# Enhancing Scientific Discovery and Decision-Making: A Knowledge Graph-based Research Support System

Liubov Kovriguina[1,*], Linn Aung[1], Peter Haase[1], Simon Heiß[1], Nicolas Heist[1] and David Lamprecht[1]

[1]*metaphacts GmbH, Germany*

**Abstract**

This work presents a neurosymbolic AI system for support of research and discovery activities. The system is powered by metaphactory and builds on the open scholarly knowledge graphs SemOpenAlex, LPWC and CS-KG. Researchers are supported through a range of AI methods (generative natural language interface to KGs, retrieval with semantic templates, neighborhood exploration with KG embeddings) offering convenient means to solve research tasks like scientific artifacts overview, publication search and getting recommendations. This work gives an overview of the system design, architecture, data landscape and supported functionalities. A concrete implementation of this system is used in a research project exploring the application of AI methods for electric traction drives design.

**Keywords**

scholarly knowledge graphs, neurosymbolic AI methods for research support, KG-based systems for scientific discovery, LLM interfaces to knowledge graphs, AI for scientific discovery

## 1. Introduction

Research and scientific discovery are based on complex cognitive, creative and scientific human-centred processes that can hardly be fully formalized, defined or verified solely by a machine. The majority of these processes are relying on methodologies and approaches that are based on learned experience, knowledge and reasoning abilities.

Nowadays, the involvement of AI in creative processes has established as a new reality, particularly in the form of AI-powered research services and assistants. While methods of AI and machine learning have been used for hypothesis testing or finding optimal solutions (e.g., in proteins, algorithms and games), the human has always been defining the methodology, application rules and evaluation. At the same time, humanity has accumulated and continues to generate a vast amount of knowledge and data that can now be explored through AI.

Integration between generative AI and knowledge graphs (KGs) has proven to be beneficial. LLMs, being black-box models, and KGs, which explicitly store rich factual knowledge, are, however, both incomplete, but in a different way. LLMs contain commonsense knowledge, but

*4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment 11/12 November 2024 - Baltimore, MD, USA co-located with The 23rd International Semantic Web Conference, ISWC 2024*

*Corresponding author.

lk@metaphacts.com (L. Kovriguina); la@metaphacts.com (L. Aung); ph@metaphacts.com (P. Haase); sh@metaphacts.com (S. Heiß); nh@metaphacts.com (N. Heist); dl@metaphacts.com (D. Lamprecht)

0000-0001-9962-1138 (L. Kovriguina); 0009-0009-4338-0983 (L. Aung); 0000-0002-7561-7000 (P. Haase); 0009-0005-3256-2694 (S. Heiß); 0000-0002-4354-9138 (N. Heist); 0000-0002-9098-5389 (D. Lamprecht)

might be based on outdated training data; they can generate an unlimited number of coherent and fluent texts, suggesting creative hypotheses, but these are not always factually correct. On the other hand, KGs can be viewed as sources of reliable and trustworthy information.

Although supporting research tasks involves some creativity, it is equally important to produce reliable and trustworthy results. We ensure a certain level of reliability and trust by developing a research support system that produces results grounded on knowledge available in established scientific knowledge graphs, and allows the user to use natural language search and conversational AI functionalities on demand.

The proposed research assistance system, further referred as **KIRA Research Support System**, has been designed and implemented in the context of the KIRA project, which focused on optimized control of electric traction drives, to support project partners in their aspirations for methods improvement in a particular research context. The motivation, therefore, was to design trustworthy scenarios for project collaboration, decision making and task solving on top of knowledge graphs, and make them accessible to the users. The system was available to about 50 users from 10 organizations during the project, and was developed from their direct needs and feedback, as well as via continuous improvement of methods, allowing extrinsic evaluation. Moreover, to allow further development of collaborative Human-AI discovery, based on reinforcement learning and multi-agent approaches, a stable research support system baseline, which can be later populated with AI agents, was required.

Our **contributions** can be thus summarized as follows:

- a KG-powered neurosymbolic AI system for research activities support;
- a data landscape, integrating several scientific KGs in a single system to provide broader explorative experience;
- a spectrum of research assistance tasks, mapped to neurosymbolic AI methods and applications, which has been developed iteratively through involvement of the users, continuous validation and feedback from them.

## 2. Supported Research Tasks

Doing research involves a lot of inspiration and planning, spread across creative and routine activities. Scientific communication is also an important part of it, promoting ideas transfer and exchange, collaborative discovery and results dissemination. In the KIRA System, we are supporting both the tasks, facilitating research activities (i.e. project collaboration), and the tasks, related to experiment planning (i.e. exploratory search of connections and ways to improve methods for particular tasks). The tasks, which are currently supported, are listed below formulated as user stories, representing requirements for the implementation of the system, as described in Sec. 5). A detailed description of using the system to explore scientific artifacts is provided in Fig. 3, Appendix A. Besides tasks and knowledge graphs access, each user of the system has an individual dashboard with a personalized overview of relevant information, covering affiliation, topics of interest, papers and team members.

**Find Publications, Topics, Authors, Datasets** As a user, I would like to ask questions in natural language about publications on particular topics, their authors, affiliations, etc., and get back results from the knowledge graph, for example, *List the ten most influential*

*authors publishing about Amyloid-beta precursor protein* or *What are the papers about holograms, calibration and augmented reality published in the last 2 years?*. For some cases, I would like to summarize the results with keeping provenance from the knowledge graph, and get statistics about paper metadata distribution.
**Foster Collaboration within the Project** As a user, I would like to have an overview over the project, including its goal, funding, runtime, outcomes, partners and involved persons, as well as existing relations between persons in the system to other entities in the system. Also, I would like to identify synergies between a certain partner and other partners, related to project-relevant tasks or methods, i.e. *Who of the partners is working on the same tasks as me?*.
**Get Recommendations** As a user I would like to get personalized recommendations for scientific key content, i.e. *If I have a specific Datasets I want recommendations which Task and Methods I can use for it.* I want to be recommended connections between them that may not yet exist, being informed that embedding-based link prediction has been used to forecast new links.
**Get Overview of Scholarly Artifacts** Scientific artifacts can be Datasets, Methods, Tasks, Models, Publications, Repositories, etc. As a user, I would like to have an overview of existing AI-Methods and obtain a detail view of a specific AI-Method. I want to see existing connections for scientific artifacts, i.e. *Which methods are improving the given method?*, *What is the state-of-the-art (leaderboard) for the given task?*, *Which metrics are used to evaluate a task or method?*
**Explore Concepts' Neighbourhood** As a user, I would like to find similar methods, tasks, datasets, and topics via neighborhood exploration and see them visualized in the vector space.
**Compare Methods** As a user, I want to compare a pair methods by its application across methods, tasks, materials, and datasets, and observe the scientific artifacts, co-occurring with each of the methods, as well as artifacts, with are specific for and shared by both methods, i.e. *What is the difference in application of recurrent neural network and long short-term memory across tasks.* Also, being aware of possible hallucinations, I want to prompt an LLM to do a featured methods comparison, based on the information retrieved from the scholarly KG.

## 3. Related Work

Neurosymbolic systems combine the power of human expert knowledge (symbolic) with the language understanding capabilities (neural) of large neural networks such as LLMs. As the authors highlight in [1], neurosymbolic systems (1) extrapolate to out-of-distribution data, e.g., constantly updating KGs or unseen KGs, (2) offer an interpretation of the underlying reasoning, e.g., explaining the grounding of variables and the choice of the underlying query structures, (3) learn from small data, e.g., reusing symbolic knowledge and exploring novel (KG) data based on exploration through agents.

The closest neurosymbolic system providing access to scholarly knowledge is the Open Research Knowledge Graph (ORKG) [2]. The project aims to provide a KG-based infrastructure for semantically capturing and representing the content of research papers. It models scientific contributions and methodology aspects with a focus on data quality and key insights of papers. The core service is a comparison mode offering a comprehensive overview of the state-of-the-art for a research question, taking into account various properties. Another noteworthy services are *ORKG Ask*, a scientific search and exploration system based on vector search, large

language models and knowledge graphs, and *AIDA-Bot 2.0*[3], a conversational agent, which answers user questions by translating them to formal queries and summarizes inormation from relevant articles. With the integration of LLMs, enhanced generative AI platforms are able to provide research support as well. *You.com*[4] is a AI-driven search engine and assistant platform designed to enhance user productivity with advanced search, summarization, and natural language processing capabilities. *You.com* recently introduced a *genius mode* which can provide research support to help researchers find and understand relevant information more efficiently. It also aids in citation management, data analysis, and collaboration, while personalizing support based on user preferences and research patterns.

## 4. Data Landscape

The data landscape of our system is populated with knowledge graphs, representing diverse scientific content. Firstly, publications metadata are covered by SemOpenAlex [5] which serves as a semantic scientific hub within the KIRA project. To bring in relevant content about the computer science and especially the machine learning domain, we leverage the Computer Science Knowledge Graph (CS-KG) [6] and Linked Papers with Code (LPWC) [7].

Furthermore, we created the KIRA Project Knowledge Graph (KIRA KG) modelling the involved partners and persons as well as their relationships, fields of expertise, project obligations and outcomes. This graph also stores user profiles that are generated from user interaction with the system artifacts, i.e. bookmarking relevant papers, ranking recommendations, adding research topics.

### 4.1. Scholarly Knowledge Graphs

**SemOpenAlex.** SemOpenAlex is an extensive RDF knowledge graph that contains over 26 billion triples about scientific publications and their associated entities from all scientific domains. Containing more than 256 million publications, it can serve as a central hub for interlinking within the scientific domain.

**Linked Papers With Code.** LPWC is an RDF knowledge graph that comprehensively models the research field of machine learning. It contains information about more than 400,000 machine learning publications and describes the tasks addressed, the datasets utilized, the methods implemented, and the evaluations conducted, along with their results. It is tightly interlinked to SemOpenAlex with more than 1,2 million links between works and authors, making it a perfect fit for the KIRA system.

**Computer Science Knowledge Graph.** CS-KG contains facts and relations about research entities automatically extracted from scientific publications. Describing relationships between *Tasks*, *Methods*, *Materials* and *Metrics*, the facts are further annotated with information about frequency, co-occurrences, extraction time and extraction method.

**KIRA Project Knowledge Graph.** Based on well-established ontologies like *FOAF* and *schema.org*, the KIRA Project KG aims to reflect the research project including its key participants, their relations and artifacts produced in the context of the project. The data of this KG is hence not focused on actual scientific concepts but rather on data that helps to foster collaboration between the relevant actors in the project.

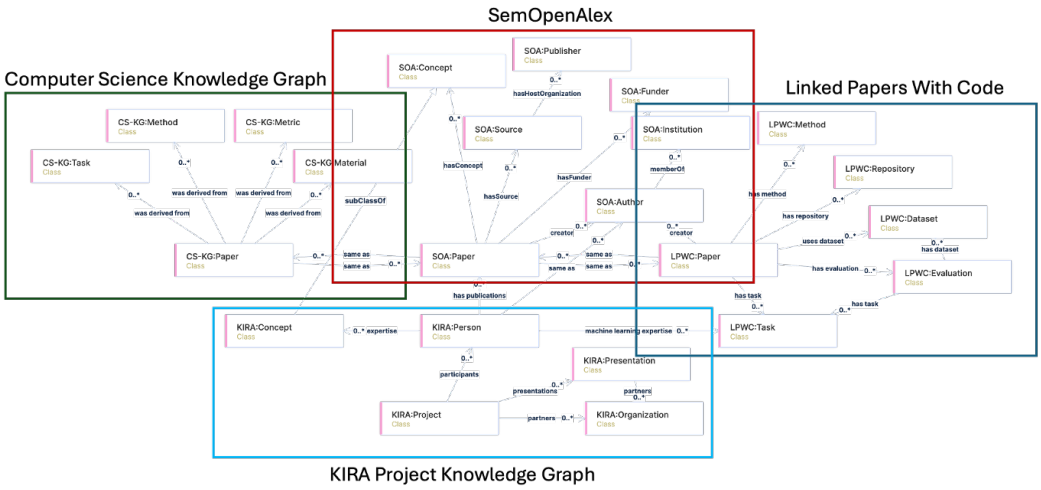


**Figure 1:** Integrated ontology for the research knowledge graph

### 4.2. Data Integration

The central hub of our data integration efforts are publications in SemOpenAlex. Both CS-KG and LPWC contain types for publications and provide references to SemOpenAlex publications either in the form of identifiers or as explicit *owl:sameAs* links (cf. Fig. 1). The facts in the KIRA Project KG are created on the basis of SemOpenAlex publications as well to ensure tight interlinking. Following these links, we can combine relevant content from multiple graphs, for example to find synergies and overlapping fields of expertise for project participants.

## 5. System Overview

### 5.1. Scope and Functionality

The main purpose of the presented system is to provide research assistance on top of scholarly knowledge graphs. This assistance encompasses several aspects:

1. **Research support**: (1) automating scientific routines, i.e. publications search, methods overview and comparison, etc., (2) facilitating research activities with AI methods;
2. **Access to relevant knowledge bases and publication databases**: (1) natural language interfaces to knowledge graphs, (2) domain exploration and methods application in a specific research context, (3) context-specific recommendations of scholarly artifacts;
3. **Personalized experience**: (1) a personal dashboard for every system user, (2) paper recommendations based on personal dashboard and user's interaction with the scholarly artifacts, (3) feedback and user activity tracking with the interaction knowledge graph;
4. **Support of collaboration through the knowledge graph**: (1) identifying synergies through partners via mapping activities in the project, (2) enabling exchange on specific topics, (3) sharing relevant content that is created or found during the project.

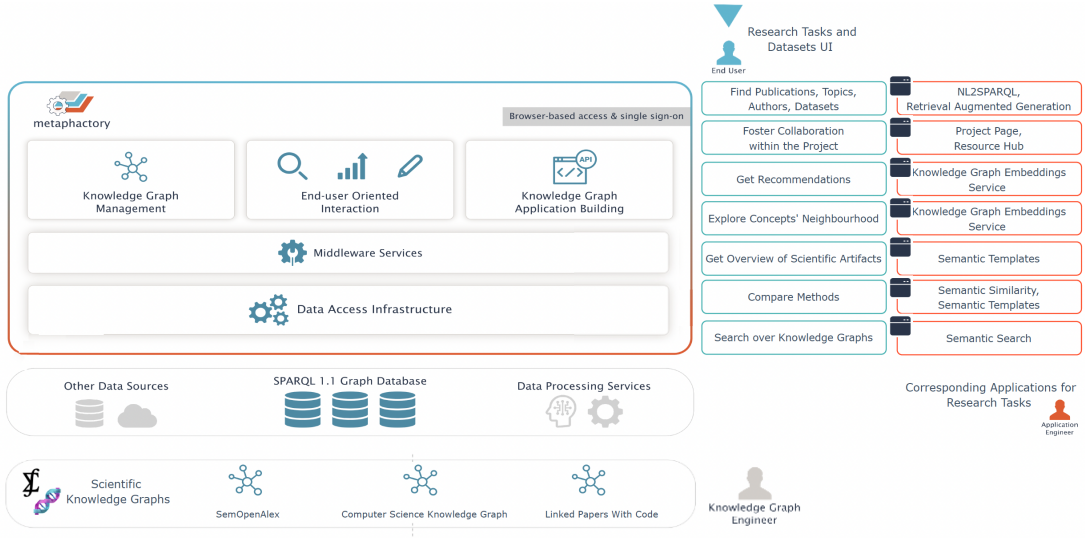


**Figure 2:** Architecture of the KIRA Research Support System

## 5.2. Architecture

The KIRA Research Support System is implemented on top of the metaphactory platform, having research tasks implemented in dedicated apps with the usage of built-in metaphactory semantic search, querying and templating functionality, as well as integrating external services (large language models, graph embeddings).

The metaphactory platform offers simple extension points to build, bundle and deploy lightweight ”apps” along with the platform without the need of changing the platform binary or re-compiling the platform. Sources of an ”app” can be put under the source control and a packaged version of an ”app” can be deployed as a docker container volume along with the platform.[8].

The architecture of the KIRA Research Support System is shown in Fig. 2. The integration of the respective knowledge graphs is provided in Sec. 4.2, and connections between tasks, datasets and methods are described in Table 1.

## 5.3. Mapping of Research Tasks, Knowledge Graphs, Neurosymbolic AI Methods, and Applications within the KIRA Research Support System

The KIRA Research Support System is a metaphactory App which hosts a plethora of resources and tools. Each of the research tasks mentioned in Sec. 2 is a component of that App, employing a combination of neurosymbolic AI methods.

To make these connections explicit, we provide a mapping between tasks, datasets and methods in Table 1.

**Table 1**
Mapping between Knowledge Graphs (Datasets), Research Tasks and Neurosymbolic AI Methods within the KIRA Research Support System

| Research Task | Knowledge Graph | Neurosymbolic AI Methods |
|---|---|---|
| *Find Publications, Topics, Authors, Datasets* | SemOpenAlex | Translation of natural language to structured queries (NL2SPARQL), In-context learning, Retrieval methods, Similarity measures |
| *Foster Collaboration within the Project* | KIRA KG | Semantic search, Semantic templates, Dashboarding |
| *Get Recommendations* | LPWC | KGEMs, Similarity measures |
| *Explore Concepts' Neighborhood* | LPWC | KGEMs, Similarity measures, t-SNE |
| *Get Overview of Scientific Artifacts* | CS-KG, LPWC, SemOpenAlex | Semantic search, Semantic templates |
| *Compare Methods* | CS-KG | Semantic search, Semantic templates, Co-occurence statistics |
| *Search Over Knowledge Graphs* | all above mentioned KGs | Semantic Search |

# 6. Overview of Neurosymbolic AI Methods Behind the "Research Routines"

## 6.1. Semantic Search and Querying

Semantic search over knowledge graphs is an advanced method for retrieving the unambiguous concepts and its relationships with a focus on concept meaning, rather than the surface form of a word. Focusing on the user's information need (intent), metaphactory semantic search framework allows the user to define complex information needs in an incremental way. The UI components supporting search definition to capture this information need and express it as a SPARQL query to be executed against the graph database. At the next step, it is possible to refine the set of search results using facets, and visualize the search results in different ways.

## 6.2. Dashboarding

Semantic templates in metaphactory handle formatting and rendering of the structured data (semantic search results) into final markup by applying features of the template language. The template language instructs where to render variables and provides basic flow control blocks. Thus, retrieved knowledge is surfaced as pages, dashboards and diagrams, presenting semantic search results to the end users.

### 6.3. Natural Language Translation to Queries

There is a plethora of methods to solve the problem of translating user's information needs, expressed in natural language, to query languages, in particularly, SPARQL. In the system, we implemented a zero-shot approach, based on augmenting LLMs with an RDFS ontology definition, required to construct the query, during prompting. This approach extends existing in-context learning approaches, proposed in [9], [10], by incorporating on-the-fly ontology conversion from SHACL to RDFS and entity detection and linking into the translation pipeline.

### 6.4. Knowledge Graph Embeddings

Knowledge graph embedding methods aim to learn dense vector representations of graph's nodes and edges to support knowledge graph completion and similarity search[11][12], as well as downstream tasks like recommendation systems and neighborhood visualization and exploration. In the KIRA system we utilize KGE models for the embedding of scientific concepts to support the search of similar concepts.

### 6.5. Similarity Measures

Similarity measures are based on a distance function and can be computed for different types of information objects (strings, numbers, tuples, vectors, images). In comparison to speech recognition, where edit-based measures (i.e. word error rate and its derivatives, based on Levenshtein distance) are more relevant, machine learning approaches (i.e. k-means algorithm in distance-based clustering, nearest neighbor search) learn similarities between objects in the vector space, actively employ token-based and hybrid measures (Euclidean distance, cosine similarity, Jaccard index, etc.)[13]. In the KIRA system we use similarity measures to retrieve methods close to a given method and for nearest neighbors search.

## 7. Conclusion and Future Work: Towards Collaborative Research Agents

The implemented knowledge graph-based system on top of the metaphactory platform aims to support researchers in their everyday research tasks and routines. This system has originated from the KIRA project, which goal was to optimize control of electric traction drives with AI methods. From this perspective, we have populated the system with the applications, leveraging AI neurosymbolic methods, in particularly, by adding *neural* (large language models and vector representations) to the metaphactory built-in *symbolic* methods (semantic search and semantic templates). The central part of the system are the scholarly knowledge graphs, and the project interaction graph, that altogether serve the users' needs in research assistance and foster collaboration through the knowledge graph.

In the future work, we aspire to design and integrate multi-agent workflows to address the challenges and limitations of deep learning models[14][15], in particularly, occurring during the in-context integration of LLMs and KGs, and extend research support tasks to human-AI collaborative discovery based on domain knowledge graphs.

## Acknowledgments

This work has been funded by the German ministry BMWK under project KIRA FKZ: 19I21030I.

## References

[1] P. Hitzler, A. Eberhart, M. Ebrahimi, M. K. Sarker, L. Zhou, Neuro-symbolic approaches in artificial intelligence, National Science Review 9 (2022) nwac035.

[2] 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.

[3] A. Meloni, S. Angioni, A. Salatino, F. Osborne, A. Birukou, D. Reforgiato Recupero, E. Motta, Aida-bot 2.0: enhancing conversational agents with knowledge graphs for analysing the research landscape, in: International Semantic Web Conference, Springer, 2023, pp. 400–418.

[4] B. McCann, R. Socher, Systems and methods for a language model-based customized search platform, 2024. US Patent App. 18/441,903.

[5] M. Färber, D. Lamprecht, J. Krause, L. Aung, P. Haase, Semopenalex: the scientific landscape in 26 billion rdf triples, in: International Semantic Web Conference, Springer, 2023, pp. 94–112.

[6] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Cs-kg: A large-scale knowledge graph of research entities and claims in computer science, in: International Semantic Web Conference, Springer, 2022, pp. 678–696.

[7] M. Färber, D. Lamprecht, Linked papers with code: the latest in machine learning as an rdf knowledge graph, arXiv preprint arXiv:2310.20475 (2023).

[8] P. Haase, D. M. Herzig, A. Kozlov, A. Nikolov, J. Trame, metaphactory: A platform for knowledge graph management, Semantic Web 10 (2019) 1109–1125.

[9] J. Sequeda, D. Allemang, B. Jacob, A benchmark to understand the role of knowledge graphs on large language model's accuracy for question answering on enterprise sql databases, in: Proceedings of the 7th Joint Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA), 2024, pp. 1–12.

[10] L. Kovriguina, R. Teucher, D. Radyush, D. Mouromtsev, Sparqlgen: One-shot prompt-based approach for sparql query generation., in: Proceedings of the Posters and Demo Track of the 19th International Conference on Semantic Systems (SEMANTiCS 2023), CEUR Workshop Proceedings, 2023.

[11] R. Biswas, L.-A. Kaffee, M. Cochez, S. Dumbrava, T. E. Jendal, M. Lissandrini, V. Lopez, E. L. Mencía, H. Paulheim, H. Sack, et al., Knowledge graph embeddings: open challenges and opportunities, Transactions on Graph Data and Knowledge 1 (2023) 4–1.

[12] P. Ristoski, H. Paulheim, Rdf2vec: Rdf graph embeddings for data mining, in: International semantic web conference, Springer, 2016, pp. 498–514.

[13] M. Van de Velden, A. Iodice D'Enza, A. Markos, Distance-based clustering of mixed data, Wiley Interdisciplinary Reviews: Computational Statistics 11 (2019) e1456.

[14] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, H. Lai, Y. Gu, H. Ding, K. Men, K. Yang, S. Zhang,

X. Deng, A. Zeng, Z. Du, C. Zhang, S. Shen, T. Zhang, Y. Su, H. Sun, M. Huang, Y. Dong, J. Tang, Agentbench: Evaluating llms as agents, 2023. URL: https://arxiv.org/abs/2308.03688. arXiv:2308.03688.

[15] X. Tang, Q. Jin, K. Zhu, T. Yuan, Y. Zhang, W. Zhou, M. Qu, Y. Zhao, J. Tang, Z. Zhang, A. Cohan, Z. Lu, M. Gerstein, Prioritizing safeguarding over autonomy: Risks of llm agents for science, 2024. URL: https://arxiv.org/abs/2402.04247. arXiv:2402.04247.

## A. ”Scholarly Artifacts Overview” Task Walkthrough

This task allows the user to see existing connections between scientific artifacts. These connections are extracted from the scholarly graphs, integrated on the KIRA system, and surfaced with semantic templates.

Firstly, from the KIRA system start page the user selects the task **Get Overview of Scientific Artifacts**, and secondly, the knowledge graph for exploration: CS-KG, LPWC, or SemOpenAlex.

Depending on the graph, the user can search across different combinations of artifacts, such as *Method*, *Task*, *Material*, *Metric*, *Repository*, *Dataset*, *Evaluation*, *Paper*, *Conference*, and some other artifacts.

Let user search for *genetic algorithm* method. After choosing the method from the search results, the user is taken to the artifact overview page. In Fig. 3, connections between scientific artifacts across CS-KG and SemOpenAlex are shown for this particular method.

## B. Abbreviations

**CS-KG** - Computer Science knowledge graph;
**KGE** - Knowledge graph embeddings;
**KGEM** - Knowledge graph embedding model;
**KIRA KG** - KIRA Project knowledge graph, describing project content, partners (both researchers and organizations), their interactions and artifacts, created during the project;
**LLM** - Large language model;
**LPWC** - Linked Papers With Code knowledge graph.


**Figure 3:** Overview of Scientific Artifacts related to the *Genetic Algorithm* Method
*Notes:* The image shows the related artifacts and its statistics for the *genetic algorithm* method. Besides exploring the connections, like authors, publications, institutions, and topics, the user can further refine interactions between a method and its related tasks. In the implemented template, *topics* and *method* and *task* relations are retrieved from CS-KG, and information about *authors*, *publications* and *institutions* is integrated from SemOpenAlex.