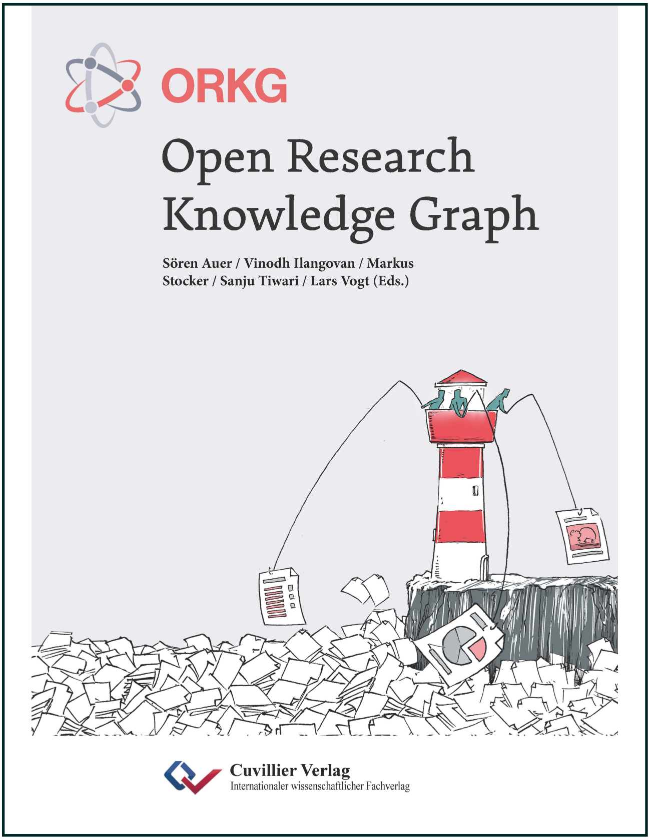
Vinodh Ilangovan (Herausgeber)
Sören Auer (Herausgeber)
Markus Stocker (Herausgeber)
Lars Vogt (Herausgeber)
Sanju Tiwari (Herausgeber)

Open Research Knowledge Graph



https://cuvillier.de/de/shop/publications/9037

1. Introduction

Sören Auer[1,2], Vinodh Ilangovan[1], Markus Stocker[1], Sanju Tiwari[3], and Lars Vogt[1]

[1]TIB - Leibniz Information Centre for Science and Technology, 30167 Hanover, Germany

[2] L3S Research Center, University of Hannover, 30167 Hannover, Germany

[3] BVICAM, New Delhi, India & UAT Mexico

In the rapidly evolving landscape of scientific research and scholarship, the dissemination and utilization of knowledge are paramount. Traditional methods of publishing and sharing scientific knowledge, while valuable, silo knowledge within dense, static documents that challenge integration, comparison, and reuse across disciplines. The Open Research Knowledge Graph (ORKG) presented in this book is a pioneering initiative that reimagines the future of scholarly communication. By leveraging the power of knowledge graph technologies, the ORKG transforms scholarly articles into a structured, interconnected web of research findings, making scientific knowledge more accessible, discoverable, and actionable. As such, the ORKG is an infrastructure that aims to support the production, curation, publication, and use of FAIR (*Wilkinson et al., 2016*) scientific knowledge with a mission to shape future scholarly publishing and communication where the contents of scholarly articles are FAIR research data (*Stocker et al., 2023*).

The inception of the ORKG is rooted in the recognition of the vast, untapped potential of digital scholarship. As researchers around the globe generate vast quantities of data and insights, the imperative to harness this wealth of knowledge becomes increasingly critical. The ORKG represents a paradigm shift, moving beyond the limitations of traditional research artifacts to a dynamic, open knowledge network. This network not only facilitates the seamless integration, comparison, reproducibility, and machine-based reuse of research findings, but also fosters new collaborations, innovations, and a deeper understanding of complex scientific questions.

This book aims to provide a comprehensive overview of the Open Research Knowledge Graph, from its conceptual foundation to its practical applications and beyond. Through a series of meticulously curated chapters, readers will embark on a journey through the architecture of the ORKG, its implementation challenges, successes, and the visionary roadmap for its future. The discussions will span the technical underpinnings of the ORKG service, including semantic web technologies and knowledge representation, as well as user-centric perspectives on how the ORKG can revolutionize research discovery, analysis, and dissemination.

Moreover, the book will explore the ORKG's impact on various stakeholders in the research ecosystem, including researchers, librarians, publishers, and policymakers. It will highlight case studies that illustrate the ORKG's transformative potential in enhancing research visibility, interoperability, and impact across diverse scientific domains.

Organizing scientific knowledge (only) as a collection of articles has been challenged for some time and the development of systems for more advanced scientific knowledge organization has received considerable attention in the literature (e.g., *Hars, 2001; Waard et al., 2009; Groth et al., 2010; Shotton et al., 2009; Iorio et al., 2015*). Research communities also routinely identify the problem when conducting systematic reviews and creating tailored databases that manage knowledge extracted from the literature. Yet, scaling and sustaining implementation remains a challenge as the systematic production of structured scientific knowledge and, thus, digitalization in scholarly communication remains elusive.

The sluggish progress in scholarly communication stands in stark contrast with the much faster digitalization we have witnessed in the past two decades in other areas, including e-commerce and web mapping platforms. Advanced knowledge organization would benefit research similarly to the benefits of modern web mapping platforms over traditional printed maps. Which technologies can support such advanced knowledge organization also in research is clear, too. How the research community and the scholarly infrastructure can ensure the systematic production of structured scientific knowledge, accurately, comprehensively, and efficiently remains unclear though.

ORKG addresses the challenge as-a-Service by providing research communities with a readily usable and sustainably governed Open infrastructure. Figure 1.1 provides a high-level illustration of the key ORKG services, namely comparisons and related visualizations, thematic reviews that leverage such knowledge products, and observatories as expert-curated virtual spaces for knowledge organization.

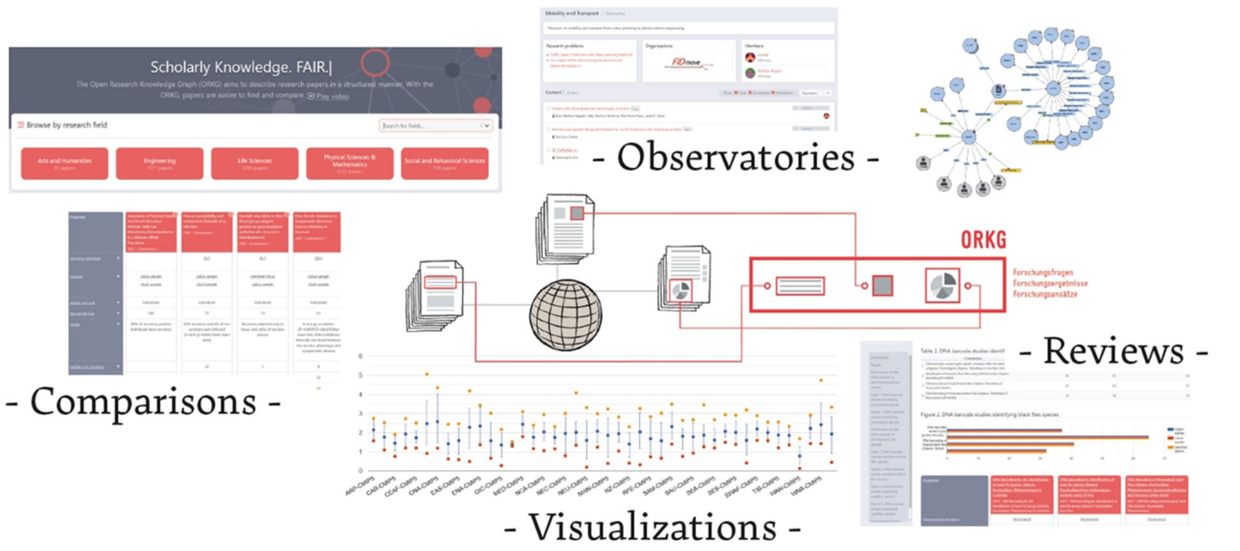


Figure 1.1 ORKG and its primary services: Tabular comparisons of scientific knowledge, visualizations of comparison data, thematic reviews, and expert-curated observatories. (Source: https://doi.org/10.3233/fc-221513)

At the core of the ORKG is a Knowledge Graph. Knowledge Graphs are not new in Artificial Intelligence, as the concept has meanwhile been used and discussed for more than a decade (*Popping, 2003*) and is grounded in the semantic web, which has a history and development spanning over a quarter of a century. Knowledge Graphs are presented as an extended form of ontology to provide richer entity descriptions at the instance level (*Schrader, 2020*). They play a significant role in data integration and semantic web technologies by providing a structured framework for organizing and connecting heterogeneous information sources. By leveraging semantic relationships and ontologies, Knowledge Graphs facilitate the discovery of meaningful relations between different data types, thereby enhancing data interoperability and enabling more effective data analysis and retrieval. Some well-known Knowledge Graphs are Google Knowledge Graphs (*Singhal, 2012*), DBpedia (*Lehmann et al., 2007*), and Bing (*Noy et al., 2019*), etc.

The ORKG initiative engages stakeholders in numerous ways. As expert-curated virtual communities and collaborative virtual spaces, ORKG observatories are community-specific entry points to the ORKG. As members of observatories, experts may support identifying and specifying ORKG templates that are relevant to the community, organize research problems in their field, and monitor the quality of observatory content. Beyond research communities, ORKG engages with publishers and conferences with the aim of integrating the ORKG into manuscript production, submission, review, and publishing processes. To develop applications beyond research, ORKG also engages with industry stakeholders, intergovernmental organizations, and the general public, e.g., to explore the role of the ORKG in evidence-based news reporting.

The journey that aims at frictionless scientific knowledge use with advanced machine processing has begun, yet considerable mileage remains to be travelled. Various initiatives in information technology have prototyped systems and in the context of (living) systematic reviews numerous disciplines have shown what conducting science with machine-reusable scientific knowledge can look like in their respective domains. ORKG contributes to further driving the required fundamental transformations by increasing productivity through generic infrastructure and services, delivering training and support, and building capacity towards a future in which scientific knowledge is FAIR research data.

References

Groth, P., Gibson, A., & Velterop, J. (2010). The anatomy of a nanopublication. In Information Services & Use (Vol. 30, Issues 1–2, pp. 51–56). IOS Press. https://doi.org/10.3233/isu-2010-0613

Hars, A. (2001). Designing Scientific Knowledge Infrastructures: The Contribution of Epistemology. Information Systems Frontiers 3, 63–73. https://doi.org/10.1023/A:1011401704862

Iorio, A. D., Lange, C., Dimou, A., & Vahdati, S. (2015). Semantic Publishing Challenge – Assessing the Quality of Scientific Output by Information Extraction and Interlinking. In Semantic Web Evaluation Challenges (pp. 65–80). Springer International Publishing. https://doi.org/10.1007/978-3-319-25518-7_6

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., & Bizer, C. (2015). DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. In Semantic Web (Vol. 6, Issue 2, pp. 167–195). IOS Press. https://doi.org/10.3233/sw-140134

Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale Knowledge Graphs: Lessons and Challenges. In Queue (Vol. 17, Issue 2, pp. 48–75). Association for Computing Machinery (ACM). https://doi.org/10.1145/3329781.3332266

Popping, R. (2003). Knowledge Graphs and Network Text Analysis. In Social Science Information (Vol. 42, Issue 1, pp. 91–106). SAGE Publications. https://doi.org/10.1177/0539018403042001798

Schrader, B. (2020). What's the difference between an ontology and a knowledge graph? https://enterprise-knowledge.com/whats-the-difference-between-an-ontology-and-a-knowledge-graph/ (Accessed: April 2024)

Shotton, D., Portwin, K., Klyne, G., & Miles, A. (2009). Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. In P. E. Bourne (Ed.), PLoS Computational Biology (Vol. 5, Issue 4, p. e1000361). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pcbi.1000361

Singhal, A. (2012). Introducing the Knowledge Graph: things, not strings. https://blog.google/products/search/introducing-knowledge-graph-things-not/ (Accessed: April 2024)

Stocker, M., Oelen, A., Jaradeh, M. Y., Haris, M., Oghli, O. A., Heidari, G., Hussein, H., Lorenz, A.-L., Kabenamualu, S., Farfar, K. E., Prinz, M., Karras, O., D'Souza, J., Vogt, L., & Auer, S. (2023). FAIR scientific information with the Open Research Knowledge Graph. In B. Magagna (Ed.), FAIR Connect (Vol. 1, Issue 1, pp. 19–21). IOS Press. https://doi.org/10.3233/fc-221513

Waard, A.D., Shum, S.B., Carusi, A., Park, J., Samwald, M., & Sándor, Á. (2009). Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims. Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009), collocated with the 8th International Semantic Web Conference (ISWC-2009), Washington DC, USA, October 26, 2009. https://ceur-ws.org/Vol-523/ (Accessed: June 2023)

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. In Scientific Data (Vol. 3, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/sdata.2016.18

2. ORKG Concepts

Allard Oelen and Vinodh Ilangovan

TIB - Leibniz Information Centre for Science and Technology, 30167 Hanover, Germany

In this chapter, we will discuss the key ORKG concepts in more detail. In order to better understand the underlying data model of the ORKG, we will start with a brief introduction of terminology from the Semantic Web. Afterwards, we continue with an in-depth explanation of ORKG specific terminology, the so-called Content Types. Finally, we present several miscellaneous tools that are implemented in the ORKG.

2.1 Graph Concepts Background

The ORKG data model is structured as a knowledge graph. The term knowledge graph comes from the Semantic Web domain. The Semantic Web is related to the World Wide Web, but instead of linking documents together, data is linked. On top of the web of linked data, semantics are added to capture the meaning of data, hence the Semantic Web. The ORKG follows the Semantic Web approach to describe data, however, regular users of the system do not have to be familiar with these concepts. The ORKG User Interface (UI) is designed in such a way that it can be operated without any Semantic Web domain knowledge. However, in order to understand some of the underlying concepts of the ORKG, a brief introduction is helpful. Therefore, we will now briefly describe some of the main Semantic Web terms.

The ORKG closely follows the specification of RDF (Resource Description Framework). In this framework, knowledge is described as triples, consisting of a subject, a predicate, and an object. A triple is also called a statement. Some of the terms of RDF are coming from the linguistics domain. The subject and object position can contain resources, properties and classes. The predicate position contains properties. In addition, the object position can also contain literals. Literals are atomic pieces of knowledge that cannot be linked to, for example, natural text, numbers, etc. The ORKG automatically assigns IDs to all the previously mentioned concepts, making it easier to refer to specific pieces of data. By assigning a class to a resource, a resource becomes an instance of that class. Although assigning classes in the ORKG is not enforced, it helps to better organize knowledge, which is one of the main goals of the ORKG.

2.2 Content Types

Frequently used concepts within the ORKG system are called Content Types. These Content Types generally have dedicated pages in the ORKG UI and adhere to a predefined data model. With this data model, it is possible for users to freely describe scholarly knowledge in structured form. The Content Types, however, ensure that data follows the same structure and is therefore more machine-actionable. In the remainder of this section, we discuss the most important ORKG-specific Content Types in more detail.

2.2.1 Papers and Contributions

ORKG Papers represent any published scholarly article. Each paper has a limited set of metadata assigned to it. Only metadata that is actually used within the ORKG is recorded. Any other metadata is ignored. Some of the metadata includes the paper title, DOI, authors, publication date, and publication venue. Furthermore, a Research Field is assigned to a paper. The Research Field is also an ORKG Content Type, which we will discuss in this chapter as well.

When a new paper is added to the ORKG, the metadata is fetched automatically via Crossref, if a DOI is provided. In case only the paper title is provided, the metadata is fetched using a lookup at Semantic Scholar by trying to find a matching paper title. A screenshot of the page to add a paper to the ORKG is displayed in Figure 2.1 below. As can be seen on the screenshot, it is also possible to upload a PDF file or to import a paper using a BibTeX entry. In case of the PDF upload, the metadata of the paper is automatically extracted from the PDF.

After a paper is added, the graph only contains the metadata of the paper. The structured *contribution data* can be entered on the View Paper page. Since the ORKG focuses on the knowledge presented within research articles, adding the contribution data is the most important step when adding papers. Structured paper data is organized in *Contributions*, which is another ORKG Content Type. Since Contributions are closely related to Papers, we will discuss them in this section.

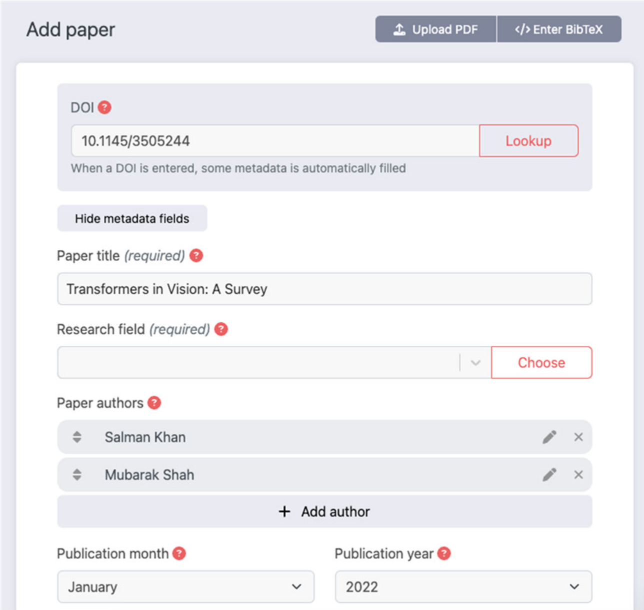


Figure 2.1 Add Paper form

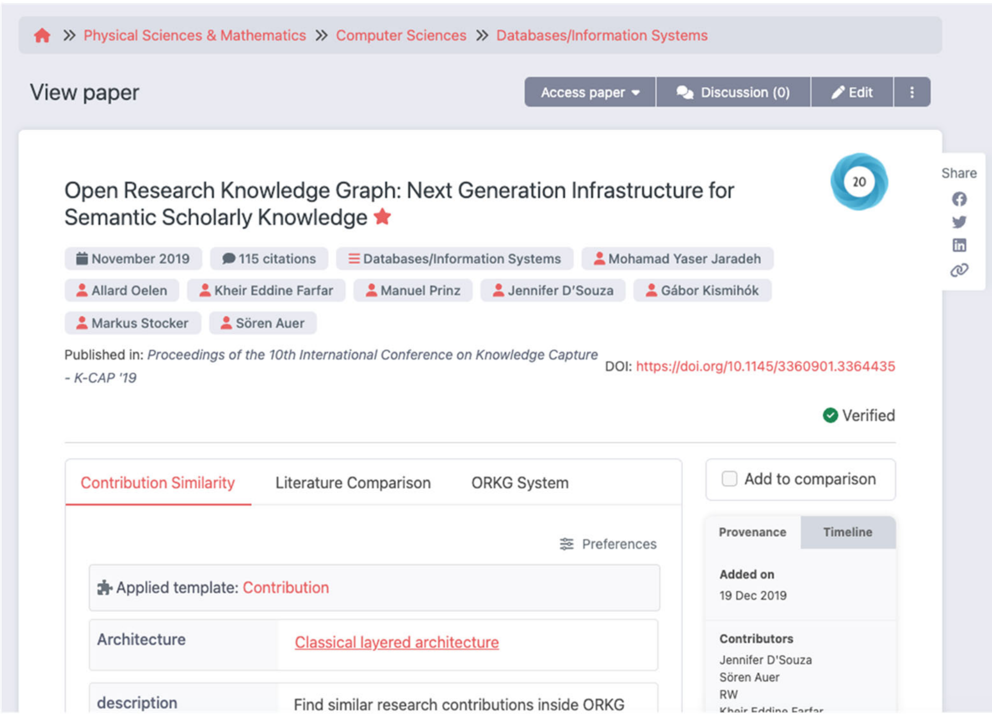


Figure 2.2 Paper page-showing contributions from a single paper

Contributions capture what a paper contributes to science, and essentially why the paper was published in the first place. All knowledge within a paper must be organized in one - or multiple - Contributions. Contributions can be considered a means to organize paper knowledge in separate, self-contained, collections. Each contribution can be described freely, but the ORKG recommends users to at least use the following properties for contributions: research problem, materials, methods, and results. The research problem describes what topic the specific paper is addressing. Figure 2.2 depicts a paper with three contributions, displayed using tabs. Each contribution contains structured data related to that contribution. Furthermore, the metadata of the paper is visible on this page, as well as the research field.

From the Paper page, users can view all the structured knowledge related to a specific paper. Furthermore, it is possible to directly access openly accessible version or preprints of a paper (if available). Users may also start a discussion about the paper.

2.2.2 Comparisons

When a set of papers is addressing the same research problem, for many cases it is interesting to see how those papers compare. For example, in case a set of Computer Science papers addresses the research problem Author Name Disambiguation (i.e. distinguishing between authors with similar or identical names), it makes sense to compare those papers to see which model performs best. Apart from ranking papers, there are many other cases in which tabular overviews of literature are useful: compiling state-of-the-art literature overviews, showing trend analysis, comparing research on geographical differences, etc. Because papers in the ORKG are described in a structured form, compiling those overviews can be done semi-automatically, using the structured paper data that is already present. Such literature overviews are called ORKG Comparisons (*Oelen et al., 2020*). In Figure 2.3 below, a Comparison is depicted.

Properties	The early phase of the COVID-19 outbreak in Lombardy, Italy *Contribution 1 - 2020*	Transmission potential of COVID-19 in Iran *Contribution 1 - 2020*	Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study *Contribution 1 - 2020*
research problem	Determination of the COVID-19 basic reproduction number	Determination of the COVID-19 basic reproduction number	Determination of the COVID-19 basic reproduction number
basic reproduction number*	3.1	3.6	2.68
location	Lombardy, Italy	Iran	Wuhan
time period/time interval			
↪ has beginning*	2020-01-14	2020-02-19	2019-12-31
↪ has end*	2020-03-08	2020-02-29	2020-01-28

Figure 2.3 Comparison visualizing three papers in tabular form