

# Advancing Scientific Knowledge Retrieval and Reuse with a Novel Digital Library for Machine-Readable Knowledge

Hadi Ghaemi<sup>1</sup>, Lauren Snyder<sup>1</sup>, Markus Stocker<sup>1</sup>

<sup>1</sup>TIB - Leibniz Information Centre for Science and Technology

Hadi.Ghaemi@tib.eu, Lauren.Snyder@tib.eu, Markus.Stocker@tib.eu

## Abstract

Digital libraries for research, such as the ACM Digital Library or Semantic Scholar, do not enable the machine-supported, efficient reuse of scientific knowledge (e.g., in synthesis research). This is because these libraries are based on document-centric models with narrative text knowledge expressions that require manual or semi-automated knowledge extraction, structuring, and organization. We present ORKG reborn, an emerging digital library that supports finding, accessing, and reusing accurate, fine-grained, and reproducible machine-readable expressions of scientific knowledge that relate scientific statements and their supporting evidence in terms of data and code. The rich expressions of scientific knowledge are published as reborn (born-reusable) articles and provide novel possibilities for scientific knowledge retrieval, for instance by statistical methods, software packages, variables, or data matching specific constraints. We describe the proposed system and demonstrate its practical viability and potential for information retrieval in contrast to state-of-the-art digital libraries and document-centric scholarly communication using several published articles in research fields ranging from computer science to soil science. Our work underscores the enormous potential of scientific knowledge databases and a viable approach to their construction.

**Comments:** Accepted for publication at SIGIR 2025.

This is the authors' accepted version (preprint) prior to ACM formatting and copyediting.

The official version of record is available in the ACM Digital Library at:

<https://doi.org/10.1145/3726302.3730134>

## 1 Introduction

Research is an iterative process in which researchers build on existing knowledge to create new knowledge. Synthesis research is a key tool for creating new knowledge, and relies on primary scientific literature (i.e., scientific findings that are published for the first time) as a data source. For example, systematic reviews and meta-analyses are based on the integration and analysis of data extracted from published literature [20, 5]. The validity of such syntheses depends on the quality of the primary scientific literature that is being synthesized [5]. To assess this quality, researchers need to understand the data and methods underlying the scientific statements made in the literature.

Digital libraries, such as the ACM Digital Library or similar, support access to scientific articles, but are document-centric and the knowledge they contain is not machine-readable, meaning that to reuse this knowledge, researchers must first extract it. Manual data extraction is time consuming and semi-automated approaches usually trade extraction accuracy for speed to enable discovery and access to related data, code, or other supplementary materials. Digital libraries often interlink with artefacts distributed on specialized platforms such as GitHub, Kaggle, or (domain-specific) data repositories. Nonetheless, the ability of machines to reliably process scientific knowledge expressed in heterogeneous formats and published as sets of distributed files continues to be very limited.

With ORKG reborn—accessible at [reborn.orkg.org](https://reborn.orkg.org)—we rethink the publication of and access to machine-readable scientific knowledge in digital libraries for research. Inspired by [8], we use a conceptual model that organizes scientific knowledge (specifically, research findings expressed in articles) as statements and supporting evidence, expressed as machine-readable data with formal syntax. The proposed system consists of three interconnected layers: 1) the data deposition and collection layer, which facilitates the harvesting of reborn article [18] data published on distributed data repositories; 2) the knowledge organization layer, which provides a centralized database for the management and retrieval of reborn article data; and 3) the presentation layer, which

renders reborn article data as structured scientific knowledge in the form of statements and supporting evidence in a user-friendly Web-based interface.

## 2 Related Work

The academic expansion of the past decades has dramatically increased the volume of published articles, scientific knowledge, and related artefacts such as data and code. The global scholarly infrastructure [3] is designed to provide reliable and comprehensive access to these artefacts [19]. Traditional systems enable discovery and access with persistent identifiers and associated standardized metadata (information about articles, authors, and datasets). Some systems, such as Zenodo<sup>1</sup> support the deposition and interlinking of data and code. With the exception of specialized systems (e.g., domain-specific data repositories) these systems are generally unable to support direct access to and retrieval of content, be it research data in data files or scientific knowledge expressed in articles [1, 4, 14].

A new generation of systems aims to enable access to the content of articles by supporting collaborative manual or semi-automated extraction of knowledge from articles and collaborative curation and organization of extracted knowledge. In addition to the Open Research Knowledge Graph (ORKG) [9], Hi-Knowledge<sup>2</sup> [11] aims to enable novel presentation and access to scientific knowledge in invasion biology and urban ecology, with the vision of creating an interactive atlas of knowledge covering other research fields. Similar systems can be found in other domains, e.g., Machine Learning<sup>3</sup>. A fundamental limitation of these systems is their reliance on post-publication knowledge extraction from articles, which is time consuming and error prone when done manually and requires tradeoffs in accuracy when completed with (semi)-automated approaches.

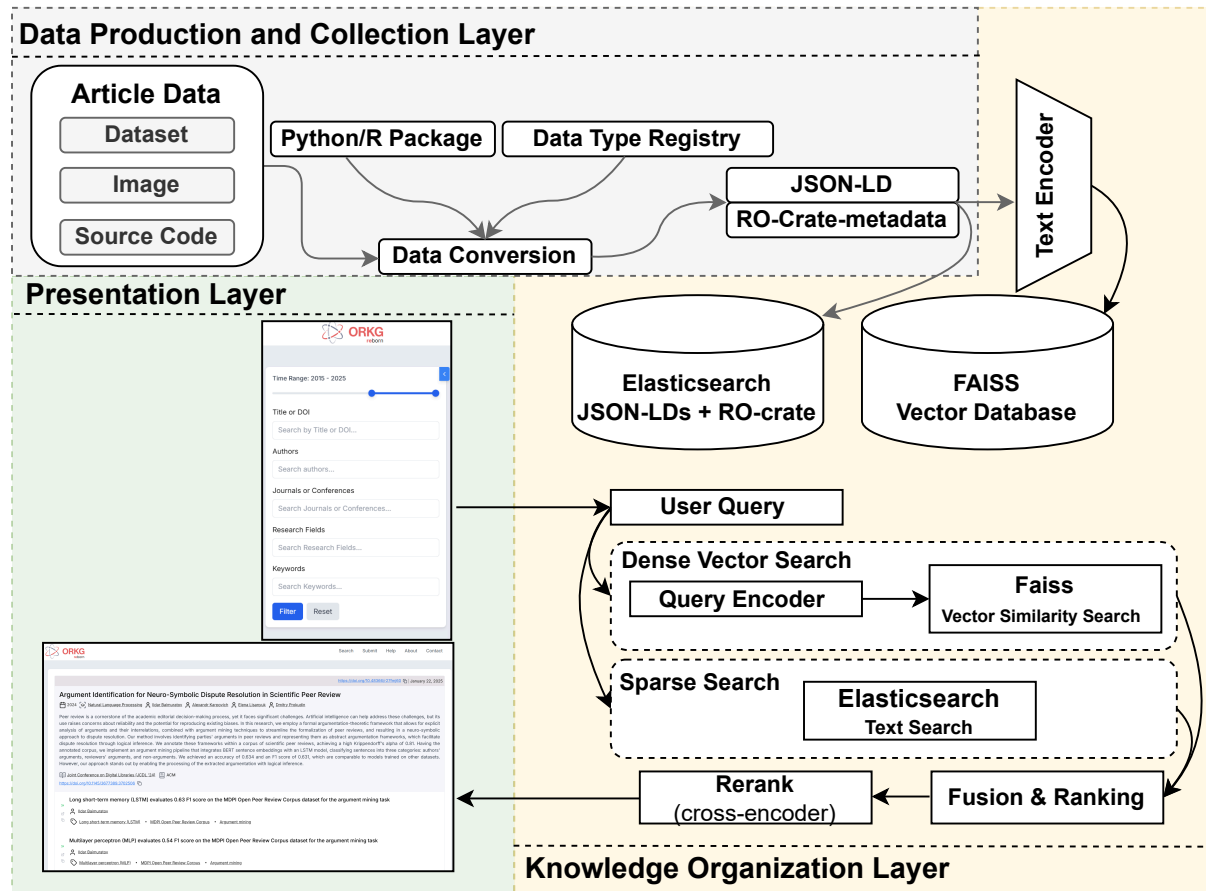


Figure 1: Proposed system architecture showing the three main layers: Data Deposition and Collection Layer, Knowledge Organization Layer, and Presentation Layer.

<sup>1</sup><https://zenodo.org>

<sup>2</sup><https://hi-knowledge.org>

<sup>3</sup><https://paperswithcode.com>

### 3 Architectural Overview

We now describe the architecture of the proposed system. As stated above, its aim is to publish machine-readable scientific statements and underlying evidence, with the aim of supporting synthesis research in particular. Toward this aim, it is necessary to provide reliable access not only to scientific findings, but also to related research data and source code. Figure 1 illustrates how we divide the overall architecture into three main layers: data deposition and collection layer, knowledge organization layer, and presentation layer.

#### 3.1 Data Deposition and Collection Layer

To efficiently reuse research data, it is necessary to make data understandable to both humans and machines. This requires formalizing data types. We leverage a Data Type Registry (DTR) to manage the standardization and registration of, as well as access to, data types required to describe scientific statements and their underlying evidence. For unambiguous identification, each data type in the DTR is assigned a DOI as a unique, resolvable persistent identifier.

Table 1: Data types and their definition.

Data type	Definition
<a href="#">Data Preprocessing</a>	Prepare datasets for further analysis
<a href="#">Descriptive Statistics</a>	Describe dataset characteristics
<a href="#">Algorithm Evaluation</a>	Algorithm performance evaluation for a specified task and dataset
<a href="#">Multilevel Analysis</a>	Multilevel data analysis with models including fixed and random effects
<a href="#">Correlation Analysis</a>	Evaluate the strength and direction of the relationship between variables
<a href="#">Group Comparison</a>	Compare the means of two or more groups
<a href="#">Regression Analysis</a>	Explore the relationship between dependent variable(s) and independent variable(s)
<a href="#">Class Prediction</a>	Predict a categorical class label given input data
<a href="#">Class Discovery</a>	Discover classes or clusters in unlabeled data
<a href="#">Factor Analysis</a>	Discern latent factors in data

Our current focus is on scientific statements where the underlying evidence is a statistical data analysis or set of data analyses, for which we have developed the data types listed in Table 1. Figure 2 illustrates the data type for ‘Data Preprocessing’, which captures information related to executed software methods, and input and output data items. For example, in this data type, software methods refer to functions in libraries that are part of (Python, R, etc.) software. Data items may be sourced from a tabular data structure or from a URL, are characterized by a matrix size and components, and may be expressed as a figure. For a full description of the ‘Data Preprocessing’ data type, refer to [doi:10.21969/37182ecfb4474942e255](https://doi.org/10.21969/37182ecfb4474942e255).

The data describing scientific statements and their underlying evidence are deposited in JSON-LD format on a standard data repository. We leverage a CKAN-based system called the Leibniz Data Manager [2]. The proposed system deposits these data as RO-Crates [16] with additional metadata about the article, authors, journals, publishers, statements, and data. Figure 3 shows the structure and graph of the RO-Crate metadata file. The root of this file consists of the `@context` and the `@graph`. The `@context` consists of the RO-Crate specification URL and the `@graph` includes different entity types described with their properties. Possible entity types include Dataset, Person, Publisher, Concept, Component, File, and Statement. By interlinking the data deposition with the original article in DOI metadata, we enable the discovery of this data by article DOI, and hence data collection in the proposed system.

#### 3.2 Knowledge Organization Layer

The Knowledge Organization Layer represents the system’s core data management and processing infrastructure. At its heart are two primary storage systems: Elasticsearch and the open source Faiss (Facebook AI Similarity Search) library [12], each serving distinct but complementary purposes. Elasticsearch stores the structured JSON-LD representations of scientific statements and supporting evidence along with their associated RO-Crate metadata, thus providing robust document storage and traditional text search capabilities. The system uses a Text

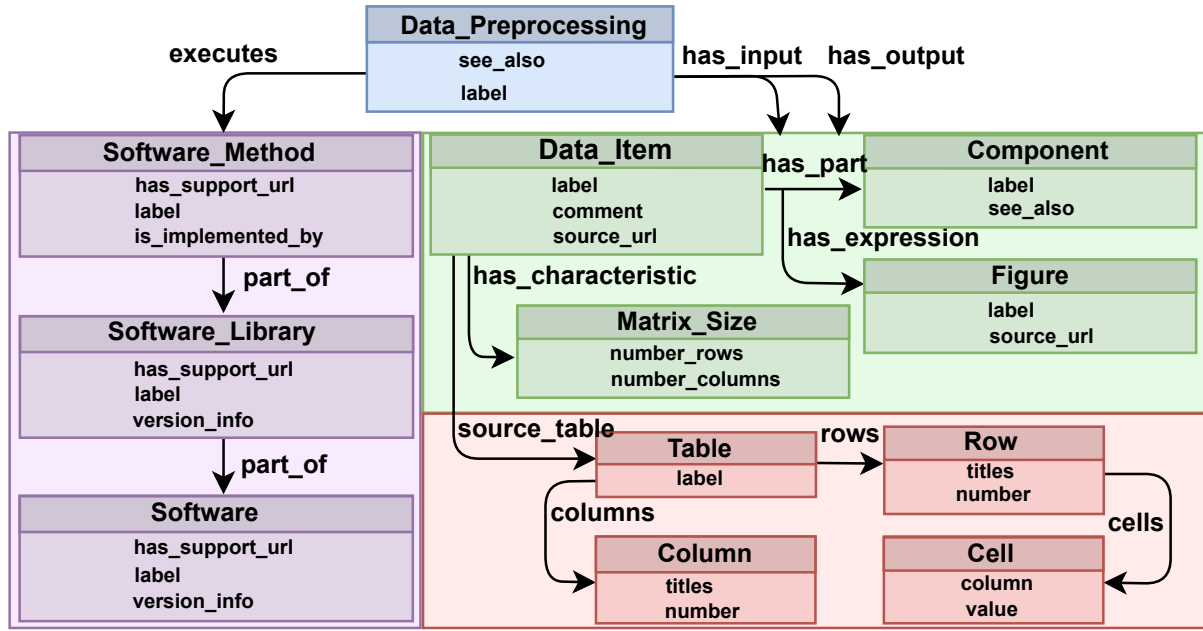


Figure 2: Diagram of the ‘Data Preprocessing’ data type describing the executed procedure, the utilized input data, and produced output data.

Encoder as a bridge between raw text and vector representations, processing incoming data and preparing data for semantic search capabilities.

### 3.3 Presentation Layer

The Presentation Layer implements a Web-based user interface and supports interacting with the scientific knowledge database as well as knowledge retrieval and reuse. Figure 4 illustrates how ORKG reborn presents a reborn article in terms of scientific statements (from here on referred to simply as statements) with supporting evidence in a structured manner that is intuitive for users. Each reborn article is presented with essential bibliographic metadata (title, authors, abstract, journal, DOI to the original PDF article) and a list of scientific statements published in the original work. Each reborn article is also assigned its own DOI, distinct from the original publication. Statements are annotated with concepts, which are described and utilized in the search interface. When statements are selected, the underlying evidence is displayed (Figure 4). Supporting evidence for statements is composed of different sections. In Figure 4, Section 1 describes the data analysis and its parts, e.g., a descriptive statistic. For each data analysis part, the system describes (a) the executed procedure (Section 2) in terms of the executed function of a specific package in Python or R languages; (b) the utilized input data (Section 4) and produced output data (Section 5), possibly associated with figures; (c) essential components such as target variables, also known as response variables (Section 3); as well as the full implementation of the data analysis in Python or R code (Section 6). As a result, the research methods and results are transparent, reproducible, and machine-readable, enabling other researchers to more easily confirm or reuse the published scientific knowledge.

## 4 Implementation

The presented system is deployed using Docker Compose and comprises a backend, a frontend, and infrastructure containers. The backend is implemented as a Web service in Python, and the frontend is implemented in Next.js. The textual information is indexed using Elasticsearch.

As suggested in Figure 1, the user interface features a search interface that handles dense and sparse natural language queries. The dense search path uses the Query Encoder (powered by all-MiniLM-L6-v2<sup>4</sup>) to convert the user’s query into a vector representation, which is then used by Faiss for vector similarity search. Concurrently, the keyword search path utilizes Elasticsearch’s text search capabilities to find relevant documents based on traditional

<sup>4</sup><https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2>

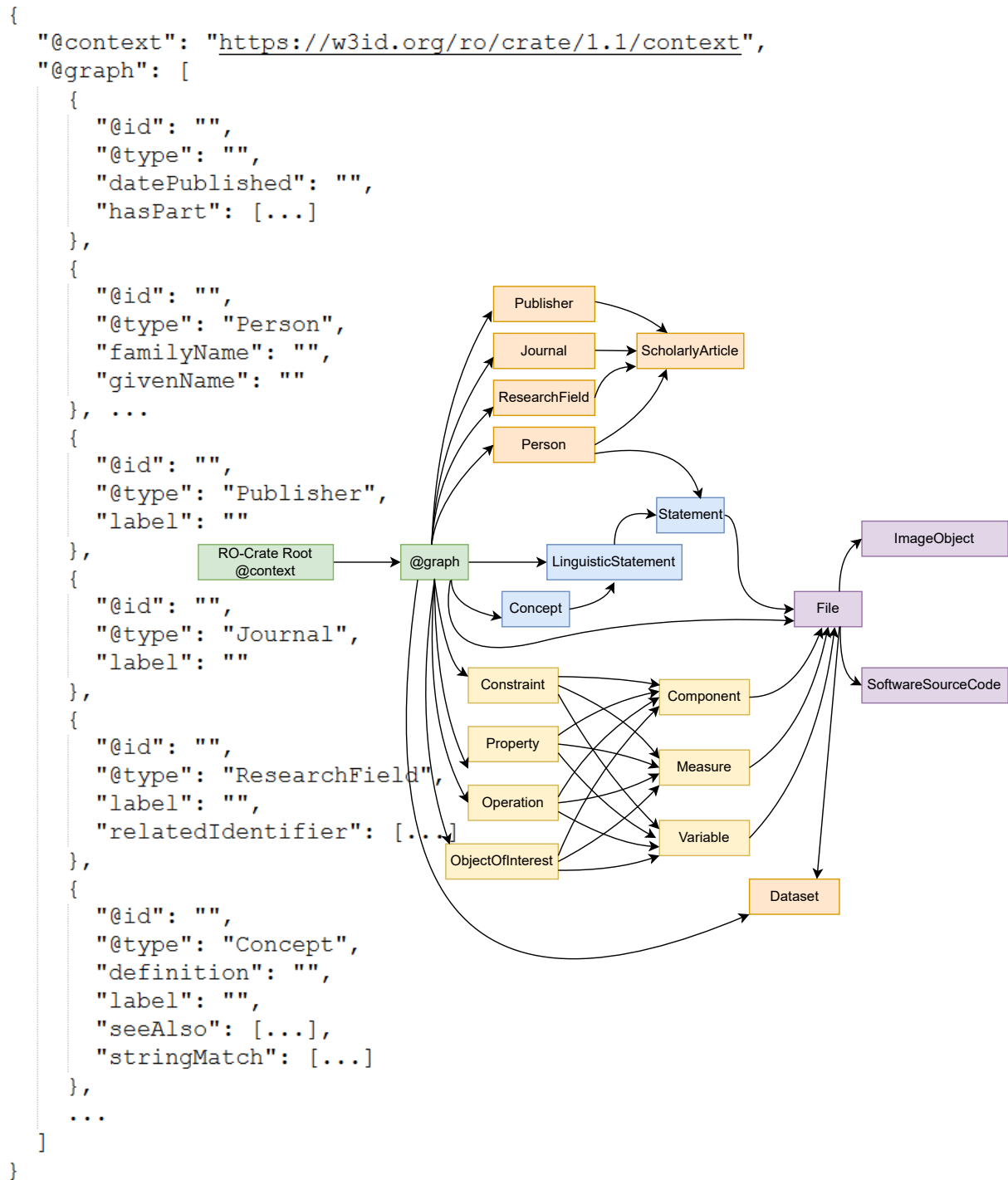


Figure 3: Overview of the RO-Crate metadata file structure.

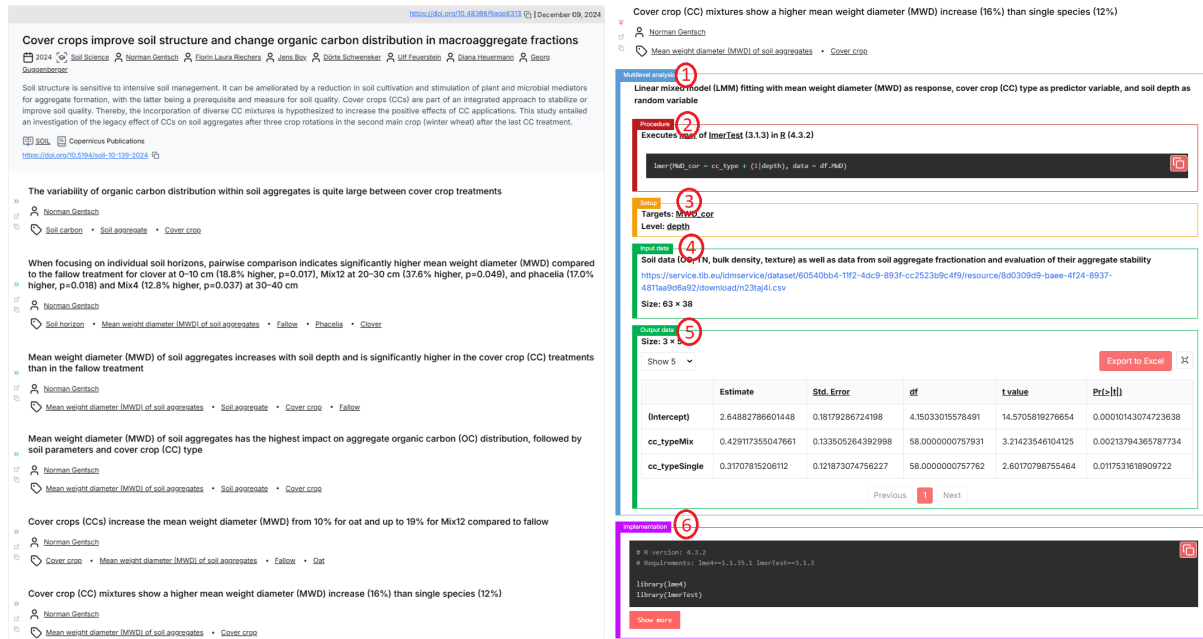


Figure 4: Scientific statements and supporting evidence as originally published by Gentsch et al. [7] presented here as a reborn article accessible in the ORKG reborn digital library. (left) A reborn article presenting the original research findings as structured scientific statements and supporting evidence (doi:10.48366/5eqe8313). (right) Display of a scientific statement and supporting evidence in terms of a data analysis described by the executed procedure, utilized input data, produced output data, and full implementation in source code.

keyword matching algorithms. The results from both search paths are then combined through the Fusion & Ranking component, which applies intelligent weighting to blend the results from both approaches. This hybrid approach ensures that users receive comprehensive results that capture both semantic relevance and keyword precision. The presentation layer then receives these ranked results and displays them in a clear and organized manner.

We utilize Faiss over Elasticsearch's dense vector retrieval because of its high performance, GPU acceleration, and flexible index options, which optimize our semantic search for large-scale, low-latency requirements. FlatIndex supports the search for the exact nearest neighbor. As data increases, we can use approximation techniques to improve scalability. For instance, Hierarchical Navigable Small World (HNSW) [13] finds nearest neighbors based on graphs, Inverted File Index (IVF) [15] partitions vector spaces, and Product Quantization (PQ) [10] compresses vectors to reduce memory usage.

For sentence embedding, we use the allmpnet-base-v2<sup>5</sup> model, based on MPNet [17] and BERT [6]. Following the embedding process, Faiss serves as the vector database, optimized for storing and retrieving high-dimensional vector embeddings. Complementing this vector-based approach, Elasticsearch provides robust full-text search. The integration of Faiss for vector storage and Elasticsearch for document storage establishes a comprehensive system that effectively handles both semantic and keyword-based queries. Finally, to perform the search, we implement a weighted score fusion methodology that integrates Elasticsearch and Faiss results. Furthermore, to enhance result quality, we use a cross-encoder (cross-encoder/ms-marco-MiniLM-L-6-v2) for re-ranking the top-10 aggregated results.

## 5 Conclusions and Future Work

This paper is a first, concise presentation of ORKG reborn, an emerging digital library that advances the machine-based retrieval and reuse of scientific knowledge. ORKG reborn is fueled by an initiative that advocates transitioning from born-digital, unstructured expressions of scientific knowledge in the form of traditional PDF articles to born-reusable and high-quality expressions of scientific knowledge that are produced machine-readable from the outset. By adopting a pre-publication approach, aka reborn articles, the proposed system and initiative distinguish

<sup>5</sup><https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

themselves from classical approaches centered around post-publication knowledge extraction from articles.

Building on a three-layered architecture, ORKG reborn supports the collection of reborn article data published in distributed data repositories, the organization of such machine-readable data in a centralized scientific knowledge database, and user-friendly presentation of the data in a Web-interface that supports data search, access, and download as packaged data collections.

In future work, we will further advance system capabilities along several directions. First, we will broaden the supported scientific knowledge types. Of particular interest are mathematical statements and supporting proofs. Second, enabled by the ZIP-packaged download of selected statements, we will develop use case projects in synthesis research that leverage ORKG reborn for scientific knowledge integration and synthesis. Third, inspired by Papers with Code, we will implement in ORKG reborn the display of algorithm evaluations as leaderboards and extend such specialized visualizations to other scientific knowledge types.

## 6 Acknowledgments

The authors gratefully acknowledge the contributions of several colleagues, in particular Olga Lezhnina, Lars Vogt, and Manuel Prinz. This work has been supported by the Leibniz-Lab “Systemic Sustainability” (GA LL-2024-SYSTAIN) funded by the Leibniz Association and the Horizon Europe project FAIR2Adapt (GA 101188256) funded by the European Union.

## References

- [1] Amir Aryani, Marta Poblet, Kathryn Unsworth, Jingbo Wang, Ben Evans, Anusuriya Devaraju, Brigitte Hausstein, Claus-Peter Klas, Benjamin Zapilko, and Samuele Kaplun. A research graph dataset for connecting research data repositories using rd-switchboard. *Scientific data*, 5(1):1–9, 2018.
- [2] Anna Beer, Mauricio Brunet, Vibhav Srivastava, and Maria-Esther Vidal. Leibniz data manager—a research data management system. In *European Semantic Web Conference*, pages 73–77, Greece, 2022. Springer, Cham.
- [3] Christine L Borgman. *Scholarship in the digital age: Information, infrastructure, and the Internet*. MIT press, Cambridge, MA, 2010.
- [4] Adrian Burton, Hylke Koers, Paolo Manghi, Markus Stocker, Martin Fenner, Amir Aryani, Sandro La Bruzzo, Michael Diepenbroek, and Uwe Schindler. The scholix framework for interoperability in data-literature information exchange. *D-Lib Magazine*, 23(1/2), 2017.
- [5] Joke Depraetere, Christophe Vandeviver, Ines Keygnaert, and Tom Vander Beken. The critical interpretive synthesis: an assessment of reporting practices. *International Journal of Social Research Methodology*, 24(6):669–689, 2021.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [7] Norman Gentsch, Florin Laura Riechers, Jens Boy, Dörte Schweneker, Ulf Feuerstein, Diana Heuermann, and Georg Guggenberger. Cover crops improve soil structure and change organic carbon distribution in macroaggregate fractions. *Soil*, 10(1):139–150, 2024.
- [8] Alexander Hars. Designing scientific knowledge infrastructures: the contribution of epistemology. *Information Systems Frontiers*, 3:63–73, 2001.
- [9] Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D’Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th international conference on knowledge capture*, pages 243–246, USA, 2019. Association for Computing Machinery.
- [10] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010.



- [11] J.M. Jeschke, M. Enders, M. Bagni, D. Aumann, P. Jeschke, M. Zimmermann, and T Heger. Hi-knowledge.org, version 2.0. <https://www.hi-knowledge.org>, 2020. Accessed: 2025-04-23.
- [12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.
- [13] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.
- [14] Afshin Sadeghi, Christoph Lange, Maria-Esther Vidal, and Sören Auer. Integration of scholarly communication metadata using knowledge graphs. In *Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings 21*, pages 328–341, Cham, 2017. Springer, Springer.
- [15] Sivic and Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings ninth IEEE international conference on computer vision*, pages 1470–1477, Nice, France, 2003. IEEE, IEEE.
- [16] Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, and Carole Goble. Packaging research artefacts with RO-Crate. *Data Science*, 5(2):97–138, January 2022.
- [17] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.
- [18] Markus Stocker, Lauren Snyder, Matthew Anfusio, Oliver Ludwig, Freya Thießen, Kheir Eddine Farfar, Muhammad Haris, Allard Oelen, and Mohamad Yaser Jaradeh. Rethinking the production and publication of machine-reusable expressions of research findings. *Scientific Data*, 12(1):677, 2025.
- [19] Shilpa Verma, Rajesh Bhatia, Sandeep Harit, and Sanjay Batish. Scholarly knowledge graphs through structuring scholarly communication: a review. *Complex & intelligent systems*, 9(1):1059–1095, 2023.
- [20] Tianqi Yu, Lifeng Lin, Luis Furuya-Kanamori, and Chang Xu. Synthesizing evidence from the earliest studies to support decision-making: To what extent could the evidence be reliable? *Research synthesis methods*, 13(5):632–644, 2022.