

# ORKG ASK: a Neuro-symbolic Scholarly Search and Exploration System

Allard Oelen<sup>1</sup>, Mohamad Yaser Jaradeh<sup>2</sup> and Sören Auer<sup>1,2</sup>

<sup>1</sup>TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany

<sup>2</sup>L3S Research Center, Leibniz University of Hannover, Hannover, Germany

## Abstract

**Purpose:** Finding scholarly articles is a time-consuming and cumbersome activity, yet crucial for conducting science. Due to the growing number of scholarly articles, new scholarly search systems are needed to effectively assist researchers in finding relevant literature.

**Methodology:** We take a neuro-symbolic approach to scholarly search and exploration by leveraging state-of-the-art components, including semantic search, Large Language Models (LLMs), and Knowledge Graphs (KGs). The semantic search component composes a set of relevant articles. From this set of articles, information is extracted and presented to the user.

**Findings:** The presented system, called ORKG ASK (Assistant for Scientific Knowledge), provides a production-ready search and exploration system. Our preliminary evaluation indicates that our proposed approach is indeed suitable for the task of scholarly information retrieval.

**Value:** With ORKG ASK, we present a next-generation scholarly search and exploration system and make it available online. Additionally, the system components are open source with a permissive license.

## Keywords

Neuro-symbolic AI, Large Language Models, Scholarly Knowledge Graphs, Scholarly Search System

## 1. Introduction

Finding scholarly articles and exploring the body of scholarly literature consumes a significant share of a researcher's time. Due to the growing number of scholarly articles, this issue only becomes more apparent [1]. Current scholarly search systems passively assist users with their information needs by providing a list of relevant articles. If instead active assistance were provided, the users' information needs, such as a research question, would be answered for them. We present ORKG ASK (Assistant for Scientific Knowledge), a new generation scholarly search and exploration system<sup>1</sup>. ORKG ASK helps researchers find relevant literature and automatically extract knowledge from the retrieved literature, actively supporting researchers with their information needs. The approach consists of three main components: 1) Semantic Search, 2) a Large Language Model (LLM), and 3) Knowledge Graphs (KGs). First, the semantic search addresses the previously discussed challenge of retrieving articles based on their relevance to a specific information need. In ORKG ASK users can formulate their information need as a

---

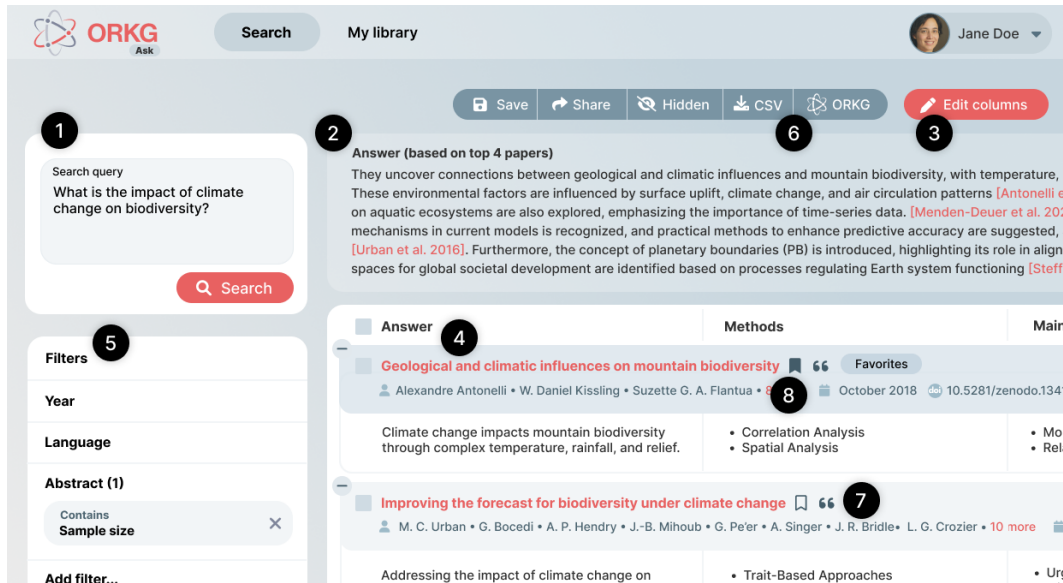
SEMANTiCS 2024 EU: 20th International Conference on Semantic Systems, September 17-19, 2024, Amsterdam, The Netherlands

✉ allard.oelen@tib.eu (A. Oelen); jaradeh@l3s.de (M. Y. Jaradeh); auer@tib.eu (S. Auer)

🆔 0000-0001-9924-9153 (A. Oelen); 0000-0001-8777-2780 (M. Y. Jaradeh); 0000-0002-0698-2864 (S. Auer)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

<sup>1</sup>Available online via <https://ask.orkg.org>



**Figure 1:** Design of the search result page of the ORKG ASK application.

research question, which is entered as a search query. Second, an LLM is leveraged to answer the research question by prompting with the context of the set of relevant articles. In addition to answers to the research question, a set of properties is extracted, among others, a summary, materials, methods, and results of the contributions described in the articles. Third, KGs are used to provide more fine-grained information extraction as well as for curating extraction results. This includes results filtering based on mentioned concepts in scholarly articles.

## 2. Background

There are various established, large, and multidisciplinary scholarly search systems, among others, Google Scholar, Semantic Scholar, and Scopus [2]. Other systems, such as PubMed and ACM Digital Library, are domain-specific. These search systems take a similar approach where articles are ranked based on relevance, but where users have to manually extract relevant information from articles. A new approach provides active support via automatic information extraction by systems such as Elicit, Consensus, and Scispace [3]. These systems are not open-source, leaving details about their approach, such as the model and dataset, to be unknown, in turn making results harder to reproduce. This makes such systems less suitable for systematic literature reviews where reproducibility is a key aspect of the approach [4].

To extract knowledge from a large set of scholarly documents, a Retrieval-Augmented Generation (RAG) approach can be used to provide the LLM with relevant context [5]. The Retrieval aspect retrieves a set of documents, commonly done using vector databases. The Augmented aspect, augments the user query with the found context. Finally, the Generation aspect creates the response. To our knowledge, the previously mentioned AI-supported scholarly search systems use this approach and are thus similar to the approach we propose with ORKG ASK.

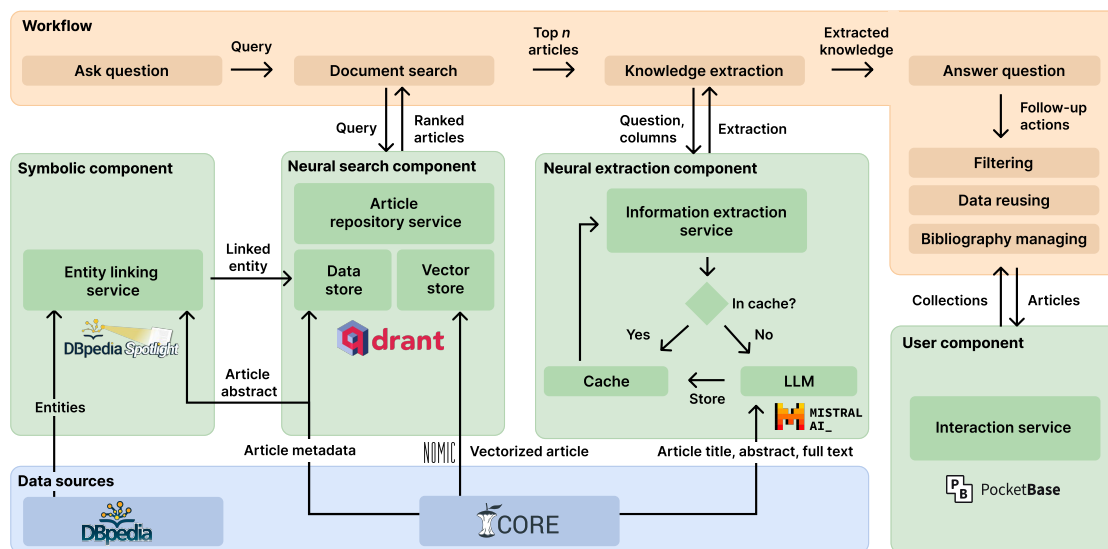


Figure 2: ORKG ASK system workflow integrating neuro-symbolic components.

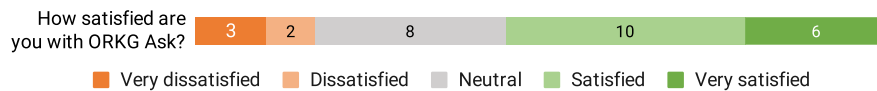
### 3. System Overview

We now discuss the ORKG ASK system in more detail. The service is designed and developed in such a way that it provides a solid foundation for a sustainable service. Additionally, we focus on accessibility by providing a dark mode (for low-light conditions), a responsive interface (for mobile usage, or high zoom levels for visually impaired users), and implementing ARIA accessibility attributes where necessary. The ORKG ASK code base is published as open source under an MIT license and available online.<sup>2</sup> Figure 1 depicts the design of the search result page for a specific research question.

#### 3.1. User-Oriented Features

The **Question Answering** feature as depicted in Node 1 in Figure 1 illustrates the input field for the research question. Nodes 2 and 4 present the answers to the question. Node 2 shows a synthesized answer of the first five displayed results. The **Information Extraction** feature extracts additional information from an article (node 4). There are several default columns displayed, but users can customize the extracted information to their needs (node 3). The **Filtering** feature enables users to filter articles based on user-provided criteria (node 5). This includes the ability to filter based on year, language, words that appear in the title or abstract, the number of citations, author names, etc. The **Bibliography Managing** featured called “My Library” provides a bibliography manager where users can store and curate a list of articles. Articles are added by clicking on the bookmark icon in the interface (node 8) or added manually via the My Library page (via DOI, title, or BibTeX). Articles from My Library can be manually added to a search query, which prepends the manually selected articles to the search results.

<sup>2</sup><https://gitlab.com/TIBHannover/orkg/orkg-ask>



**Figure 3:** Results for user satisfaction evaluation indicating relatively satisfied users.

The **Data Reuse** feature supports citing articles in APA, Vancouver, Harvard, citation styles, and exports to BibTeX, RIS, and CLS-JSON (node 7). The export button is displayed in node 8. Furthermore, there is an option to export the entire search result table to CSV and ORKG [6] CSV (node 6). Finally, the **Entity Linking** feature links entities in article abstracts to their respective DBpedia entries. This provides the ability to filter articles based on semantically identical concepts, providing an additional means to more targeted information retrieval.

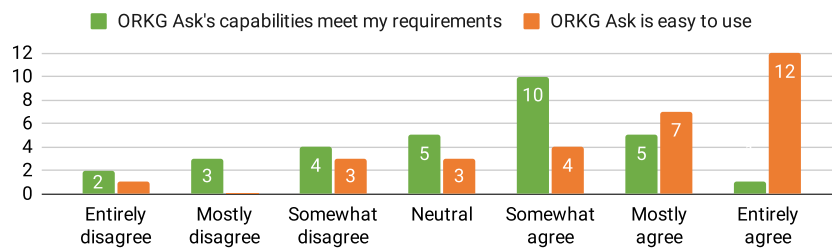
### 3.2. System Workflow

Figure 2 depicts the system workflow. It starts with a user asking a research question. A set of relevant documents for this question is retrieved. The neural search component uses vectorized representations of the query and articles via the Nomic embeddings model to retrieve a set of relevant documents. Optionally, the search space can be narrowed down by filtering specific metadata or linked entities. Qdrant<sup>3</sup> is used as a vector and data store. The symbolic component processed article abstracts offline and stored these linked entities in the data store. The entity linking is conducted using DBpedia Spotlight [7]. The CORE dataset [8], containing article metadata, abstracts, and full-text (in the case of open-access articles), is used as a data source for the vector store. Next, knowledge is extracted using an LLM from the top  $n$  articles, resembling the RAG approach. Currently, we use the Mistral Instruct 7B v0.2 model for the information extraction. To reduce system resource usage, the LLM is only prompted if the answer does not yet exist in the cache. Finally, the information is presented to the user.

## 4. Evaluation

As a preliminary system evaluation, we aim to assess the usability of the system. We did this using a 5-point user satisfaction assessment and the Usability Metric for User Experience lite (UMUX-lite) [9] evaluation. Participants were recruited via the ORKG ASK production system, via a non-intrusive tooltip asking real-world system users for their opinion. To keep participation efforts as low as possible, no participant demographics were requested from users. In total, 30 participants took part in this evaluation. As Figure 3 shows, users are relatively satisfied with ORKG ASK. The UMUX-Lite evaluation displayed in Figure 4 results in an overall score of 65.2. As the individual results show, most participants agree that ORKG ASK is easy to use, but the system does not always meet their requirements. This could be explained by users' search behavior, as logs of asked questions revealed that not all questions are valid and answerable, leaving the user's specific search requirement unmet. Further evaluation is needed to determine what is needed to understand the user's expectations and needs better.

<sup>3</sup><https://qdrant.tech>



**Figure 4:** Results for UMUX-Lite evaluation with a total score of 65.2.

## 5. Conclusion

The introduction of ORKG ASK serves as a starting point for a neuro-symbolic approach to finding and exploring scholarly articles. The preliminary evaluation indicates that our approach is easy to use. In the future, we plan to extend the system by providing provenance information to highlight the source of extracted information. Furthermore, we plan to extend the KG part significantly, growing the KG automatically while the system is being used.

## References

- [1] E. Landhuis, Scientific literature: Information overload, *Nature* 535 (2016) 457–458.
- [2] M. Gusenbauer, N. R. Haddaway, Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources, *Research Synthesis Methods* 11 (2020) 181–217. doi:10.1002/jrsm.1378.
- [3] F. Bolanos, A. Salatino, F. Osborne, E. Motta, Artificial Intelligence for Literature Reviews: Opportunities and Challenges, arXiv preprint arXiv:2402.08565 (2024). arXiv:2402.08565.
- [4] A. MacFarlane, T. Russell-Rose, F. Shokrane, Search strategy formulation for systematic reviews: Issues, challenges and opportunities, *Intelligent Systems with Applications* 15 (2022) 200091. doi:10.1016/j.iswa.2022.200091.
- [5] 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.
- [6] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D’Souza, K. E. Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving access to scientific literature with knowledge graphs, *Bibliothek Forschung und Praxis* 44 (2020) 516–529.
- [7] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, DBpedia spotlight: Shedding light on the web of documents, in: *7th International Conference on Semantic Systems*, 2011, pp. 1–8.
- [8] P. Knoth, D. Herrmannova, M. Cancellieri, L. Anastasiou, N. Pontika, S. Pearce, B. Gyawali, D. Pride, CORE: A global aggregation service for open access papers, *Nature Scientific Data* 10 (2023) 366.
- [9] J. R. Lewis, B. S. Utesch, D. E. Maher, UMUX-LITE: When there’s no time for the SUS, in: *SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2099–2102.