

# 4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment

Angelo Salatino<sup>1</sup>, Andrea Mannocci<sup>2</sup>, Francesco Osborne<sup>1</sup>, Georg Rehm<sup>3</sup> and Sonja Schimmler<sup>4</sup>

<sup>1</sup>The Open University, Milton Keynes (UK)

<sup>2</sup>Italian Research Council (CNR), Pisa (IT)

<sup>3</sup>Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin (DE)

<sup>4</sup>TU Berlin, Fraunhofer FOKUS, Berlin (DE)

## 1. Forewords

The International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment (Sci-K 2024) is now running its fourth edition and will be co-located with the 23rd International Semantic Web Conference, the premier conference in the *Semantic Web*. The Sci-K workshop is a venue that brings together researchers and practitioners from different disciplines (including, but not limited to, Digital Libraries, Information Extraction, Machine Learning, Semantic Web, Knowledge Engineering, Natural Language Processing, Scholarly Communication, Science of Science, Scientometrics and Bibliometrics), as well as professionals from the industry, to explore innovative solutions and ideas for the production and consumption of Scientific Knowledge Graphs (SKG) and assessing the research impact. The workshop has called for high-quality submissions around the three main themes of research related to scientific knowledge: *representation*, *discovery*, and *assessment*.

In response to the call for papers, the workshop has received 13 outstanding submissions from researchers in 8 different countries: United States of America, Germany, United Kingdom, Brazil, Italy, France, Austria, and Portugal. Each paper was reviewed at least by three members of the program committee. Given the quality and the interesting topics covered by the submissions, we accepted 9 papers. The full program can be found on the Sci-K website: <https://sci-k.github.io/2024/>, whereas in the next section, we briefly introduce the accepted papers.

Sci-K 2024 builds on three previous successful editions and keeps attracting a combined pool of attendees. The first edition (Sci-K 2021) was held on 13 April 2021 in conjunction with The Web Conference 2021. Its program consisted of two keynote talks and the presentation of 11 research papers. The second edition (Sci-K 2022) took place on 26 April 2022 at The Web Conference 2022. The program included the presentation of 5 long papers, 4 short papers, 2 vision papers, 2 keynote speeches and a panel on “What’s next after Microsoft Academic Graph?” [1]. The third edition (Sci-K 2023) took place on 30 April 2023 at The Web Conference 2023. The program included the presentation of 4 long papers, 6 short papers, and a keynote speech from Matt Buys, Executive Director of DataCite, on the topic “Scaling the Global Data Citation Corpus: An International Collaboration” [2].

Moreover, the Sci-K series<sup>1</sup> spawned from the collaboration of two former workshops, namely, the *Scientific Knowledge Graphs Workshop (SKG 2020)*, and the *Workshop on Assessing Impact and Merit in Science (AIMinScience 2020)*, held in conjunction with the 2020 edition of the *International Conference*

---

Sci-K 2024: 4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment

✉ [angelo.salatino@open.ac.uk](mailto:angelo.salatino@open.ac.uk) (A. Salatino); [andrea.mannocci@isti.cnr.it](mailto:andrea.mannocci@isti.cnr.it) (A. Mannocci); [francesco.osborne@open.ac.uk](mailto:francesco.osborne@open.ac.uk) (F. Osborne); [georg.rehm@dfki.de](mailto:georg.rehm@dfki.de) (G. Rehm); [sonja.schimmler@tu-berlin.de](mailto:sonja.schimmler@tu-berlin.de) (S. Schimmler)

🌐 <https://salatino.org> (A. Salatino); <https://andremann.github.io> (A. Mannocci); <https://people.kmi.open.ac.uk/francesco/> (F. Osborne); <http://georg-rehm.com> (G. Rehm); <https://sites.google.com/view/sonjaschimmler/> (S. Schimmler)

🆔 0000-0002-4763-3943 (A. Salatino); 0000-0002-5193-7851 (A. Mannocci); 0000-0001-6557-3131 (F. Osborne); 0000-0002-7800-1893 (G. Rehm); 0000-0002-8786-7250 (S. Schimmler)



© 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>Sci-K History – <https://sci-k.github.io/history>

on *Theory and Practice of Digital Libraries (TPDL)* on the 25th of August 2020 [3]. Accepted papers from these two workshops have been invited to submit extended versions to a special issue we guest-edited at the open-access journal *Quantitative Science Studies (QSS)*, a leader in the field of *Science of Science* and *Scientometrics*. The special issue, titled “Scientific Knowledge Graphs and Research Impact Assessment”, provides an overview of recent advances in scholarly knowledge representation and research assessment featuring 10 peer-reviewed articles [4]. The synergy between the SKG 2020 and AIMinScience 2020 workshops and this QSS special issue helped define the focus and goals of the Sci-K workshop series.

## 2. Accepted Papers

Here we briefly introduce the accepted papers according to the main themes of the workshop.

### 2.1. Representation

Tsaneva et al. [5] in “Enhancing Scientific Knowledge Graph Generation Pipelines with LLMs and Human-in-the-Loop” examine methods to enhance the quality of automatically generated Scientific Knowledge Graphs. The authors argue that while automated approaches are efficient, they often compromise on accuracy. By employing an LLM in the validation process, they were able to boost the precision of the generated knowledge graph by 12%. Additionally, combining LLM validation with a human-in-the-loop approach further refined the results, leading to an improved F1 score. These findings suggest a promising path towards creating more accurate and reliable SKGs.

In “Knowledge Graph Enabled Scientific Data Repositories”, Pinheiro et al. [6] address the limitations of current data repositories in integrating and harmonising data. They propose a Human-Aware Data Acquisition Infrastructure (HADatAc) that implements FAIR principles to systematically represent scientific knowledge. By employing metadata templates and ontologies, HADatAc facilitates data browsing, search, summarisation, and dataset generation. Its extensive use across various institutions demonstrates its effectiveness in scientific data management and sharing.

Aggarwal et al. [7], in “Identifying Semantic Relationships Between Research Topics Using Large Language Models in a Zero-Shot Learning Setting” explore the use of LLMs to automatically identify semantic relationships between research topics, a task crucial for building efficient Knowledge Organisation Systems. Their analysis of six open and lightweight LLMs reveals Dolphin2.1-OpenOrca-7B as a strong performer, achieving a 0.853 F1-score in identifying broader, narrower, same-as, and other relationships. These encouraging findings advance the development of tools for automating KOS curation, facilitating easier exploration of scientific literature.

### 2.2. Discovery

In “Enhancing Scientific Discovery and Decision-Making: A Knowledge Graph-based Research Support System”, Kovriguina et al. [8] introduce a neurosymbolic AI system designed to enhance research and discovery activities. Leveraging metaphactory and open scholarly knowledge graphs like SemOpenAlex, LPWC, and CS-KG, the system offers researchers various AI-powered tools. These include a generative natural language interface for seamless interaction with knowledge graphs, semantic template-based retrieval, and neighbourhood exploration through KG embeddings. These capabilities facilitate research tasks such as gaining overviews of scientific artefacts, searching for publications, and receiving recommendations. The paper provides a comprehensive overview of the system’s design, architecture, data landscape, and functionalities, highlighting its practical application in an ongoing research project that focuses on utilising AI methods for electric traction drive design.

Jacyszyn et al. [9] in “DiTraRe: AI on a Spider’s Web. Interweaving Disciplines for Digitalisation” examine the wide-ranging impacts of research digitalisation on various levels. The project focuses on four interconnected research clusters and dimensions, each with unique perspectives and research questions, aiming to create both practical solutions and generalised knowledge. The paper highlights

the ‘Exploration and knowledge organisation’ dimension, which employs AI techniques like natural language processing and ontology engineering to dive into research transformation.

In “Federated Querying of Scholarly Communication Infrastructures”, Haris et al. [10] address the challenge of accessing exponentially growing scholarly knowledge spread across different infrastructures. They propose a federated interface that seamlessly retrieves inter-related data from various platforms like ORKG, DataCite, OpenAIRE, and SemanticScholar. The authors demonstrate the practicality and added value of their approach by showcasing how it can cater to the diverse information needs of different stakeholders.

### 2.3. Assessment

Schuster et al. [11] in “Assessing the Reliability and Scientific Rigor of References in Wikidata” investigate the usage, types, and scientific value of references in Wikidata, a growing open knowledge graph. They conduct a heuristic evaluation of Wikidata references, focusing particularly on DOIs, which link to scientific publications. The analysis shows extensive use of DOIs within Wikidata, particularly in the social and physical sciences. To assess the quality of these scholarly resources, the authors use percentile metrics and h-index indicators from OpenAlex. Interestingly, their findings reveal a broader representation of scholarly contributions in Wikidata compared to OpenAlex, with mid-tier and emerging journals receiving higher citation rates.

In “Ensuring FAIRness in Machine Learning Projects”, Efeoglu et al. [12] highlight the achievements of subsymbolic AI in tasks like question answering and ontology matching while pointing out the limitations due to the lack of transparency in LLMs’ training datasets and source codes. They propose a comprehensive schema for ML model metadata aligned with the FAIR principles to address this issue, aiming to facilitate the evaluation and promotion of FAIRness in ML development.

Spahiu et al. [13] in “Skills and Expertise in Large Organisations: An Enterprise Knowledge Graph Approach” address the challenge of representing complex enterprise information in a way that is both human and machine-readable. They focus specifically on the skills and expertise of individuals within the organisation, using publicly available datasets on US patents and scholarly publications to extract skills and represent them in an Enterprise Knowledge Graph (EKG). The resulting EKG enables querying and analysing employees’ skills, facilitating the identification of experts in specific domains.

## 3. Keynotes

For this edition, we are excited to have Natasha Noy from Google, who will talk to us about the importance of structured data and, more specifically, scholarly data.

Natasha Noy is a research scientist at Google Research, focusing on making structured data more accessible and usable. She is known for her work in the Semantic Web field, particularly with the development of the Protégé ontology editor and the Prompt alignment tool. Noy has also made significant contributions to ontology building and alignment, collaborative ontology engineering, and the development of Google Dataset Search.

She has received several awards and recognitions for her work, including the AAAI Classic Paper award (2018), and has been elected as an AAAI Fellow (2020) and an ACM Fellow (2023).

We are honoured to have her at Sci-K and look forward to her keynote!

## 4. Program Committee

We would like to express our heartfelt gratitude to the members of the programme committee, who, in the middle of August, dedicated time to reviewing papers.

- Miriam Albers (ZB MED Information Centre for Life Sciences)
- Miriam Baglioni (ISTI – CNR)

- Francisco Bolanos (KMi, The Open University)
- Leyla Jael Castro (ZB MED Information Centre for Life Sciences)
- Serafeim Chatzopoulos (Athena Research Center)
- Patricia Feeney (CrossRef)
- Daniel Garijo (Universidad Politécnica de Madrid)
- Mohamad Yaser Jaradeh (L3S Research Center, Leibniz University)
- Alireza Javadian Sabet (University of Pittsburgh)
- Mostafa Karami (University of Connecticut)
- Paris Koloveas (Athena Research Center)
- Anastasia Krithara (NCSR “Demokritos”)
- Shubhanshu Mishra (University of Illinois at Urbana-Champaign)
- Allard Oelen (L3S Research Center, Leibniz University)
- David Pride (KMi, The Open University)
- Jodi Schneider (University of Illinois at Urbana-Champaign)
- Christos Tryfonopoulos (University of Peloponnese)

## References

- [1] P. Manghi, A. Mannocci, F. Osborne, D. Sacharidis, A. Salatino, T. Vergoulis, Sci-k 2022-international workshop on scientific knowledge: Representation, discovery, and assessment, in: Companion Proceedings of the Web Conference 2022, 2022, pp. 735–738.
- [2] A. A. Salatino, Y. Bu, Y. Ding, Á. Horvát, Y. Huang, M. Liu, P. Manghi, A. Mannocci, F. Osborne, D. Romero, et al., 3rd international workshop on scientific knowledge representation, discovery, and assessment (sci-k 2023), in: Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 783–783.
- [3] L. Bellatreche, F. Bentayeb, M. Bieliková, O. Boussaid, B. Catania, P. Ceravolo, E. Demidova, M. Halfeld Ferrari, M. T. G. Lopez, C. S. Hara, et al., Databases and information systems in the ai era: Contributions from adbis, tpd and eda 2020 workshops and doctoral consortium, in: ADBIS, TPD and EDA 2020 Common Workshops and Doctoral Consortium: International Workshops: DOING, MADEISD, SKG, BBIGAP, SIMPDA, AIMinScience 2020 and Doctoral Consortium, Lyon, France, August 25–27, 2020, Proceedings 24, Springer, 2020, pp. 3–20.
- [4] P. Manghi, A. Mannocci, F. Osborne, D. Sacharidis, A. Salatino, T. Vergoulis, New trends in scientific knowledge graphs and research impact assessment, Quantitative Science Studies 2 (2021) 1296–1300.
- [5] S. Tsaneva, D. Dessì, F. Osborne, M. Sabou, Enhancing scientific knowledge graph generation pipelines with llms and human-in-the-loop, CEUR Workshop Proceedings Vol-3780 (2024). URL: <https://ceur-ws.org/Vol-3780/paper1.pdf>.
- [6] P. Pinheiro, H. Santos, J. Masters, M. Johnson, J. A. Stingone, S. Bengoa, M. Bax, D. L. McGuinness, Knowledge graph enabled scientific data repositories, CEUR Workshop Proceedings Vol-3780 (2024). URL: <https://ceur-ws.org/Vol-3780/paper2.pdf>.
- [7] T. Aggarwal, A. Salatino, F. Osborne, E. Motta, Identifying semantic relationships between research topics using large language models in a zero-shot learning setting, CEUR Workshop Proceedings Vol-3780 (2024). URL: <https://ceur-ws.org/Vol-3780/paper3.pdf>.
- [8] L. Kovriguina, L. Aung, P. Haase, N. Heist, D. Lamprecht, S. Heiß, Enhancing scientific discovery and decision-making: A knowledge graph-based research support system, CEUR Workshop Proceedings Vol-3780 (2024). URL: <https://ceur-ws.org/Vol-3780/paper4.pdf>.
- [9] A. Jacyszyn, H. Sack, D.-S. Group, M. Razum, F. Bach, Ditrare: Ai on a spider’s web. interweaving disciplines for digitalisation, CEUR Workshop Proceedings Vol-3780 (2024). URL: <https://ceur-ws.org/Vol-3780/paper5.pdf>.
- [10] M. Haris, S. Auer, M. F. Stocker, Querying of scholarly communication infrastructures, CEUR Workshop Proceedings Vol-3780 (2024). URL: <https://ceur-ws.org/Vol-3780/paper6.pdf>.

- [11] H. Schuster, A. Anjomshoaa, A. Polleres, Assessing the reliability and scientific rigor of references in wikidata, CEUR Workshop Proceedings Vol-3780 (2024). URL: <https://ceur-ws.org/Vol-3780/paper7.pdf>.
- [12] S. Efeoglu, Z. Chen, S. Schimmler, Ensuring fairness in machine learning projects, CEUR Workshop Proceedings Vol-3780 (2024). URL: <https://ceur-ws.org/Vol-3780/paper8.pdf>.
- [13] B. Spahiu, A. Lisa Gentile, C. De Luca, A. Maurino, Skills and expertise in large organizations: An enterprise knowledge graph approach, CEUR Workshop Proceedings Vol-3780 (2024). URL: <https://ceur-ws.org/Vol-3780/paper9.pdf>.