5200 (D. Symeonidou); 0000-0002-8048-8960 (J. Guérin); 0000-0002-1287-3269 (I. Mougenot); 0000-0002-4142-5289 (J. Desconnets)

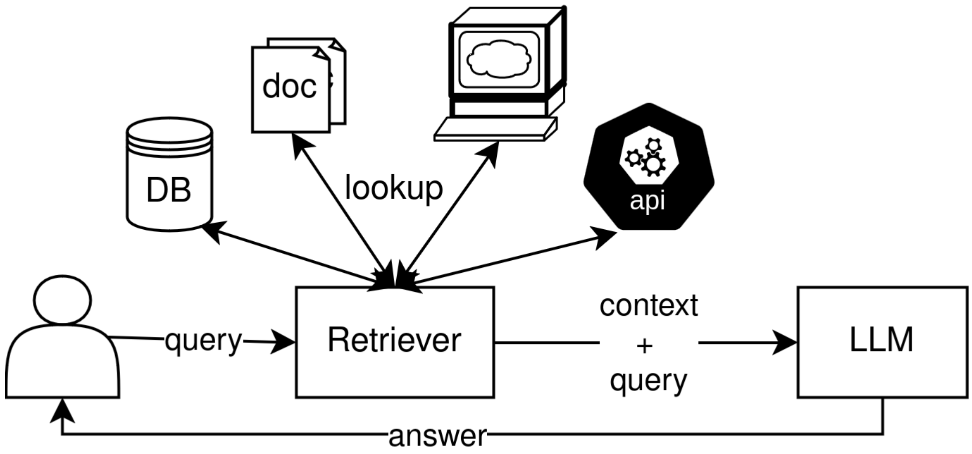



**Figure 1:** Retrieval-Augmented Generation (RAG): General System Architecture

and interoperability. It's important to distinguish our proposed research question from the commonly addressed "Why" and "Why not" explanations [5, 6] which focus on missing results and user trust. Our focus here is the "How" question, aiming to enhance user confidence by elucidating the semantic connections between retrieved results and the user's intent. This approach seeks to improve user understanding and trust by clearly explaining how the results semantically relate to their initial queries.

## 2. Background

To address the dual needs of efficiency and explanations, RAG systems [7, 8, 9] have gained significant interest in both research and industry. These systems combine the efficiency of vector-based information retrieval with the natural language explanation capability of LLMs. RAG's development was driven by the need to overcome limitations inherent in LLMs, such as inaccurate answers ("hallucinations") and the lack of up-to-date, domain-specific, or private information.

Figure 1 illustrates the general architecture of RAG systems [10]. A user initiates the process by submitting a query. The retriever then searches various data sources, including local or external databases, private documents, web applications, or APIs. The LLM then creates a context by merging the retrieved information with the initial user query. This context serves as a one-shot learning experience for the LLM, enabling it to leverage this additional knowledge to potentially generate more accurate answers for the specific query. The accuracy of the constructed context directly influences the quality of the answers returned by the LLM, as well as the likelihood of hallucinations. We propose a novel system architecture that explains the semantic relationships between user queries and retrieved resources, building upon this framework.

## 3. Prospective System Architecture

This work envisions a prospective workflow that utilizes RAG. In the sequel, we present the system architecture, which aims to automatically construct user-centric explanations that connect a user query to the retrieved resources. We then discuss the implementation challenges of the prospective architecture. Figure 2 shows the prospective system architecture intended to explain to users how their queries semantically relate to the retrieved resources. At the beginning of the pipeline, the user inputs a search term (basic user query) and a profile.

The system aims to provide both textual and graphical explanations that describe the relationship between the user's query and the retrieved resources.

1. The Semantic Relationship Retrieval component follows a well-established Semantic Web Search workflow. This workflow involves enriching the user query with related concepts before eval-

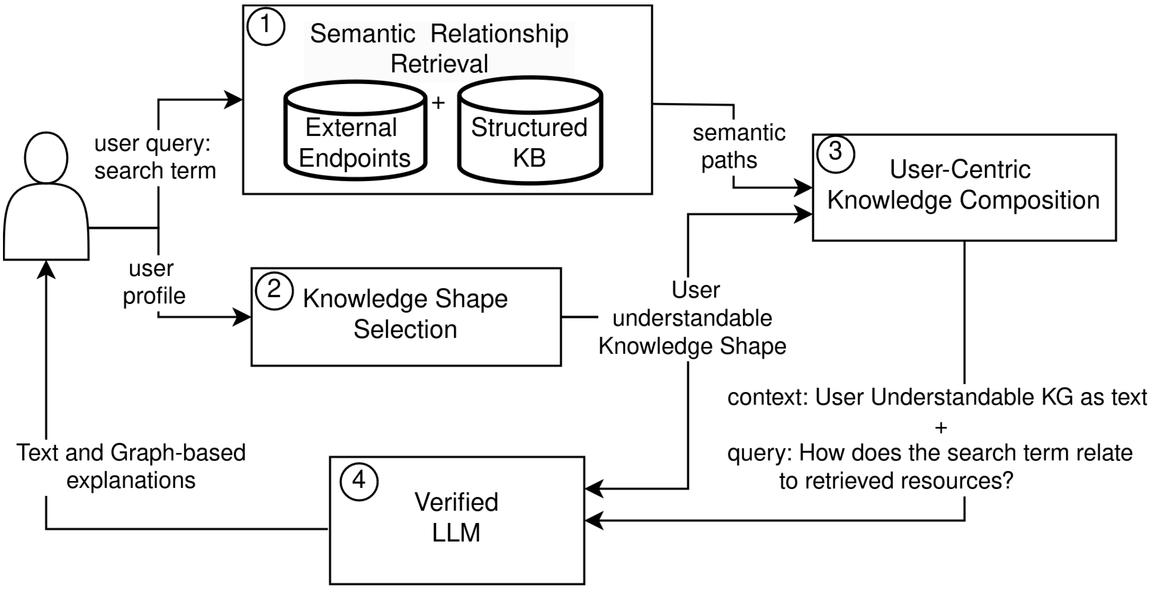



**Figure 2:** Explaining "How": Prospective System Architecture

uating it against a structured knowledge base. However, unlike the classical approach, which solely focuses on retrieving relevant resources, this component identifies a set of paths. Each path represents a sequence of semantic relationships connecting the user's search term to the retrieved resources. The combined set of these paths forms a knowledge graph, integrating relevant knowledge extracted from both the knowledge base and external resources.

2. The Knowledge Shape Selection component takes the user profile (i.e., general user, domain expert) as input and outputs a knowledge shape. This knowledge shape represents the set of sentences that the user profile is likely to understand. We use the knowledge shape to format both the semantic relationships from the Knowledge Composition component and the output from the verified LLM.
3. The User-Centric Knowledge Composition component utilizes the user's knowledge shape to transform the retrieved semantic paths into a knowledge graph that is comprehensible to the user. This resulting knowledge graph is then verbalized using appropriate techniques, creating the context for the LLM.
4. The Verified LLM component receives three inputs: the verbalized, user-understandable knowledge graph as context, and the question "How does the user query relate to the retrieved resources?" as the query. It provides explanations to the user in both textual and graphical formats. This component goes beyond the traditional RAG pipeline by incorporating an answer verification setup. This verification process ensures that the returned graph is consistent with the previously determined knowledge shape.
5. At the end of the pipeline, we expect a significant decrease in the occurrence of hallucinations within the explanations. This prioritizes verifiability and user comprehension, potentially leading to fewer expression explanations. This trade-off captures the natural conflict in LLM answers between expressiveness and consistency.

## 4. Some Challenges

**Semantic Relationship Retrieval**: Unlike the traditional semantic web workflow, which only returns retrieved resources, the Semantic Relationship Retrieval component requires additional computational effort. This challenge involves designing fast, federated approaches capable of querying external endpoints and composing semantic paths efficiently to enable real-time user interactions.

**Knowledge Shape Selection**: Choosing an appropriate language to model knowledge that is comprehensible to diverse user profiles presents a key challenge. For instance, the SHACL language offers an expressive framework for modeling authorized knowledge using SHACL shapes. In this context, a user profile's knowledge shape might represent complex constraints restricting vocabulary usage or limiting nested knowledge representation.
**The User-Centric Knowledge Composition**: Rewriting the retrieved semantic path, extracted from an open knowledge repository and conforming to a knowledge shape, presents a significant challenge. The success of this process depends on the difference between the complexity of the retrieved semantic path (the presence of nested structures and blank nodes) and the simplicity of the user-understandable knowledge shape.
**The Verified LLM**: Given a context and a query, recent advances in LLMs and RAG technologies (e.g., ChatGPT 4.0) already enable contextual text and graph-based explanations. This last feature partially mitigates the recurring criticism by minimizing the occurrence of hallucinations. In this direction, our prospective pipeline not only envisions feeding these tools with structured and connected knowledge contexts, but it also envisions verifying that the outcome will be both consistent and understandable by the user. With this last challenge, we hope to drastically reduce the number of hallucinations.
**Evaluation and Comparison of Approaches**: A significant challenge lies in comparing explanation generation approaches, given the subjective nature of interpretability. Evaluating the effectiveness of proposed systems without relying on large cohorts of human participants to assess responses is not straightforward. This subjectivity makes it difficult to establish objective metrics for measuring the quality and usefulness of explanations. To properly assess the KG + RAG approach, an experimental protocol should involve both an evaluation of the semantic accuracy and relevance of explanations by domain experts and a user study with participants from various backgrounds to gauge comprehension and perceived usefulness of the explanations.

## 5. Conclusion

This vision paper addresses the challenge of explaining the semantic relationship between user queries and retrieved resources in search engine results. First, we introduce the strengths and limitations of traditional keyword-based search engines. Next, we delve into the latest explanation capabilities of RAG. This comprehensive framework addresses common problems with LLMs like hallucinations, outdated knowledge, and a deficiency in domain-specific expertise. To address the challenges of truthfully and accurately explaining to users how their queries relate to the retrieved resources, we present a prospective system that leverages Semantic Web and knowledge graphs. This system aims to improve the context provided by the RAG system and verify the accuracy of its explanations.

## Acknowledgments

This study was made possible thanks to discussions initiated during the kick-off meeting of the MOSAIC Horizon Europe project (grant number: 101137398).

## References

[1] C. Gormley, Z. Tong, Elasticsearch: the definitive guide: a distributed real-time search and analytics engine, " O'Reilly Media, Inc.", 2015.
[2] N. Kathare, O. V. Reddy, V. Prabhu, A comprehensive study of elasticsearch, International Journal of Science and Research (IJSR) (2020).
[3] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L.-D. Ibáñez, E. Kacprzak, P. Groth, Dataset search: a survey, The VLDB Journal 29 (2020) 251–272.
[4] K. Balog, Entity-oriented search, Springer Nature, 2018.

[5] A. Chapman, H. Jagadish, Why not?, in: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, 2009, pp. 523–534.
[6] N. Bidoit, M. Herschel, K. Tzompanaki, Query-based why-not provenance with nedexplain, in: Proceedings of the 17th International Conference on Extending Database Technology, EDBT 2014, Athens, Greece, March 24-28, 2014, 2014, pp. 145–156. doi:10.5441/002/EDBT.2014.14.
[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.
[8] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, W. Chen, Generation-augmented retrieval for open-domain question answering, arXiv preprint arXiv:2009.08553 (2020).
[9] G. Izacard, P. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, E. Grave, Atlas: Few-shot learning with retrieval augmented language models, Journal of Machine Learning Research 24 (2023) 1–43.
[10] A. Kimothi, Retrieval Augmented Generation - A Simple Introduction, 2024. URL: https://abhinavkimothi.gumroad.com/l/RAG.

# A. Appendix

In this appendix, we provide an illustrative example utilizing ChatGPT-4. The sections with a grey background represent the prompts we used, while the sections with a white background display the responses generated by ChatGPT-4.

## A.1. Explaining Ocean Data Properties

The dataset D1 has the title “Observations of the Pacific Ocean”. The Observation Collection O1 has aggregated results the dataset D1. O1 has members Observation Collections O3 and O4. O3 has the Observed Property “Water Conductivity”, and O4 has the Observed Property “Temperature at 5-meter depth”. “Water Conductivity” and “Temperature at 5-meter depth” are both Observed properties of the Sea. Water Conductivity can only be understood by experts while ”Temperature at 5-meter depth” can be understood by non-experts. Can you explain to non-experts how D1 relates to Sea-observed properties?

To explain the relationship between the dataset D1 and the observed properties of the sea in a way that non-experts can understand, let's break it down:

1. **Dataset D1: ”Observations of the Pacific Ocean”** - This is a collection of data gathered about various aspects of the Pacific Ocean. Think of it like a big book that records different measurements and observations made in the ocean.
2. **Observation Collection O1** - This is like a chapter in the book (D1) that focuses on specific aggregated results. It's a grouping of related data.
3. **Observation Collections O3 and O4** - These are sub-chapters under O1. They focus on specific topics or types of measurements:
   - **O3** deals with ”Water Conductivity.” This is a measure of how well the water conducts electricity, which can tell experts a lot about the water's salt content and overall quality.
   - **O4** focuses on ”Temperature at 5-meter depth,” which is simply measuring how warm or cool the water is at a depth of 5 meters below the surface.
4. **Observed Properties of the Sea** - These are the specific characteristics of the sea that are being measured. In this case, the key observed properties are:
   - **Water Conductivity:** Important for understanding the salt content and other chemical properties of seawater. This might be more technical and usually understood by experts in oceanography.
   - **Temperature at 5-meter depth:** This is a straightforward measurement of how warm the water is at a certain depth. It's similar to measuring air temperature and can be easily understood by anyone.

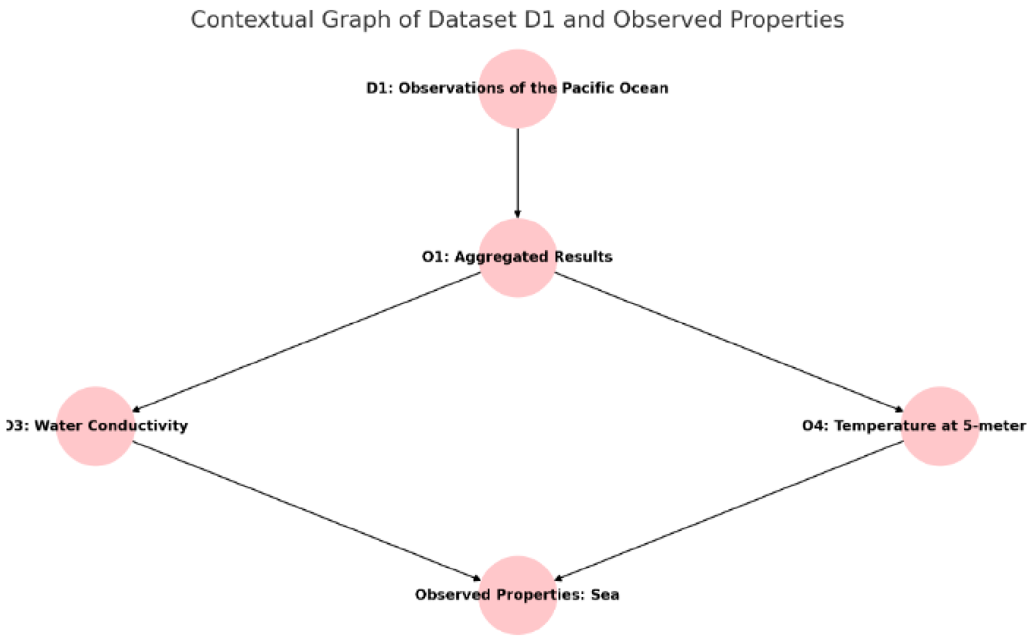
### Simplified Explanation for Non-Experts

Imagine Dataset D1 as a comprehensive report on the Pacific Ocean, detailing various characteristics of the sea. Observation Collection O1 is a specific section of this report that pulls together data on particular topics. Within O1, there are two important parts:

1. One part (O3) looks at how salty or conductive the water is, which is more technical and mainly useful for scientists.
2. The other part (O4) simply checks the water temperature at a certain depth, which is more straightforward and can be understood by anyone.

These measurements help scientists understand the state of the ocean, such as how it changes over time and what might be causing those changes.

Can you draw the graph of the context?




Here's the contextual graph illustrating the relationship between the dataset D1, the observation collections, and the observed properties of the sea:

1. **D1: Observations of the Pacific Ocean** - The main dataset containing comprehensive observations.
2. **O1: Aggregated Results** - A collection of specific results derived from D1.
3. **O3: Water Conductivity and O4:** Temperature at 5-meter depth - Specific observation collections under O1, focusing on different aspects of the sea.
4. **Observed Properties:** Sea - The overarching category for properties like water conductivity and temperature, indicating the specific aspects of the ocean being measured.

The arrows indicate the hierarchical structure and how each part contributes to understanding the properties of the sea.