ORKG

Open Research Knowledge Graph

Sören Auer / Vinodh Ilangovan / Markus Stocker / Sanju Tiwari / Lars Vogt (Eds.)

Cuvillier Verlag
Internationaler wissenschaftlicher Fachverlag

[May 2024, First Edition]

Open Research Knowledge Graph

Open Research Knowledge Graph

Editors

Sören Auer

Vinodh Ilangovan

Markus Stocker

Sanju Tiwari

Lars Vogt

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.
1. Aufl. - Göttingen : Cuvillier, 2024

Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

ISBN eBook OA 978-3-68952-038-0

Acknowledgements

We are immensely grateful to a remarkable group of individuals whose expertise and attention to detail greatly enriched this publication. Firstly, we thank all the authors of this book for preparing their chapters to a wide readership. We also extend our sincere thanks to Marina Wurzbacher, Anna-Lena Lorenz, Ildar Baimuratov, Julia Evans, Olga Lezhnina, Qurat-ul Ain Aftab, Pallavi Karanth, and Akhilesh Vyas for their invaluable help with reviewing and proofreading this manuscript. Their keen insights and suggestions were essential in shaping the final product.

We express our gratitude to Nadine Klöver for her exceptional cover design, which beautifully encapsulates the essence of our work.

Our publication journey was made smoother thanks to the professionalism and support of Cuvillier Verlag. We are particularly thankful to Ms. Annette Jentzsch-Cuvillier for her patience and understanding throughout the publishing process. Her guidance was instrumental in bringing this project to fruition.

Thank you all for your dedication and hard work. This book would not have been possible without your collective efforts.

Sincerely,

Sören Auer, Vinodh Ilangovan, Markus Stocker, Sanju Tiwari, and Lars Vogt

Prologue

As we mark the fifth anniversary of the alpha release of the Open Research Knowledge Graph (ORKG), it is both timely and exhilarating to celebrate the significant strides made in this pioneering project. We designed this book as a tribute to the evolution and achievements of the ORKG and as a practical guide encapsulating its essence in a form that resonates with both the general reader and the specialist.

The ORKG has opened a new era in the way scholarly knowledge is curated, managed, and disseminated. By transforming vast arrays of unstructured narrative text into structured, machine-processable knowledge, the ORKG has emerged as an essential service with sophisticated functionalities. Over the past five years, our team has developed the ORKG into a vibrant platform that enhances the accessibility and visibility of scientific research. This book serves as a non-technical guide and a comprehensive reference for new and existing users that outlines the ORKG's approach, technologies, and its role in revolutionizing scholarly communication. By elucidating how the ORKG facilitates the collection, enhancement, and sharing of knowledge, we invite readers to appreciate the value and potential of this groundbreaking digital tool presented in a tangible form.

Looking ahead, we are thrilled to announce the upcoming unveiling of promising new features and tools at the fifth-year celebration of the ORKG's alpha release. These innovations are set to redefine the boundaries of machine assistance enabled by research knowledge graphs. Among these enhancements, you can expect more intuitive interfaces that simplify the user experience, and enhanced machine-learning models that improve the automation and accuracy of data curation.

We also included a glossary tailored to clarifying key terms and concepts associated with the ORKG to ensure that all readers, regardless of their technical background, can fully engage with and understand the content presented. This book transcends the boundaries of a typical technical report. We crafted this as an inspiration for future applications, a testament to the ongoing evolution in scholarly communication that invites further collaboration and innovation. Let this book serve as both your guide and invitation to explore the ORKG as it continues to grow and shape the landscape of scientific inquiry and communication.

Contents

Acknowledgements 5
Prologue 7
1. Introduction 15
2. ORKG Concepts 21
2.1 Graph Concepts Background 21
2.2 Content Types 22
2.2.1 Papers and Contributions 22
2.2.2 Comparisons 24
2.2.3 Visualizations 25
2.2.4 Reviews 26
2.2.5 Lists 28
2.2.6 Research Fields 29
2.2.7 Other Content Types 30
2.3 Miscellaneous Tools 31
2.3.1 Observatories and Organizations 31
2.3.2 Statement Browser 31
2.3.3 Templates 32
2.3.4 Contribution Editor 33
2.3.5 CSV Importer 33
2.3.6 Survey Importer 34
2.3.7 Smart Suggestions 35
3. Guidelines for creating Comparisons 37
3.1 Understanding the value of Comparisons in the ORKG 37
3.2 Important characteristics of a Comparison 39
3.3 Creating high quality Comparisons 40
3.3.1 Human- and machine-actionable elements 40
3.3.2 Knowledge Graph Structure 41
3.4 Ensuring data quality of Comparisons 42
3.5 Discoverability of ORKG Comparisons 45
3.6 Conclusion 46
4. ORKG Benchmarks 49

4.1 Definitions 51
4.2 Guide to Creating a Benchmark in the ORKG 52
4.3 The Workflow Dynamics of ORKG Benchmarks 53
4.4 Conclusion 55
5. Modeling and Quality Assurance through Templates 57
5.1 Need for a template system 58
5.2 Overview 59
5.2.1 The Role of Domain Experts in Template Creation 59
5.2.2 Integration with an Ontology Lookup Service 60
5.2.3 Template System's Role in Creating Input Forms 60
5.2.4 Data Validation Process 61
5.2.5 Community Contribution and Data Addition 61
5.3 SHACL Shapes 61
5.3.1 Template editor 62
5.3.2 Formatted Labels 63
5.3.3 Template Visualization Diagram 63
5.4 Import/Export Functionality 65
5.4.1 Managing Existing Templates 65
5.4.2 The Import Tool Workflow and Process 65
5.4.3 Exporting Templates to SHACL Files 66
5.5 Future Perspectives 66
5.5.1 Advanced SHACL Constraints Implementation 66
5.5.2 Improved SHACL Shapes Support 66
5.5.3 Interactive Template Visualization Diagram Editing 66
5.5.4 Evolution of Formatted Labels 67
5.6 Conclusion 67
6. Natural Language Processing for the ORKG 69
6.1 ORKG Natural Language Processing Facets 70
6.2 Evolution of NLP Services with Large Language Models 72
Traditional Machine Learning Objectives 72
6.3 LLMs' Comprehensive Capabilities 75

6.4 LLM-based ORKG Smart Suggestions 75
6.5 Scholarly Question Answering with the ORKG 77
6.6 JarvisQA and Beyond 77
6.7 LLMs in Scholarly QA 78
6.8 Conclusion and Outlook 78
7. Energy Systems Analysis as an ORKG Use Case 83
7.1 Motivation 83
7.2 Research Question 84
7.2.1 Comparison 85
7.2.2 Visualizations 86
7.3 Conclusion and Outlook 87
7.3.1 Conclusion 87
7.3.2 Outlook 89
8. Harnessing the potential of the ORKG for synthesis research in agroecology 93
8.1 Motivation 93
8.2 Research Question 95
8.3 ORKG Comparison 95
8.4 Visualizations 98
8.5 Conclusions 100
9. Knowledge synthesis in Invasion Biology: from a prototype to community-designed templates 105
9.1 The prototype with Hi Knowledge data 105
Motivation 105
Approach and results 106
9.2 The ecologist community gets more involved 108
Motivation 108
Method 109
What we learned 110
9.3 Engaging with the broader community of invasion biologists 110
Motivation 110
Method 111
9.4 Further use of ORKG in the context of invasion biology 113

ORKG for teaching in ecology 113
A tool for publishers to collect structured information about submissions .. 114
Smart searches 114
9.5 Conclusion 115
10. Data to Knowledge: Exploring the Semantic IoT with ORKG 117
10.1 Motivation 117
10.1.1. Research highlights and contribution 117
10.2 Background 119
10.2.1. Web and Semantic Web of Things (WoT/SWoT) 119
10.2.2. IoT Ontologies 119
10.2.3. IoT Knowledge Graphs 119
10.2.4. IoT in Digital Twins 120
10.3 Semantic IoT in Specific Domains 120
10.3.1. Semantic IoT in Water 120
10.3.2. Semantic IoT in Healthcare 121
10.3.3. Semantic IoT in Industry 4.0 and Manufacturing 121
10.3.4. Semantic IoT in Energy Efficient Building 122
10.3.5. Semantic IoT in Agriculture 122
10.4 Major Sources of IoT Ontologies 122
10.4.1 Semantic IoT Frameworks 124
10.5 Conclusion 124
11. Food Information Engineering for a Sustainable Future 129
11.1 Motivation 129
11.2 Food Information Engineering 130
11.2.1. Collecting food information 130
11.2.2 Organizing Food Information 131
11.2.3 Food information processing 132
11.2.4 Using of food information 132
11.3 Food Information Engineering Observatory 133
11.4 Summary and conclusion 137
Afterword 141

Glossary.....143

List of Figures

Figure 1.1 ORKG and its primary services:.....17
Figure 2.1 Add Paper form.....23
Figure 2.2 Paper page-showing contributions from a single paper.....23
Figure 2.3 Comparison visualizing three papers in tabular form.....24
Figure 2.4 Visualization of R0 estimates for COVID-19 from a Comparison.....26
Figure 2.5 ORKG Review.....27
Figure 2.6 ORKG List showing three related papers.....29
Figure 2.7 Workflow for structured literature reviews using the ORKG.....29
Figure 2.8 Author page.....30
Figure 2.9 ORKG Statement Browser.....32
Figure 2.10 Simultaneous editing of papers with Contribution Editor.....33
Figure 2.11 CSV Importer.....34
Figure 2.12 ORKG Survey Importer.....35
Figure 2.13 Smart Suggestions for possibly relevant properties.....35
Figure 3.1 KGGM Level 2 suggestions to improve the properties description.....43
Figure 3.2 KGGM Level 3 to address conciseness of resource labels.....44
Figure 3.3 KGGM Level 5 linking external resources.....45
Figure 4.1 A contrastive view of Task-Dataset-Metric information.....50
Figure 4.2 Leaderboard template.....53
Figure 4.3 The dynamic frontend for the ORKG Benchmarks feature.....54
Figure 5.1 ORKG template system.....59
Figure 5.2 Template-based input form.....60
Figure 5.3 Template diagram with a zoom in on one of the entities.....64
Figure 6.1 A pie chart of the ORKG NLP facets.....74
Figure 6.2 Smart Suggestions (AI) guide users (Humans).....76
Figure 6.3 Depiction of JarvisQA.....77
Figure 7.1 ORKG comparison of 25 scenarios from GHG reduction studies.....86
Figure 7.2 Reported installed capacity aggregated in the 25 studies.....86
Figure 7.3 Reported energy supplyin the 25 studies.....87
Figure 7.4 Visualized results from the SPARQL query.....90
Figure 8.1 ORKG Add paper function.....96
Figure 8.2 The ORKG contribution editor.....96
Figure 8.3 Example of an ORKG template.....97
Figure 8.4 Partial view of our ORKG comparison on cereal-legume intercrops.....98
Figure 8.5 Visualisation created using agroecology comparison.....99
Figure 9.1 Comparison for Hi Knowledge data.....107

Figure 9.2 Share of contributions that support, question, or are undecided about the propagule pressure hypothesis. 108
Figure 9.3 Visualization created with Hi Knowledge data 108
Figure 9.4 Screenshot of an R Shiny app. 109
Figure 10.1 Semantic IoT workflow with the ORKG 118
Figure 11.1 An overview of food information engineering observatory 134
Figure 11.2 Food composition tables 134
Figure 11.3 Food ontologies 135
Figure 11.4 Food Knowledge Graph 135
Figure 11.5 Food Question Answering 136

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.

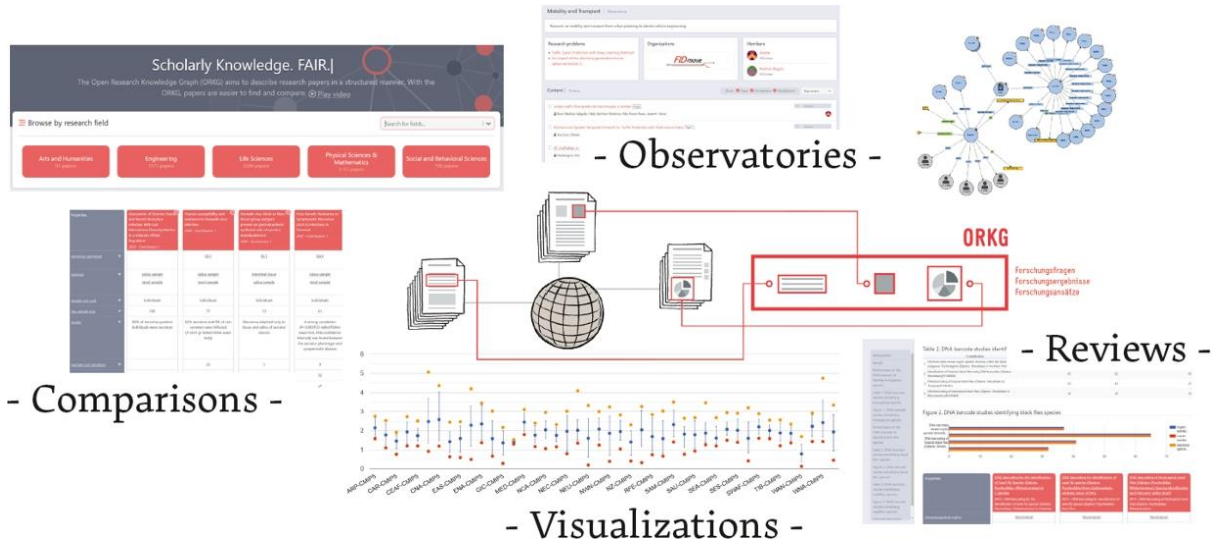

Figure 1.1 ORKG and its primary services: Tabular comparisons of scientific knowledge, visualizations of comparison data, thematic reviews, and expert-curated observatories. (Source: https://doi.org/10.3233/fc-221513)

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.

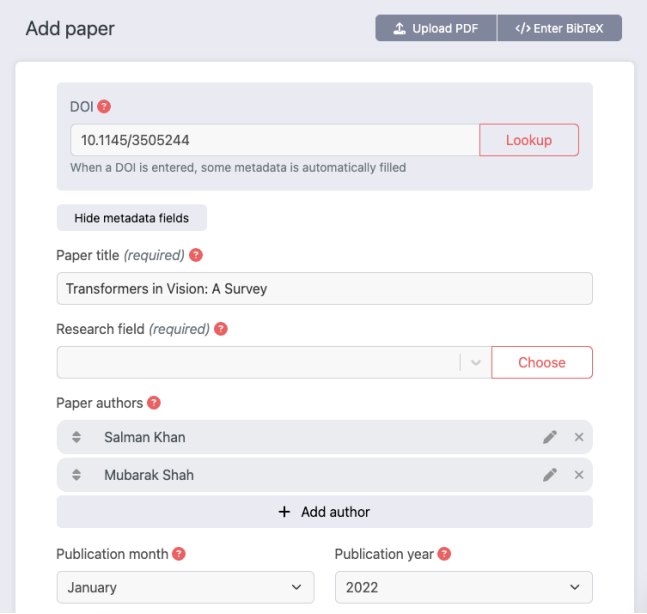

Figure 2.1 Add Paper form

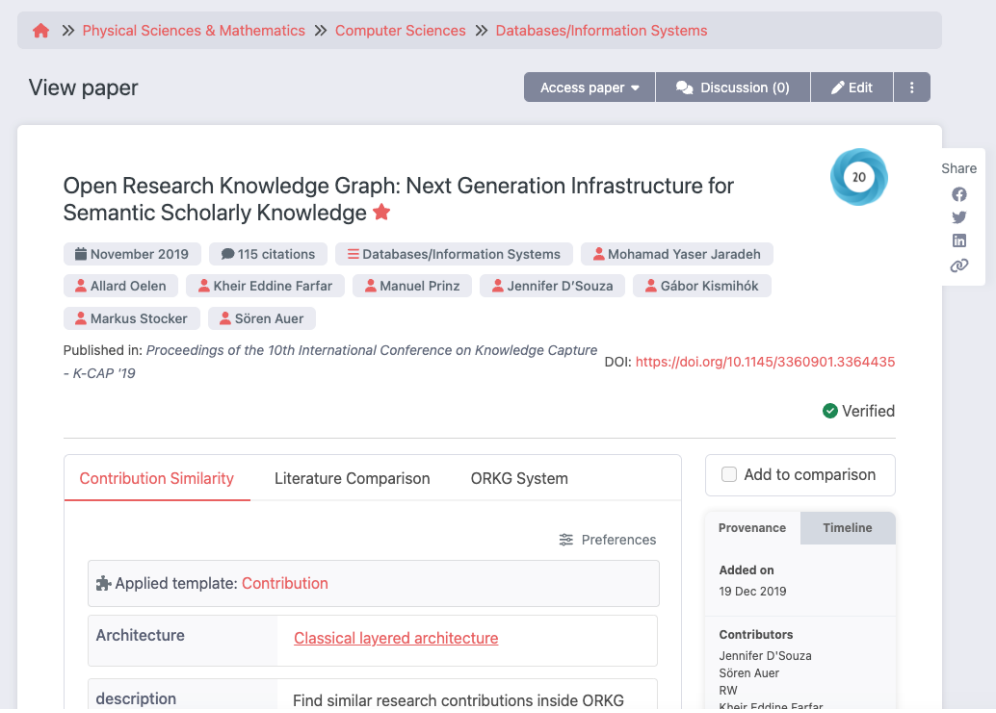

Figure 2.2 Paper page-showing contributions from a single paper

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

It is part of a larger comparison that comprises 31 papers in total[1]. In our specific example, three papers are displayed that are all addressing the same research problem. These papers all report basic reproduction numbers of COVID-19, measured at different locations and for different period.

Comparisons are one of the key features of the ORKG and a detailed discussion is in chapter 3. It is possible to publish Comparisons, which captures a snapshot of the comparison and stores this in a persistent manner. Additionally, a DOI can be assigned to the comparison, making it suitable to be used within the related work section of research articles. The generated comparison can be properly cited using the DOI. Furthermore, comparison can be created in a collaborative manner, after publishing a comparison, new versions can be created of the same comparison. This means comparison becomes dynamic, and can be updated as soon as new literature becomes available. Finally, comparisons can be exported into various formats to further enhance the machine-actionability of the data, e.g. SPARQL, RDF, CSV, LaTeX, and PDF.

Comparisons are generated semi-automatically by matching similar properties across different papers. As previously mentioned, Papers organize the structured data into Contributions. Therefore, Comparisons are in fact comparisons of Contributions, and not of the papers themselves. This makes it possible to compare different Contributions from the same paper.

2.2.3 Visualizations

Tabular visualizations such as Comparisons are particularly suitable for non-numeric data. For numeric data, generally other types of visualizations are more suitable, such as bar charts or scatter plots. Once a comparison is published, it becomes possible to add additional types of visualizations. ORKG Visualizations provide an alternative method of visualizing comparison data and are displayed on top of comparisons. In Figure 2.4 a visualization is displayed for the entire COVID-19 comparison that we discussed previously. As can be seen, this visualization provides a better summary of the data than the textual comparison table. Similar to comparisons, visualizations can be published, which ensures the data is persistent and cannot be changed. Since data is available as structured data, creating alternative visualizations is relatively simple as there is no need to do data cleaning. Generally, data can be used as-it-is, and can be directly used to create visualizations. Additionally, updating visualizations when additional data becomes available is therefore also straightforward.

[1] https://orkg.org/comparison/R44930/

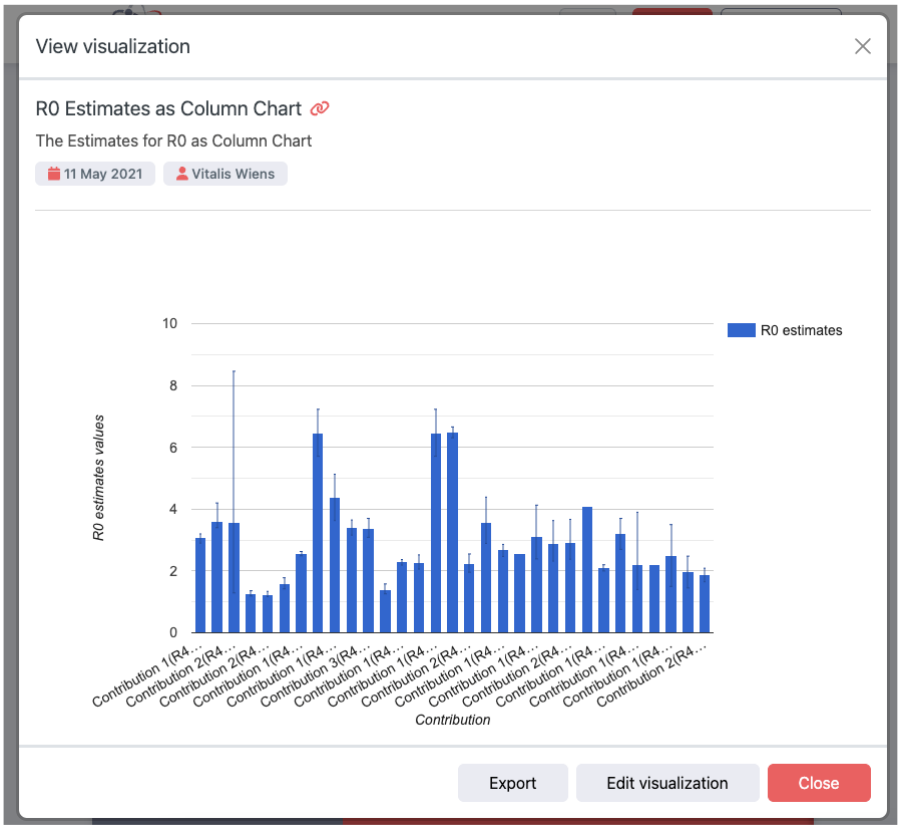

Figure 2.4 Visualization of R0 estimates for COVID-19 from a Comparison

2.2.4 Reviews

Review articles have a crucial role to organize scholarly knowledge for specific domains. However, the current practice of publishing review articles suffers from weaknesses. For example, when a review is published, it is generally not updated anymore, rendering them outdated soon after it is published. Furthermore, the underlying data used to author the review remains hidden, and can get lost over time. Also, reviews are created by a select set of authors, and therefore may not represent the opinions from communities as a whole. ORKG Reviews try to address these issues by providing a community-maintained collaborative review authoring platform. Existing ORKG Content Types form the foundation of ORKG Reviews, specifically ORKG Comparisons. Generally, an ORKG Review consists of a set of comparisons (between three and five comparisons). Furthermore, visualizations and other structured graph data can be added to the review. Finally, natural text is used as glue to ensure the Review is a human comprehensible document.

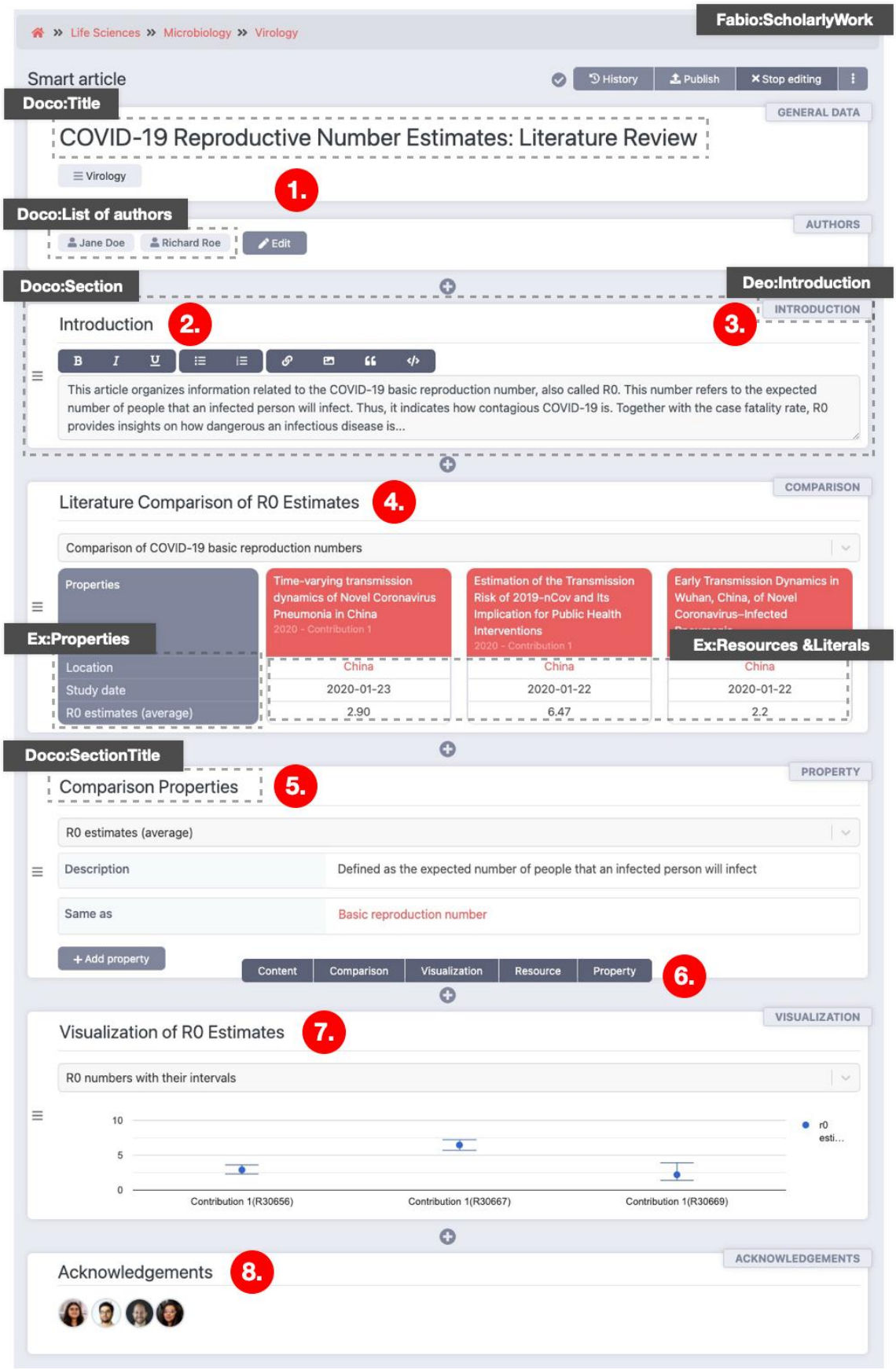

Figure 2.5 ORKG Review : Authoring interface showing a natural text section Comparison, Visualization, and additional structured data (Oelen et.al., 2021)

Similar to ORKG Comparisons, Reviews can be published to make them persistent over time. Furthermore, it is possible to assign a DOI to the review, facilitating citing the review in other research articles. After a Review is published, it is possible to modify the Review only by publishing a new version. All of the underlying data to generate the ORKG Review is machine-actionable, meaning that it is possible to create custom tools to further analyze the data. This addresses one of the weaknesses of the existing review authoring practices, where underlying data remains hidden.

The authoring interface of ORKG Reviews is displayed in Figure 2.5. Item 1 shows the title of the article. Item 2 shows the text authoring interface. The interface is supported by Markdown and allows for creating in-text references to other ORKG Content Types and to citations, which are managed using a built-in BibTeX manager. Item 3 shows the type selector for the text section. This provides some additional knowledge regarding the contents of the section. Item 4 shows the comparison sections, showing a similar comparison to the previously mentioned COVID-19 example. Item 5 shows a description of a single property from the ORKG graph. This section also supports displaying arbitrary ORKG resources. Item 6 shows the menu to add additional sections. Item 7 shows a visualization of the comparisons displayed above. Finally, item 8 is the acknowledgements. These acknowledgements are automatically generated based on provenance data stored in the graph.

2.2.5 Lists

ORKG Lists provide a means to organize scholarly articles without the need to provide any structured data for them. With a List, it is possible to group related articles together. The dynamic and collaborative nature of Lists makes sure that organized lists of literature can be published and updated when necessary. An example of a List is depicted in Figure 2.6. The displayed list contains three papers. By clicking on the paper title, the Paper page is opened, from where it becomes possible to add structured data to the paper. However, to use ORKG lists, structured data is not required.

Lists can serve as a starting point when using the ORKG for conducting structured literature reviews. If all related literature is organized in a List, structured data can be added for those papers. Once the structured data is present, it can be used to generate an ORKG Comparison. It is then possible to add Visualizations to the comparison. Finally, all the generated Content Types can be used to form an ORKG Review. All ORKG Content Types can also be used individually, without following the workflow. However, to provide guidance to users, the workflow helps

to understand how different ORKG Content Types are related to each other. This workflow is depicted in Figure 2.7.

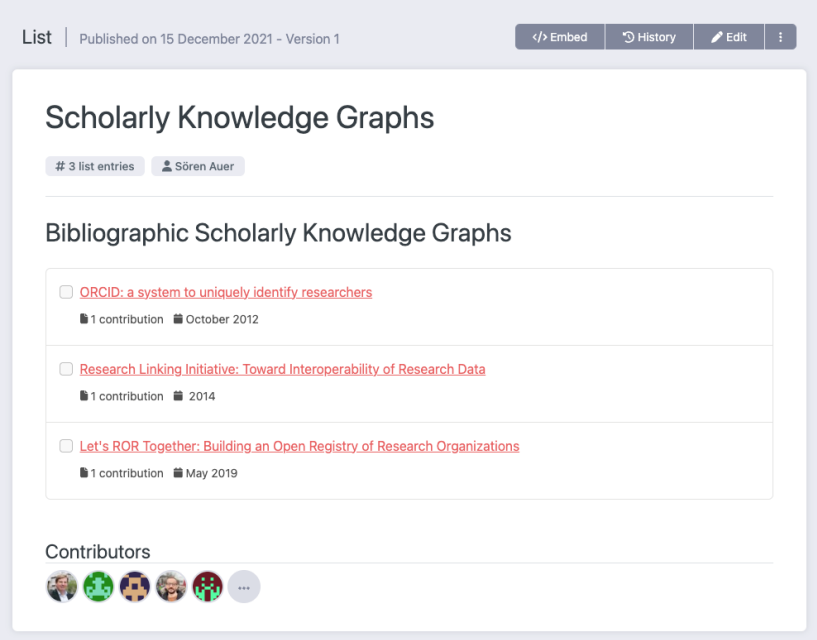

Figure 2.6 ORKG List showing three related papers

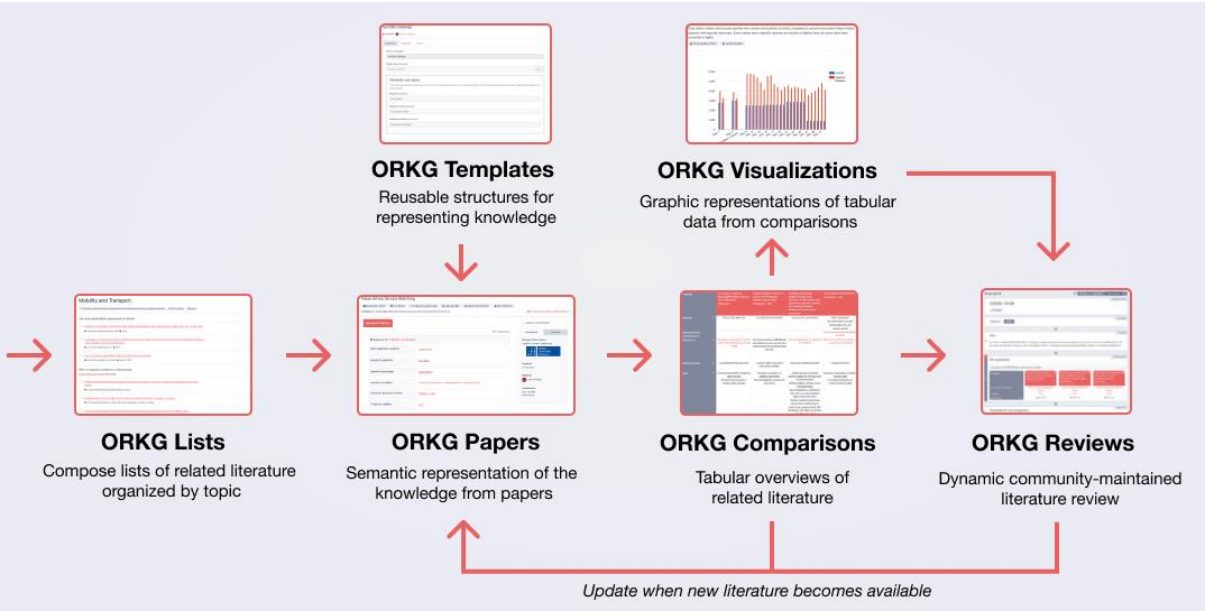

Figure 2.7 Workflow for structured literature reviews using the ORKG

2.2.6 Research Fields

As previously mentioned, Papers are assigned to Research Fields. ORKG Research Fields can be assigned to all Content Types. This helps to further organize

knowledge within the graph, and provides a means to view content for specific domains. The ORKG Research Field taxonomy is based on a taxonomy created by the National Academy of Sciences[2]. ORKG Research Fields are curated by the ORKG team. When users want to add a new field, they have to contact the ORKG team. Then it is decided whether indeed a new field is required, or whether it is possible to use one of the existing research fields.

2.2.7 Other Content Types

Finally, there are various other Content Types. This includes Author and Venue. Those Content Types have dedicated displayed pages in the UI as well. The author page shows the ORCID of an author (when available), and all the related content within the ORKG from a specific author. An example of an Author page is depicted below in Figure 2.8. The Venue page shows the Papers that are associated with a specific Venue.

In addition, we have several Content Types without dedicated pages. These types are considered relevant for specific use cases, and are listed on the ORKG page. This includes the Dataset and Software content types. They can be described using templates that provide a structure for describing the respective data. In the end, these Content Types can be used within papers, and form links between different literature using the same materials.

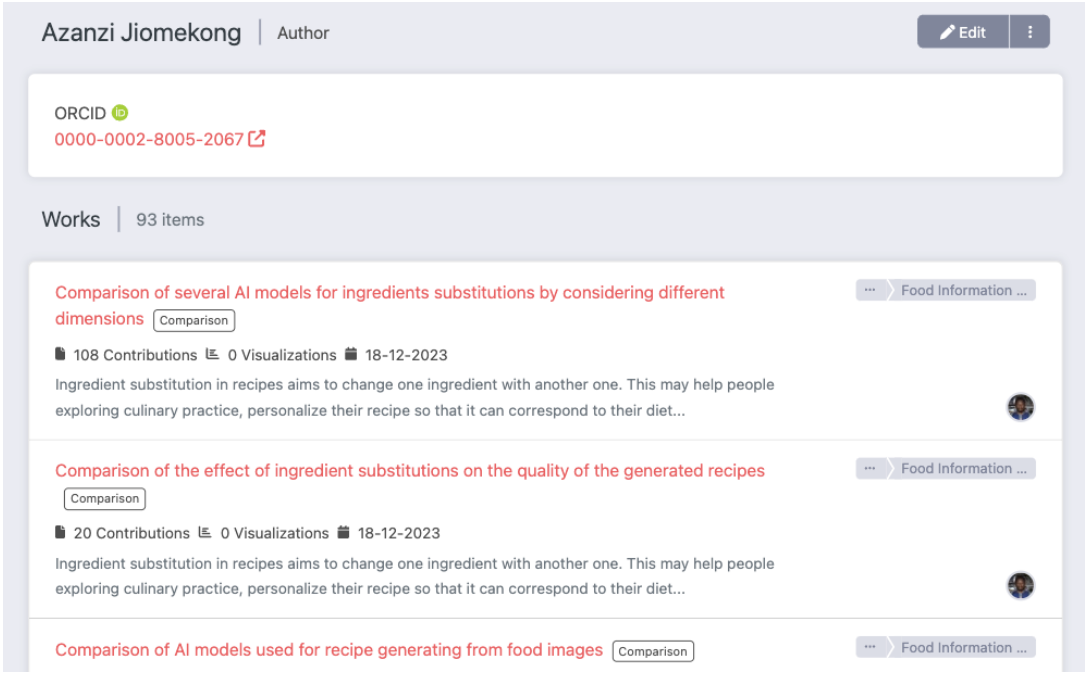

Figure 2.8 Author page showing all associated Content Types from a specific author

[2] https://www.nationalacademies.org/our-work/an-assessment-of-research-doctorate-programs#sl-three-columns-aa4e3585-5bac-4198-9e7b-eadc98de85cb

2.3 Miscellaneous Tools

2.3.1 Observatories and Organizations

Different communities and domains require different types of structured data to support their use cases. For example, for machine learning, structured data comparing the performance of different algorithms is of interest (i.e., benchmarks) while for virology, structured data regarding different measurements is of interest. To support such a variety of use cases and domains, ORKG users can join specific interest groups, called observatories. An observatory consists of a group of domain experts that collaboratively identify ORKG use cases for their field of expertise. Furthermore, they support these use cases by creating ORKG Templates and by curating knowledge from their field. Observatories are manifested by the ORKG observatory page, which shows a textual description of the objectives of the respective observatory. Additionally, research problems and observatory members are listed. A list of content shows various ORKG content types that are associated with the observatory, making it possible to explore content for a specific domain.

Organizations are responsible for managing observatories. Generally, organizations are real-world institutions (such as universities, research institutes, or libraries). Organization logos and member names are prominently listed on several ORKG pages, to appropriately acknowledge the creators and curators of ORKG content. Observatories and Organizations play a crucial role in the crowdsourcing strategy of the ORKG. For crowdsourcing to take off, it is important that actual domain experts are able to organize themselves to determine which use cases are relevant and leverage structured scholarly knowledge for their domains.

2.3.2 Statement Browser

The statement browser shows a list of statements (or RDF triples) displayed as property value pairs, and therefore lists the available structured data for a specific concept. It is possible to navigate from one page to the next inline within the statement browser. Although the statement browser plays a crucial role in the ORKG, the term ‘statement browser’ itself is never used, as the tool forms an integrative part of the user interface. In addition to statements, the statement browser shows classes and gives the ability to further describe properties. Several tools are integrated within the statement browser that provide guidance for users to structure their data. One of these tools is the Lookup functionality, which helps users in finding the most appropriate resources and predicates for their structured data. This is done by performing both a lookup into the ORKG and in external systems. These

external systems contain, among others, Wikidata, Geonames, and a variety of popular ontologies provided by the TIB ontology service[3]. Users are encouraged to reuse existing data instead of creating new predicates and resources to increase the interoperability of their knowledge descriptions. The statement browser is tightly integrated with the ORKG Template system, which is discussed in chapter 5. The statement browser provides users with several options to show more detailed information, including the classes and data types. By default, this data is hidden from the users, in order to hide information that can be distracting and is not strictly required to describe a paper.

Figure 2.9 below shows an example of the statement browser, as displayed on an ORKG Paper page. In total, four predicates are displayed with their corresponding resources. Additionally, the applied templates are displayed. To further explore the data, it is possible to click on the displayed links, to navigate further in the graph.

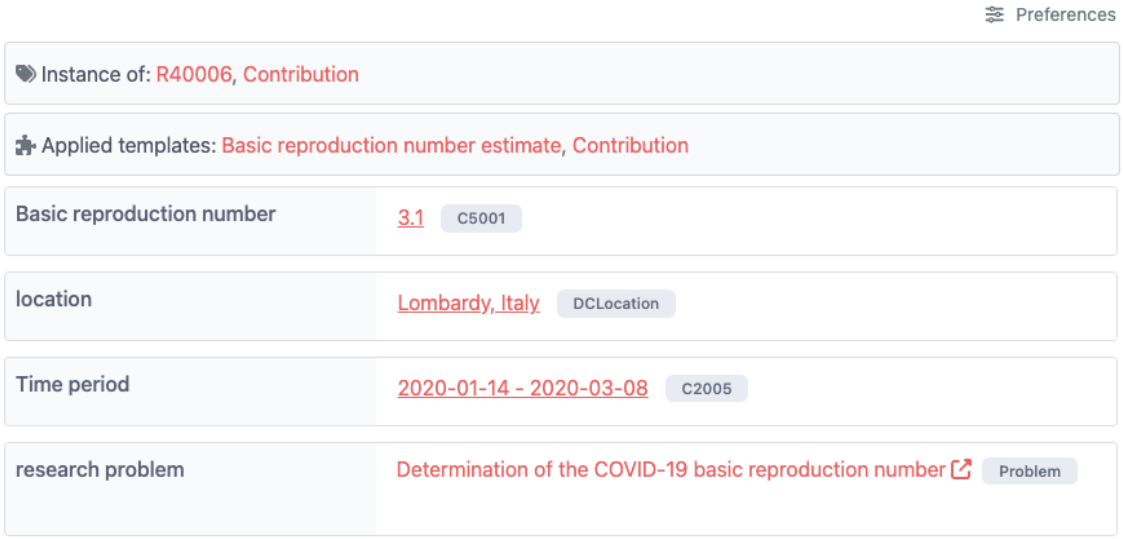

Figure 2.9 ORKG Statement Browser showing four properties and the resources with their respective classes

2.3.3 Templates

Templates help users to create reusable data models to describe their data. This fosters reusability of the data, as similarly modeled data enhances interoperability. A template defines the properties of the data described, and lets users specify the values that these properties accept (i.e., the range). ORKG Templates are an important tool for power users and therefore discussed separately in chapter 5 of this book.

[3] https://terminology.tib.eu/ts/ontologies

2.3.4 Contribution Editor

As previously described, Comparisons are one of the main features of the ORKG. In order to create and edit comparisons, users can either decide to edit individual papers used for the comparison, or to edit comparisons in bulk. Bulk editing of paper data is possible within the Contribution Editor, which serves as a grid editor juxtaposing multiple papers. Papers are displayed in the columns, and individual properties of the papers in the rows. The contribution editor shows only data directly associated with a paper contribution, nested data is not displayed within the table. Although it is possible to click on individual resources to further explore them and see the nested data, the contribution editor is mainly targeting simple comparison building, with a flat (i.e., non-nested) data structure. All cells within the table can be edited by double clicking on them. Furthermore, it is possible to apply templates to the data. Finally, when a user is satisfied with the entered data, it is possible to click the ‘Create comparison’ button, which opens a new comparison window, listing the papers that were used in the contribution editor. An example of the contribution editor, showing three papers and five properties, is displayed in Figure 2.10 below.

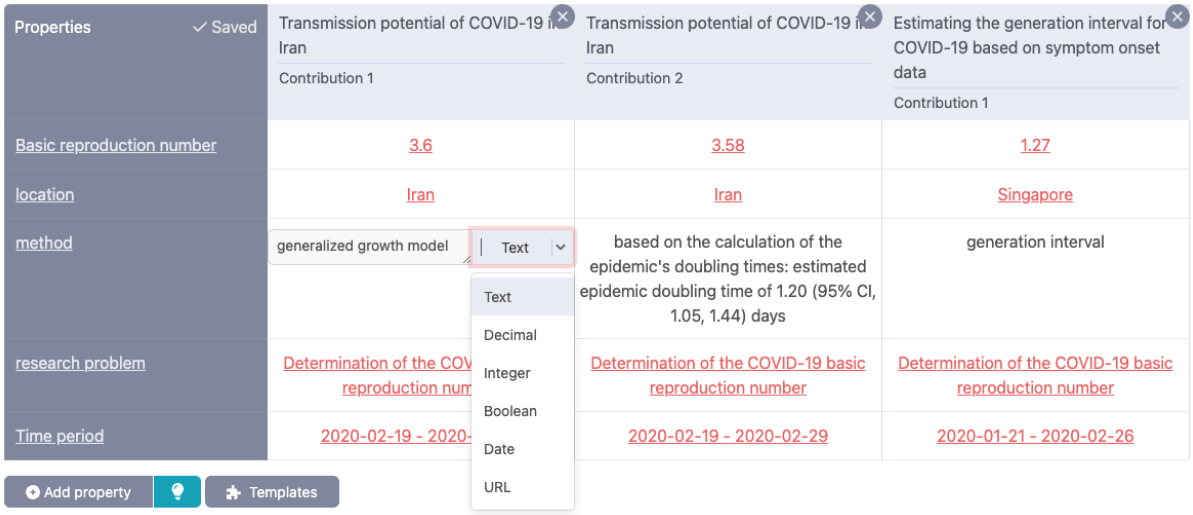

Figure 2.10 Simultaneous editing of papers with Contribution Editor

2.3.5 CSV Importer

Another method to get started with the ORKG is using the CSV Import functionality. This makes it possible to import paper data, described in the rows of the CSV file, with their respective properties, listed in the columns of the CSV file. A set of predefined properties can be used to describe the paper’s metadata. Furthermore, any other arbitrary properties can be used to describe the contents (contribution data) of a paper. When importing a CSV file, it is possible to either use IDs of

entities, or to let the system try to automatically determine what data should be reused and what data should be newly created. The CSV importer furthermore contains checks to determine whether the provided CSV file indeed follows the required format. In our experience, many researchers already keep track of topic-related research in some sort of spreadsheet. Therefore, the CSV import functionality provides an entry point for those researchers to easily get started with structured ORKG data. Naturally, the CSV format has its limitations due to the simple, but limited, syntax. Therefore, we recommend using the CSV importer only for simple use cases, and using ORKG REST API for other cases. An example of the CSV importer is displayed in Figure 2.11 below. As can be seen, in the second step, the syntax and data of the CSV file is validated to ensure it can be imported into the ORKG without issues.

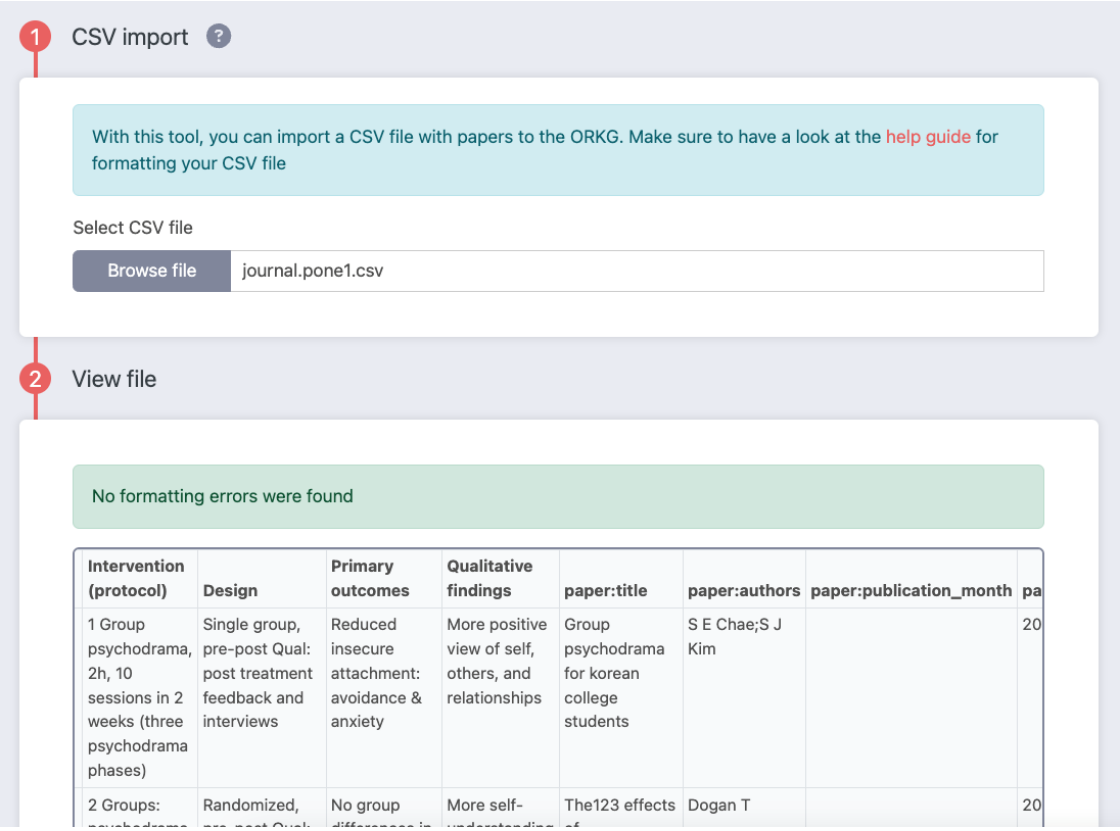

Figure 2.11 CSV Importer showing the first two steps of the CSV Importer, including the validation

2.3.6 Survey Importer

ORKG Comparisons are comparable to traditional tables in articles, where related literature is compared. They are especially common in review articles, to provide a summarized overview of related literature. ORKG supports importing those review tables from articles via a tool called the Survey Importer. A screenshot of the tool is displayed in Figure 2.12.

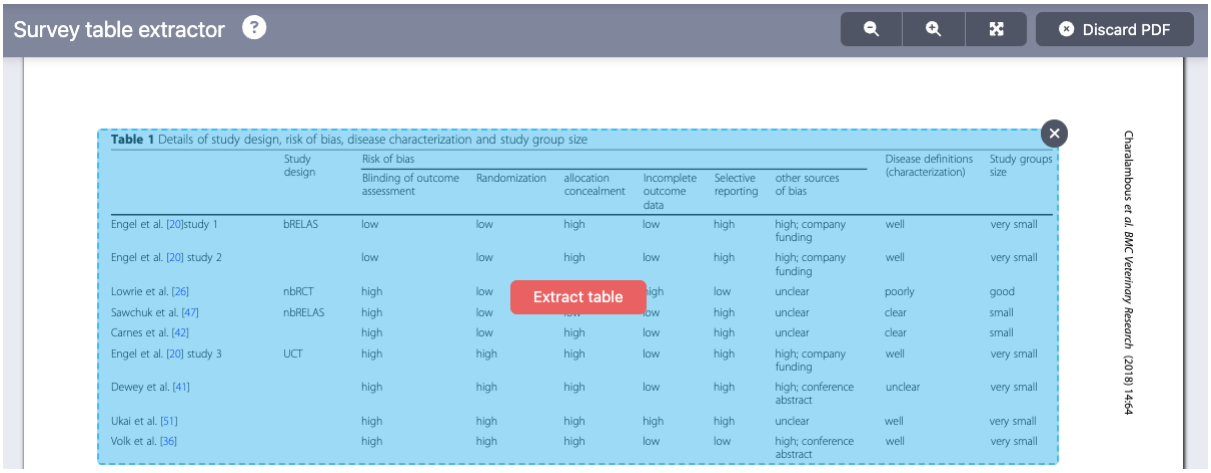

Figure 2.12 ORKG Survey Importer showing the selection of a review table for import (Oelen et al., 2020a)

The workflow of importing a survey is as follows. First, a user has to upload the PDF file that contains literature tables. We define literature tables as tables that display information from a specific paper, and include a reference to that paper in each row. Second, a user has to select the table region (see the blue area in the screenshot). The selected area will be extracted. Third, the user has to manually fix extraction errors within the built-in spreadsheet editor. Fourth, the user has to convert the data to ORKG data (linking to existing resources or creating new resources). Finally, the data can be imported into the graph. When imported, it is possible to create a Comparison from the imported data (*Oelen et al., 2020a*). Compared to the original format in which the data was presented, the data within the ORKG is more machine-readable and reusable.

2.3.7 Smart Suggestions

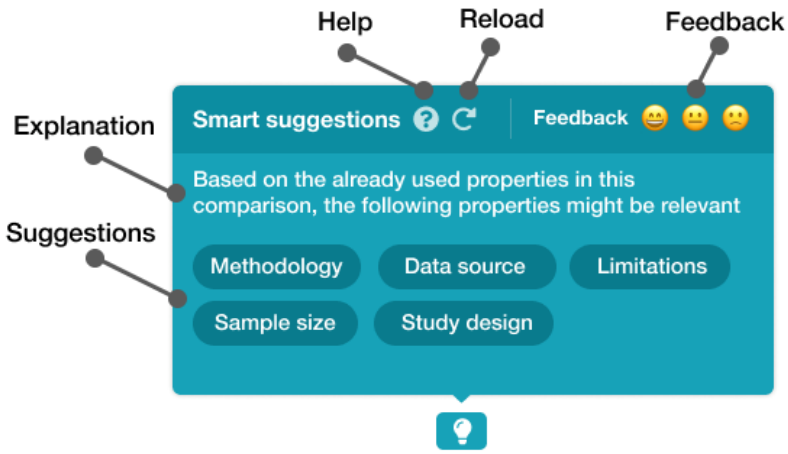

Figure 2.13 Smart Suggestions for possibly relevant properties

With the recent developments of Large Language Models (LLMs), the ORKG also focuses on the automatic extraction of knowledge from papers. One of the key elements of the ORKG is the manual verification of data, and therefore automatically extracted data is not added to the graph without human verification. Instead, we leverage LLMs to provide intelligent user interfaces that actively support users in creating structured knowledge. Within the ORKG, this becomes apparent by the same light bulb button that is displayed wherever Smart Suggestions are available. Smart Suggestions are integrated in several parts of the UI, including for recommending relevant properties for paper descriptions, recommending resources for specific properties, determining the relevance of metadata descriptions, and assessing the correctness of specific graph structures (*Oelen and Auer, 2024*). An example of is Smart Suggestions is displayed in Figure 2.13. Details are discussed in chapter 6.

References

Oelen, A. and Auer, S. 2024. Leveraging Large Language Models for Realizing Truly Intelligent User Interfaces. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11--16, 2024, Honolulu, HI, USA* (Honolulu, HI, USA, 2024). https://programs.sigchi.org/chi/2024/program/content/150511

Oelen, A., Jaradeh, M.Y., Stocker, M. and Auer, S. 2020. Generate FAIR literature surveys with scholarly knowledge graphs. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (2020), 97–106. https://doi.org/10.1145/3383583.3398520

Oelen, A., Stocker, M. and Auer, S. 2020. Creating a scholarly knowledge graph from survey article tables. *International Conference on Asian Digital Libraries* (2020 a), 373–389. https://doi.org/10.1007/978-3-030-64452-9_35

Oelen, A., Stocker, M. and Auer, S. 2021. SmartReviews: towards human-and machine-actionable reviews. *Linking Theory and Practice of Digital Libraries: 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13–17, 2021, Proceedings 25* (2021), 181–186. https://doi.org/10.1007/978-3-030-86324-1_22

3. Guidelines for creating Comparisons

Muhammad Haris, Hassan Hussein, and Vinodh Ilangovan

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

The ORKG supports describing scholarly articles in the form of research contributions, which represent scientific results, obtained using particular materials and methods that address a research problem. In addition, the ORKG allows comparison of these research contributions and thus supports knowledge synthesis. Comparisons are a primitive form of synthesis, may be useful as a dataset organizing data, which is then, processed for a specific synthesis objective. Serving as a fundamental ORKG feature, comparisons offer users a unique and powerful tool for navigating and understanding the evolving landscape of research in a specific field. Researchers can utilize these comparisons to quickly grasp the current landscape of a field and identify key developments. One notable feature of ORKG's comparative analyses is their living/ dynamic nature.

An ORKG comparison allows users to conduct a dynamic and comprehensive analysis that provides a consolidated overview of the state-of-the-art scholarly literature for a specific research topic (*Oelen et al. 2020*) Comparisons in ORKG have the option to be continuously updated and thus can complement traditional literature surveys to reflect the latest advancements in the field. This is crucial for researchers and enthusiasts seeking to stay abreast of the most recent advancements and emerging trends. Moreover, comparisons in the ORKG adhere to the FAIR principles (*Wilkinson et al., 2016*), emphasizing that the data is Findable, Accessible, Interoperable, and Reusable. This commitment to the FAIR principles enhances the usability and reliability of the scholarly knowledge, making the ORKG a trusted resource for a diverse audience, including researchers, educators, policymakers, and industry professionals, promoting transparency and facilitating the exchange of knowledge within the research community.

3.1 Understanding the value of Comparisons in the ORKG

The key benefits of the ORKG's comparisons include their ability to offer a consolidated view of existing literature, identify gaps in research, discover emerging trends, facilitate interdisciplinary collaboration, and build upon existing research. Researchers can leverage this feature to make informed decisions about their work, and educators can use it to enhance their teaching materials with the latest

insights from the academic community. Overall, ORKG's comparisons stand as a pivotal component of the platform, embodying a commitment to dynamic, FAIR, and accessible knowledge dissemination in the realm of open research. While the ORKG's comparative analyses present a powerful tool to support synthesizing and navigating scholarly knowledge, challenges such as data quality, standardization, and ensuring real-time updates need continuous attention.

Having comparisons in the ORKG is crucial for several reasons

1. **Comprehensive Overview:** Comparisons in ORKG provide users with a comprehensive overview of the state-of-the-art literature on a particular topic. This allows researchers, educators, and other stakeholders to quickly understand the current landscape of a field, identify key contributions, and gain insights into the evolution of knowledge.
2. **Dynamic and Updated Information:** The dynamic nature of comparisons ensures that the information presented can be updated in real-time. This feature is invaluable in rapidly evolving research fields, enabling users to stay current with the latest developments, trends, and breakthroughs.
3. **Identification of Trends and Gaps:** By aggregating and comparing various research findings, ORKG's comparisons enable users to identify emerging trends and gaps in existing literature. This is essential for researchers planning new studies, helping them make informed decisions about where their contributions can have the most significant impact.
4. **Interdisciplinary Collaboration:** Comparisons facilitate interdisciplinary collaboration by allowing users to present a consolidated view of research from various disciplines. Researchers from different fields can use this feature to identify common ground, potential collaborations, and opportunities for cross-disciplinary exploration.
5. **Educational Value:** Comparisons serve as valuable educational tools, offering educators and students a curated and up-to-date resource for understanding the current state of knowledge in a specific field. This aids in the development of educational materials that reflect the latest advancements and perspectives.

In essence, comparisons in the ORKG contribute significantly to the mission of organizing and disseminating open research knowledge. They empower users with timely and relevant information, foster collaboration, and support evidence-based decision-making in research and education.

3.2 Important characteristics of a Comparison

The effectiveness of a comparison in the ORKG depends on several key elements that contribute to its comprehensiveness, usability, and reliability. We discuss important characteristics of a comparison in the ORKG and outline guidelines for structuring a comparison.

1. **Relevance and Scope:** A successful comparison is defined by its relevance to a specific topic or research area. The scope of the comparison should be clearly defined to ensure that the included literature and data are directly related to the user's area of interest. Before diving into the comparison process, clearly define the purpose of your research. Understand the specific aspects you want to compare, the variables involved, and the research questions you seek to answer. This foundational step will guide the structure and focus of your comparison.

2. **Clearly Documenting Literature Search Process:** Similar to classical meta-analyses and systematic reviews, the literature search process underlying an ORKG comparison should be clearly documented, so other ORKG users can assess the literature search effort. A robust comparison considers a diverse range of sources, including academic papers, conference proceedings, books, and other scholarly outputs. This inclusivity helps present a holistic view of the research landscape and avoids bias towards specific types of publications. A curated approach to adding and updating content within the comparison is crucial for maintaining high-quality information. Quality control measures, such as peer review or automated checks help ensure the accuracy and reliability of the included literature

3. **Structured and Clear Presentation:** The presentation of the comparison should be structured and clear, facilitating easy navigation for users. A well-organized layout with clear headings, categories, and labels enhances the user experience and makes it simpler to extract relevant information. Create a taxonomy or classification system to organize and classify the resources involved in your comparison. A well-structured taxonomy improves the clarity and readability of your comparison, making it easier for others to comprehend and build upon your work. Selecting the relevant properties for your comparison is crucial. Resources represent the objects of study, such as datasets, methods, or concepts. Ensure that the properties you choose align

with your research objectives and contribute meaningfully to the overall comparison.

4. **Metadata and Contextual Information:** Each entry in the comparison should be accompanied by metadata and contextual information, providing details about the publication, authors, publication date, and other relevant information. This metadata enhances the transparency and trustworthiness of the comparison.

5. **FAIR Principles Compliance:** Adherence to the FAIR principles ensures that the data in the comparison is easily discoverable, accessible to a wide range of users, interoperable with other systems, and reusable for various purposes.

6. **Utilise Visualisations:** Visual representations enhance the readability and interpretation of your comparisons. Use graphs, charts, and diagrams to illustrate relationships, trends, and variations within your data. ORKG supports various visualisation tools, making it easy to create convincing graphics to complement the textual content.

7. **Dynamic Updating:** The crucial aspect is the comparison's ability to dynamically update in real-time when the user edits it. This ensures that the information can remain current and reflect the latest developments and contributions in the field, providing users with the most up-to-date insights at the edit time. Interactive elements, such as the ability for users to contribute feedback, comments, or additional references, enhance collaboration within the research community. These features foster a sense of community engagement and allow for continuous improvement of the comparison.

3.3 Creating high quality Comparisons

Creating a comparison involves understanding and highlighting the differences between different knowledge graph approaches, particularly with a focus on human- and machine-actionable elements. In this conceptual comparison, we explore the considerations of human accessibility and machine actionability, as well as the role of data modelling and the use of resources in contrast to literals.

3.3.1 Human- and machine-actionable elements

Knowledge Graphs are not inherently user-friendly for human editing. They are primarily designed for machine processing through structured data modelling,

while also incorporating select fields to enhance human readability. In this approach, the emphasis is on utilizing linked resources instead of direct data literals.

Human Accessibility

Traditional knowledge graphs are usually designed with a focus on machine readability. Editing may require expertise in query languages or specialized tools (e.g., SPARQL), making it less accessible for non-experts. On the other hand, the ORKG approach emphasizes a user-friendly interface, permitting researchers and domain experts to contribute and edit the knowledge graph without deep technical knowledge. In addition, ORKG comparison incorporates visualizations and interactive features to enhance human understanding and collaboration.

Machine Actionability

Traditional Knowledge Graphs are primarily designed for computerized processing by machines. On the other hand, the ORKG balances machine actionability with human readability. The ORKG utilizes semantic structures that machines can interpret while maintaining a genuine language expressiveness for human understanding.

3.3.2 Knowledge Graph Structure

Data Modeling for Machine Actionability

Traditional Knowledge Graphs prioritize rigorous schema adherence and normalized structures for efficient machine processing, often favoring ontologies and taxonomies. The ORKG adjusts a flexible data model, accommodating diverse data structures and evolving research needs. Supports semantically rich metadata, enabling subtle representation of knowledge.

Human Readability and Select Fields

In traditional Knowledge Graphs, the focus leans heavily on structured data modelling, potentially sacrificing human readability for machine interpretability. Contrarily, the ORKG integrates select fields and formats to prioritize human readability and machine interpretability, enhancing accessibility for users of varying expertise levels.

Resource Usage vs. Literals

Traditional Knowledge Graphs often rely on literals to represent data directly within the graph structure, prioritizing simplicity and efficiency. In contrast, the ORKG

incorporates linked resources alongside literals, enriching the depth and contextuality of the data representation. This approach enables users to explore complex relationships between entities, enhancing overall understanding of data.

3.4 Ensuring data quality of Comparisons

Crowdsourcing is currently the main approach for populating the ORKG. While crowdsourcing offers many benefits to create an ORKG comparison, it also comes with its set of challenges. Any user can make edits of the existing comparison and save a new version of the comparison. It should be noted that each version of a comparison should be saved to avoid losing data. Crowdsourced data may reflect the biases or subjective perspectives of contributors. Clearly documenting the literature search process is one important mechanism for mitigating bias for the objectivity of the knowledge graph. Ensuring consistency in the representation of information is crucial for the overall quality of the knowledge graph. Crowdsourced contributions might introduce semantic ambiguity, where different contributors interpret concepts differently.

ORKG users may have varying levels of expertise in data modelling and knowledge graph representation. Thus, it can be challenging to maintain a high level of data quality across the ORKG. We propose a graded framework for Knowledge Graph Maturity Model (KGMM) underlining joint and evolutionary curation of knowledge graphs (*Hussein et al., 2022*). The model comprises five maturity levels and emphasizes 20 quality measures. We categorize them into three priority levels within each maturity level. This structured approach enhances the model's practicality. Drawing inspiration from the FAIR data principles (*Wilkinson et al., 2016*), the Linked Open Data star scheme by Berners-Lee[4], and the Linked Data Quality Framework (*Zaveri et al., 2016*). We tailored and expanded the model to suit scholarly knowledge graphs, with a particular focus on facilitating human-machine collaboration. Specifically designed to support the realization and implementation of the FAIR principles, making data Findable, Accessible, Interoperable, and Reusable. The model guides knowledge graph developers and curators, offering a principled framework for ensuring quality in knowledge graph applications.

The framework utilizes the quality measures as an instrument to enrich the data quality in comparisons. The inherent nature of comparisons allows users to edit

[4] https://www.w3.org/DesignIssues/LinkedData.html

and refine them, contributing to ongoing refinements in data quality. The comparisons implementation encourages users to engage in a feedback mechanism with other researchers, fostering collaborative efforts to enhance data quality. At each level, the model focuses on specific data quality factors as outlined below:

Level 1: We set the priority at this level towards the scrutiny of the system's infrastructure and its responsiveness. The ORKG inherently shows respectable performance due to the adoption of a high-performance graph database, Neo4j, for data storage.

Level 2: positions a detailed focus on enhancing data completeness within the framework. In this regard, we designed the interface to guide and prompt the users to complete the missing resources, properties, and essential descriptions as shown in Figure 3.1. The systematic approach of the interface directs the users towards areas where information may be insufficient, thereby fostering a more comprehensive and well-rounded dataset. By actively engaging users in the process of completing missing elements, level two contributes to the overall robustness and thoroughness of the data available within the system. This particular attention to data completeness aligns with the intention to promote a more comprehensive and accurate representation of information in the knowledge graph.

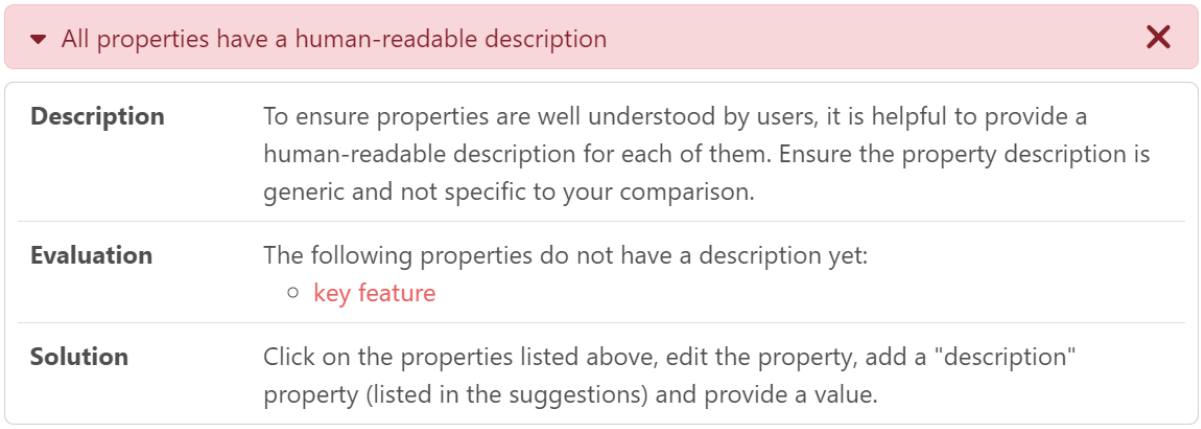

Figure 3.1 KGGM Level 2 suggestions to improve the properties description

Level 3: We dedicate Level 3 to enhancing the data reusability aspect of comparison, encompassing several measurements. Firstly, we give meticulous attention to the metadata reuse. Data reusability ensures thorough documentation of the metadata associated with the knowledge graph, promoting clarity and transparency. Furthermore, the commitment to data reusability extends to the metadata publication with a clearly defined and accessible utilization agreement. A critical aspect of level 3 is the integration of comprehensive provenance information within

the metadata. By associating the data with detailed provenance, the framework ensures a transparent record of the origin and evolution of the information, contributing to increased trust and reliability. Conciseness is another element at this level, supporting the absence of duplicated entities and relations within the knowledge graph. Conciseness minimizes redundancy, promoting a more efficient and manageable system. Finally, we address the data representation, emphasizing that the knowledge graph should present data in a suitable language and unit with explicit data definitions. Data representation ensures that the information is not only accessible but also interpretable, catering to diverse user needs and preferences. Level 3, through its multifaceted approach, underscores the importance of data reusability, precision, and clarity within the scholarly knowledge graph.

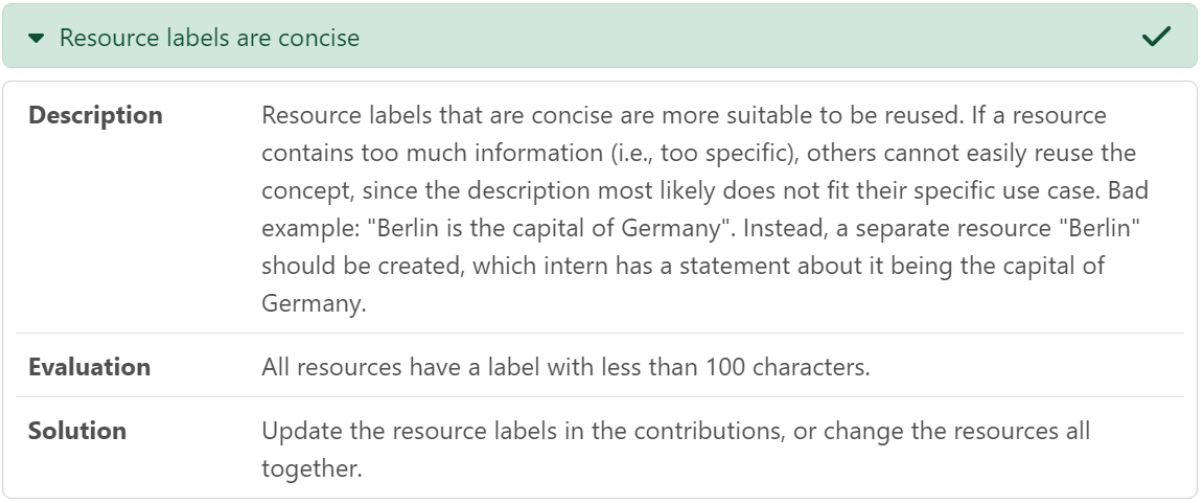

Figure 3.2 KGGM Level 3 to address conciseness of resource labels

Level 4: Level 4 directs its focus on aspects that contribute to the stable and reliable nature of the knowledge graph. A primary consideration within level 4 is trackability, which involves using Uniform Resource Identifiers (URIs) as distinctive identifiers for real-world objects. ORKG assigns a unique URI for each created resource. This practice ensures a consistent and traceable link between entities in the knowledge graph and their real-world counterparts. Moreover, we highlight identifier stability, emphasizing the importance of utilizing URIs as stable and persistent identifiers. This choice enhances the reliability of the comparison by providing a consistent and unchanging means of reference over time. In parallel, queryability takes priority at Level 4 to involve the provision of SPARQL, GraphQL, and API endpoints, which simplifies the process for data consumers to retrieve information from the knowledge graph. This accessibility enhances the usability of the data, making it straightforward for researchers and stakeholders to interact with and extract relevant insights efficiently. ORKG provides SPARQL, an API endpoint for data integration, which makes the comparison data available for consumers in

a machine-actionable format. By prioritizing trackability and identifier stability through using URIs and by emphasizing queryability through accessible query endpoints, Level 4 of the framework ensures the stability and accessibility of the knowledge graph. In turn, it contributes to the reliability and long-term utility of the scholarly information encapsulated within the system.

Level 5: The primary focus lies in the capacity to dereference resources based on Uniform Resource Identifiers (URIs). Another crucial objective at this level is to ensure linkability, representing the extent to which instances within the data set are interconnected. This measure underscores the collaborative nature of human-machine interaction. This approach aligns with the overarching goal of creating a highly linked and interconnected scholarly knowledge graph, where human input complements automated processes to enhance the overall coherence and reliability of the dataset.

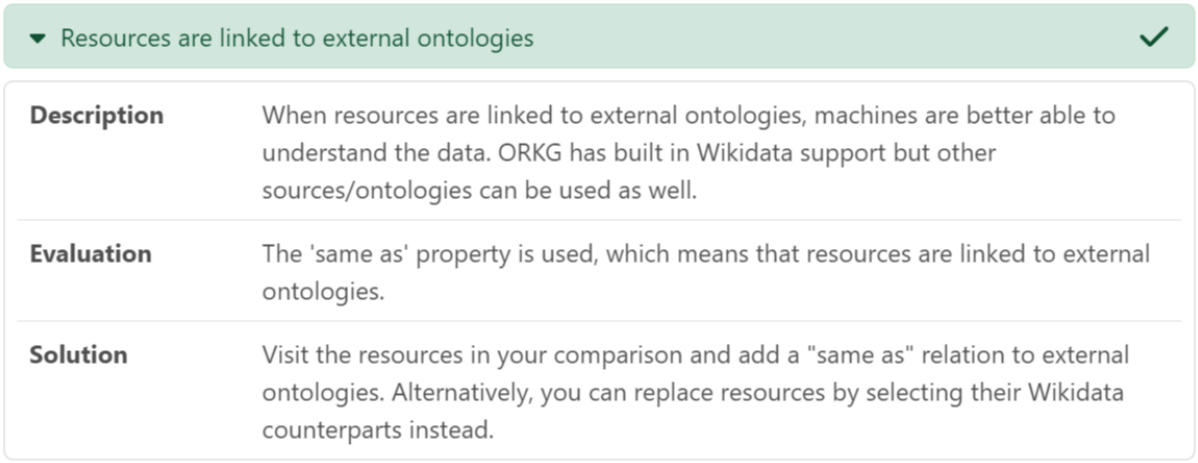

Figure 3.3 KGGM Level 5 linking external resources

3.5 Discoverability of ORKG Comparisons

The indexing of ORKG comparisons is crucial in global scholarly communication infrastructures for the discovery of the diverse information resources they contain. Publication of comparisons is essential for their global discovery, as well as enhancing data interoperability and streamlining research workflows. ORKG supports the DOI-based persistent identification of its comparisons to make these artefacts citeable and findable in global scholarly communication infrastructures

(e.g., DataCite, OpenAIRE, ORCID). A DOI is assigned to a comparison by leveraging DataCite services and publishing metadata following the DataCite metadata schema. While publishing the ORKG comparison, it is ensured that the metadata contains links between the ORKG comparison and articles described in the comparison. Other persistent identifiers (for example, contributor ORCID IDs and organization IDs) are also specified in the metadata. This rich and interlinked metadata is shared with DataCite, which in turn shares it with scholarly communication infrastructures. With this ORKG comparisons become discoverable in global scholarly communication infrastructures. With the publication of ORKG comparisons, researchers can discover the descriptions of articles and their comparisons in summarized and structured form.

3.6 Conclusion

Creating comparisons in the Open Research Knowledge Graph is a powerful way to synthesize and present information from multiple research sources. ORKG facilitates the development of scholarly communication by enabling machine-readable descriptions of research contributions. This makes research outputs more transparent and comparable, thereby improving information needs for readers (*Jaradeh et al., 2019*). Additionally, the iterative refinement process, involving regular updates with new information and peer feedback incorporation, ensures that comparisons can remain current and accurate. This dynamic approach aligns with the evolving nature of scientific knowledge. The goal is to aid in understanding complex research landscapes and to provide clear, accessible comparisons that advance knowledge across scientific fields.

References

Oelen, A., Jaradeh, M.Y., Stocker, M. and Auer, S. 2020. Generate FAIR literature surveys with scholarly knowledge graphs. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (2020), 97–106. https://doi.org/10.1145/3383583.3398520

Hussein, H., Oelen, A., Karras, O., Auer, S. 2022. KGMM - A Maturity Model for Scholarly Knowledge Graphs based on Intertwined Human-Machine Collaboration. International Conference on Asian Digital Libraries 2022 https://doi.org/10.48550/arXiv.2211.12223

Jaradeh, M.Y., Oelen, A., Farfar K.E., Prinz, M., D'Souza, J., Kismihók,G., Stocker, M., and Auer, S. 2019. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. In Proceedings of the 10th International Conference on Knowledge Capture (K-CAP '19). Association for Computing Machinery, New York, NY, USA, 243–246. https://doi.org/10.1145/3360901.3364435

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

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web 7(1), 63–93 (2016) https://doi.org/10.3233/SW-150175

4. ORKG Benchmarks

Jennifer D'Souza, Salomon Kabongo, Manuel Prinz, Yaser Jaradeh, and Kheir Eddine Farfar

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

As alluded to in earlier chapters of this book, the relentless expansion of scientific literature, with an estimated two million articles published annually across 30,000 journals (*Bornmann et al., 2021, Altbach and De Wit, 2019*), presents an overwhelming challenge for researchers striving to stay abreast of developments in their fields. This deluge of unstructured text necessitates innovative solutions for efficient knowledge navigation and assimilation. A pivotal approach to addressing this issue is the representation of research contributions buried in the discourse of unstructured text into structured formats, as introduced in the earlier chapters of this book. Structured representations of research contributions, by distilling complex information into comprehensible and machine-actionable formats, not only aid in managing the vast scientific corpus but also enhance the accessibility and utility of research findings.

In the realm of Artificial Intelligence (AI) research, the primary focus often revolves around developing new models capable of achieving state-of-the-art (SOTA) performance, a process traditionally encapsulated through the reporting of four integral elements: Task, Dataset, Metric, and Score (TDMS). In this context, a novel manifestation of structured knowledge representation focuses only on the TDMS facet and is captured through benchmarks. Benchmarks, traditionally community-curated on public websites, present a distilled view of the AI research landscape by tracking models' performances across various tasks offering additional functionalities such as sorting models' scores from highest to lowest, and vice versa, in leaderboards or computing performance trendlines. Renowned examples include websites like NLP-Progress (http://nlpprogress.com/), AI-metrics (https://www.eff.org/ai/metrics), SQUaD explorer (https://rajpurkar.github.io/SQuAD-explorer/), and more recently, Papers with Code (https://paperswithcode.com/). These platforms efficiently track and display the performance of various AI models across different tasks, datasets, and metrics, offering a clear and concise overview of the state-of-the-art advancements. This enables researchers to quickly determine the leading models and methodologies in their field.

The ORKG represents a significant leap forward in this arena with its "Benchmarks" feature (https://orkg.org/benchmarks). The ORKG Benchmarks feature, while also a community curation endeavor, diverges from the aforementioned platforms by incorporating these AI model scores into a knowledge graph (KG). This transition from mere website listings to a structured, graph-based representation aligns with the FAIR principles (Findable, Accessible, Interoperable, and Reusable), thereby enhancing the utility, visibility, and interoperability of this information. Furthermore, the ORKG's approach to representing benchmarks as part of a semantic web-based KG ensures that the data is grounded in a universally accessible format specified in RDF or OWL. Researchers can easily compare models based on standardized metrics, view state-of-the-art results for specific tasks, and access additional resources such as source code URLs. ORKG Benchmarks, with its streamlined and user-friendly interface, stands in contrast to traditional search engines' document-heavy approach.

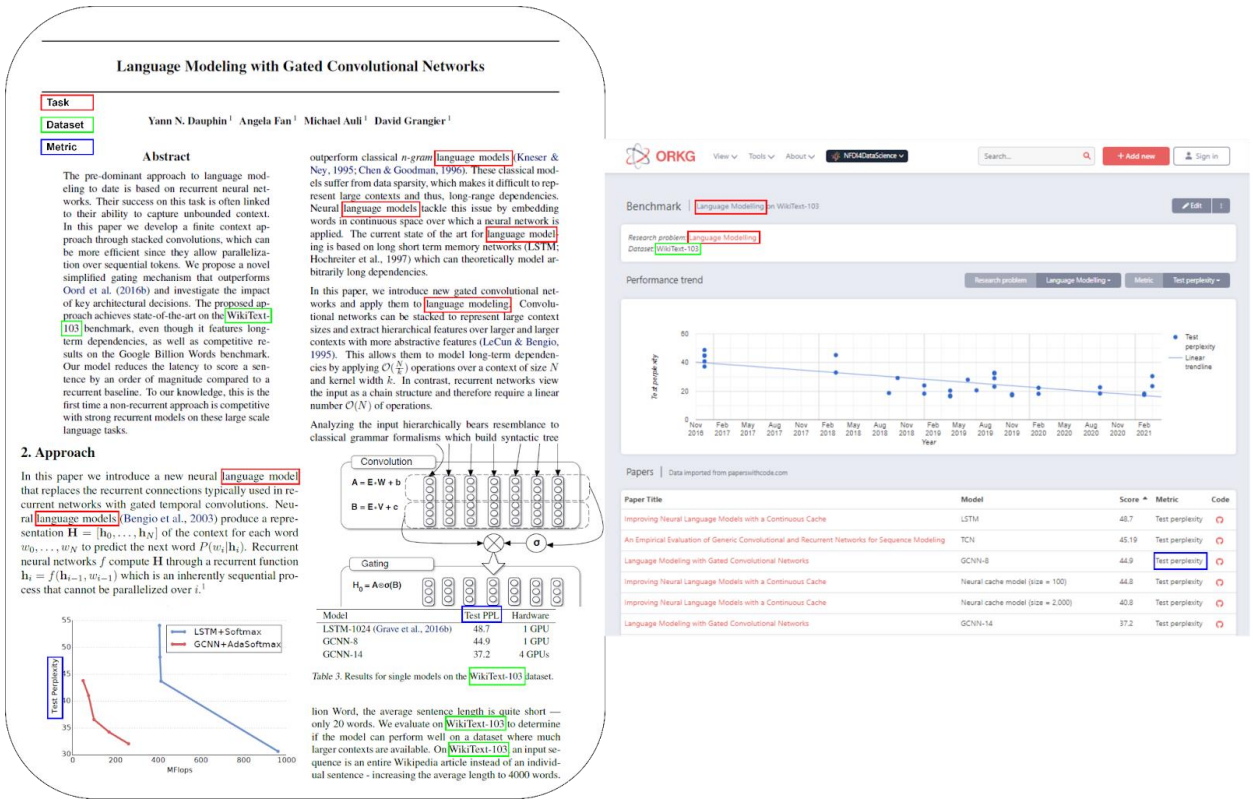

Figure 4.1 A contrastive view of Task-Dataset-Metric information in the traditional PDF format of publishing as non-machine-actionable data (on the left) versus as machine-actionable data as the Open Research Knowledge Graph (ORKG) Benchmarks (on the right).

This platform not only enables efficient tracking of AI advancements but also promotes strategic reading, community engagement, and collective curation. Its representation method is vital in an era of rapid AI progress, adeptly addressing the

crucial question, “What’s the current state-of-the-art result for task XYZ?” and keeping pace with evolving benchmarks. Figure 4.1 contrasts the predominant discourse-based publishing of model scores with information scattered and buried within the text versus as a machine-actionable benchmark in the ORKG.

As we delve deeper into the details and implications of the ORKG Benchmarks in this chapter, we invite the AI research community to engage with this innovative feature. The ORKG Benchmarks feature exemplifies the potential of semantic web technologies in transforming how we capture, compare, and communicate scientific advancements in AI. The aim is to enhance the dissemination and accessibility of AI research, fostering a collaborative environment that keeps pace with the rapid advancements in the field.

4.1 Definitions

In this section, we define the ORKG Benchmarks’ structured information capture facets.

Task. A task, in scholarly articles, signifies the central research objective, crucial for machine learning model development. Commonly highlighted in the Title, Abstract, Introduction, or Results sections, it can vary across domains like question answering, image classification, and drug discovery.

Dataset. A dataset, as referenced in empirical scholarly articles, represents a specific collection of data tailored for a particular Task in machine learning experiments. An article may discuss one or multiple datasets, with mentions typically located in the same sections as Task mentions. E.g., HellaSwag (*Zellers et al., 2019*) or Winogrande (*Sakaguchi et al., 2021*).

Metric. A metric in scholarly articles is a standard measurement for assessing machine learning model performance, aligned with specific Tasks and Datasets. Articles may evaluate models using various metrics, typically discussed in Results sections and in Tables. Examples include BLEU (bilingual evaluation understudy) for machine translation tasks (*Papineni et al., 2002*), F-score (*Sasaki, 2007*) for classification tasks, and MRR (mean reciprocal rank) (*Voorhees, 1999*) for information retrieval or question answering tasks.

Model. An AI model in scholarly articles refers to a computational framework executing specific Tasks with chosen Datasets and evaluated via Metrics. Model references are typically found in the Methodology section, where the model's design and implementation are detailed, and in the Results section, where its performance is evaluated. E.g., BERT (*Devlin et al., 2019*) or GPT-1 (*Radford et al., 2018*).

Benchmark. ORKG Benchmarks (https://orkg.org/benchmarks) systematically categorizes state-of-the-art empirical research within ORKG research fields (https://orkg.org/fields). Each benchmark comprehensively details elements such as Task, Dataset, Metric, Model, and source code for a specific research field. For example, an ORKG Benchmark on "Language Modelling" may involve evaluation on the WikiText-2 dataset, using the "Validation perplexity" metric, and include a compilation of various models with their corresponding scores.

Leaderboard. Depicted on ORKG Benchmark pages, leaderboards are a dynamically computed chart that depict the performance trend-line of models developed over time based on specific evaluation metrics.

4.2 Guide to Creating a Benchmark in the ORKG

In the process of contributing benchmarks to the ORKG, the user's starting point is the 'Leaderboard' template, accessible at https://orkg.org/template/R107801. The 'Leaderboard' template is an instrumental feature, akin to the role of templates in Wikipedia, for standardizing and facilitating the addition of benchmarks. In Wikipedia, templates have been instrumental in ensuring consistency, quality, and ease of information curation. Drawing from this concept, the ORKG Leaderboard template provides a similar structure, with predefined properties like Task, Dataset, Model, Metric, Score, and Source Code. This template not only streamlines the benchmark submission process but also aligns with the FAIR principles, ensuring that the contributions are findable, accessible, interoperable, and reusable.

The Leaderboard template, adhering to the best practices of semantic modeling, is described as a graph in a five-level specification to capture the intricate interrelationships and detailed aspects of reporting benchmarks. Figure 4.2 depicts the diagrammatic view or schematic representation of the Leaderboard template. Despite the complexity inherent in its multi-level specification, the ORKG frontend interface significantly simplifies the process of data entry. It presents the Leaderboard template as a user-friendly form, making it accessible and manageable for researchers to input their data. This design choice effectively bridges the gap, on the one hand, between the depth required for detailed and meaningful semantic representation, and on the other, the practical ease of data submission.

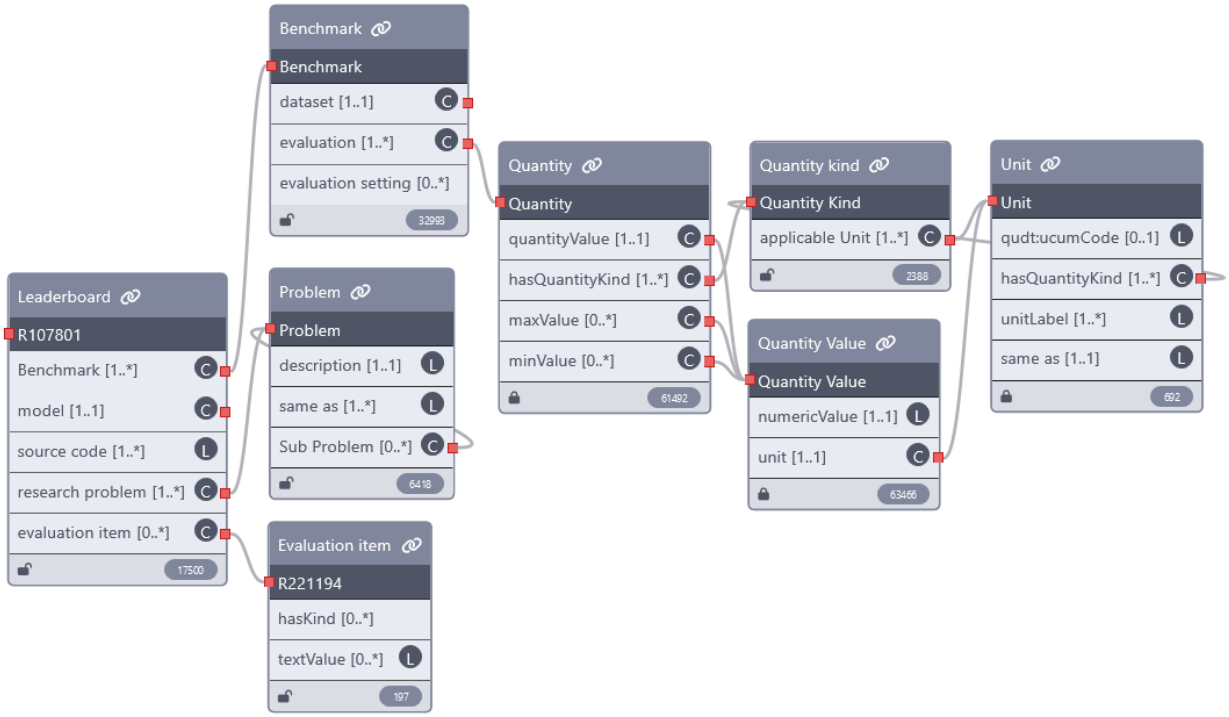

Figure 4.2 Leaderboard template diagrammatic view https://orkg.org/template/R107801

Benchmarks can be reported via the “Add paper” workflow followed to add a paper’s structured contribution description in the ORKG. Except in the case of reporting a Benchmark, the contribution structure is predetermined by the Leaderboard template. Our database of templates including the Leaderboard template can be searched and selected at the time of adding a paper. The information the user must have at hand are the following: the model name, the research problem or task addressed, the name of the dataset used in the evaluation, the evaluation metric, the score reported by the model for the metric, and the source code of the model, if available. All these properties are automatically specified when the leaderboard template is selected as a form in the frontend that the user can use to submit their respective benchmarks. A video demonstrating the process of creating a benchmark in the ORKG as a step-by-step guide can be accessed online here https://www.doi.org/10.5446/56183.

4.3 The Workflow Dynamics of ORKG Benchmarks

In addressing the challenge posed by the proliferation of scientific publications, generally, the ORKG platform presents a next-generation solution of representing research contributions as structured FAIR information. But the benefits of the platform are also concerned with the representation of the structured data in smart

frontend interfaces which facilitate more efficient navigation and filtering of research findings. This section discusses the ORKG Benchmarks user-centric workflow that enables exploring the benchmarks subgraph data part of the ORKG. The ORKG platform's workflow dynamically interacts with the (T, D, M, S) quadruple – Task, Dataset, Metric, Score – representing the empirical AI research landscape. The overall exploration workflow, depicted in Figure 4.3, is thoughtfully designed to guide users through a series of intuitive stages, ensuring an effortless and informative experience.

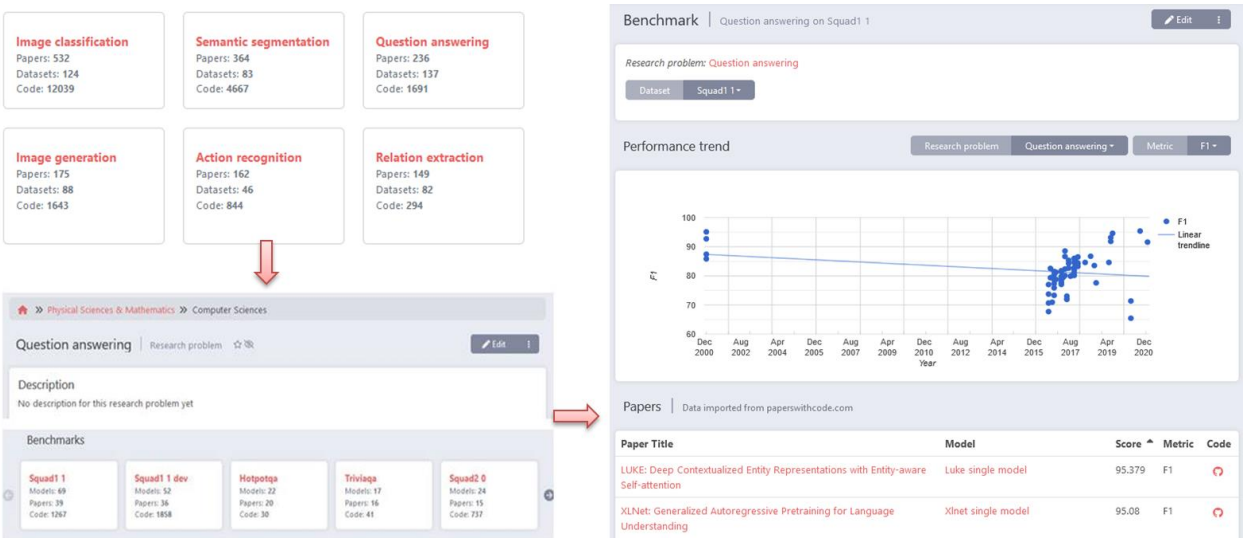

Figure 4.3 The dynamic frontend exploration workflow of various AI benchmarks reported as w.r.t. standardized properties as task, dataset, model, metric, score, and source code for the ORKG Benchmarks feature.

In the first stage of the workflow, users are presented with a comprehensive display of all Tasks addressed within the platform. This overview allows users to quickly grasp the breadth of research areas covered. Upon selecting a Task of interest, the user is then led to the second stage: a carousel of Datasets that address the chosen Task. This stage not only highlights the diverse datasets employed in AI research but also assists users in pinpointing the specific context of their interest. The final stage culminates in a leaderboard display, showcasing all models that address the Task on the selected Dataset. Here, evaluation Scores are presented in relation to the relevant Metric. This information is further enriched with a performance trendline, offering users a visual representation of model performance over time. Such a detailed and structured presentation of information empowers researchers to rapidly assimilate key findings, compare model performances, and make informed decisions about which papers to delve into for further reading.

In conclusion, the ORKG Benchmarks workflow exemplifies a significant advancement in the realm of scholarly communication. It offers a meta-analysis alternative that not only displays a list of papers but also provides vital filtering and comparison

tools. These tools assist researchers in making more informed decisions, thus addressing the critical need for smart, structured, and user-friendly platforms in the ever-evolving field of AI research.

4.4 Conclusion

In conclusion, the future of ORKG Benchmarks is geared towards integrating automated text mining solutions, including the use of human-in-the-loop AI approaches and Large Language Models (LLMs), as highlighted in ongoing research endeavors (*Kabongo et al., 2021, Kabongo et al., 2023 & Kabongo et al., 2023a*). This integration aims to address the challenge of converting unstructured scholarly texts, predominantly in PDF format, into structured, machine-readable formats, thus enhancing the efficiency of knowledge discovery. Central to this endeavor is the advancement of Research Knowledge Graphs (RKGs), which organize information into graph structures, aligning with FAIR principles and facilitating downstream applications like search engines and recommender systems. These developments promise to significantly advance the structuring and accessibility of scientific knowledge, contributing to a more efficient and navigable scientific research landscape.

References

Bornmann, L., Haunschild, R. and Mutz, R., 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1), pp.1-15.

Altbach, P.G. and De Wit, H., 2019. Too much academic research is being published. *International Higher Education*, (96), pp.2-3.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. and Choi, Y., 2019, July. HellaSwag: Can a Machine Really Finish Your Sentence?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4791-4800).

Sakaguchi, K., Bras, R.L., Bhagavatula, C. and Choi, Y., 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, *64*(9), pp.99-106.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Sasaki, Y., 2007. The truth of the F-measure. *Teach tutor mater*, *1*(5), pp.1-5.

Voorhees, E.M., 1999, November. The trec-8 question answering track report. In *Trec* (Vol. 99, pp. 77-82).

Devlin, J., Chang, M., Lee K., Toutanova K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT*. 2019.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I., 2018. Improving language understanding by generative pre-training. *OpenAI Blog* https://openai.com/research/language-unsupervised

Kabongo, S., D'Souza, J. and Auer, S., 2021, December. Automated Mining of Leaderboards for Empirical AI Research. In *International Conference on Asian Digital Libraries* (pp. 453-470).

Kabongo, S., D'Souza, J. and Auer, S., 2023. ORKG-Leaderboards: a systematic workflow for mining leaderboards as a knowledge graph. *International Journal on Digital Libraries*, pp.1-14.

S. Kabongo, J. D'Souza and S. Auer, "Zero-Shot Entailment of Leaderboards for Empirical AI Research," *2023 a ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Santa Fe, NM, USA, 2023, pp. 237-241, doi: 10.1109/JCDL57899.2023.00042.

5. Modeling and Quality Assurance through Templates

Kheir Eddine Farfar, Markus Stocker, and Lars Vogt

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

As we continue to explore the ORKG, it is important to address a critical aspect of scientific research: the challenge of ensuring high data quality and effective data modeling. To support comprehensive analysis and comparison of this data, especially in domains like climate change and biodiversity loss, researchers need a way to ensure their research data adheres to Findability, Accessibility, Interoperability, and Reusability (FAIR) principles (*Wilkinson et al., 2016*). This is where the concept of semantic interoperability comes in – the ability for machines to understand and compare data from different sources. These challenges stem from the varied ways in which research data is traditionally recorded and shared. Without standardized schemas, data can be inconsistent, making it difficult to compare and analyze across different studies.

This is where the ORKG's template system plays a crucial role. The system offers a structured way for researchers to record their contributions, ensuring that data is not only consistently formatted but also easily understandable. By using templates based on the Shapes Constraint Language (SHACL) (*Knublauch and Kontokostas, 2017*), the ORKG adopts a standardized language to data modeling, enhancing consistency and quality across diverse research fields.

In this part of our discussion, we will dive deeper into the significance of the template system in addressing these data quality and modeling challenges. We will explore how this system contrasts with previous methods and how it simplifies the process for researchers, especially those who may not be experts in data modeling. The role of these templates in enhancing data quality and their impact on the broader scientific community will be examined in detail.

The ORKG template system empowers domain experts to define the structure of contributions, which is crucial in standardizing and validating research data. This not only facilitates data entry through user-friendly input forms but also ensures consistency and quality control through constraints on properties like data types

(e.g., text, boolean, date, ontology term) and cardinality (i.e., how many inputs of this type the template allows).

Through this chapter, you will gain a clearer understanding of how the ORKG's template system contributes to solving key issues in data management within scientific research. This understanding is essential for recognizing the value of the ORKG in the broader context of advancing and streamlining scientific knowledge sharing.

5.1 Need for a template system

In the context of rapidly expanding scientific data, a structured method for data curation is essential. The template system in the ORKG addresses this necessity through several technical strategies:

1. **Standardizing Data Input**: In the diverse field of scientific research, ORKG templates play a crucial role in simplifying data curation. The ORKG provides scientists with ready-to-use templates that alleviate the complexities of formatting and describing data from their papers. This is particularly valuable for standard elements like reporting results with values and units. Researchers can choose from pre-existing templates or create new ones to suit their specific needs, thereby streamlining the curation of their paper's data. This method ensures both clarity and consistency, enhancing the efficiency of the data curation process.
2. **Facilitating Data Validation**: While researchers curate their data, the ORKG template system automatically checks the data types and how many times a piece of data appears (cardinalities). This ensures that the data meets the specific rules and structures (schemas and constraints) set by the system. Because of this, the data in the ORKG is more comparable, as it follows these predefined standards.
3. **Enhancing Interoperability**: The ORKG template system uses a SHACL-based design, making it easier to share and use data with other systems. This is important for connecting ORKG data with other datasets that also follow SHACL. With this setup, data can be more easily shared between different platforms. The template system supports a subset of SHACL shapes and has tools for both importing and exporting data, further aiding in this seamless data exchange.
4. **Adhering to FAIR Principles**: The ORKG template system is designed to align with the FAIR principles (*Stocker et al., 2023*). This alignment ensures

that research content within the ORKG is more findable, accessible, interoperable and reusable both for humans and machine-based systems, thereby fostering a more collaborative and efficient scientific research environment.

5. **Improving Research Efficiency**: By abstracting the complexities of data formatting, the ORKG template system allows researchers to focus more on the substantive content of their work rather than on the intricacies of data modeling and presentation. This efficiency in managing data not only saves time but also enhances the quality of the research output.

5.2 Overview

To understand how the ORKG template system works, Figure 5.1 illustrates the process of creating templates by domain experts, linked to ontologies, and turned into forms for the community to fill out. This section breaks down each step, explaining how it all comes together.

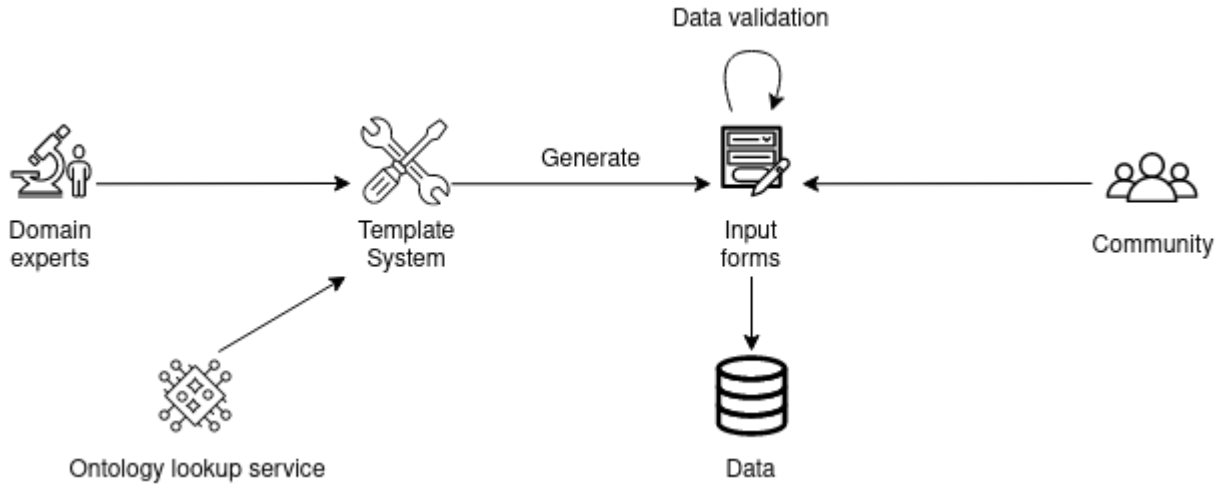

Figure 5.1 ORKG template system.

5.2.1 The Role of Domain Experts in Template Creation

Domain experts in various research fields play a crucial role in the ORKG template system. They leverage their expertise to create and refine templates that accurately represent the data structures needed in their respective fields. This process is facilitated by the template editor, a user-friendly interface where experts can define the graph pattern for a specific type of data, specifying concepts (nodes) and the relationships between them (edges), similar to a mind-map. Each template translates into an input form with input fields that allow only specific types of input, thus constraining the input (e.g. only a float value for the value node of a measurement and a unit resource for the unit node).

5.2.2 Integration with an Ontology Lookup Service

A key feature of the template system is its integration with an ontology lookup service, more specifically the TIB Terminology Service[5]. This feature allows domain experts to connect their templates to existing ontologies, enriching the templates with standardized vocabularies and classifications. This connection ensures that the data captured in the templates is consistent with broader semantic frameworks, enhancing the interoperability and reusability of the data.

5.2.3 Template System's Role in Creating Input Forms

Once a user selects a template from the ORKG template gallery, the 'Statement Browser' - a component used for browsing and editing data comes into play. It parses the chosen template, identifying and setting up the right properties along with their constraints. This results in the creation of input forms that are aligned with the template's specifications, as depicted in Figure 5.2 These specialized forms thus collect all data and relationships, as defined by the template, and ensure that the data entered is consistent.

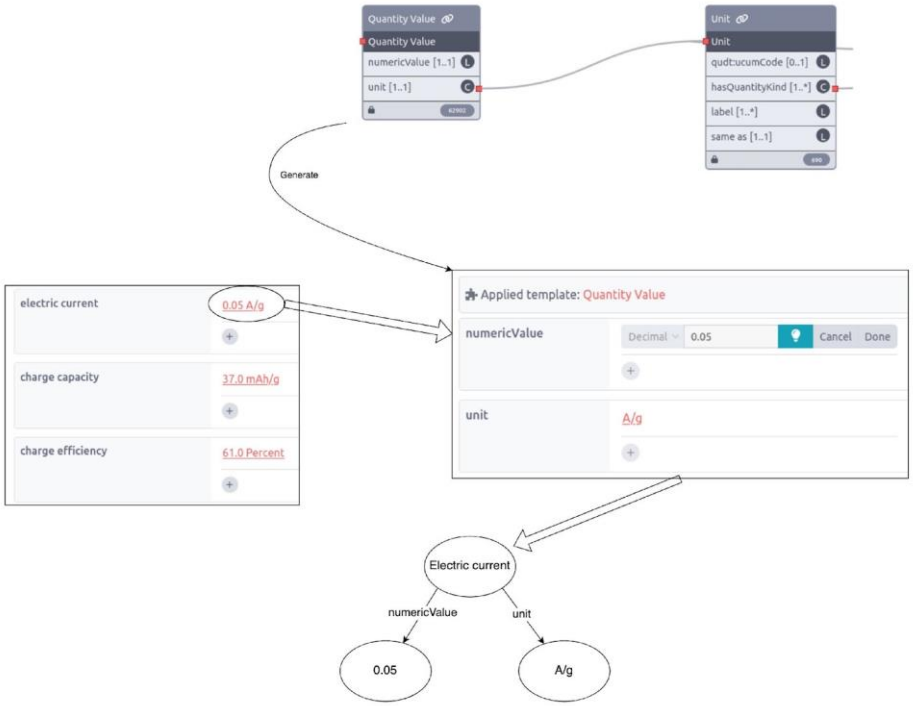

Figure 5.2 Template-based input form (middle), derived from the template shown in the upper part, and corresponding graph (bottom).

[5] https://terminology.tib.eu/ts

5.2.4 Data Validation Process

Before any new data is stored in the system, the Statement Browser performs a data validation process. This process checks the incoming data against the constraints and requirements specified in the template. It ensures that all data entries are consistent with the expected data types, formats, and relationships, thus maintaining the integrity and quality of the data in the ORKG.

5.2.5 Community Contribution and Data Addition

When researchers and community contributors add data for a new paper to the ORKG, they begin by selecting an appropriate template from the template gallery. The Statement Browser guides them through the data entry and performs validation checks to confirm adherence to the template's specifications. Once the data clears these checks, it is stored in the ORKG, making it available for access and use by the wider research community.

This workflow, from template creation by domain experts to data addition by the community, represents a streamlined and efficient process for managing and structuring research data. By integrating ontology lookup services, automating form creation, and enforcing data validation, the ORKG template system helps maintain a good level of data quality and usability in scientific research.

The ORKG includes an NLP system that supports users in choosing a template. When a researcher adds a paper, this system suggests a template based on the paper's research field, title, and abstract. This means users do not always have to search through the template gallery. The NLP system's suggestions help ensure that the data is organized in a template that fits the paper's content, making the process quicker and more straightforward.

5.3 SHACL Shapes

SHACL is a standard from the World Wide Web Consortium (W3C) for validating data in RDF (Resource Description Framework) graphs. RDF is a model used for data exchange on the Web and many knowledge graphs use RDF. SHACL sets up rules (shapes) to ensure that data adhering to these shapes is interoperable across all graphs that apply this shape.

In the ORKG, while SHACL is not implemented in its entirety, its vocabulary is essential for creating input forms and performing data validation. Instead of applying SHACL directly to the RDF graph, the ORKG uses SHACL's structure and

terms to guide the setup of input forms and validate data before it is stored. This approach ensures that the data adheres to a consistent format and meets the necessary standards, maintaining data quality and structure without the need for full SHACL implementation on the graph.

By utilizing a subset of SHACL shapes, the ORKG aligns with other SHACL-based systems. This shared format allows the ORKG to both export its data structures as SHACL shapes and import data structures from other SHACL-compliant systems. This facilitates easier exchange and integration of knowledge across these systems.

5.3.1 Template editor

To understand how the ORKG template system functions, we present a concrete example template for documenting engineering experiments, including the attributes Experiment ID, Substance, and Temperature. What makes the ORKG template system user-friendly is that users don't need to deal with the technical intricacies of SHACL or template construction. Instead, they can effortlessly define and edit templates using the template editor.

The template creation process revolves around three fundamental questions:

1. What are the use cases of the template?
2. Which properties can be used to describe a specific type of entity (or an instance of a specific concept)?
3. How do we refer to its instances (i.e., all the graphs created using the template)?

Each of these questions aligns with a specific tab in the template editor: Description, Properties, and Format.

- **Description Tab**: In this tab, we provide essential information such as the template name ("Experiment"), target class, and properties linking the contribution resource to the graph created using the template. We also specify relevant research fields and problems.
- **Properties Tab**: Here, we detail the input fields required, such as "Experiment ID" (Text type), "Substance" (Text type), and "Temperature" (Number type). You can set cardinalities and even define nested templates for specific types, e.g. "Substance".
- **Format Tab**: This tab specifies how the information from the underlying graph created using the template can be presented in a user-friendly way in the UI. For example, an experiment with ID "123", involving "Water", and

conducted at "25°C" would be presented in the UI as "Experiment 123: Water at 25°C". More details are given in the next section.

- **Instances Tab**: This tab allows for exploration and browsing of graphs that has been created using the template. It provides a convenient overview of all instances associated with the template, but it is not used during the editing process.

With these intuitive tabs, the ORKG template editor simplifies the task of creating and managing templates, making it accessible to both users and domain experts in scientific research.

5.3.2 Formatted Labels

The ORKG's Formatted Label feature represents a notable improvement in data representation and user convenience. Automating the generation of human-readable labels for resources, it emulates the functionality of Python's f-string formatting[6]. This feature not only simplifies data entry, but also enhances readability by auto-filling placeholders with relevant property values.

In scientific data, users establish formats with placeholders for properties in a resource's data structure. For example, a dataset might use:

`Experiment {Experiment_ID}: {Substance} at {Temperature}°C`

Here, `{Experiment_ID}`, `{Substance}`, and `{Temperature}` are placeholders that will be replaced with actual information from the underlying graph.

This feature streamlines data entry, ensuring consistency and readability across resources. It abstracts complexity by using properties instead of raw data, creating a user-friendly interface. Applied in scientific contexts, like chemical experiments, the feature allows for automated, descriptive labels such as "Experiment 00123: Sodium Chloride at 25°C". This facilitates quick identification and categorization of resources, making datasets more accessible and easier to navigate.

5.3.3 Template Visualization Diagram

The Template Visualization Diagram is a feature of the ORKG template system, offering a UML-like representation to visually map the structure and relationships of templates. This diagram aids in comprehending complex template designs, allowing users to grasp the overall framework and connections more intuitively. Figure 5.3 presents the different components of the diagram, which we discuss next.

[6] https://docs.python.org/3/tutorial/inputoutput.html

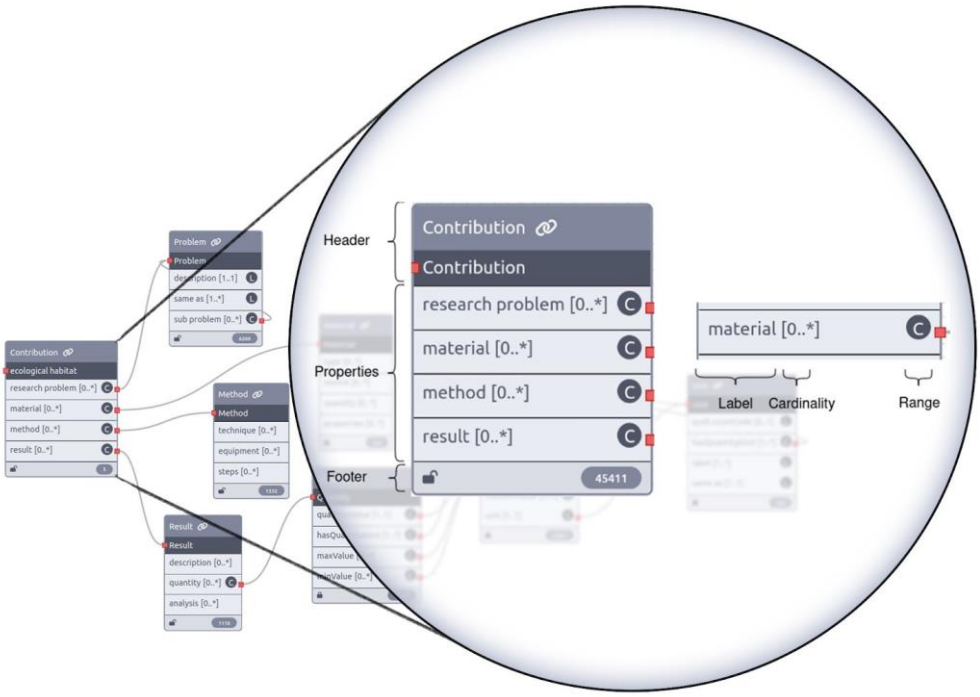

Figure 5.3 Template diagram with a zoom in on one of the entities.

Entity Representation: Each entity in the diagram represents a template in the ORKG. It comprises three main sections: the header, the properties section, and the footer.

1. **Header**: The header prominently displays the template name, followed by its associated target class. The template name acts as a descriptive identifier, while the target class denotes the intended class or type for the template. This layout offers immediate recognition of the template's function and its classification context.
2. **Properties Section**: This central part of the entity outlines the template's properties. Each property is depicted through its label, cardinality, and range.
    a. **Label**: Indicates the property's name, signifying the kind of information it captures.
    b. **Cardinality**: Shows the number of values a property can have, ranging from optional (0..1), mandatory (1..1), to multiple values (0..*).
    c. **Range**: Defines the expected data type or class for the property's value, marked with a "C" for class or an “L” for literal.
3. **Property Relationships**: When a property's range is a class, an edge links the property to the respective template entity, visually representing their interconnection.

4. **Footer**: Displays a small icon indicating whether the template is closed (no further properties can be added) or open (allows additional properties). And also the number of instances of the template.

The Template Visualization Diagram is a powerful tool for quickly grasping the structure of templates, the properties they encompass, and the interrelations among different templates as entities. This visual aid streamlines the process of analyzing, creating, and modifying templates, making it more accessible and efficient for users.

5.4 Import/Export Functionality

The ORKG platform simplifies the import process of SHACL shapes, featuring a user-friendly interface that begins with uploading an N-Triples[7] file. Prior to the actual import, users are provided with a preview option to ensure data accuracy. It is important to note, however, that the ORKG currently supports only one class as the target class for each SHACL shape.

5.4.1 Managing Existing Templates

A key functionality of ORKG's import process is its efficient management of existing templates. When a SHACL shape being imported targets a class already existing in the ORKG, the system uses the class's URI to verify if it matches an existing template. To maintain data integrity and avoid conflicts, the system disregards any imported template that targets an already templated class.

5.4.2 The Import Tool Workflow and Process

The import tool operates by:

1. Comparing the target class of the incoming SHACL shape with those in existing templates.
2. Ignoring shapes targeting classes that are already templated to prevent data duplication.
3. Importing shapes with unique target classes, adding them as new templates to the system.

This process ensures the preservation of the integrity of existing templates in the ORKG. After the data has been previewed and validated, users can initiate the import. This step integrates the SHACL shapes into the ORKG, making them avail-

[7] https://www.w3.org/TR/n-triples/

able for use in research contributions. The import and export functionalities together enhance the ORKG's capability in handling complex research data, thereby facilitating more effective and efficient research data management.

5.4.3 Exporting Templates to SHACL Files

In addition to importing, ORKG offers the capability to export templates as SHACL shapes in N-Triples format. This feature enhances the platform's versatility, allowing users to not only bring in new SHACL shapes, but also to extract and reuse existing templates in other contexts or systems. This export functionality is instrumental for data sharing and collaboration in research environments.

5.5 Future Perspectives

As we look ahead, the ORKG system is poised for significant enhancements that will further transform how research contributions are shared and utilized. We briefly present key areas of development.

5.5.1 Advanced SHACL Constraints Implementation

A major focus will be the implementation of SHACL constraints as a background task within the ORKG. This approach will enable the system to continuously run validation checks on existing data, identifying and reporting any inconsistencies or errors. By providing users with real-time feedback on validation issues, we aim to empower them to correct and align their data with the defined SHACL shapes. This ongoing validation process will be instrumental in maintaining data consistency and reliability across the ORKG platform.

5.5.2 Improved SHACL Shapes Support

We are committed to expanding our support for more properties and features within SHACL shapes. This expansion will enable the ORKG system to handle a broader range of data complexities and variations, further enhancing its flexibility and utility in diverse research contexts.

5.5.3 Interactive Template Visualization Diagram Editing

Another exciting development will be the introduction of editing capabilities within the Template Visualization Diagram. This feature will facilitate the modification of complex templates directly through an intuitive, visual interface. By simplifying the editing process, we aim to make template management more accessible and efficient for users.

5.5.4 Evolution of Formatted Labels

Moving beyond the current implementation of formatted labels, we envision evolving this feature into a more sophisticated template engine. This engine will be capable of abstracting complex graph structures into simple, human-readable sentences. This advancement will not only enhance data viewing but also transform data input methods. By using formatted labels as an input form, we will significantly streamline data entry processes, making them more intuitive and user-friendly.

5.6 Conclusion

The ORKG template system significantly streamlines the process of managing and curating research data, making it both easier to handle and more reliable. By providing a structured framework, it assists researchers in organizing their data effectively, ensuring that all contributions added using a template are consistently formatted and aligned with standardized norms. This system plays a crucial role in checking for errors and inconsistencies, which is vital in maintaining the integrity and trustworthiness of research content. By implementing these checks, the ORKG template system not only enhances the quality of the data but also bolsters the confidence of the scientific community in the results presented. Moreover, its user-friendly design makes it accessible to a wide range of users, from experienced researchers to those new to data modeling. Overall, this system holds the potential to become an invaluable tool in the quest for clear, correct, and reliable research data. Its adoption and further development could contribute to advancing scientific knowledge and promoting collaboration within the research community.

References

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).

Knublauch, H., Kontokostas, D. Shapes constraint language (SHACL).W3C Recommendation https://www.w3.org/TR/shacl/ (2017)

Stocker, M., et al. "FAIR scientific information with the open research knowledge graph." *FAIR Connect* 1.1 (2023): 19-21.

6. Natural Language Processing for the ORKG

Jennifer D'Souza, Yaser Jaradeh, and Allard Oelen

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

The ORKG represents a paradigm shift in the way scholarly knowledge is published and accessed. By leveraging the tools of the semantic web, the ORKG transforms traditional, narrative scientific contributions into structured, machine-actionable descriptions. This innovative platform facilitates a more efficient dissemination and retrieval of research findings, with the potential to accelerate the pace of scientific discovery. However, the complexity and diversity of research contributions present significant challenges for structuring and integrating this knowledge. Natural Language Processing (NLP), the technology that enables computers to understand, interpret, and generate human language, emerges as a pivotal technology in semi-automatically designing these workflows as a recommendation engine with humans-in-the-loop, offering a spectrum of services tailored to the unique facets of the ORKG.

In general, the development of NLP services has undergone a transformative evolution with the advent of Large Language Models (LLMs) such as ChatGPT[8]. Before the significant uptake of LLMs, traditional machine learning paradigms were employed to address the different NLP tasks that comprised the ORKG as a whole. These different NLP tasks can alternatively be seen as candidate ORKG NLP facets that need to be addressed with custom-tailored machine learning systems in the context of traditional learning paradigms. The transition to LLMs has not only enhanced the performance of these services, but also expanded the possibilities–i.e., the promise of artificial general intelligence (AGI) offered by LLMs means there is no longer a need for custom-tailored machine learning systems addressing the different ORKG NLP facets. Instead, one model can handle all ORKG NLP tasks (*Radford et al., 2019 & Raffel et al., 2020*)–for automating knowledge extraction and structuring within the ORKG.

This chapter unfolds in two main parts. First, we delve into the foundational NLP technologies necessary for constructing the ORKG. We embark by introducing the

[8] https://openai.com/blog/chatgpt

diverse array of ORKG NLP facets, illuminating the traditional machine learning objectives essential for their realization. This exploration contrasts with the conventional machine learning paradigm, necessitating bespoke solutions for each facet, with the potential of leveraging the broader intelligence afforded by recent advancements in LLMs, and its implications for the ORKG. Second, we transition to the application of the ORKG in scholarly question answering (SQA), elucidating the novel opportunities it presents within this domain. Herein lies a focus on leveraging the ORKG's repository of FAIR scholarly contributions. Thus, the concluding segment is dedicated to elucidating the SQA problem landscape and how the ORKG can serve as a unique benchmark for evaluating NLP systems' performance in the realm of scholarly question answering.

6.1 ORKG Natural Language Processing Facets

The comprehensive NLP problem for the ORKG platform comprises a diverse array of meticulously examined NLP facets, otherwise known as objectives. Specifically, ten distinct ORKG NLP facets have been identified. Each of these facets aim to tackle distinct challenges inherent in the organization and dissemination of scholarly knowledge in the ORKG platform. They holistically encapsulate the platform's essential requirement as that of automating the categorization and extraction or generation of research contributions. It is imperative for readers delving into the ORKG NLP chapter to grasp the breadth of facets outlined next from a conceptual perspective. The ORKG has not yet addressed all NLP facets, and some are part of ongoing research and development. An outline of the NLP facets are nevertheless offered to the reader as it serves not only as a panoramic perspective, but also to ignite inspiration within the NLP community at large to contribute to the ORKG NLP ecosystem with the development of novel solutions.

1. **Research Field Classification**: To organize scholarly knowledge in the ORKG, over 700 research fields from a comprehensive taxonomy (https://orkg.org/fields) are employed (https://orkg.org/stats). Within the ORKG's NLP set of facets, the research field classification objective is to automatically classify incoming papers based on contextual information like titles, abstracts, and metadata such as keywords, aiming for single-class classification into fine-grained taxonomy.
2. **Template Recommendation**: As introduced in earlier chapters, the ORKG utilizes templates to standardize the structure of its contribution sub-graphs. An associated NLP facet would involve automatically recommending templates from the ORKG's knowledge base for structuring the contributions of new papers. This automated system would compute the similarity between

the new paper's content, primarily its title and abstract, and each template in the ORKG's collection.

3. **Predicates Recommendation**: This facet can be seen as the counterpart to template recommendation. Template recommendation suggests (a) template(s), which is a human-expert-based predefined collection of predicates as a form. In contrast, a predicates recommendation service would look at the whole ORKG collection of predicates and suggest one or many of those it thinks are relevant to describing the contribution of an incoming paper. Unlike templates, which have a narrower scope, the ORKG's collection of predicates is broader and more versatile in its application across research fields[9]. This NLP facet could indirectly support template construction by proposing groups of predicates for potential template formation. Complementing the template suggestion, this NLP facet would handle suggesting predicates for structuring the contributions of papers in research themes that lack predefined templates but have ample structured descriptions.
4. **Template Population**: This facet's objective is the automatic filling of values for a chosen template on the ORKG platform, akin to automated form filling tasks. It would utilize contextual information, such as paper title, abstract, and full-text, to directly extract or infer values. In essence, a system for this facet would help streamline the ORKG "Add paper" or "Add comparison" workflows.
5. **Predicate Value Completion**: Similar to template population, except valid only for single predicates at a time.
6. **Similar Paper Retrieval**: ORKG Comparisons compile papers addressing the same research problems, requiring periodic updates with new findings. The similar paper retrieval facet aims to identify and suggest external papers resembling those in ORKG's Comparison collections, facilitating easy updates.
7. **Comparison Completion**: Assuming a new paper has been added to a comparison, this NLP facet would automatically populate its values for the given set of the comparison's properties. Similar principles applied to the development of the aforementioned template population facet also address this facet.
8. **Leaderboard Extraction**: Empirical AI papers often release models, benchmarked on a dataset, that provide an evaluation using a specific metric and report a performance score. Inspired from the Papers with Code platform (https://paperswithcode.com/), the ORKG implemented a benchmarks feature (https://orkg.org/benchmarks) that captures only the task, dataset, met-

[9] As of the current writing, the ORKG includes 487 unique templates (https://orkg.org/templates) and 10,065 properties (https://orkg.org/properties).

ric, score (T,D,M,S) values from AI papers which in turn powers performance trend lines allowing users to see the best or lowest performing models on a task dataset over time with regard to different metrics. As an NLP facet, leaderboard extraction is defined as an automatic extraction task of the T, D, M, S tuples objective given an empirical AI paper *(Hou et al., 2019*).

9. **Custom Knowledge Extraction Pipelines:** An alternative objective to constructing knowledge graphs (KGs) comprises a modular pipelined framework of two main tasks to which there may be additional supplementary tasks. The first task is named entity extraction (NER), which identifies and categorizes specific entities, such as the names of people, organizations, and locations in the general knowledge domain or within science disease or treatment names. The second task is relation extraction (RE), which identifies connections between entities. The NER and RE objectives can be addressed via fully supervised trained models. This setting entails training models based on gold-standard human annotated data with example NER and RE annotation targets that the model learns and then generalizes to new incoming data.
10. **Scholarly Knowledge Embeddings**: Scholarly knowledge embeddings distill intricate academic concepts and relationships into numerical representations, creating semantic vector spaces for computational comprehension and analysis. The extensive coverage of the ORKG platform, spanning over 700 research fields, enables the creation of multidisciplinary embeddings. This advancement could push the state-of-the-art in current embedding methods using language models such as SciBERT(*Beltagy et al., 2019*), which are limited to specific scientific domains like Computer Science and Biomedicine.

6.2 Evolution of NLP Services with Large Language Models

This section explores the ORKG's implemented NLP services, tracing the shift from traditional machine learning methods to the integration of Large Language Models (LLMs), marking a significant advancement in scholarly knowledge structuring and analysis automation. Divided into two parts, it first examines traditional machine learning objectives for specific ORKG NLP facets, followed by a discussion of more recent LLM-based implementations.

Traditional Machine Learning Objectives

1. **Sequence Labeling**: Traditional methods relied heavily on sequence labeling (*Ma and Hovy, 2016*) for scholarly NER, annotating abstracts, and identifying mentions of datasets (*Heddes et al., 2021*) and software (*Schindler et al., 2021*) in scientific literature. These techniques required

extensive training on domain-specific datasets to accurately identify and label the relevant entities. With the sequence labeling objective, our internally implemented services addressed the predicate value completion NLP facet generically in a multidisciplinary manner with STEM-ECR (*Brack et al., 2020*) and specifically for the domains of Computer Science (*D'Souza and Auer, 2022*) and Agriculture (*D'Souza, 2024*) with domain-specific NER types.

2. **Clustering**: The template and predicate recommendation facets of ORKG NLP have been addressed using traditional clustering machine learning (*Oghli et al., 2022*). The implemented recommendation services can be accessed through the ORKG NLP python package (https://gitlab.com/TIBHannover/orkg/nlp/orkg-nlp-api).This approach leverages unsupervised methods to define semantic clusters for the templates or predicates based on the existing ORKG papers structured by the recommendations. Thus for a new incoming paper, it was simply measured similarly to any of the existing clusters and thus recommended templates or predicates from that cluster. Based on a predefined threshold for the similarity measure, if the computed new paper similarity was higher than the threshold, it could also receive no template or predicate recommendation.
3. **Sentence Completion**: For the template and comparison completion facets, a language model's, i.e., BERT's (*Devlin et al., 2019*), ingrained sentence completion language modeling objective was employed. The task was designed by using the predicate as the prompt to be completed with the paper's abstract given as context from which the value for completion could be extracted (*D'Souza et al., 2023*).
4. **Pattern Extraction**: This method was crucial for extracting specific information, such as the R_0 number and Case Fatality Rate estimates for infectious diseases, through predefined patterns or regular expressions. However, it lacked flexibility and required extensive manual curation of patterns (*D'Souza and Auer, 2021*).
5. **Natural Language Inference (NLI)**: NLI techniques were applied to address the ORKG's leaderboard extraction NLP facet (*Kabongo et al., 2023, Kabongo et al., 2023 a*), requiring models to deduce relationships and extract structured information from unstructured text, a task that demanded significant understanding of the text's implicit meanings. A preliminary implementation of the leaderboard extraction can be found as the ORKG NLP python package online (https://orkg-nlp-pypi.readthedocs.io/en/latest/services/services.html#tdm-extraction-task-dataset-metric)
6. **Classification**: The ORKG research fields classification facet was addressed using the method of computing semantic embeddings between an incoming paper and a reduced set of the 700 ORKG research fields. The

most similar research field to the incoming paper determined the classification outcome. Our codebase is publicly released (https://gitlab.com/TIBHannover/orkg/nlp/experiments/orkg-research-fields-classifier) and the service is also available via the ORKG python package at this link (https://orkg-nlp-pypi.readthedocs.io/en/latest/services/services.html#research-fields-classification).

7. **Semantic Embeddings:** The similar paper recommendation ORKG NLP facet was implemented based on a method of semantic embeddings computed for existing papers and then applied to new incoming papers (*Nechakhin and D'Souza, 2023*). For the embeddings themselves, we relied on the Semantic Scholar API (https://www.semanticscholar.org/product/api).

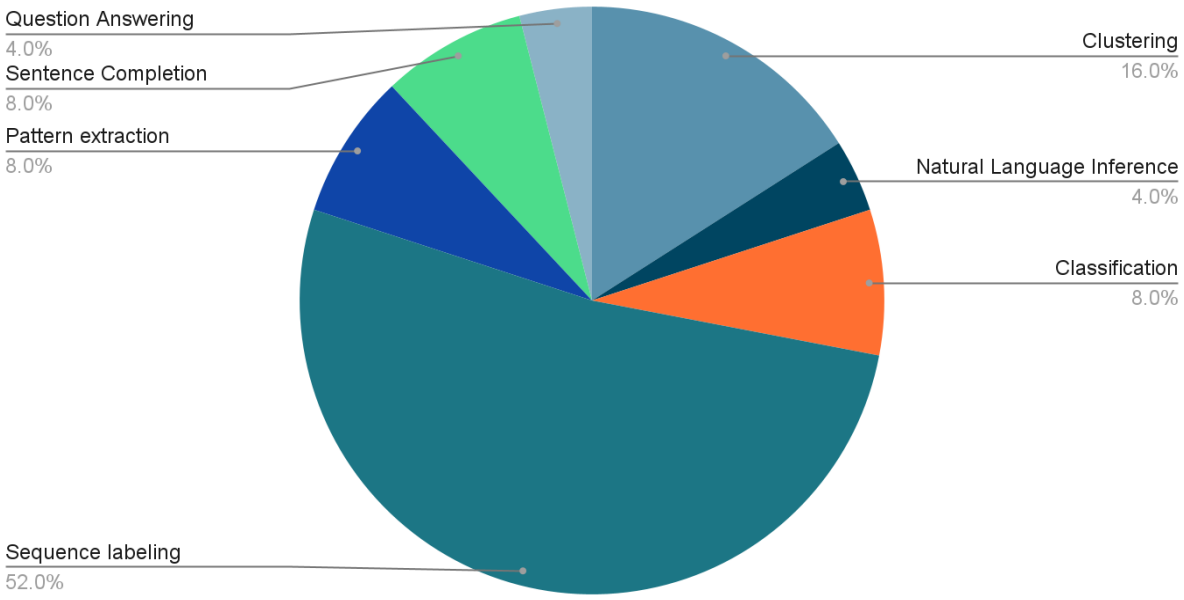

Figure 6.1 A pie chart of the ORKG NLP facets and the overall ratio of traditional machine learning objectives needed to address them

Each of these objectives required distinct models, datasets, and training regimens, making the processes, while extremely precise (*Ming et al., 2020*) and optimal in terms of computing resources, time-intensive to the broad and evolving needs of the ORKG. The code base for NLP services developed for the ORKG is open-sourced with the MIT license here https://gitlab.com/TIBHannover/orkg/nlp/experiments. Furthermore, the services are available via the ORKG NLP python package https://orkg-nlp-pypi.readthedocs.io/ or via the REST API https://orkg.org/nlp/api/docs#/ .

6.3 LLMs' Comprehensive Capabilities

LLMs, with their extensive pre-training on diverse text corpora (*Radford et al., 2019 & Raffel et al., 2020*), offer a unified solution to the multifaceted NLP tasks required by the ORKG. The key to harnessing the versatility of LLMs lies in prompt engineering, a method that involves crafting queries or instructions in natural language to guide the model's response towards a desired outcome. This approach allows LLMs to:

- **Simultaneously Address Multiple Objectives**: A single LLM, through tailored prompts, can perform a variety of tasks—from NER and clustering to sentence completion and pattern extraction—without the need for separate, specialized models.
- **Adapt with Minimal Effort**: Prompt engineering enables rapid adaptation to new tasks or changes in task requirements, bypassing the extensive retraining processes associated with traditional models.
- **Reduce Diverse Resource Requirements**: By leveraging a single LLM for multiple tasks, the ORKG can streamline its NLP services, reducing the diverse computational and data resource configurations needed for maintaining several task-specific models.
- **Enhance Accuracy and Contextual Relevance**: LLMs bring a deep understanding of context and language nuances, improving the quality and relevance of NLP outputs across the ORKG's diverse services.

Prompt engineering marks a transformative approach to NLP tasks in the ORKG. By consolidating various objectives into flexible and intelligently crafted prompts, LLMs provide a scalable, efficient, and adaptable solution. However, it is important to note that LLMs out of the box may not fully suffice for domains requiring high expertise due to inherent limitations, such as their general lack of domain specificity, challenges in fine-grained understanding of specialized concepts, and limited grounded knowledge of real-world entities and relationships. Additionally, the availability of high-quality, domain-specific training data for fine-tuning LLMs may be limited, hindering their performance in specialized domains. Despite these challenges, leveraging prompt engineering streamlines NLP service deployment and advances the ORKG's goal of enhancing the accessibility, interpretability, and actionability of scientific knowledge. This shift toward LLMs, driven by prompt engineering, showcases innovative utilization of cutting-edge NLP technologies for improving scholarly knowledge structuring and dissemination.

6.4 LLM-based ORKG Smart Suggestions

Smart Suggestions are a specific implementation of LLM-support within the ORKG user interface (*Oelen and Auer, 2024*). They guide users through the data creation

and curation process. They are implemented in an assistive, non-intrusive manner, and are displayed on demand. The existing workflows and content creation and curation remain unaltered. Figure 6.2 illustrates the workflow and user interface (UI). Smart Suggestions only provide additional guidance to users, and can be completely ignored when deemed irrelevant. They are implemented via a recognizable blue color palette and light bulb icon button. Upon clicking the lightbulb button, a tooltip appears, containing the recommendation from the LLM. This non-intrusive approach makes it possible to more rapidly implement LLM assistance throughout the interface, and allows the improvement of prompts over time, even if the recommendations are not always fully correct or relevant.

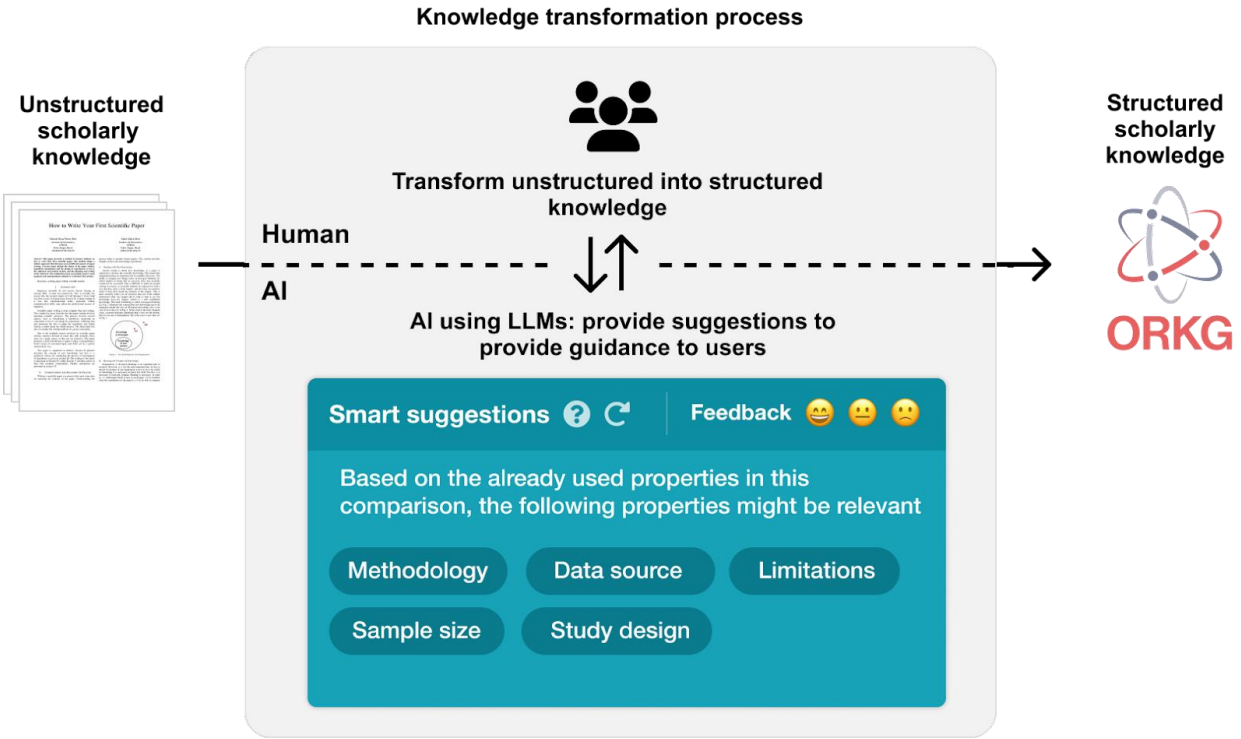

Figure 6.2 Smart Suggestions (AI) guide users (Humans) through the transformation process from unstructured to structured scholarly knowledge.

Specifically, Smart Suggestions are implemented for six different tasks in the UI. They can be categorized as "Closed recommendations" and "Open feedback". The closed recommendations are implemented in two use cases, and provide interactive suggestions that can be activated by clicking on the suggested values. The first use case relates to recommending related predicates, based on a set of existing predicates. The recommendation is based on a set of already used predicates for a specific paper or comparison. The second use case recommends resources for the object position, for a predefined set of predicates. Currently, values are recommended for the predicates "research problem", "method", and "approach". The set is limited to ensure a suitable prompt is used to recommend values for the

respective predicates. The remaining four open feedback use cases provide textual feedback instead of providing interactive buttons. The third use case is related to determining whether an object value should be a literal or resource. The fourth use case aims to assess whether a resource can be decomposed into smaller components (and thus increase the reusability of the data). The fifth use case determines whether a predicate label is sufficiently generic for reuse. Finally, the sixth use case evaluates whether a comparison description is sufficiently descriptive, or whether more context is required.

6.5 Scholarly Question Answering with the ORKG

Beyond the core NLP Services aimed at structuring and extracting knowledge from scholarly literature, the ORKG platform also presents unique opportunities for Question Answering (QA) systems. QA systems represent a pivotal application of KGs, offering sophisticated means for information retrieval and accurate data extraction (*Bhavya et al., 2019*). In the realm of scholarly communication, the ORKG serves as a prime example of how KGs can underpin QA services to enhance academic inquiry and exploration. The ORKG, with its structured representation of scholarly data, particularly through comparison tables, provides a fertile ground for developing advanced QA systems like JarvisQA (*Jaradeh et al., 2020*).

6.6 JarvisQA and Beyond

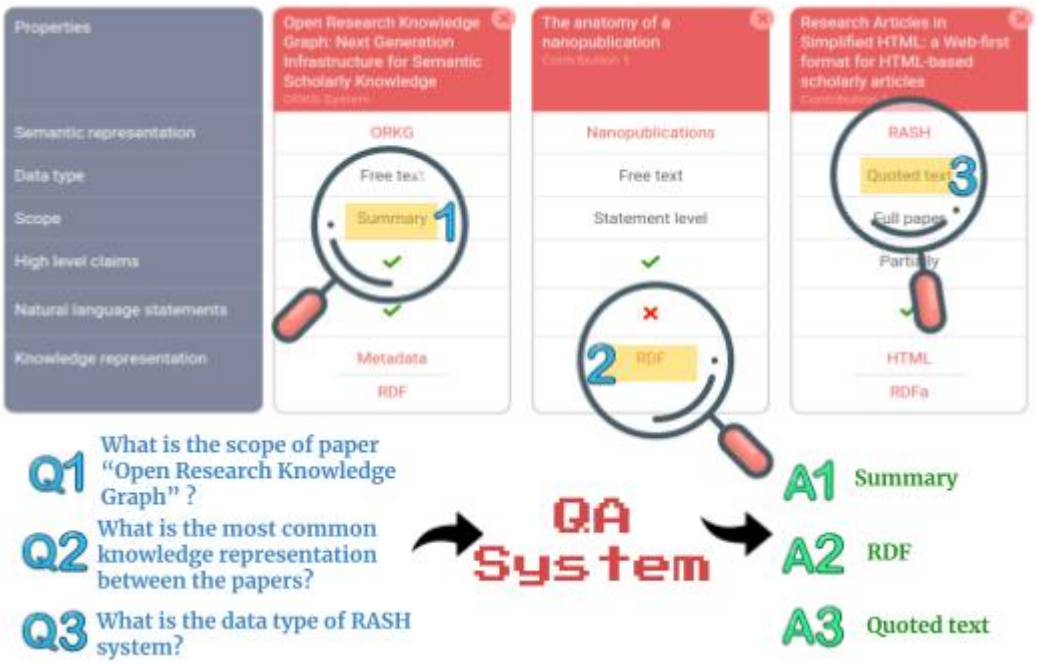

Figure 6.3 Depiction of the types of questions that JarvisQA can answer from the tabular views of structured data within a knowledge graph.

JarvisQA, specifically designed for the ORKG, employs Transformer models to understand the complex semantics of scholarly tables. Figure 6.3 highlights the

main types of questions that can be poised to JarvisQA from an ORKG comparison table. Despite its innovative approach, JarvisQA faces challenges inherent to scholarly KGs, including navigating diverse knowledge representations and adapting to the dynamic nature of scientific discourse. These issues underscore the need for continuous learning and sophisticated analysis techniques.

In response to the limitations of existing QA systems (*Singh et al., 2019*) and the absence of scholarly domain benchmarks, the SciQA benchmark (*Auer et al., 2023*) was developed. It features a mix of manually and automatically generated questions, tailored for the scholarly communication domain, covering a wide variety of subjects and incorporating SPARQL for queries. The vision for SciQA is not a one and done case, rather it serves as a stepping stone for the community to build upon and keep expanding and augmenting the benchmark's content. Thus, the benchmark was part of the open competitions at the 22nd internal Semantic Web Conference (ESWC) 2023 at the Scholarly Question Answering over Linked Data (QALD) Challenge[10].

6.7 LLMs in Scholarly QA

In the context of QA, LLMs like GPT-4[11] and PaLM2[12] present a promising solution to the complexities of scholarly QA. Fine-tuned on domain-specific datasets such as SciQA, these models can bridge the gap between the structured knowledge of KGs and the nuanced inquiries of researchers. By interpreting both formal graph structures and natural language, LLMs enhance the precision and depth of responses, facilitating a richer academic literature search and recommendation experience.

The integration of LLMs and QA systems within the ORKG framework represents a promising direction for advancing scholarly inquiry. While challenges persist, as highlighted by the JarvisQA and SciQA initiatives, the potential of LLMs to bridge the gap between structured knowledge and natural language queries offers a path towards more accessible and insightful academic exploration.

6.8 Conclusion and Outlook

The integration of NLP in the ORKG signifies a transformative step forward in the way scholarly knowledge is structured, accessed, and disseminated. By gradually transitioning to the adoption of LLMs, the ORKG has significantly enhanced its

[10] https://kgqa.github.io/scholarly-QALD-challenge/2023/
[11] https://openai.com/gpt-4
[12] https://ai.google/discover/palm2/

array of NLP services. These advancements not only streamline the process of integrating and analyzing vast amounts of scholarly data but also pave the way for more intuitive and efficient research workflows. As we continue to refine and expand these NLP capabilities, particularly in addressing the challenges highlighted by initiatives like the SciQA benchmark, the ORKG is set to offer unprecedented support to the academic community, fostering a more connected and accessible landscape of scientific knowledge. The journey of NLP within the ORKG, marked by continuous innovation and collaboration, promises to unlock new horizons in scholarly communication and research discovery.

To further this endeavor, we envision hackathons as a vibrant platform for collective creativity and technical prowess, where participants can explore the vast potential of NLP services within the ORKG framework. Hackathons, centered around the ORKG, present an invaluable opportunity for developers, researchers, and enthusiasts to collaborate on enhancing and expanding the NLP capabilities of the ORKG. By focusing on real-world challenges in scholarly communication and leveraging the open-source nature of the ORKG, these events can spur the development of novel NLP services that address the nuanced needs of the academic community. These facets could cover knowledge extraction, refinement of embedded scientific semantic representations, ORKG-knowledge-augmented context-based search, challenges based on prompt engineering strategies, and benchmarking of the inference capabilities of LLMs. These collaborative events not only contribute to the technological advancement of the ORKG, but also foster a community of practice that is deeply invested in the future of scholarly communication. By inviting community participation in the development of NLP services, we can ensure that the ORKG remains at the cutting edge of research technology, offering a dynamic and responsive tool that evolves to meet the changing landscape of academic research.

Acknowledgements

The authors thank Gollam Raby and reviewers for helpful comments.

References

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), p.9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), pp.5485-5551.

Hou, Y., et al. "Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

Beltagy, I, Kyle L, and Arman C. "SciBERT: A Pretrained Language Model for Scientific Text." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Ma, X., and Hovy E., "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016.

Heddes, J., et al. "The automatic detection of dataset names in scientific articles." *Data* 6.8 (2021): 84.

Schindler, D., et al. "Somesci-a 5 star open data gold standard knowledge graph of software mentions in scientific articles." *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021.

Brack A., D'Souza J., Hoppe A., Auer S., and Ewerth, R. (2020). Domain-Independent Extraction of Scientific Concepts from Research Articles. *In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12035*. Springer, Cham. https://doi.org/10.1007/978-3-030-45439-5_17

D'Souza J and Auer S (2022). Computer Science Named Entity Recognition in the Open Research Knowledge Graph. In: Tseng, YH., Katsurai, M., Nguyen, H.N. (eds) From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries. ICADL 2022. Lecture Notes in Computer Science, vol 13636. Springer, Cham. https://doi.org/10.1007/978-3-031-21756-2_3

D'Souza J (2024). Agriculture Named Entity Recognition—Towards FAIR, Reusable Scholarly Contributions in Agriculture. *Knowledge*, 4, no. 1: 1-26. https://doi.org/10.3390/knowledge4010001

Oghli, O.A, D'Souza J , and Auer S. "Clustering Semantic Predicates in the Open Research Knowledge Graph." *International Conference on Asian Digital Libraries*. Cham: Springer International Publishing, 2022.

Devlin J et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT*. 2019.

D'Souza J, Hrou M, and Auer S (2023). Evaluating Prompt-Based Question Answering for Object Prediction in the Open Research Knowledge Graph. In: Strauss, C., Amagasa, T., Kotsis, G., Tjoa, A.M., Khalil, I. (eds) *Database and Expert Systems Applications*. DEXA 2023. Lecture Notes in Computer Science, vol 14146. Springer, Cham. https://doi.org/10.1007/978-3-031-39847-6_40

D'Souza J and Auer S (2021). Pattern-Based Acquisition of Scientific Entities from Scholarly Article Titles. *Ke HR., Lee C.S., Sugiyama K. (eds) Towards Open and Trustworthy Digital Societies*. ICADL 2021. Lecture Notes in Computer Science, vol 13133. Springer, Cham. https://doi.org/10.1007/978-3-030-91669-5_31

Kabongo S, D'Souza J, & Auer S (2023). ORKG-Leaderboards: a systematic workflow for mining leaderboards as a knowledge graph. International Journal on Digital Libraries (2023). https://doi.org/10.1007/s00799-023-00366-1

Kabongo S, D'Souza J, & Auer S (2023 a). Zero-Shot Entailment of Leaderboards for Empirical AI Research. In: *ACM/IEEE Joint Conference on Digital Libraries*. JCDL 2023.Santa Fe, NM, USA, 2023, pp. 237-241. https://doi.org/10.1109/JCDL57899.2023.00042

Nechakhin V and D'Souza J (2023). Similar Papers Recommendation for Research Comparisons. In: *Joint Workshop Proceedings of the 5th International Workshop on A Semantic Data Space For Transport (Sem4Tra) and 2nd NLP4KGC: Natural Language Processing for Knowledge Graph Construction*. SEMANTiCS 2023. CEUR Workshop Proceedings, vol 3510. https://ceur-ws.org/Vol-3510/paper_nlp_5.pdf

Jiang M, D'Souza J, Auer S, and Downie S.J. (2020). Targeting precision: A hybrid scientific relation extraction pipeline for improved scholarly knowledge organization. *Proc Assoc Inf Sci Technol.* 57:e303. https://doi.org/10.1002/pra2.303

Oelen, A. and Auer, S. 2024. Leveraging Large Language Models for Realizing Truly Intelligent User Interfaces. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11--16, 2024, Honolulu, HI, USA (Honolulu, HI, USA, 2024). https://doi.org/10.1145/3613905.3650949

Bhavya K., Fan H., Nithin H., Suhail B., Zihua L., Lucile C., Matthias G., Anthony T.. 2019. Question answering via web extracted tables. In Proceedings of the Second International Workshop on Exploiting Artificial Intelligence Techniques for Data Management (aiDM '19). Association for Computing Machinery, New York, NY, USA, Article 4, 1–8. https://doi.org/10.1145/3329859.3329879

Jaradeh, M.Y., Stocker, M., Auer, S. (2020). Question Answering on Scholarly Knowledge Graphs. In: Hall, M., Merčun, T., Risse, T., Duchateau, F. (eds) Digital Libraries for Open Knowledge. TPDL 2020. Lecture Notes in Computer Science, vol 12246. Springer, Cham. https://doi.org/10.1007/978-3-030-54956-5_2

Singh K., Saleem, M., et al. (2019). QaldGen: Towards Microbenchmarking of Question Answering Systems over Knowledge Graphs. In: Ghidini, C., et al. The Semantic Web – ISWC 2019. Lecture Notes in Computer Science(), vol 11779. Springer, Cham https://doi.org/10.1007/978-3-030-30796-7_18

Auer, S., Barone, D.A.C., Bartz, C. et al. The SciQA Scientific Question Answering Benchmark for Scholarly Knowledge. Sci Rep 13, 7240 (2023). https://doi.org/10.1038/s41598-023-33607-z

7. Energy Systems Analysis as an ORKG Use Case

Oliver Karras[1], Felix Kullmann[2], Jan Göpfert[2], Patrick Kuckertz[2], Jann M. Weinand[2], Detlef Stolten[2,3], Sören Auer[1,4]

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

[2]*Forschungszentrum Jülich GmbH, Institute of Energy and Climate Research – Techno-economic Systems Analysis (IEK-3), 52425 Jülich, Germany*

[3]*RWTH Aachen University, Chair for Fuel Cells, Faculty of Mechanical Engineering, 52062 Aachen, Germany*

[4]*L3S Research Center, University of Hannover, 30167 Hannover, Germany*

7.1 Motivation

One of the greatest challenges of our time is curbing man-made climate change, which requires a massive reduction in greenhouse gas (GHG) emissions. GHGs are emitted through the combustion of fossil fuels, but also through industrial processes or food production. Fossil fuels must be replaced by alternative, renewable energy sources, and CO_2-neutral technologies must be further developed and implemented to achieve globally and nationally set climate targets. The expansion of renewable energies is a key factor in achieving these targets. However, further mitigation measures in the demand sectors are required as well. It is essential that not only the energy sector but also the industry, transport, and building sectors are considered.

Due to the interrelationships and interactions between these sectors, an evaluation of strategies to reduce GHG emissions only makes sense in the context of the entire energy system. Because of the complexity, computer models are used in such analyses and solved computationally. Various models are used to answer national energy industry and climate policy questions. These models differ, among other things, in terms of the model structure, the spatial and temporal resolution, and the observation horizon. Depending on the level of detail, different energy system framework data is required. This includes, for example, transport services, production volumes in industry, the total living space, or the energy requirements of the sectors. These framework data are in turn forecasted by other models. A

vast landscape of values is therefore available for the respective framework data, which directly influences the model results. It is crucial to know the used energy system framework data to better understand and classify an energy system's result. Depending on which values are selected, the boundary conditions in the energy, industry, buildings, and transport sectors change, as do the technologies and energy sources used. The selection of framework data therefore directly influences the optimal transformation path of the energy system. Depending on how great the influence of the input data is, a stronger focus must be placed on the data of the respective sector.

Our work aimed to investigate the influence of energy system framework data on GHG reduction strategies using model-based scenario calculations. With the help of existing studies on the German energy system, we identified extreme values for framework data and calculated scenarios based on these values. By comparing the results of the calculated scenarios, we determined the influence of the framework data and investigated the significance of energy system modeling.

For our use case, we used the ORKG to publish our review of 25 existing studies on the German energy system regarding their scenarios and energy system framework data. In this way, we provide a reusable and expandable database of this scientific knowledge and data for other energy system researchers. Furthermore, we ensure the transparency of the input data used for our model-based scenario calculations (*Giesen, 2020*). The following sections detailing our use case are based on a conference presentation and expand on the accompanying abstract (*Karras et al., 2024*).

7.2 Research Question

Given both the variety of forecast values in the framework data and their large impact on the solution of energy system analyses, we examine the sensitivity of the choice of framework data on individual sectors. In particular, we ask the following research question:

> Which sectors are particularly sensitive to changes in the framework data when designing future energy systems?

The answer to this question is of great importance for the selection of framework data. Depending on how great the influence of the input data is, a stronger focus must be placed on the data of the respective sector.

With the structural collection of framework datasets, we wanted to analyze the previously unexamined influence of framework data from the energy, industry,

building, and transport sectors on a future energy system and its transformation path (*Giesen, 2020*).

7.2.1 Comparison

In this use case, we organized scientific knowledge and data from 25 GHG reduction studies for Germany on their scenarios and energy system framework data. The studies were selected by *Robinius et al., 2020* and contrasted in terms of their reported energy supplies and installed capacities for various energy sources and their respective scenario goals. Based on the work by *Robinius et al., 2020*, we described all 25 studies as ORKG contributions regarding the corresponding scientific knowledge and data of interest to create and publish a corresponding ORKG comparison.

For the semantic description of the studies, we defined a set of comparison criteria and embedded them in ORKG templates to support the extraction and uniform representation of information (*Hussein et al., 2023*). Similar to data structures specified by the Shape Constraint Language, ORKG templates define the metadata profiles of ORKG contributions (*Knublauch, 2017*). ORKG templates integrate the Terminology Service (*Stroemert et al., 2023*) and thereby facilitate the usage of ontologies, like the Open Energy Ontology[13] (OEO) of the energy research domain (*Booshehri et al., 2021*). This allows the semantically distinct description of scientific knowledge and data and a consistent comparison across all studies under consideration. For example, we developed ORKG templates for the scenario goal[14] and the energy supply[15] and used the term definitions of the OEO to ensure the accurate interpretation of different energy sector types[16].

We created an ORKG comparison of all 25 GHG reduction studies for Germany to contrast their scenario and framework data (Figure 7.1). We published the comparison under a DOI to provide a sustainable, referenceable, citable, and standalone literature review (*Kullmann et al., 2021*). This ORKG comparison consists of typical metadata, including a title, publication month and year, authors, a description, a DOI, as well as the scientific knowledge and data of the 25 studies in tabular form. This tabular form is interactive, allowing navigation and filtering of its content for exploration and more detailed consideration. In addition, we enriched the ORKG comparison by creating and adding 18 visualizations based on the data

[13] https://openenergy-platform.org/ontology/
[14] https://orkg.org/template/R153118/
[15] https://orkg.org/template/R152170/
[16] https://terminology.tib.eu/ts/ontologies/oeo/terms?iri=http\%3A\%2F\%2Fopenenergy-platform.org\%2Fontology\%2Foeo\%2FOEO_00000367&subtab=graph

contained in it. In the following section Visualizations, we present and explain some of these visualizations in more detail. In contrast to the traditional dissemination of such a study reviews as text publications, ORKG comparisons enable their versioning and continuous (re)use, updates, and expansions by any user of the ORKG. The ORKG comparison serves as the central access point to provide the intended reusable and expandable database of this scientific knowledge and data for other energy system researchers, but also for every ORKG user.

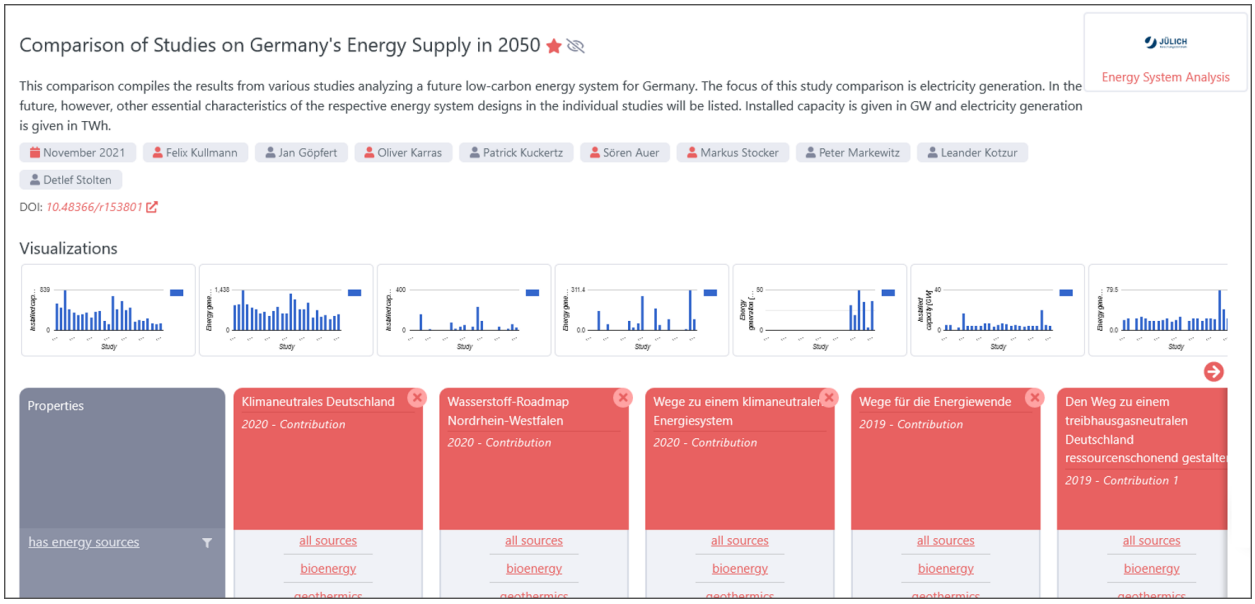

Figure 7.1 ORKG comparison of 25 scenarios from GHG reduction studies for Germany.

7.2.2 Visualizations

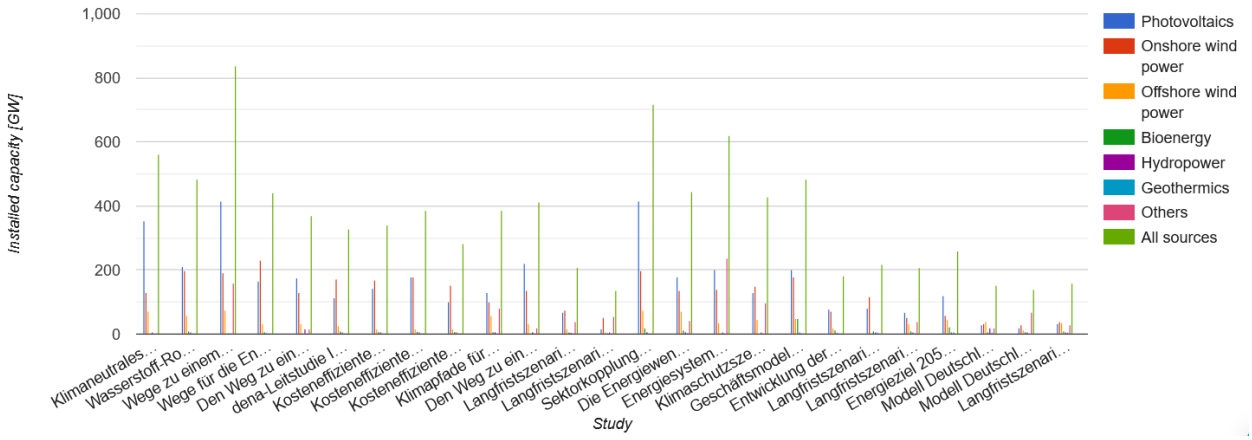

Figure 7.2 Reported installed capacity in gigawatts for all energy sources individually and aggregated in the 25 studies compared https://orkg.org/resource/R153804/preview

The ORKG supports the supplementation of ORKG comparisons by creating and adding visualizations based on the knowledge and data contained therein. These visualizations serve to present specific contents of the ORKG comparisons in a focused manner to facilitate access for users and enable them to gain a good understanding of the contents. Overall, we created 20 different visualizations with the

web frontend of the ORKG to provide custom overviews for the energy supplies and the installed capacities for all energy sources under consideration. Below, we present two examples of these visualizations.

Figure 7.2 presents an overview of all reported installed capacities for all energy sources individually and aggregated from the 25 studies compared. Compared to the interactive tabular form of the ORKG comparison, this visualization provides a quicker and simpler overview of all studies, such as identifying studies with extreme values in the framework data at a glance.

As can be seen in Figure 7.2, the overall view of all energy sources makes it difficult to take a closer look at individual values or to compare the values of individual energy sources. For this reason, we also created visualizations for the individual energy sources, such as Figure 7.3, which shows the reported energy supplies for the energy source onshore wind power in the 25 studies compared.

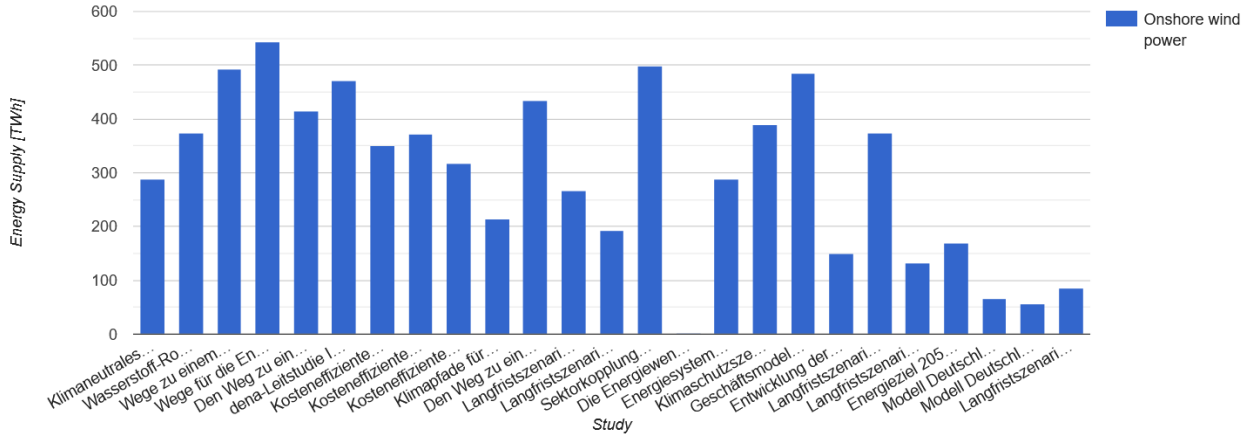

Figure 7.3 Reported energy supply in terawatt hours for the energy source onshore wind power in the 25 studies compared https://orkg.org/resource/R153807/preview

7.3 Conclusion and Outlook

7.3.1 Conclusion

Energy systems analysis uses optimization and simulation models as well as statistical analyses to investigate relationships and interactions between and within individual energy sectors. General developments such as population and settlement development, transport performance, or industrial demand are essential framework data that energy system analyses require as input data. Based on the collected framework data, it was found that the data and assumptions used to calculate the forecasts were not always clearly identified or even made publicly available (*Robinius et al., 2020*). Furthermore, some framework data is given without comprehensible documentation of its calculation. At the same time, the selection

of framework data has a significant influence on the results of energy systems analyses and can therefore have a decisive impact on the scientific knowledge process and political recommendations for action.

Against this background, care was taken to store the collected framework data in clearly defined structures within the ORKG and to annotate them unambiguously utilizing the OEO. In addition to the framework data collected in the ORKG, further data was used as input for the energy system model (e.g. technology costs, market entry points, etc.). These were not changed during the assessment of the impact of framework data on energy system design. The framework data from the 25 studies examined were assigned to 15 clearly defined parameters, and their values were compared in tabular form in an ORKG comparison. To further facilitate the analysis of the data, the ORKG user interface was then used to prepare subsets of the data in 18 visualizations. Based on this data preparation, the data analysis and selection could be carried out methodically. While deviating and contradictory assumptions could be quickly recognized, questioned, and revised, those that were supported by many studies could be considered relatively robust and incorporated into the input data set to be used for the study. In this way, three data sets were derived for each of the parameters, in which minimum, average, and maximum development scenarios were mapped. Their respective effects on the design of energy systems were analyzed in a series of model calculations (*Giesen, 2020*). The answer to the research question can be summarized in simplified terms as follows:

> Using those three distinct data sets (minimum, maximum, average) as input for the ETHOS.NESTOR model (Kullmann.2022) to generate synthetic energy system design futures, it could be determined that the assumptions regarding the input data used in the industrial sector had the greatest influence on the design of the energy system. Depending on which framework data set was used, the total energy generation capacity of the energy system varied between 120 GW and 830 GW (see Figure 7.2). This significant difference was also reflected in the conversion costs.

Overall, the consistent use of the ORKG features contributed significantly to the quality and efficiency of the research process. The preparation and labeling of the collected and used framework data contribute to the transparency and traceability of the study. The independent publication of the data under a permanently referenced DOI enables its sustainable reuse in future energy systems analyses and expandability by other research groups.

7.3.2 Outlook

Transparency is particularly important in research on energy systems, as results guide decision makers and influence public debate (*Pfenninger et al., 2017*). However, energy systems research lags behind other fields in this respect (*Pfenninger et al., 2017*). The use of complex models and a large amount of heterogeneous data makes research data management in energy research particularly challenging (*Niesse et al., 2022*). The National Research Data Infrastructure (NFDI) consortia NFDI4Energy (*Niesse et al., 2022*) and NFDI4Ing (*Schmitt et al., 2020*) highlight the need for infrastructures and services to improve research data management in energy systems research.

Listing 1: SPARQL query for the competency question: "*What is the average energy supply for each energy source considered in 5-year intervals in Greenhouse Gas Reduction Scenarios for Germany?*" by *Auer et al., 2023*

```
SELECT ?range ?srcLabel AVG(?val) AS ?avgVal
WHERE {
 r:R153801  p:compareContribution  ?contrib.
 ?paper     p:hasContribution      ?contrib;
            p:hasPublicationYear   ?year.
 BIND(xsd:int(?year) AS ?y).
 VALUES(?range ?min ?max) {
  ("2001-2005" 2001 2005)
  ("2006-2010" 2006 2010)
  ("2011-2015" 2011 2015)
  ("2016-2020" 2016 2020)
 } FILTER(?min <= ?y && ?y <= ?max).
 ?contrib    p:hasEnergySources  ?energySrc.
 ?energySrc  rdfs:label          ?srcLabel;
             p:hasGeneration     ?energyGen.
 ?energyGen  p:hasValue          ?genVal.
 BIND(xsd:float(?genVal) AS ?val).
} ORDER BY ASC(?range)
```

With our use case, we demonstrate how researchers can use the ORKG infrastructure to organize scientific knowledge and data in the field of energy systems analysis. The ORKG supports FAIR research data management and open science by providing a platform to collaboratively curate scientific knowledge in a way that is both human-readable and machine-actionable. By providing features such as version histories and unique identifiers for comparisons, as well as integration with ontologies, ORKG facilitates the reuse and further extension of curated

knowledge. *Auer et al., 2023* have already reused our scenario comparison from studies on the transformation of the German energy system and exemplifies how structured, openly accessible data facilitates its reuse. They answer the question ``*What is the average energy supply for each energy source considered in 5-year intervals in Greenhouse Gas Reduction Scenarios for Germany?*'' by formulating a SPARQL query (see Listing 1) and executing it on the SPARQL endpoint of the ORKG. The results are visualized in Figure 7.4 and indicate a fourfold increase in the average energy supply from photovoltaics and onshore wind power between the period of 2006 -2010 to the period of 2016 -2020.

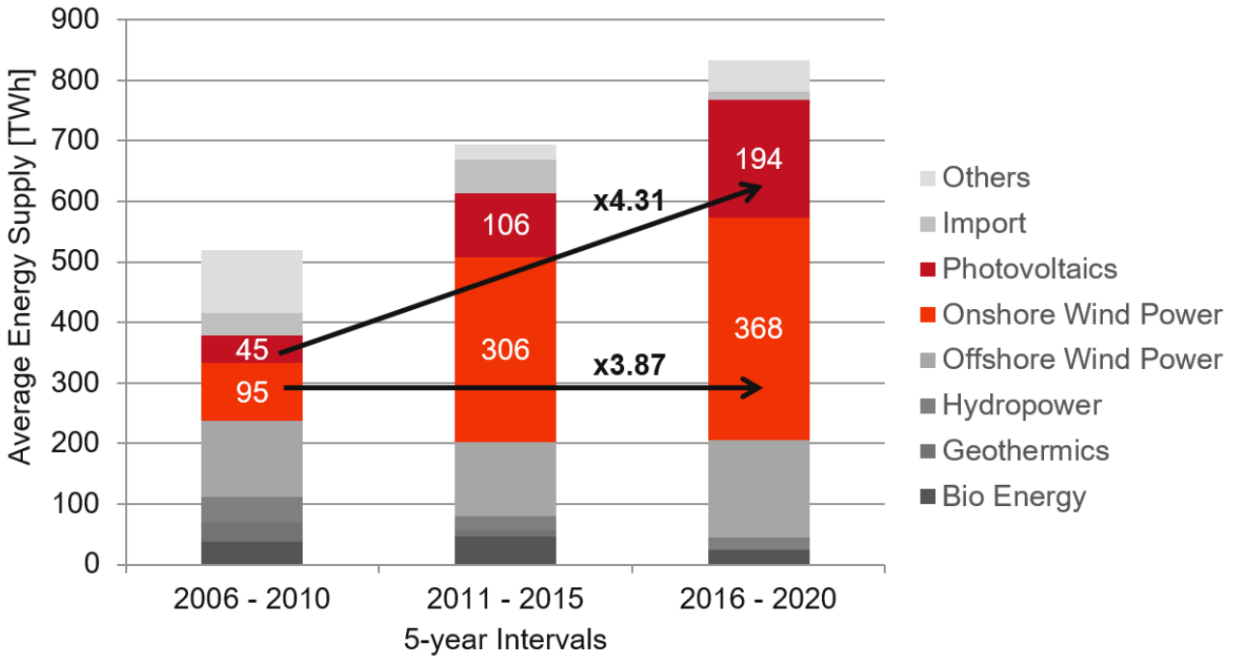

Figure 7.4 Visualized results from the SPARQL query by Auer et al., 2023

We hope to encourage researchers in the field of energy systems analysis to use the ORKG to make their research more transparent, to build on previously curated data and knowledge, and to make their own more reusable. We further support this goal by establishing an ORKG observatory on energy system research[17] that lists all curated content (publications, datasets, software, comparisons, and visualizations) related to energy systems analysis in one place. Given these measures and results, we continue to work purposefully on positioning the ORKG as a promising infrastructure for the sustainable organization of FAIR scientific knowledge and data in energy system research.

[17] https://orkg.org/observatory/Energy_System_Research

Data availability statement

All data used are openly available in the Open Research Knowledge Graph (https://orkg.org/) and in particular in the ORKG observatory on Energy System Research (https://orkg.org/observatory/Energy_System_Research)

Acknowledgements

The authors thank the Federal Government, the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the NFDI4Ing, NFDI4DataScience, and NFDI4Energy consortia. This work was funded by the German Research Foundation (DFG) -project numbers 442146713, 460234259, 501865131, by the European Research Council for the project ScienceGRAPH (Grant agreement ID: 819536), by the TIB - Leibniz Information Centre for Science and Technology, and by the Helmholtz Association as part of the program “Energy System Design”.

References

Matteo Giesen. Einfluss von Szenario-Rahmendaten auf Treibausgasminderungsstrategien. Master thesis. TU Braunschweig, 2020.

Oliver Karras et al. Organizing Scientific Knowledge From Energy System Research Using the Open Research Knowledge Graph. Version 1.0. Jan. 2024. doi: 10.5281/zenodo.10560077.

Martin Robinius et al. WEGE FUR DIE ENERGIEWENDE Kostenef- ¨ fiziente und klimagerechte Transformationsstrategien für das deutsche Energiesystem bis zum Jahr 2050. Vol. 499. Schriften des Forschungszentrums Jülich Reihe Energie & Umwelt / Energy & Environment. Forschungszentrum Jülich GmbH Zentralbibliothek, Verlag, 2020, VIII, 141 S. isbn: 978- 3-95806-483-6. url: https://juser.fz-juelich.de/record/877960.

Hassan Hussein et al. “Increasing Reproducibility in Science by Interlinking Semantic Artifact Descriptions in a Knowledge Graph”. In: Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration. Springer. 2023. doi: 10.1007/978-981-99-8088-8_19.

Holger Knublauch and Dimitris Kontokostas. Shapes Constraint Language (SHACL). W3C Recommendation. W3C, 2017. url: https://www.w3. org/TR/2017/REC-shacl-20170720/.

Philip Strömert et al. “Towards a Versatile Terminology Service for Empowering FAIR Research Data: Enabling Ontology Discovery, Design, Curation, and Utilization Across Scientific Communities”. In: Knowledge Graphs: Semantics, Machine Learning, and Languages. IOS Press, 2023. doi: 10.3233/SSW230005.

Meisam Booshehri et al. “Introducing the Open Energy Ontology: Enhancing Data Interpretation and Interfacing in Energy Systems Analysis”. In: Energy and AI 5 (2021). doi: 10.1016/j.egyai.2021.100074.

Felix Kullmann et al. Comparison of Studies on Germany’s Energy Supply in 2050. Open Research Knowledge Graph. 2021. doi: 10.48366/R153801.

Felix Kullmann et al. “The Value of Recycling for Low-Carbon Energy Systems - A Case Study of Germany’s Energy Transition”. In: Energy 256.124660 (2022). doi: 10.1016/j.energy.2022.124660.

Stefan Pfenninger et al. “The Importance of Open Data and Software: Is Energy Research Lagging Behind?” In: Energy Policy 101 (2017). doi: 10.1016/j.enpol.2016.11.046.

Astrid Nieße et al. NFDI4Energy – National Research Data Infrastructure for the Interdisciplinary Energy System Research. 2022. doi: 10.5281/ zenodo.6772013.

Robert H. Schmitt et al. NFDI4Ing - The National Research Data Infrastructure for Engineering Sciences. 2020. doi: 10.5281/zenodo.4015201.

Sören Auer et al. “The SciQA Scientific Question Answering Benchmark for Scholarly Knowledge”. In: Nature Scientific Reports 13.7240 (2023). doi: 10.1038/s41598-023-33607-z.

8. Harnessing the potential of the ORKG for synthesis research in agroecology

Lauren D. Snyder[1], Ricardo Perez-Alvarez[2], Emily A. Martin[2]

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

[2]Institute of Animal Ecology and Systematics, Justus Liebig University of Gießen, Germany

8.1 Motivation

Agrobiodiversity plays a key role in supporting valuable ecosystem services such as pest suppression and crop productivity, which in turn provide economic and nutritional benefits to humans (*Snyder et al., 2020* and references therein). However, there is no one-size-fits-all approach for leveraging agrobiodiversity to promote ecosystem services. Rather, enhancing agrobiodiversity at local and landscape scales can produce positive, negative, and neutral outcomes (*Kleijn et al., 2019; Seufert and Ramankutty, 2017*). This context-dependency makes it challenging for researchers, farmers, and other practitioners to determine when and how to increase agrobiodiversity for optimal effect.

Synthesis research emerges as a pivotal tool in unravelling this complexity by providing a framework to identify patterns and processes across space and time (*Halpern et al., 2020*). For instance, meta-analysis - a common form of synthesis used in agroecological research - support cross disciplinary connections and play a critical role in driving, modifying, and resolving core questions to guide policy and practice *(Díaz et al., 2015; Dicks et al., 2014; Halpern et al., 2020*). Yet, despite its value, conducting synthesis research is increasingly challenging due to the increasing volume of scientific publications. This challenge is further compounded by the tedious nature of extracting information from unstructured narrative PDF articles, and the need to regularly update existing syntheses with new information as it becomes available.

The growing flood of scientific articles poses a formidable obstacle to staying up-to-date on the latest findings, and to identifying clear trends and patterns that have broad-scale applicability. This dilemma is exemplified in the field of agroecology, where close to 800 publications are produced annually (*Mason et al., 2021*). In the

absence of effective tools and standards for facilitating knowledge sharing, the proliferation of academic publications presents major challenges for reproducibility and the peer-review process, and ultimately leads to the loss of knowledge. Indeed, some estimates suggest that approximately 10% of research papers remain uncited after five years of publication, despite advances in the internet era that make it easier to find and cite relevant papers (*Van Noorden, 2017*).

A second key challenge is the way we formally communicate scientific research. While scientific knowledge expressed in articles is now largely pseudo-digitalized as PDF publications that can be easily shared electronically, their unstructured narrative text format is unintelligible to computers (i.e., not machine-reusable). This mode of communication represents a major limitation in our current approach to knowledge sharing, as it prevents us from taking full advantage of computer support tools like intelligent search, filter, or other processing functions that enable machine-supported knowledge organization and reuse *(Auer et al., 2020; Stocker et al., 2023*). Rather, data destined for reuse must be harvested manually by scientific experts, dramatically slowing down the research lifecycle. For example, conducting synthesis research in the form of a meta-analysis or systematic review requires manually extracting data from multiple publications and organising this information into a new database, a process that could entail three to six months of full-time work (*Li et al., 2023)*

Moreover, publishing synthesis research as static PDF articles makes updating existing syntheses virtually impossible (*Shackelford et al., 2021*). Instead, as research on a particular topic advances, a novel synthesis must be conducted, resulting in yet another publication and dataset (*Heberling et al., 2021; Culina et al., 2018; Feng et al., 2022*) for the scientific community to track. This perpetual reinvention of the wheel hinders progress and renders the relevance of syntheses fleeting in the wake of new findings. Ultimately, the inherent inefficiency of manually retrieving information from static PDF publications diminishes the utility and relevance of systematic reviews in informing policy and management decisions.

Given the growth, complexity, and societal relevance of ecological research, there is a critical need for tools and standards that facilitate sharing, synthesizing, and reproducing this knowledge for a range of stakeholders (*Dicks et al., 2014*). Here, we present a use case in the ORKG to evaluate how the platform can support these goals in the field of agroecology. Specifically, we describe our experience using the platform to create an ORKG comparison, a tabular summary of research contributions that helps researchers summarize the state-of-the-art around a particular research topic. Based on our experience, we share our vision for how the platform could help address some of the current challenges associated with transferring ecological knowledge and outline opportunities for continued development.

8.2 Research Question

For our ORKG use case, we compiled information from a set of peer-reviewed articles evaluating the yield effects of intercropping cereal crops with legumes—plants that make atmospheric nitrogen available to other plants. We selected this research question as it represents a timely topic within the field of agroecology and sits within the broader framework of developing agroecological solutions to meet global food security needs amidst growing constraints on the availability of arable land, a rising global human population, and increased food demands (*Pérez-Escamilla, 2017*).

Articles included in our comparison were selected using a Web of Science search, which is described in the ORKG comparison itself. Our goal was to test the capabilities of the ORKG platform in scoping articles related to our research question and to evaluate its usability for this purpose, rather than to conduct a robust synthesis on the topic.

8.3 ORKG Comparison

The ORKG platform provides an online interface that allows researchers to digitalise scientific knowledge, ensuring it is readable and usable by humans and computers (i.e., machine reusable) (Open Research Knowledge Graph, 2023). By organising research contributions from scientific articles alongside one another in tabular format, an ORKG comparison helps researchers examine scientific results across publications to gain a quick overview of a specific research question (*Auer et al., 2020; Oelen et al., 2020*). This approach to scientific knowledge curation results in highly structured descriptions of research contributions published in scientific articles, such as the research problem, methodology, and results (*Stocker et al., 2023*).

Implementation

To create our ORKG comparison on legume-cereal intercrops, we first added the meta-data associated with each article identified in the Web of Science search to the ORKG platform using the ORKG “Add paper” function (Figure 8.1). This was a straightforward process that involved using a simple title or DOI search to locate the desired article.

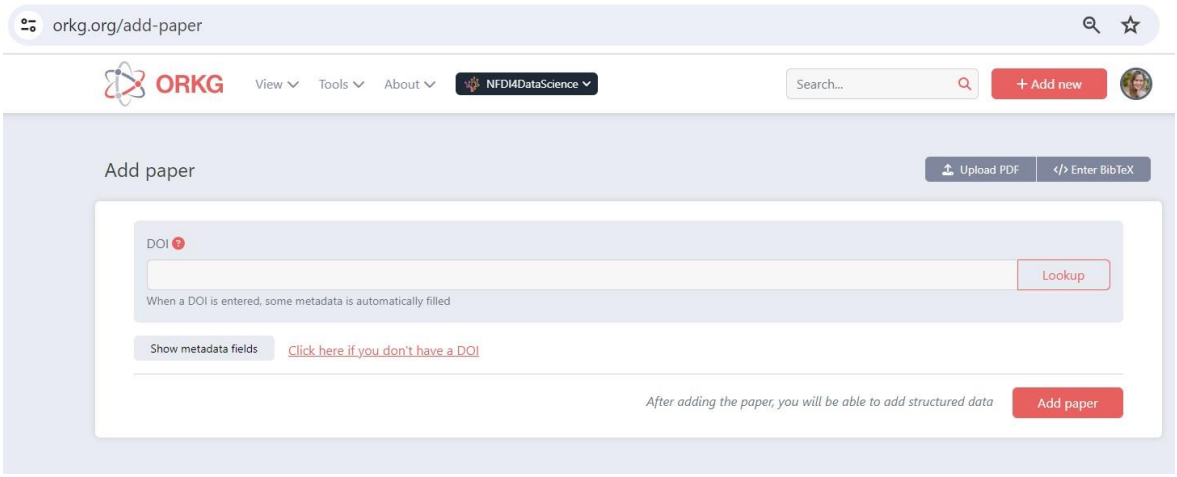

Figure 8.1 ORKG Add paper function.

With the relevant articles entered into the platform, we began building our ORKG comparison using the ORKG contribution editor (Figure 8.2). This tool allows users to identify a paper that has already been added to the ORKG using a title or DOI look-up function, or to add a new paper entry. After using the contribution editor to select a particular article, we then entered additional information about specific research contributions associated with that paper. Key research contributions included information about the study location, experimental methods, experimental control and treatment, quantitative yield measurements, etc. We repeated this process for each paper in our comparison.

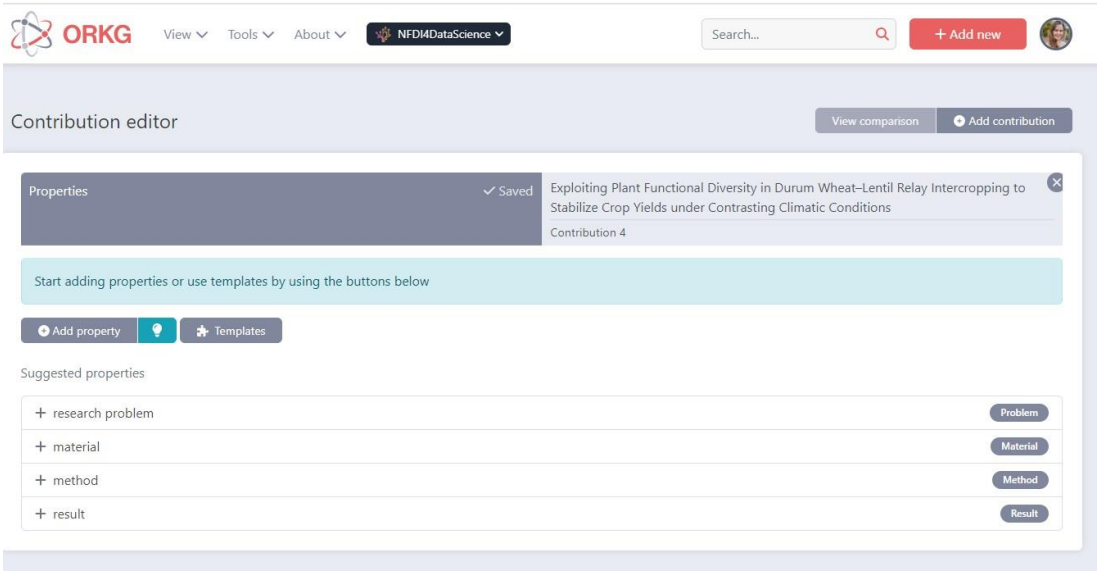

Figure 8.2 The ORKG contribution editor suggests potential properties (e.g., research problem, methods, etc.) to help researchers highlight key scientific findings.

The process of creating an ORKG comparison is analogous to compiling and coding primary data for a meta-analysis or systematic review. However, rather than organising this information in an Excel or CSV file, data is organised directly in the ORKG platform. As we built our comparison, we had control over how we structured the data and information associated with each research contribution. This process can also be guided by ORKG templates, which are similar to a fill-in-the-blank form that guides users in developing a structured and semantic description of research contributions by suggesting the kinds of information (referred to as *properties* in the ORKG) that should be provided to adequately describe a specific kind of research contribution (Figure 8.3). For example, a template describing a linear regression could include properties related to the input and output datasets, the independent and dependent variables, and an output figure. Templates also make explicit which category a property falls into (referred to as a *property type* in the ORKG), such as text, decimal, URL, table, etc.

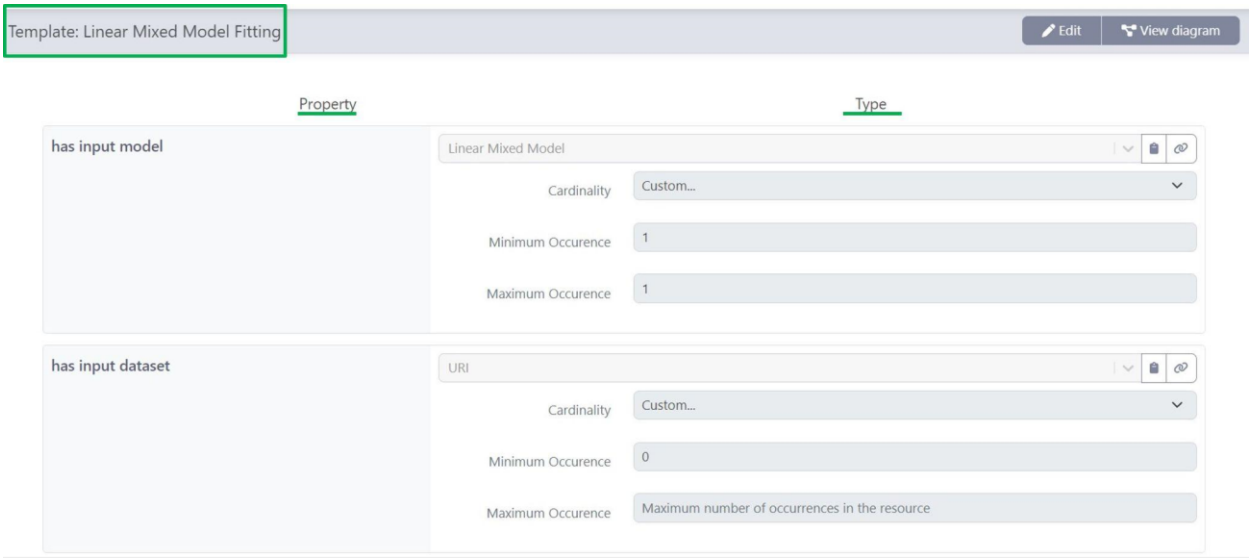

Figure 8.3 Example of an ORKG template that guides the user through providing information related to a linear mixed effects model, for example input model and input dataset.

When we created this comparison, the available ORKG templates did not provide a suitable structure for our agroecology research contributions. Given the complexity of the agroecological data we wanted to describe, we had to develop our own approach to modelling the research contributions. While building our featured comparison, L. Snyder was participating in an ORKG curation grant, which provided valuable training and guidance on the best practices for generating ORKG comparisons. Without these resources, we could envision a lack of suitable templates as a barrier to new ORKG users, especially those who are unfamiliar with

semantic modelling. At the same time, the flexibility the ORKG offers in terms of modelling and structuring research contributions makes it adaptable across disciplines and allows researchers to tailor comparisons to suit their specific research needs.

Properties	Intercropping corn with soybean, lupin and forages: yield component responses *Soybean Intercrop—Corn Grain Yield, Macdonald 1993 - 2000* Research Contribution 1	Intercropping corn with soybean, lupin and forages: yield component responses *Lupin Intercrop—Corn Grain Yield, Macdonald 1993 - 2000* Research Contribution 2	Intercropping corn with soybean, lupin and forages: yield component responses *Red Clover Intercrop—Corn Grain Yield, Macdonald 1993 - 2000* Research Contribution 3
research problem	Effect of legume intercrops on cereal grain yield	Effect of legume intercrops on cereal grain yield	Effect of legume intercrops on cereal grain yield
cropping system	Annual	Annual	Annual
study year	1993	1993	1993
experimental design	Randomized complete block design	Randomized complete block design	Randomized complete block design
experimental setup	field experiment	field experiment	field experiment
cereal crop	maize	maize	maize
control	Corn monoculture	Corn monoculture	Corn monoculture
research intervention	Soybean and corn intercrop	Lupin and corn intercrop	Red clover and corn intercrop
has effect	Neutral	Neutral	Neutral

Figure 8.4 Partial view of our ORKG comparison on cereal-legume intercrops.

Our interactive comparison (Figure 8.4) can be viewed in full form on the ORKG platform: https://orkg.org/comparison/R655553/. This comparison exists in an open-access environment that allows other experts in the field to expand upon the search criteria we used to generate the original comparison and incorporate additional studies. This dynamic approach to scientific knowledge curation provides a comprehensive, living resource for the agroecology community that can be regularly updated with new and relevant research findings; we found this to be one of the most useful features of the ORKG platform.

8.4 Visualizations

The ORKG platform also offers a visualization tool that allows users to visualize content from a comparison in the form of a table or bar, column, line, or scatter chart. Once we coded the information into the ORKG, creating the visualization was relatively straightforward and took a matter of minutes. The ORKG platform guided us through the process of creating the visualization, which suggests that

users with even a low-level of scientific expertise could create meaningful visualizations from existing ORKG comparisons. Moreover, multiple kinds of visualizations can be created without the need to recode the data. In other words, depending on the specific kind of data included in a comparison, ORKG users could quickly create a table and bar chart to visualize the same data set.

Importantly, the original data and information used to create this visualization can be readily extracted and exported, for example as a CSV file or directly to a programming language like R, once it is integrated into the ORKG platform. Subsequently, it can be reused on other platforms with enhanced visualization capabilities (e.g., R or Python), enabling the creation of custom-made plots that provide a more nuanced approach to visualizing the data and trends, and allow for more complex analyses of the compiled data.

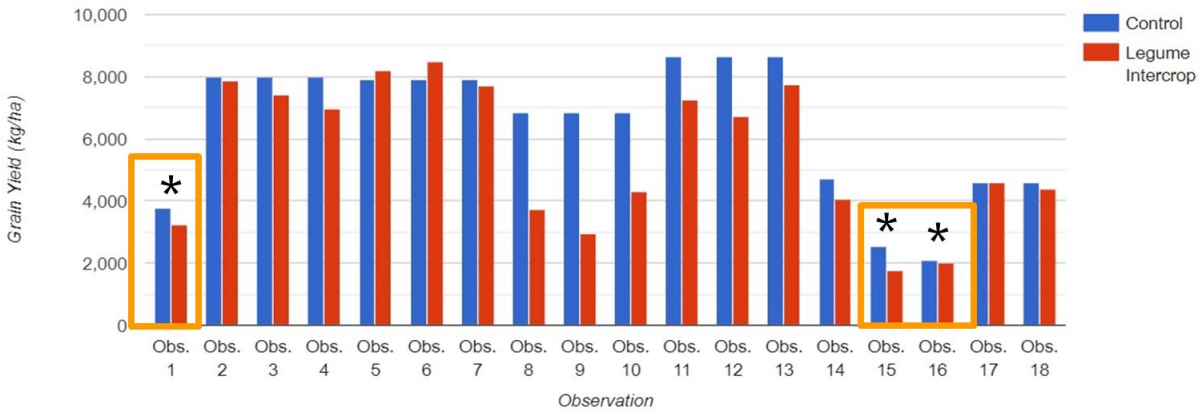

Figure 8.5 Visualisation created using content from our agroecology comparison. The y-axis indicates grain yields in kg/ha reported in the original research articles. The x-axis shows the individual observations (i.e., research contributions) included in the comparison; each paper can include multiple observations. Blue bars represent cereal grain yield from cereal monocultures (controls). In red are the cereal grain yields from the legume intercropping systems (treatments). Significant differences between the yield of the controls and treatments reported in the original article are represented with a star and orange box.

Figure 8.5 demonstrates how the platform allows researchers to rapidly visualise trends across studies, providing an important overview of the state of a research field. While ORKG visualisations do not bring statistical rigour to their summaries, they allow researchers to quickly obtain a superficial sense of the papers available for scoping. This is particularly useful for systematic mapping in preparation for a review (*James et al., 2016*). In the specific example above, the visualisation allows

researchers to look more closely at the specific instances where intercrops underperformed, potentially leading to more targeted research questions focused on improving the performance of legume intercropping systems.

View visualization

Summary Table of Literature Search

This table summarizes the study location, cereal crop of interest, data analysis method, and number of replicates associated with each contribution (observation) included in the comparison.

05 December 2023 Lauren Snyder

	Contribution	study location	cereal crop	has effect	statistical_methods	is about/crop yield control/has quality/crop grain yield/number of replicates
1	Subclover Intercrop—Durum Wheat Yield(R229191)	Italy	Triticum durum	Negative	analysis of variance	4
2	Soybean Intercrop—Corn Grain Yield, Macdonald 1993(R229235)	E. Lods Agronomy Research Centre, Canada	maize	Neutral	analysis of variance	4
3	Lupin Intercrop—Corn Grain Yield, Macdonald 1993(R229235)	E. Lods Agronomy Research Centre, Canada	maize	Neutral	analysis of variance	4
4	Red Clover Intercrop—Corn Grain Yield, Macdonald 1993(R229235)	E. Lods Agronomy Research Centre, Canada	maize	Neutral	analysis of variance	4
5	Soybean Intercrop—Corn Grain Yield, L'Assomption 1993(R229235)	L'Assomption Field Station of Agriculture & Agri-foods, Canada	maize	Neutral	analysis of variance	4
6	Lupin Intercrop—Corn Grain Yield, L'Assomption 1993(R229235)	L'Assomption Field Station of Agriculture & Agri-foods, Canada	maize	Neutral	analysis of variance	4
7	Red Clover Intercrop—Corn Grain Yield, L'Assomption 1993(R229235)	L'Assomption Field Station of Agriculture & Agri-foods, Canada	maize	Neutral	analysis of variance	4
8	Soybean Intercrop—Corn Grain Yield, Macdonald 1994(R229235)	E. Lods Agronomy Research Centre, Canada	maize	Neutral	analysis of variance	4
9	Lupin Intercrop—Corn Grain Yield, Macdonald 1994(R229235)	E. Lods Agronomy Research Centre, Canada	maize	Neutral	analysis of variance	4
10	Red Clover Intercrop—Corn Grain Yield, Macdonald 1994(R229235)	E. Lods Agronomy Research Centre, Canada	maize	Neutral	analysis of variance	4
11	Soybean Intercrop—Corn Grain Yield, L'Assomption 1994(R229235)	L'Assomption Field Station of Agriculture & Agri-foods, Canada	maize	Neutral	analysis of variance	4
12	Lupin Intercrop—Corn Grain Yield, L'Assomption 1994(R229235)	L'Assomption Field Station of Agriculture & Agri-foods, Canada	maize	Neutral	analysis of variance	4
13	Red Clover Intercrop—Corn Grain Yield, L'Assomption 1994(R229235)	L'Assomption Field Station of Agriculture & Agri-foods, Canada	maize	Neutral	analysis of variance	4
14	2019(R237944)	Italy	Triticum durum	Neutral	Generalized linear mixed model	4
15	2020(R237944)	Italy	Triticum durum	Negative	Generalized linear mixed model	4
16	2021(R237944)	Italy	Triticum durum	Negative	Generalized linear mixed model	4
17	Lucerne Intercrop(R237990)	Sweden	Avena sativa	Neutral	analysis of variance	4
18	Red Clover Intercrop(R237990)	Sweden	Avena sativa	Neutral	analysis of variance	4

Export Edit visualization Close

Table 8.1: *Summary table of the retrieved literature, including study location, cereal crop of interest, data analysis method, and number of replicates associated with each contribution (observation) included in the comparison. As in classical meta-analyses and systematic reviews, clearly documenting the literature search process underlying an ORKG comparison is a foundational step in creating a comparison.*

8.5 Conclusions

The ORKG provides researchers with a powerful platform in which to visualise trends and identify knowledge gaps that could be addressed with future research. As with a traditional meta-analysis or systematic review, extracting and organising the data (i.e., research contributions) in our ORKG comparison was a time consuming process. Learning how to develop the models/templates needed to structure the data also required an upfront time investment. Developing templates to structure data and information related to common ecology methods and analyses are key to addressing this issue, and we expect this hurdle to lessen rapidly as appropriate templates become available for users in the field.

Because the scientific data we included in the comparison was published in PDF format, we had to manually extract and add it to the ORKG platform. This approach to populating the ORKG with scientific knowledge comes at a high temporal cost and is prone to error. To scale the use of the ORKG across the field of agroecology and other disciplines, we foresee automating knowledge extraction from articles as an important objective for the ORKG. This could even be in the form of a semi-

automated process in which experts are needed to manually review and improve automatically extracted knowledge to ensure richness, quality, and accuracy.

The ORKG is moving in this direction with new tools like SciKGTeX and born-reusable scientific knowledge that enable researchers to produce scientific knowledge in a machine-reusable format from the outset of knowledge production. Widespread implementation of these approaches would ensure new research findings could automatically be harvested in machine-reusable form by the ORKG, or other knowledge bases, every time a paper is published, thereby eliminating the need for laborious post-publication manual or semi-automated knowledge extraction and making this knowledge available for immediate reuse by researchers anywhere in the world. We envision such capabilities could dramatically reduce the high time costs currently associated with synthesis research, for example by facilitating the automatic integration of new research into existing ORKG comparisons resulting in a continually updated living resource that informs research, policy, and management decisions.

Such approaches would also enable easy access to data and information that is hard to extract when represented in the format of a figure. When creating our agroecology comparison, our objective was to report the data exactly as it was presented in the paper, so we were limited to including data that was presented in narrative text or tabular format. Extracting data from a figure would have required the use of a data extraction tool, which is time consuming and prone to error, so data presented in this format was not included in our comparison. This limitation further highlights the importance of publishing scientific data in a machine-reusable format from the outset of knowledge production to ensure that data and information underlying figures is transparent and easily available for reuse. In addition to ORKG-specific initiatives to promote machine-reusable scientific knowledge, the platform could also bolster and facilitate other initiatives moving in this direction (e.g. *Nüst and Eglen, 2021*).

While efficiently populating the ORKG with scientific knowledge is a current challenge, one of the most powerful aspects of the ORKG is the ease with which data and information can be exported and reused once it is in the platform. Given the ease of accessing data once it is in machine-reusable format in the ORKG platform, it is easy to envision a well-populated ORKG drastically accelerating the process of conducting synthesis research.

The ability to efficiently compile and organize data and information in the ORKG will rely on the use of standardized language as authors code their data into the platform. This could be a challenge for agroecologists, as the lack of a cohesive

vocabulary to articulate methods and results often impedes effective communication and collaboration (*Herrando-Pérez et al., 2014*). Currently, terms used in ORKG templates do not necessarily map to formalized ontologies, so ensuring the use of consistent language within a scientific discipline remains a challenge. As It is likely not the role of the ORKG to act as an ontology provider, to fully leverage the potential of the ORKG platform, a key goal for the ecological community is to develop an agreed upon ontology that resolves this linguistic gap. We foresee this as one of the biggest hurdles to synthesis research and encourage continued discussion around how to address it.

Creating additional agroecology use cases in the ORKG will be critical to promoting the broadscale adoption of the platform within the ecology community. As with other FAIR data initiatives (e.g., making field data and programming code available upon publication), training and outreach efforts to advertise the benefits of the ORKG platform and normalize its use for ecological research are important next steps.

References

Li T, Higgins JPT, Deeks JJ (editors). Chapter 5: Collecting data. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). *Cochrane Handbook for Systematic Reviews of Interventions* version 6.4 (updated August 2023). Cochrane, 2023. Available from www.training.cochrane.org/handbook.

Auer, S., Oelen, A., Haris, M., Stocker, M., D'Souza, J., Farfar, K.E., Vogt, L., Prinz, M., Wiens, V., Jaradeh, M.Y., 2020. Improving Access to Scientific Literature with Knowledge Graphs. Bibl. Forsch. Prax. 44, 516–529. https://doi.org/10.1515/bfp-2020-2042

Díaz, S., Demissew, S., Carabias, J., Joly, C., Lonsdale, M., Ash, N., Larigauderie, A., Adhikari, J.R., Arico, S., Báldi, A., Bartuska, A., Baste, I.A., Bilgin, A., Brondizio, E., Chan, K.M., Figueroa, V.E., Duraiappah, A., Fischer, M., Hill, R., Koetz, T., Leadley, P., Lyver, P., Mace, G.M., Martin-Lopez, B., Okumura, M., Pacheco, D., Pascual, U., Pérez, E.S., Reyers, B., Roth, E., Saito, O., Scholes, R.J., Sharma, N., Tallis, H., Thaman, R., Watson, R., Yahara, T., Hamid, Z.A., Akosim, C., Al-Hafedh, Y., Allahverdiyev, R., Amankwah, E., Asah, S.T., Asfaw, Z., Bartus, G., Brooks, L.A., Caillaux, J., Dalle, G., Darnaedi, D., Driver, A., Erpul, G., Escobar-Eyzaguirre, P., Failler, P., Fouda, A.M.M., Fu, B., Gundimeda, H., Hashimoto, S., Homer, F., Lavorel, S., Lichtenstein, G., Mala, W.A., Mandivenyi, W., Matczak, P., Mbizvo, C., Mehrdadi, M., Metzger, J.P., Mikissa, J.B., Moller, H., Mooney, H.A., Mumby, P., Nagendra, H., Nesshover, C., Oteng-Yeboah, A.A., Pataki, G., Roué, M., Rubis, J., Schultz, M., Smith, P., Sumaila, R., Takeuchi, K., Thomas, S., Verma, M., Yeo-Chang, Y., Zlatanova, D., 2015. The IPBES Conceptual Framework — connecting nature and people. Curr. Opin. Environ. Sustain. 14, 1–16. https://doi.org/10.1016/j.cosust.2014.11.002

Dicks, L.V., Walsh, J.C., Sutherland, W.J., 2014. Organising evidence for environmental management decisions: a '4S' hierarchy. Trends Ecol. Evol. 29, 607–613. https://doi.org/10.1016/j.tree.2014.09.004

Halpern, B.S., Berlow, E., Williams, R., Borer, E.T., Davis, F.W., Dobson, A., Enquist, B.J., Froehlich, H.E., Gerber, L.R., Lortie, C.J., O'connor, M.I., Regan, H., Vázquez, D.P., Willard, G., 2020. Ecological Synthesis and Its Role in Advancing Knowledge. BioScience biaa105. https://doi.org/10.1093/biosci/biaa105

Herrando-Pérez, S., Brook, B.W., Bradshaw, C.J.A., 2014. Ecology Needs a Convention of Nomenclature. BioScience 64, 311–321. https://doi.org/10.1093/biosci/biu013

James, K.L., Randall, N.P., Haddaway, N.R., 2016. A methodology for systematic mapping in environmental sciences. Environ. Evid. 5, 7. https://doi.org/10.1186/s13750-016-0059-6

Kleijn, D., Bommarco, R., Fijen, T.P.M., Garibaldi, L.A., Potts, S., Putten, W.H. van der, 2019. Ecological Intensification: Bridging the Gap between Science and Practice. Trends Ecol. Evol. 34, 154–166. https://doi.org/10.1016/j.tree.2018.11.002

Li, T., Higgins, J.P., Deeks, J.J. (Eds.), 2023. Chapter 5: Collecting data, in: Cochrane Handbook for Systematic Reviews of Interventions.

Mason, R.E., White, A., Bucini, G., Anderzén, J., Méndez, V.E., Merrill, S.C., 2021. The evolving landscape of agroecological research. Agroecol. Sustain. Food Syst. 45, 551–591. https://doi.org/10.1080/21683565.2020.1845275

Nüst, D., Eglen, S.J., 2021. CODECHECK: an Open Science initiative for the independent execution of computations underlying research articles during peer review to improve reproducibility. https://doi.org/10.12688/f1000research.51738.2

Oelen, A., Jaradeh, M.Y., Farfar, K.E., Stocker, M., Auer, S., 2019. Comparing Research Contributions in a Scholarly Knowledge Graph. Presented at the Proceedings of the Third International Workshop on Capturing Scientific Knowledge, co-located with the 10th International Conference on Knowledge Capture (K-CAP 2019), Marina del Rey, California.

Open Research Knowledge Graph. Retrieved April 4, 2023, from https://orkg.org/

Pérez-Escamilla, R., 2017. Food Security and the 2015–2030 Sustainable Development Goals: From Human to Planetary Health. Curr. Dev. Nutr. 1, e000513. https://doi.org/10.3945/cdn.117.000513

Seufert, V., Ramankutty, N., 2017. Many shades of gray—The context-dependent performance of organic agriculture. Sci. Adv. 3, e1602638. https://doi.org/10.1126/sciadv.1602638

Snyder, L.D., Gómez, M.I., Power, A.G., 2020. Crop Varietal Mixtures as a Strategy to Support Insect Pest Control, Yield, Economic, and Nutritional Services. Front. Sustain. Food Syst. 4.

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. FAIR Connect 1, 19–21. https://doi.org/10.3233/FC-221513

Van Noorden, R., 2017. The science that's never been cited. Nature 552, 162–164. https://doi.org/10.1038/d41586-017-08404-0

9. Knowledge synthesis in Invasion Biology: from a prototype to community-designed templates

Maud Bernard-Verdier[1,2], Kamel Fadel[3], Tina Heger[1,4], Jonathan M. Jeschke[1,2], Markus Stocker[3], Lars Vogt[3]

[1] *Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany*
[2] *Freie Universität Berlin, Institute of Biology, Berlin, Germany*
[3] *TIB Leibniz Information Centre for Science and Technology, Hannover, Germany*
[4] *Technical University of Munich, Restoration Ecology, Freising, Germany*

9.1 The prototype with Hi Knowledge data

Motivation

Biological invasions, i.e. the spread of organisms outside their native distributional range as a consequence of human activities, are one of the leading causes of global biodiversity decline. Invasion biology is a subfield of ecological research which has shown an exponential increase in publications in the past 25 years. The Hi Knowledge initiative[18], which was started around 2010 by Jonathan Jeschke and Tina Heger, aims to tackle this by synthesizing and visualizing knowledge in the field of invasion biology and beyond. In a collaborative book by Jeschke & Heger published in 2018, they reviewed the evidence for a set of 12 major hypotheses in invasion biology theory, which predict mechanisms favoring the introduction, spread and impact of species outside their native range. This resulted in a curated dataset assembling information from over 1000 articles testing at least one of these hypotheses.

The collaboration between Hi Knowledge and the ORKG started in Fall 2019. It was quickly clear that the Hi Knowledge dataset could demonstrate the capabilities of ORKG as a service. Ingesting community data into the ORKG, and using ORKG services such as Comparisons to demonstrate what is possible, was an invaluable activity, and with Hi Knowledge the first of this kind.

The SARS-CoV-2 pandemic had postponed more concrete activities towards these aims. However, they were resumed in 2021 in the context of a Master thesis

[18] https://hi-knowledge.org/

by Kamel Fadel (*Fadel, 2021*). In this work, we were able to ingest the Hi Knowledge data into ORKG, build an ORKG Observatory[19] for the community, create ORKG Comparisons[20] for the 10 individual Hi Knowledge hypotheses, and leverage the ORKG integrations with Jupyter to test whether computing environments / dashboards could support the production of tailored visualizations for the community. The Hi Knowledge network of hypotheses was a good objective for our ORKG prototype.

For this prototype with Hi Knowledge data, the research questions were thus of technical nature. Specifically, the work was motivated by the question whether Scientific Knowledge Graphs and ORKG in particular can be exploited in data science and with what technical approaches.

Approach and results

The activity consisted of the following key tasks: (1) Hi Knowledge data ingestion into the ORKG; (2) Create ORKG Comparisons; (3) Data science using the ingested data.

Hi Knowledge data ingestion. The starting point is data that was extracted from articles and published on the Hi Knowledge website[21] in separate files, one file per hypothesis. This data relates to 10 of the 12 hypotheses addressed in the 2018 book, as data on 2 hypotheses were structured in a different way. Both article metadata and extracted essential data as structured content were ingested for these 10 hypotheses, e.g.:

- Article's stance towards the hypothesis: Indicating whether it supports, is undecided, or questions the hypothesis
- The investigated taxa in the article, e.g., plants, birds, mammals, etc.
- Number of investigated taxa in the article
- The continent in which the study was conducted
- Used research method: Experimental or observational/correlational
- If the study was done in the lab, enclosures, or field

This data was first preprocessed to meet the syntax of ORKG CSV file import[22]. We created one CSV file per hypothesis, which thus amounted to a minor transformation of the original Hi Knowledge data to prepare the data for ingestion into ORKG.

19 https://orkg.org/observatory/Invasion_Biology?sort=combined&classesFilter=Paper,Comparison,Visualization

20 https://orkg.org/comparison/R58002/

21 https://hi-knowledge.org

22 https://orkg.org/help-center/article/16/Import_CSV_files_in_ORKG

ORKG Comparisons. Following ingestion, we created ORKG Comparisons, one for each hypothesis[23]. For this, we used the existing ORKG feature and its approach to create comparisons. Figure 9.1 exemplifies the Comparison for the enemy release hypothesis, also available online at https://orkg.org/comparison/R58002/.

Properties	The invertebrate fauna on broom, Cytisus scoparius, in two native and two exotic habitats *Contribution 2 - 2000*	A Comparison of Herbivore Damage on Three Invasive Plants and Their Native Congeners: Implications for the Enemy Release Hypothesis *Contribution 1 - 2000*	Can enemy release explain the invasion success of the diploid Leucanthemum vulgare in North America? *Contribution 1 - 2000*	Incorporation of an invasive plant into a native insect herbivore food web *Contribution 1 - 2000*
Continent	Europe Oceania	North-America	North-America	Europe
Habitat	Terrestrial	Terrestrial	Terrestrial	Terrestrial
has research problem	Testing the enemy release hypothesis in invasion biology	Testing the enemy release hypothesis in invasion biology	Testing the enemy release hypothesis in invasion biology	Testing the enemy release hypothesis in invasion biology
hypothesis	Enemy release	Enemy release	Enemy release	Enemy release
Indicator for enemy release	Infestation	Damage	Damage	Infestation
Investigated species	Plants	Plants	Plants	Plants
Number of species	1	3	1	1
Release of which kind of enemies?	Specialists	no differentiation	no differentiation	no differentiation

Figure 9.1 Comparison for Hi Knowledge data on the enemy release hypothesis.

Data science. An additional aim for this prototype with the Hi Knowledge community was to test if ORKG and its integrations with computing environments such as Jupyter could be used to perform specific analyses of the ingested data, including tailored visualizations that are meaningful for the community. We tested this by performing basic data science tasks with Jupyter Notebooks and web applications that use the ingested data and replicate the Hi Knowledge network of hypotheses. With the ORKG Python library[24], researchers can easily read the data constituting a comparison into a Python data frame and use the powerful scripting environment to implement and execute data science and analysis tasks. With such a setup, we can tackle simple and more advanced data science tasks. For instance, we can easily compute how many contributions support, are undecided, or question a specific hypothesis. Figure 9.2 visualizes the answer to this question for the propagule pressure hypothesis. Thanks to the flexibility of Python data frames, it is possible

[23] https://orkg.org/search/invasion?types=Comparison
[24] https://orkg.readthedocs.io

to slice and dice the data in an arbitrary manner. Figure 9.3 shows the distribution of Hi Knowledge studies across continents. While the approach requires some level of programming, it also shows how the versatility of a computing environment can support much more than predefined visualizations of data on a website. To address the requirement of programming skills, we also created an R Shiny application which, contrary to the Jupyter Notebooks, creates interactive dashboard-style web applications accessible to all users.

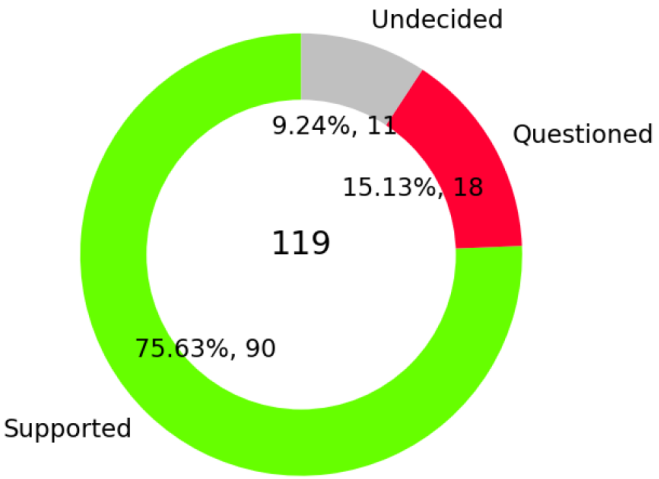

Figure 9.2 Share of contributions that support, question, or are undecided about the propagule pressure hypothesis.

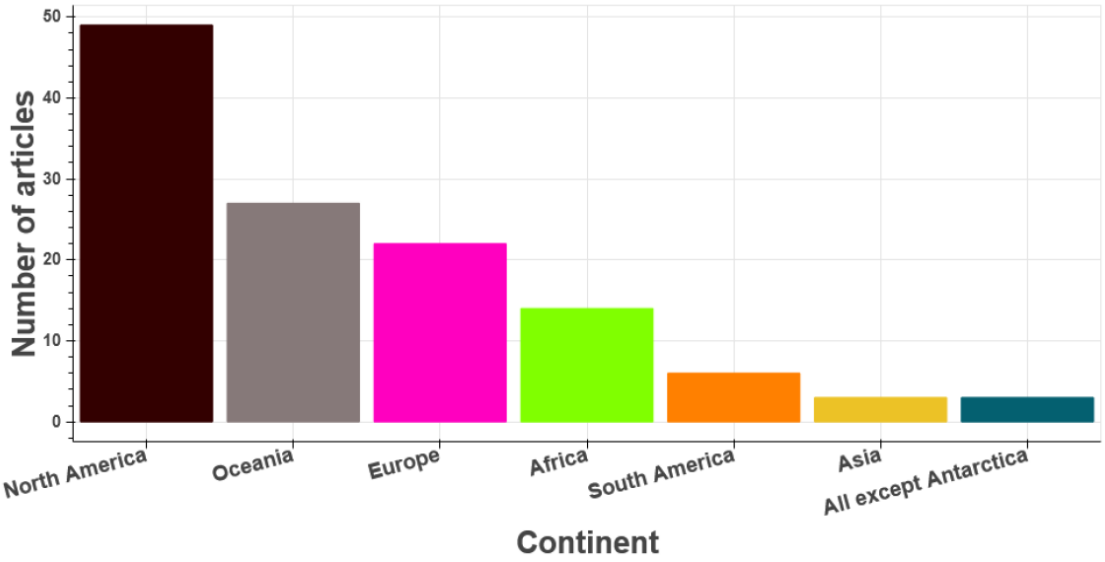

Figure 9.3 Visualization of the number of studies about the propagule pressure hypothesis across continents created with Hi Knowledge data ingested into ORKG using a computing environment.

9.2 The ecologist community gets more involved

Motivation

From 2021 to 2024, the enKORE project (*Jeschke et al., 2021*) within the Hi Knowledge initiative took further steps towards an atlas of knowledge for invasion biology. This project brought together ecologists and data scientists to work on organizing, extracting, synthesizing and visualizing literature in the field of invasion biology. The ORKG was used as a platform in this project to synthesize and visualize current scholarly literature on invasion biology. The effort was led by ecologist Maud Bernard-Verdier, in collaboration with Lars Vogt and Markus Stocker from the ORKG, with the goal first to revisit the existing data on 10 hypotheses in invasion biology.

Method

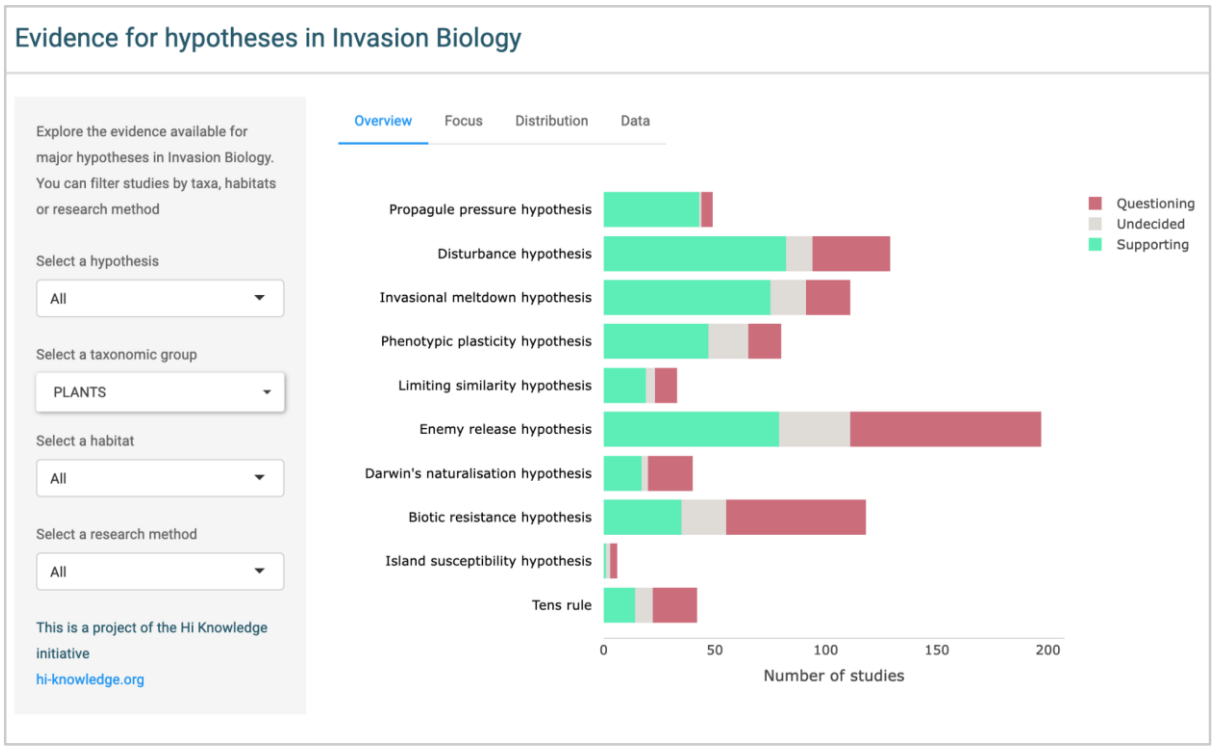

Figure 9.4 Screenshot of an R Shiny app[25] *offering an interactive visualization and summary of evidence for 10 hypotheses in invasion biology, combining 10 ORKG Comparison tables. Studies can be filtered by hypotheses, taxonomic groups, habitats or research methods. The Comparison tables (see Figure 9.1) were obtained by extracting existing published tables for synthetic reviews of hypotheses in invasion biology. The current view presents the distribution of evidence across 10 hypotheses for studies on invasive plants.*

As R is currently the preferred programming language for ecologists (*Lai et al., 2019*), the goal was to develop an R Shiny app for interactive visualization and exploration of the data, building upon the first Jupyter notebooks created by Kamel

[25] Visit the beta app: https://maudbernardverdier.shinyapps.io/Hypothesis-evidence-explorer/; R code accessible on github: https://github.com/maudbv/Hypothesis-evidence-explorer.

Fadel (see above). Using the ORKG package for Python (the R ORKG package was not yet finalized), Maud exported (as .csv) the 10 comparison tables summarizing support for the 10 hypotheses in invasion biology, and used them to create an R Shiny app, aiming first for a proof of concept on static data.

The app (Figure 9.4) presents a small number of curated figures and summary statistics relevant for ecologists to gain an overview of the state of knowledge concerning each hypothesis. Filtering options based on relevant properties annotated in ORKG Comparison tables allow for a customized exploration of the data, as well as data exports.

What we learned

Despite the careful data extraction by Kamel, substantial data cleaning and homogenization were necessary before the app could be created, mainly because the data tables from the original multi-author book (*Jeschke & Heger, 2018*) were themselves not perfectly standardized. For instance, the terms used to designate taxa groupings or habitats were not always comparable across hypothesis tables and had to be manually homogenized. This highlighted early on the need for better quality control (e.g. correcting typographic mistakes) and also standardized vocabulary, in which each term has a unique identifier, if we aim for seamless automatic synthesis. Guiding future ORKG annotations to re-use only pre-determined existing concepts in ORKG, published ontologies, or Wikidata, was identified as a solution to this problem in future steps.

Once data processing was completed, the task of creating visualizations benefited from the specialist perspective of the invasion biology community. While many figures and statistics were possible to compute, the visualizations included in the R Shiny app were selected to address basic questions in ecology concerning the current knowledge gaps and biases existing in the literature, and whether hypotheses are found to be better supported for some species or habitats. The app provides interactive versions of those static figures typically found in published systematic reviews, and one can imagine that systematic reviews could greatly benefit from being accompanied by such additional interactive material.

9.3 Engaging with the broader community of invasion biologists

Motivation

The Hi Knowledge dataset mentioned above is static and had not been updated since the publication of Jeschke & Heger, 2018. Such datasets are the product of an enormous synthesis effort by individual authors, which cannot be realistically reproduced on a regular basis. As mentioned above, the dataset was also not perfectly standardized and reusable, and, importantly, had not been fully semantically

modeled in ORKG (i.e. properties had no link to existing ontologies, Wikidata items or other semantic models).

We decided to use the ORKG as a platform to update the Hi Knowledge dataset, aiming for invasion biologists to contribute data following a comparable structure. The underlying idea is that invasion biologists who published a given study would be motivated to feed information about their study to ORKG, so that it is part of a growing database.

In the first attempts of invasion biologists in the team to add their own papers to ORKG, it quickly became clear that more guidance was needed. Invasion biologists do not typically know about semantic modeling or understand the rules, good practice and constraints associated with semantic annotations as is practiced in ORKG. If we want to motivate invasion biologists to spend time adding their work, and if we want the annotations to be comparable and valuable for automatic synthesis (e.g. in an R Shiny app), a tailored template is needed to guarantee interoperability across their contributions.

Method

Lars and Maud worked together on designing a tailored template for invasion biology that allows the annotation of basic ecological information about a study, as well as information about hypothesis testing following the Hi Knowledge dataset. This collaborative work relied on the input of invasion biologists, providing a list of example statements for Lars to build a first prototype of a semantic model. An online workshop in 2022 with over 70 invasion biologists[26] further identified a list of key concepts relevant to filter literature searches or organize meta-analyses. Building iteratively on this first graph, a first version of the template was implemented by Maud, and further tested and revisited following trial tests during a 2023 in-person workshop in Berlin[27].

We created several templates (Table 9.1): one main template for general scoping of any contribution in ecology and evolution, and five sub-templates, with three specific to invasion biology. It turned out that most of the key information we are interested in in invasion biology is common to the larger field of ecology, and we therefore seized on the opportunity to create a more general template for ecology (**#1**). After several iterations, we decided to simplify the initial template to make it more accessible, and move more complex information, such as descriptions of study design, datasets[28] or study systems, to sub-templates (**#4** and **#5**).

[26] Workshop report: https://zenodo.org/records/8421054

[27] Published workshop report: https://riojournal.com/article/115395/

[28] pre-existing ORKG template: https://orkg.org/template/R178304

Table 9.1: ORKG templates created for the field of invasion biology, and ecology in general.

#	Template name	Purpose	ORKG ID
1	Study in Ecology and Evolution (main template)	General template for any study in the field of ecology (*sensu largo)*	R593657
2	Invasion biology study research question	Annotate theme, research question, hypotheses and invasive taxa, following scheme by Musseau et al.	R593830
3	Hypothesis test in invasion biology	Annotate whether the study supports or not a major hypothesis	R646660
4	Ecological study system description	Describe the properties of a specific ecological study system, which can be shared by multiple studies	R593670
5	Ecological study design description	describe the study design (sample size, treatment, etc.) in an invasion biology study	R593806
6	Hypotheses in invasion biology template	Template for describing major theoretical hypotheses in invasion biology	R602693

Two sub-templates specific to the Hi Knowledge approach to invasion biology were designed. The first (#**2**) is a general description of the main theme, research questions, hypotheses and invasive taxa investigated, following our current conceptual scheme for invasion biology (*Musseau et al., in preparation*). The second (#**3**) describes the testing of major hypotheses in the field (described by template #**6**). It provides information about support or rebuttal of those hypotheses, in the same way as the Hi Knowledge data provided.

To create these templates, not only did new properties have to be modeled in ORKG, reusing as much as possible existing ontologies and Wikidata properties, but also new instance-resources to guide and limit the choices of template users. For instance, we wanted to allow the users to choose from a short list of research approaches, such as observational approaches, experimental approaches or conceptual approaches, and had to model those instances as well as the class to which they belong (class: “research approaches”[29]). We also created classes and instance-resources to describe all items of the conceptual scheme for invasion

[29] https://orkg.org/class/C65001

biology (5 themes, 10 research questions and 64 major hypotheses in invasion biology).

The templates then restricted the possible entries for these fields to only those belonging to the class. Of course, ORKG being fully flexible meant that users could still (and did!) create their own instances of research approach or hypotheses, which in most cases did not fit with what we had intended (e.g. too detailed, redundant with existing instance-resources, etc.). This great freedom in ORKG annotations is here a challenge for better standardization and automated knowledge synthesis.

9.4 Further use of ORKG in the context of invasion biology

ORKG for teaching in ecology

ORKG appeared as a great platform to teach students how to extract information from papers in a systematic way, and provide a published outcome for the class (published ORKG list[30]). In December 2023, we used the ORKG platform to teach (remotely) an introduction to invasion biology to a class of fourth year ecology students at Rhode Island University (USA) with Prof. Laura Meyerson, who had been part of previous workshops of the Hi Knowledge initiative. About 60 students were asked to annotate invasion biology papers using the ORKG templates described above, and with minimal guidance from us.

The pedagogical goals were the following:

1. Learn to extract key ecological information from a scientific paper in a systematic way.
2. Gain an overview of the different themes, research questions and hypotheses in invasion biology.
3. Contribute to community-curated tools for open knowledge synthesis in science.
4. Become familiar with notions of semantic graph modeling.

The students collectively annotated over 100 papers in two 3-hour sessions. The first session provided uneven results, and revealed a steep learning curve for the students to familiarize themselves with ORKG as a tool, as well as with the templates. At the end of the second session, though, most student groups had provided detailed annotations of two to five papers, spending roughly 30-60 mins per paper. This was highly encouraging regarding the usability of the templates, as well as a great learning experience for the students, who reported that they had felt “empowered” as students to actively participate in knowledge extraction rather

[30] https://orkg.org/list/R671240

than passive reading. This highlights the high pedagogical potential of such exercises with ORKG templates, and more ambitious versions of this class could even be designed as small systematic review projects.

Preliminary investigation of the data contributed by students nevertheless revealed a number of pitfalls in the template use, which need to be further analyzed. These might in part be avoided with clearer instructions (with a manual and demonstration) and better modeling. However, the inherent modeling freedom of ORKG means that we should always expect heterogeneities in data quality, and data cleaning strategies will need to be put in place for future data synthesis.

A tool for publishers to collect structured information about submissions

One clear challenge of our approach is to reach out and motivate a large portion of the community of invasion biologists to annotate papers, even their own work. One possibility to tackle this challenge could be to make such annotations part of the normal publication process in scientific journals. It is important, however, to design the process in a way that does not waste the time of authors in the publication submission process. In this perspective, semantic annotations could become a new standard for publishers at the submission level, replacing the current role of article keywords. Such annotations would make all new papers easier to search, group and filter by key ecological criteria. They would also allow dashboard-style automatic syntheses and overviews of the literature, representing the scope and possible research gaps on a given topic (similar to our R Shiny app for Hi Knowledge data), for publishers themselves, as well as any other users if the data is openly published and harvestable with each article.

Whether publishers would want to use ORKG as a platform is uncertain, but we could imagine that the platform could at least be used for preliminary tests and as a proof of concept. Partnerships with publishers willing to invest in open science and technology would be a great boost to the ORKG project. The modeling involved in designing custom templates for a given field should be published in itself as an open resource, and updated by the community around a consensus approach, to allow standardization and interoperability of annotations across journals and publishers and promote FAIR science.

Smart searches

Knowledge graphs allow us in theory to create smart searches with complex scoping and filtering based on statements or class hierarchies. Such smart searches are missing in ORKG, but many invasion biologists and other ecologist users would be interested in it. A good test case for that in ecology would be taxa (species) recognition which, due to the inherently hierarchical organization of taxonomies,

would lend itself particularly well to hierarchical grouping. Users would ideally like to be able to give the Latin name of a species, and it being recognized as a concept with all the known synonyms and taxonomic hierarchy, in such a way that studies could be grouped based on a higher taxonomic level (e.g. plants, insects, birds, etc.). Smart searches would then allow us to search for a certain taxonomic level, no matter the granularity, like “mammals” or “flowering plants”, and filter articles accordingly. While this is not yet possible in ORKG, it is something that would be a real asset to develop in the future.

9.5 Conclusion

Domain-specific templates are necessary for getting community engagement in ORKG, and partnership with scientists from different fields via collaborative projects like enKORE are a good way to build these resources. Outstanding issues are in the difficulty of scaling up engagement of the ecologist community, and data quality control. Data quality and interoperability within a field will depend on the quality of existing domain ontologies and other semantic models for a given field, which in the case of ecology still remain insufficiently developed. Potential solutions to be pursued include guiding “naive” users with better tutorials and explicit templates, engaging in teaching projects to curate certain topics, better workflows to connect with other open knowledge graph projects like Wikidata, and finally getting publishers involved.

References

Jeschke, J.M., & Heger, T. (2018). Invasion biology: hypotheses and evidence. CABI, Wallingford.

Fadel, K. (2021). Data Science with Scholarly Knowledge Graphs. Hannover : Gottfried Wilhelm Leibniz Universität Hannover. https://doi.org/10.15488/11535

Jeschke, J.M., Heger, T., Kraker, P., Schramm, M., Kittel, C., & Mietchen, D. (2021). Towards an open, zoomable atlas for invasion science and beyond. NeoBiota 68:5–18. https://doi.org/10.3897/neobiota.68.66685

Lai J, Lortie CJ, Muenchen RA, Yang J, Ma K (2019) Evaluating the popularity of R in ecology. Ecosphere 10: e02567. https://doi.org/10.1002/ecs2.2567

Jeschke, J.M., & Heger, T. (2018). Invasion biology: hypotheses and evidence. CABI, Wallingford.

Musseau, C., Bernard-Verdier, M., Heger, T., Skopeteas, L., Strasiewsky, D., Mietchen, D., & Jeschke, J. M. A conceptual classification scheme of invasion science. (in preparation)

10. Data to Knowledge: Exploring the Semantic IoT with ORKG

Sanju Tiwari

BVICAM, New Delhi, India & UAT Mexico

10.1 Motivation

Recently, the Internet of Things (IoT) has experienced substantial growth, facilitating the emergence of various applications like smart buildings, healthcare, transportation, and cities. A vast amount of unprocessed data generated by diverse IoT devices exhibits heterogeneity in terms of various types and formats. Consequently, the sharing and reuse of this raw IoT data poses a significant challenge for IoT applications [1] and highlights the need to improve the semantic aspects of IoT for better interoperability and understanding.

The Semantic IoT embodies a vision within information and communication technology that harmonizes two essential paradigms of the decade: the Semantic Web and the IoT. The necessity for interoperability in the IoT, particularly in terms of semantics, serves as a crucial driving force behind the progress of the Semantic IoT. The Semantic IoT involves incorporating semantic technologies into the IoT, with the goal of providing data with meaning and context. Conventional IoT systems typically depend on standardized communication protocols but may lack the capability to comprehend the semantics or significance embedded in the exchanged data. The purpose of the Semantic IoT is to overcome this limitation by introducing a layer of semantic interpretation to the data. To facilitate robust reasoning and inference, it is essential to offer semantic interoperability and effective data modeling, along with promoting the reuse and sharing of knowledge. Achieving these objectives is crucial without a comprehensive understanding of data semantics. Distributed, varied, and heterogeneous raw data sources, coupled with a substantial volume of crowded and incomplete data transmitted in diverse formats, give rise to challenges related to scalability, heterogeneity, and numerous interoperability issues [2].

10.1.1. Research highlights and contribution

Semantic technologies, encompassing standards of the semantic web and ontologies, facilitate the representation of data in a manner comprehensible to ma-

chines. The incorporation of semantic elements in the Semantic IoT seeks to improve the understanding, interoperability, and integration of data within the IoT. Figure 10.1 has presented a workflow of semantic IoT and the ORKG to represent the relation among end users, IoT Devices and the ORKG.

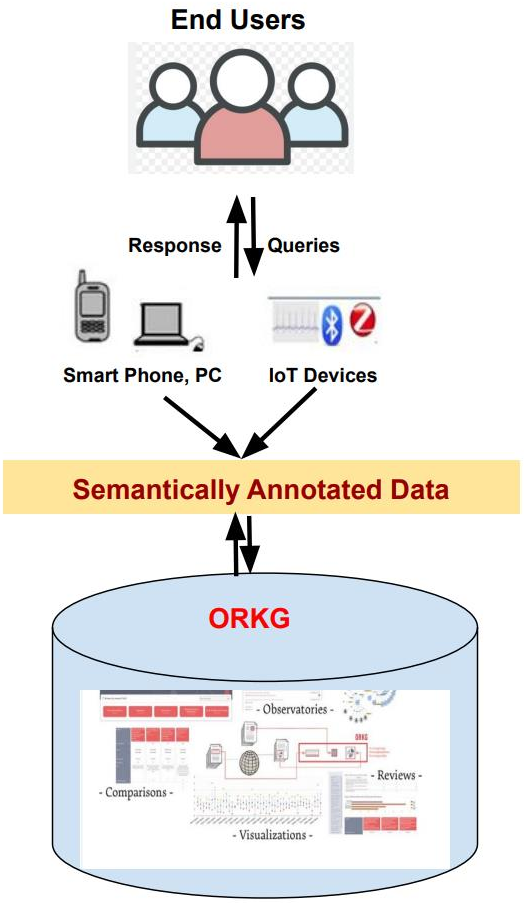

Figure 10.1 Semantic IoT workflow with the ORKG

ORKG serves as an infrastructure designed to represent, explore, and curate scholarly knowledge in a format that is machine-actionable [3]. This chapter contributes to presenting a survey on Semantic IoT, representing an emerging area within the ORKG research domain [4]. The ORKG does not merely consist of bibliographic metadata (such as information about authors, articles, and institutions), but also includes semantic descriptions of scholarly knowledge, making it machine-actionable [5]. A smart review [6] on the Semantic IoT has been conducted by using the ORKG, which included 4 different comparisons on the Semantic IoT and also compared relevant contributions on IoT in Knowledge Graphs and IoT in Digital Twins. This chapter explores the Semantic IoT by using the ORKG for different domains such as healthcare, building, industry 4.0, manufacturing and edge computing etc.

10.2 Background

The Internet of Things (IoT) is a system in which physical devices are integrated into electronic systems, enabling them to connect to the internet. These devices can be monitored, controlled, discovered, and interact with one another through diverse network interfaces. However, the absence of a universal application protocol in IoT poses a challenge, impeding the seamless integration of devices from different manufacturers into a unified application [7]. Web of Things (WoT) [8] has been introduced as an extension of the IoT to address IoT challenges. This section will highlight the Semantic IoT in different aspects such as ontologies, knowledge graphs, digital twins etc.

10.2.1. Web and Semantic Web of Things (WoT/SWoT)

The proliferation of the IoT has introduced the WoT as open web standards aimed at facilitating machine interoperability and the exchange of information [9]. The convergence of Semantic Web Technologies (SWT) with the domains of Internet of Things (IoT) or Web of Things (WoT) gives rise to a new concept known as the Semantic Web of Things (SWoT) [10]. It addresses diverse issues in the IoT, including interoperability, scalability, deep heterogeneity, security, incomplete or inaccurate metadata, and conflict resolution.

10.2.2. IoT Ontologies

IoT devices acquire a huge amount of data through the integrated system within them. The nature of acquired data is multi-modal and heterogeneous as it is collected in different formats. It is challenging to manage such large-scale heterogeneous data in smart applications. Semantic approaches, particularly ontologies, have been employed to address challenges associated with extensive heterogeneity. IoT ontologies can be categorized based on context, location, time, security and IoT applications such as SSN/SOSA [11], SAREF [12], STAC [13], IoT-O[14] and IoT-Lite [15]. The SSN (semantic sensor network) ontology[16] is among the IoT ontologies used to describe sensor resources and the data acquired by these sensors. Its primary concepts include sensor, device, and observation.

10.2.3. IoT Knowledge Graphs

Knowledge graphs are closely connected to ontologies, and there is, in fact, no unanimous agreement on definitions that distinctly differentiate the former from the latter. Knowledge Graphs are applied in several related contexts of industry 4.0 and IoT concepts [17]. Liu et. al. [18] proposed an approach to represent data for IoT-enabled cognitive manufacturing using a knowledge graph. Xie et. al. [19] has

introduced a multilayer IoT middleware based on a knowledge graph which incorporates an additional layer to address the communication protocol disparities among IoT devices. An ORKG comparison [20] has compared 8 different articles to explore the features of industry 4.0 and manufacturing domain with IoT Knowledge Graphs.

10.2.4. IoT in Digital Twins

The digital twin acts as a virtual representation to a physical object or system, requiring constant integration and updating of data to adapt to dynamic environments. Jarabo [21] has proposed a knowledge graph to effectively store and query an extensive amount of IoT devices within a sophisticated logical framework. It employs rule-based reasoning to deduce novel information and seamlessly incorporates unforeseen devices into the pre existing logical structure. An ORKG comparison [22] has compared 6 different articles to explore the role of knowledge graphs in digital twin Models.

10.3 Semantic IoT in Specific Domains

The Semantic IoT is a conceptual framework that merges the functionalities of the IoT with Semantic Web technologies. This integration facilitates effective data retrieval, knowledge extraction, and seamless integration while promoting interoperability. This section presents how the Semantic IoT strengthens different domains such as water, healthcare, industry 4.0/manufacturing/IoT, energy efficient building, and agriculture by integrating IoT concepts with semantic web techniques such as OWL/RDF, SPARQL, SWRL etc. Table 10.1 shows the popular semantic models/ontologies of each domain and also shares the ORKG comparison source to see the structured information of all ontologies. Ten ontologies are compared in the water domain for their different characteristics such as number of classes, properties, reused ontologies, online status etc; Healthcare also compared 10 ontologies for their different features; industry 4.0/manufacturing/IoT has compared 11 ontologies; 15 ontologies are compared in energy efficient building domain while 2 ontologies are considered in agriculture domain.

10.3.1. Semantic IoT in Water

Various IoT-based semantic models have been designed to depict different facets of water resources, including entities like water bodies, water types, water pipes, water meters, reservoirs, catchments, pumps, and sensors. A study [23, 24] has been presented to discuss various existing water ontologies such as Water-Nexus

Ontology, DSHWS, EU WEFNexus etc and also compared in the ORKG framework to explore the IoT-based water ontologies. SAREF4WATER [25] offers an ontology designed for applications related to water, encompassing elements like meters, infrastructure for the distribution of drinking water, and an illustrative example of a key performance indicator.

Table 10.1 Semantic IoT-based Ontologies for Specific Domains.

Specific Area	Ontology	ORKG Source
Water	SAREF4WATR, DSHWS, WaterNexus Ontology, EUWEFNexus, OntoWAWO, SurfaceWater, xLMINWS.owl	https://orkg.org/comparison/R217545/
Healthcare Ontology	HealthIoT, SAREF4Health, e-Health, SHCO, Linked Health Resource, IFO	https://orkg.org/comparison/R223002/
Industry 4.0/Manufacturing/IoT	I4.0-Onto, AMLO, ExtruOnt, SAREF4INMA, CROS, OCRA, Saref, SSN/SOSA	https://orkg.org/comparison/R659252/
Energy Efficient Building Ontologies	W3CBOT, SAREF4BLDG, Topo, EM-KPI, IoT-O, SEAS, OEMA, EEPSA	https://orkg.org/comparison/R214164/
Agriculture	saref4agri, Agri-IoT	NA

10.3.2. Semantic IoT in Healthcare

In the healthcare context, a semantic IoT framework [26, 27, 28] integrates IoT devices with semantic web technologies to enhance the management and analysis of healthcare data. This system facilitates the collection and analysis of information from diverse IoT healthcare devices such as sensors, wearables, and home monitoring systems, offering a comprehensive overview of a patient's health.

10.3.3. Semantic IoT in Industry 4.0 and Manufacturing

In the realm of Industry 4.0 and manufacturing [29, 2], Semantic IoT entails incorporating semantic technologies into the IoT landscape to augment the intelligence, interoperability, and efficiency of industrial processes. The SAREF4INMA ontology [30] was recently developed to expand upon the SAREF framework, specifically for the purpose of describing the domain of Smart Industry and Manufacturing. The

ExtruOnt [31] ontology is composed of terms designed to depict a category of manufacturing machinery utilized in extrusion processes, specifically referring to an extruder.

10.3.4. Semantic IoT in Energy Efficient Building

The application of semantic technologies within the IoT context in the Energy Efficient Building domain aims to enrich the intelligence and efficiency of building management systems. The SAREF4BLDG [32] ontology is an expansion of the SAREF (Smart Appliance Reference Ontology) specifically tailored for the building domain and aligned with the Industry Foundation Classes (IFC) standard. There are various related ontologies such as BOT, TOPO, EM-KPI, IoT-O, SEAS, OEMA, EEPSA etc. are compared in ORKG [33] framework.

10.3.5. Semantic IoT in Agriculture

The application of Semantic IoT in agriculture entails incorporating semantic technologies into the IoT landscape. This integration aims to improve data interoperability, represent knowledge in a structured manner, and enhance decision-making processes within agricultural practices. SAREF4AGRI [34] is an OWL-DL ontology designed to extend SAREF (Smart Appliance Reference Ontology) [35] specifically for the Smart Agriculture and Food Chain domain. The primary objective of SAREF4AGRI is to establish connections between SAREF and other developed ontologies, such as W3C SOSA, W3C SSN, GeoSPARQL, and various standardization initiatives and ontologies within the Smart Agriculture and Food Chain domain. Agri-IoT [36] is a semantic framework designed for intelligent farming applications based on the IoT. It facilitates real-time reasoning over diverse sensor data streams. Agri-IoT has the capability to seamlessly integrate multiple data streams from different domains, establishing a comprehensive semantic processing pipeline.

10.4 Major Sources of IoT Ontologies

Ontology can play a significant role to assist strategic and operational decision-making situations that can enhance the efficiency of IoT systems. The Linked Open Vocabularies for the Internet of Things (LOV4ToT) https://www.lov4iot.appspot.com/ [37] is a major source to find the IoT ontologies in different domains and plays a pivotal role in acquiring existing IoT ontologies to reuse. By providing a curated collection of linked vocabularies and ontologies related to the Internet of Things, LOV4IoT facilitates the identification and selection

of suitable ontological resources. Some other sources such as Linked Open Vocabularies (LOV) [38], Ontology Lookup Service https://www.ebi.ac.uk/ols4 and dataset sources (https://coggle.it/diagram/WXiSLnz3AAABhI89/t/how-to-find-ontologies-and-datasets) are also providing existing ontologies in the related field. Schema.org (https://schema.org/) is a collaborative community effort dedicated to developing, promoting and maintaining structured data schemas across the internet, encompassing electronic messages, web pages, and more.

Table 10.2 Semantic IoT Frameworks

Name	Description	Source
BiG-IoT	The BiG-IoT framework addresses IoT interoperability issues by utilizing semantic addressing with the development of the BiG-IoT API.	[39] [40]
FIESTA IoT	The FIESTA project enables the reuse of data across various IoT testbeds, employing semantic technologies for enhanced interoperability	[41]
VICINITY	The primary objective of VICINITY is to enhance semantic interoperability, achieved by leveraging the standard W3C Web Ontology Language	[42]
INTER-IoT	The primary objective is to realize, execute, and validate a framework facilitating interoperability among diverse IoT platforms.	[43]
Open-IoT	The aim of the Open-IoT project is to enhance semantic interoperability and achieve semantic integration across differ ent IoT systems by using the Sensor Network (SSN) ontology.	[44]
SymbIoTe (Symbiosis of Smart Objects Across IoT Environments)	SymbIoTe offers a semantic IoT search engine tailored for smart objects that are registered by platform providers and connected to the network.	[45]
M3 (Machine-to-Machine Measurement) Framework	The M3 Framework project is focused on addressing the issue of semantic interoperability within the IoT.	[46]

10.4.1 Semantic IoT Frameworks

Semantic IoT frameworks are presented as a layer's set that are responsible for persistence, aggregation, serving of data, and analytics [47]. Fatima et. al. [1] has discussed some existing IoT-related frameworks (BiG-IoT, VICINITY, FIESTA IoT, Open-IoT, INTER-IoT, M3, SymbIoTe) supporting semantic interoperability in IoT systems, highlighted in Table 10.2.

10.5 Conclusion

This chapter has a pivotal role in presenting and disseminating our unique perspective on the application of Semantic IoT across a spectrum of domains including water, healthcare, industry 4.0 and manufacturing, energy efficient building, and agriculture. Within these domains, we intricately explore the role of the Semantic IoT, leveraging the ORKG to explore various IoT-based ontologies. Through this comparative analysis, we delve into the diverse properties and classes encapsulated within existing studies. Moreover, our chapter meticulously addresses the substantial sources of IoT Ontologies, while also covering Semantic IoT Frameworks. By providing comprehensive coverage of these foundational elements, we aim to facilitate a deeper understanding of the landscape of Semantic IoT implementation, empowering readers with the knowledge required to navigate and innovate within this promising field.

References

[1] Fatima Zahra Amara, Mounir Hemam, Meriem Djezzar, and Moufida Maimour. Semantic web technologies for internet of things semantic interoperability. In *Advances in Information, Communication and Cybersecurity: Proceedings of ICI2C'21*, pages 133–143. Springer, 2022.

[2] Fatima Zahra Amara, Meriem Djezzar, Mounir Hemam, Sanju Tiwari, and Mo hamed Madani Hafidi. Unlocking the power of semantic interoperability in industry 4.0: A comprehensive overview. In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, pages 82–96. Springer, 2023.

[3] Mohamad Yaser Jaradeh, Allard Oelen, Manuel Prinz, Markus Stocker, and Sören Auer. Open research knowledge graph: a system walkthrough. In *Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23*, pages 348–351. Springer, 2019.

[4] Sören Auer and Sanjeet Mann. Towards an open research knowledge graph. *The Serials Librarian*, 76(1-4):35–41, 2019.

[5] 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, 2019.
[6] Sanju Tiwari. Exploring the Semantic IoT. https://orkg.org/review/R659310. [On line; accessed 2024-01-23].
[7] Dominique Guinard, Vlad Trifa, Friedemann Mattern, and Erik Wilde. From the internet of things to the web of things: Resource-oriented architecture and best practices. *Architecting the Internet of things*, pages 97–129, 2011.
[8] Dominique Guinard and Vlad Trifa. Towards the web of things: Web mashups for embedded devices. In *Workshop on Mashups, Enterprise Mashups and Lightweight Composition on the Web (MEM 2009), in proceedings of WWW (International World Wide Web Conferences), Madrid, Spain*, volume 15, page 8, 2009.
[9] Sanju Tiwari, Fernando Ortiz-Rodriguez, and MA Jabbar. Semantic modeling for healthcare applications: an introduction. *Semantic Models in IoT and eHealth Applications*, pages 1–17, 2022.
[10] Ahlem Rhayem, Mohamed Ben Ahmed Mhiri, and Faiez Gargouri. Semantic web technologies for the internet of things: Systematic literature review. *Internet of Things*, 11:100206, 2020.
[11] Krzysztof Janowicz, Armin Haller, Simon JD Cox, Danh Le Phuoc, and Maxime Lefrançois. Sosa: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*, 56:1–10, 2019.
[12] Laura Daniele, Frank den Hartog, and Jasper Roes. Created in close interaction with the industry: the smart appliances reference (saref) ontology. In *Formal Ontologies Meet Industry: 7th International Workshop, FOMI 2015, Berlin, Germany, August 5, 2015, Proceedings 7*, pages 100–112. Springer, 2015.
[13] Amelie Gyrard, Christian Bonnet, and Karima Boudaoud. The stac (security tool box: attacks & countermeasures) ontology. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 165–166, 2013.
[14] Nicolas Seydoux, Khalil Drira, Nathalie Hernandez, and Thierry Monteil. Iot-o, a core-domain iot ontology to represent connected devices networks. In the European *knowledge acquisition workshop*, pages 561–576. Springer, 2016.
[15] Maria Bermudez-Edo, Tarek Elsaleh, Payam Barnaghi, and Kerry Taylor. Iot-lite: a lightweight semantic model for the internet of things and its use with dynamic semantics. *Personal and Ubiquitous Computing*, 21:475–487, 2017.
[16] Kerry Taylor, Armin Haller, Maxime Lefrançois, Simon JD Cox, Krzysztof Janowicz, Raul Garcia-Castro, Danh Le Phuoc, Joshua Lieberman, Rob Atkinson, and Claus Stadler. The semantic sensor network ontology, revamped. In *JT@ ISWC*, 2019.
[17] Claudia Diamantini, Alex Mircoli, Domenico Potena, and Emanuele Storti. Process-aware iiot knowledge graph: A semantic model for industrial iot integration and analytics. *Future Generation Computer Systems*, 139:224–238, 2023.

[18] Mingfei Liu, Xinyu Li, Jie Li, Yahui Liu, Bin Zhou, and Jinsong Bao. A knowledge graph-based data representation approach for iiot-enabled cognitive manufacturing. *Advanced Engineering Informatics*, 51:101515, 2022.
[19] Cheng Xie, Beibei Yu, Zuoying Zeng, Yun Yang, and Qing Liu. Multilayer internet of-things middleware based on knowledge graph. *IEEE Internet of Things Journal*, 8(4):2635–2648, 2020.
[20] Sanju Tiwari. Exploring semantic interoperability with iot based knowledge graphs in industry 4.0. https://orkg.org/comparison/R656106/. [Online; accessed 2024-03- 09].
[21] Alejandro Jarabo Peñas. Digital twin knowledge graphs for iot platforms: Towards a virtual model for real-time knowledge representation in iot platforms, 2023.
[22] Sanju Tiwari. Exploring digital twin models based on knowledge graphs. https://orkg.org/comparison/R659134/. [Online; accessed 2024-03-09].
[23] S Tiwari and R Garcia-Castro. A systematic review of ontologies for the water domain. *ISTE Book*, 2022.
[24] Shikha Mehta, Sanju Tiwari, Patrick Siarry, and MA Jabbar. *Tools, Languages, Methodologies for Representing Semantics on the Web of Things*. John Wiley & Sons, 2022.
[25] Raul Garcia-Castro. Saref extension for water. https://saref.etsi.org/saref4watr/v1.1.1/. [Online; accessed 2024-03-09].
[26] Ahlem Rhayem, Mohamed Ben Ahmed Mhiri, and Faiez Gargouri. Healthiot on tology for data semantic representation and interpretation obtained from medical connected objects. In *2017 IEEE/ACS 14th international conference on computer systems and applications (AICCSA)*, pages 1470–1477. IEEE, 2017. [27] Joao Moreira, Luís Ferreira Pires, Marten van SINDEREN, and Laura Daniele. Saref4health: Iot standard-based ontology-driven healthcare systems. In *FOIS*, pages 239–252, 2018.
[28] TITI Sondes, Hadda Ben Elhadj, and Lamia Chaari. An ontology-based health care monitoring system in the internet of things. In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pages 319–324. IEEE, 2019.
[29] Fatima Zahra Amara, Meriem Djezzar, Mounir Hemam, and Sanju Tiwari. A real time semantic based approach for modeling and reasoning in industry 4.0. *International Journal of Information Technology*, pages 1–9, 2023.
[30] Mike de Roode, Alba Fernández-Izquierdo, Laura Daniele, María Poveda-Villalón, and Raúl García-Castro. Saref4inma: a saref extension for the industry and manufacturing domain. *Semantic Web*, 11(6):911–926, 2020.
[31] Víctor Julio Ramírez-Durán, Idoia Berges, and Arantza Illarramendi. Extruont: An ontology for describing a type of manufacturing machine for industry 4.0 systems. *Semantic Web*, 11(6):887–909, 2020.
[32] Raúl Garcia-Castro María Poveda-Villalón. Saref extension for building. https://saref.etsi.org/saref4bldg/v1.1.2/. [Online; accessed 2024-03-09].
[33] Sanju Tiwari. Existing smart building domain ontologies comparison. https://orkg.org/comparison/R214164/. [Online; accessed 2024-03-09].
[34] García-Castro-Raúl Laura Daniele Mike de Roode Poveda-Villalón, María. Saref4agri: an extension of saref for the agriculture and food domain. https://saref.etsi.org/saref4agri/v1.1.2/. [Online; accessed 2024-03-09].

[35] Raúl García-Castro, Maxime Lefrançois, María Poveda-Villalón, and Laura Daniele. The etsi saref ontology for smart applications: a long path of development and evolution. *Energy Smart Appliances: Applications, Methodologies, and Challenges*, pages 183–215, 2023.
[36] Andreas Kamilaris, Feng Gao, Francesc X Prenafeta-Boldu, and Muhammad Intizar Ali. Agri-iot: A semantic framework for internet of things-enabled smart farming applications. In *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, pages 442–447. IEEE, 2016.
[37] Amelie Gyrard, Christian Bonnet, Karima Boudaoud, and Martin Serrano. Lov4iot: A second life for ontology-based domain knowledge to build semantic web of things applications. In *2016 IEEE 4th international conference on future internet of things and cloud (FiCloud)*, pages 254–261. IEEE, 2016.
[38] Linked open vocabularies (lov). https://lov.linkeddata.es/dataset/lov/vocabs/. [Online; accessed 2024-03-09].
[39] George Hatzivasilis, Ioannis Askoxylakis, George Alexandris, Darko Anicic, Arne Bröring, Vivek Kulkarni, Konstantinos Fysarakis, and George Spanoudakis. The interoperability of things: Interoperable solutions as an enabler for iot and web 3.0. In *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 1–7. IEEE, 2018.
[40] Thomas Jell, Claudia Baumgartner, Arne Bröring, and Jelena Mitic. Big iot: inter connecting iot platforms from different domains—first success story. In *Information Technology-New Generations: 15th International Conference on Information Technology*, pages 721–724. Springer, 2018.
[41] Martin Serrano, Amelie Gyrard, Michael Boniface, Paul Grace, Nikolaos Georgan tas, Rachit Agarwal, Payam Barnagui, Francois Carrez, Bruno Almeida, Tiago Teixeira, et al. Cross-domain interoperability using federated interoperable semantic iot/cloud testbeds and applications: The fiesta-iot approach. In *Building the Future Internet through FIRE*, pages 287–321. River Publishers, 2022.
[42] Yajuan Guan, Juan C Vasquez, Josep M Guerrero, Natalie Samovich, Stefan Vanya, Viktor Oravec, Raúl García-Castro, Fernando Serena, María Poveda-Villalón, Carna Radojicic, et al. An open virtual neighborhood network to connect iot infrastructures and smart objects—vicinity: Iot enables interoperability as a service. In *2017 Global Internet of Things Summit (GIoTS)*, pages 1–6. IEEE, 2017.
[43] Giancarlo Fortino, Claudio Savaglio, Carlos E Palau, Jara Suarez de Puga, Maria Ganzha, Marcin Paprzycki, Miguel Montesinos, Antonio Liotta, and Miguel Llop. Towards multi-layer interoperability of heterogeneous iot platforms: The inter-iot approach. *Integration, interconnection, and interoperability of IoT systems*, pages 199–232, 2018.
[44] John Soldatos, Nikos Kefalakis, Manfred Hauswirth, Martin Serrano, Jean-Paul Calbimonte, Mehdi Riahi, Karl Aberer, Prem Prakash Jayaraman, Arkady Za slavsky, Ivana Podnar Žarko, et al. Openiot: Open source internet-of-things in the cloud. In *Interoperability and Open-Source Solutions for the Internet of Things: International Workshop, FP7 OpenIoT Project, Held in Conjunction with SoftCOM 2014, Split, Croatia, September 18, 2014, Invited Papers*, pages 13–25. Springer, 2015.

[45] Ivana Podnar Žarko. Bridging iot islands: the symbiote project. *E & I. Elektrotechnik und Informationstechnik*, 133(7):315–318, 2016.
[46] Amelie Gyrard, Soumya Kanti Datta, Christian Bonnet, and Karima Boudaoud. Cross-domain internet of things application development: M3 framework and evaluation. In *2015 3rd International conference on future Internet of Things and Cloud*, pages 9–16. IEEE, 2015.
[47] Amarnath Palavalli, Durgaprasad Karri, and Swarnalatha Pasupuleti. Semantic internet of things. In *2016 IEEE tenth international Conference on Semantic Computing (ICSC)*, pages 91–95. IEEE, 2016.

11. Food Information Engineering for a Sustainable Future

Azanzi Jiomekong

Department of Computer Science, University of Yaounde 1, Yaounde, Cameroon

11.1 Motivation

According to the World Health Organization (WHO), every country in the world is affected by one or more forms of malnutrition (*WHO Malnutrition Factsheet*). However, adequate nutrition is an essential catalyst for economic and human development as well as for achieving Sustainable Development Goals (SDGs). If well organized and disseminated, food information may be used to make relevant decisions and achieve a healthy and sustainable food future. Food information engineering involves the acquisition, organization, storage, processing and diffusion of up-to-date food information to different stakeholders (*Jiomekong, 2023*). This allows food information to be used for providing sufficient and healthy food to people while ensuring sustainable impact on both environment, economic and social systems that surround food. These consist of sustainable agricultural practices (*Kassie et al., 2009*), food distribution systems, food quality (*Bortolini et al., 2016*), diets (*Meybeck and Gitz, 2017*), etc.

A huge number of research papers have been published in the domain of food information engineering, each paper covering different aspects. These papers may constitute reliable sources of food knowledge. This research suggests extracting and organizing food information embedded into scientific papers in a scholarly knowledge graph (KG) so as to provide to stakeholders quick access to relevant food knowledge. Unlike state of the art on the subject (*Jiomekong, 2023, Min et al., 2019 & Min et al., 2022*) which provide static resources in the form of HTML or PDF documents, this research aims to provide dynamic resources stored in a KG which will be continuously updated by the researchers of the domain. This chapter presents how this work is being done using the ORKG (*Auer et al., 2020*).

The extraction and organization of food information from scientific papers follow the following main steps: (1) Extraction of knowledge from scientific and organizing this knowledge into classes, properties and relation, (2) Use of classes, properties

and relations to build ORKG templates. The latter constitute conceptual models for describing several research problems, (3) Organization of knowledge extracted from scientific papers into research contributions. During this task, the templates created are used to create class instances, (4) Creation of comparisons tables and smart reviews, (5) To allow the food information engineering community to collaborate to organize the domain and ensure high quality standard, scientific papers, templates, comparisons tables and smart reviews are organized into the "Food Information Engineering"[31].

11.2 Food Information Engineering

This section presents how food information is collected, organized, processed and used.

11.2.1. Collecting food information

Thanks to the deployment of the internet, various smart devices, Internet of Things (IoT), and networks such as social network, mobiles networks, a great amount of food data is being recorded from different sources and in various modalities such as text, images, videos, and sound. These sources can be organized into:

(1) **Human sources:** Humans are the principal source of food data. They may play different roles during food information acquisition such as domain experts, recorders of food information using tools such as food log (*Metwally et al., 2021*). The acquisition of food data from human sources is always manual because people from which information is coming from should provide these information by observation, talking or writing. Manual acquisition can be used for instance, to annotate food images by a human who identifies the food and labels the visible food ingredients. It should be noted that data acquisition through human sources is time-consuming, laborious and hard to achieve at large-scale.

(2) **Structured sources:** Structured sources (e.g., CSV, JSON, XML, relational databases etc.) provide information using a standardized schema. In the domain of food, spreadsheet (*Food Composition Database*) databases, ontologies, Knowledge Graphs (*Min et al., 2022 & Jiomekong, 2023 a*) are used to organize food data and can constitute relevant food data sources to automatically extract food information from these sources, specialized tools exploit the structure description of data.

[31] https://orkg.org/observatory/Food_Information_Engineering

Semi-structured sources: Many food information is embedded in web pages and tables in pdf documents. These follow a structure that makes it easy to build automatic tools to extract food information. For instance, web scraping can be used to extract food information from web pages and the table structure of food composition tables stored in scientific papers make it easy to build automatic Optical Character Recognition (OCR) tools for their extraction (*Jiomekong et al., 2023*)

Unstructured sources: Information extraction from unstructured sources such as text, images, videos is the most difficult and challenging due to the nature of these information. For instance, food images are different from the other types of images (*Min et al., 2019*). Many of them do not have a distinctive spatial layout, they have deformable food appearance and thus lack rigid structures. Once acquired, food information is organized into datasets for different purposes. Given the multimodal nature of food information, we consider two types of food datasets: unimodal datasets and multimodal datasets (*Jiomekong, 2023 b*). Unimodal datasets such as Recipe 1M (*Marín et al., 2021*), TSOTSATable dataset (*Jiomekong et al., 2023 a)*, food image dataset Food 101 (*Bossard et al., 2014*), etc. contain data of only one type such as image or text. Multimodal datasets (*Yagcioglu et al., 2018*) may contain structured data such as symbolic representation of some food in form of knowledge graph, ontology, etc. and images (or videos) of the corresponding food.

11.2.2 Organizing Food Information

The main way currently used to organize food information is tabular organization. This organization uses tools such as databases and spreadsheets to organize and store food information. For instance, many food related software such as FoodLog Apps use relational databases to store food data (*Metwally et al., 2021*) and follow the nutrition of people in diet. FCT organizes food and its composition using spreadsheets and relational databases (*Food Composition Database*). To organize food information, symbolic organization uses symbols to represent background food knowledge. To this end, food information are linked together forming either food classification systems (*Jiomekong, 2023 a*), food ontologies (*Jiomekong, 2022*), food knowledge graphs (*Jiomekong, 2022 a*) or food linked data such as TSOTSATable dataset (*Jiomekong et al., 2023 a*).

Connectionist organization (*Jiomekong, 2023 a*) of food information consists of learning associations from food data and storing these information in the form of connections between nodes. It uses a large amount of data to adjust the strength of the connections (weights) between its nodes (or neurons). When there is not enough data, existing models such as VGG-19, AlexNet, GoogLeNet, Resnet-50, DenseNet, MobileNets, ShuffleNets, trained on food images are fine-tuned. Neuro symbolic organization can be done for many reasons including multimodal nature

of food information, explainability. Although several works have been done on representing food information using symbolic and connectionist AI, neuro-symbolic AI is still rarely used.

11.2.3 Food information processing

Information/data processing can intervene at any step of the food information engineering workflow (*Jiomekong, 2022 b*). In many cases, after data collection, preprocessing should be done. It may consist of cleaning the data by eliminating bad, inaccurate and unnecessary data (redundant, incomplete, or incorrect data, miscalculation, etc.) and having the data in a more readable format. The dataset obtained may be analyzed using statistical tools and the results disseminated to different stakeholders to help them understand and interpret information. In this case, the data visualization such as charts, graphs, dashboard, tables or reports are used. Symbolic methods process symbols by using logic-based programming where rules and axioms are used to make inferences and deductions. Concerning food information engineering, inference engines are used to generate new facts from symbolic representation of knowledge such as food ontologies and food knowledge graphs.

Connectionist models (e.g., CNN, GoogLeNet, Resnet-50, AlexNet network, etc.) have proven their superiority in several task (*Jiomekong, 2022 b*) such as food recognition, ingredient detection, food segmentation, food volume estimation, food recommendation, food calorie estimation from food image, etc. Neuro-symbolic methods may be used to infer ingredient and/or food composition from a dish image. A deep neural network can take as input the food image and return as output the food name. Thereafter, the food name can be used as input to a knowledge based system which uses an inference mechanism to infer the food ingredients and food components.

11.2.4 Using of food information

All the people in the world are involved in the production, processing, and use of food information. These information should allow for a planet-friendly diet, and a healthy and sustainable food future (*Parody et al., 2018*). Given to different usage (*Jiomekong, 2022 c*), we classified them into the following categories:

(1) The general population: Food information is generally used by all people around the world to choose their food given to food perception, taste, preferences or their health status. The increasing instances of obesity and related diseases are making consumers more healthy-conscious. Their demand for food information may concern food and beverage products that are natural and low in fat and calorie content.

(2) Health professionals: Health professionals generally use food information for identifying the origin of a health problem such as allergy, foodborne diseases, etc. To trace back and understand the origins of certain symptoms, health professionals generally ask questions on the eating history of the patients.
(3) Nutritionists: This category of stakeholders uses food information for nutrition advice. Many people suffer health problems that need nutrition monitoring such as diabetes, overweight, cardiovascular diseases, etc. Nutritionists are specialists that generally follow the diet of these people by using tools such as food logs. Food information may help them to increase consumer education on the importance of a healthy diet and active lifestyle.
(4) Decision makers: Decision makers use food information to ensure that the population has safe and enough food. For instance, information on food production may allow decision makers to put in place a system to afford the population with sufficient food.
(5) Food manufacturers, distributors and retailers: Knowledge on the eating behavior of the population can help this category of users to identify in a geographical area, which kind of food can be proposed to customers. In addition, a better understanding of the process used by people to assess the acceptability and flavor of new food products may be used by food manufacturers to produce acceptable food.
(6) Researchers: Food information engineering is a multi-disciplinary research domain in which many types of researchers are found. It involves researchers from food science and nutrition, food chemistry, microbiology, computer science, agriculture, etc. These researchers make use of food information to draw and/or validate hypotheses, build AI models, make predictions, etc.

11.3 Food Information Engineering Observatory

Food information engineering observatory aims to allow the food information engineering community to collaborate to organize the domain and ensure high quality standards. Additionally, it provides a unique view to different users. Currently, around 230 scientific papers of the domain, 11 templates, 65 comparisons tables, 11 visualizations and 9 smart reviews are organized in this observatory. Figure 11.1 presents an excerpt of resources stored in this observatory.

The templates are used to document the research contributions of the authors. It should be noted that in one paper, many research contributions can be found. The templates are made as generic as possible to facilitate their reuse for other purposes. Figure 11.1 presents some templates for documenting papers related to retrieval systems, recognition systems, methodologies, methods and tools for ontologies and knowledge graph construction, image datasets and questions answering. These templates were used to describe food recognition systems, food

retrieval systems, ontologies and knowledge graph construction, food images datasets and food question answering.

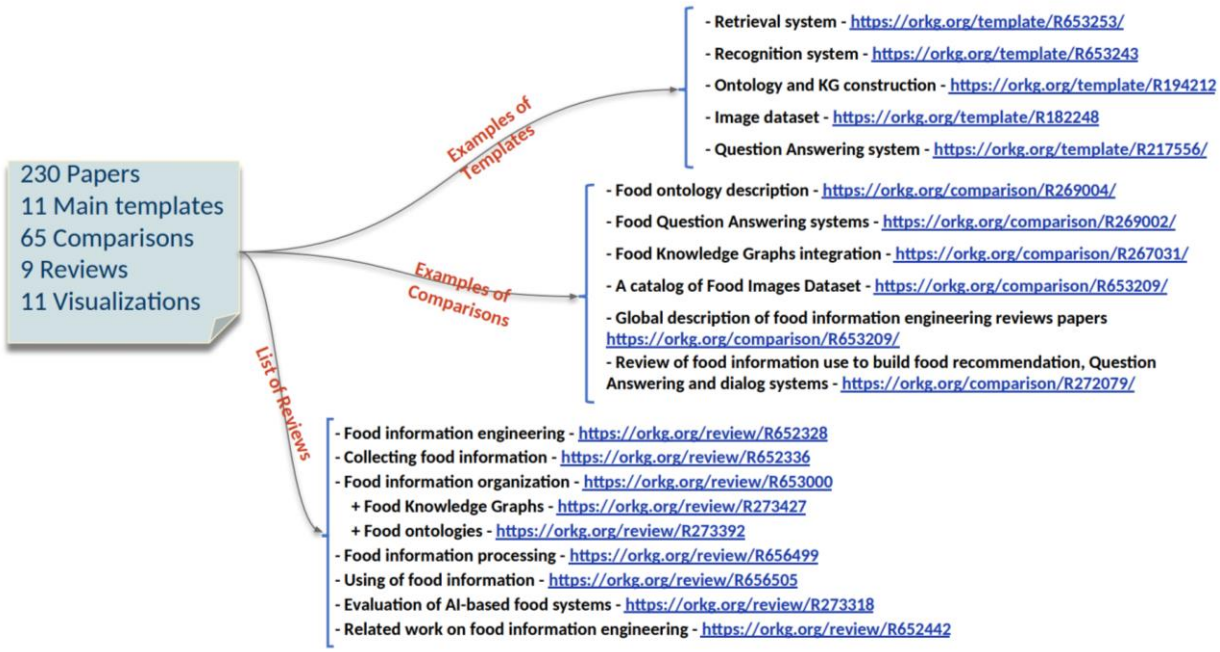

Figure 11.1 An overview of food information engineering observatory

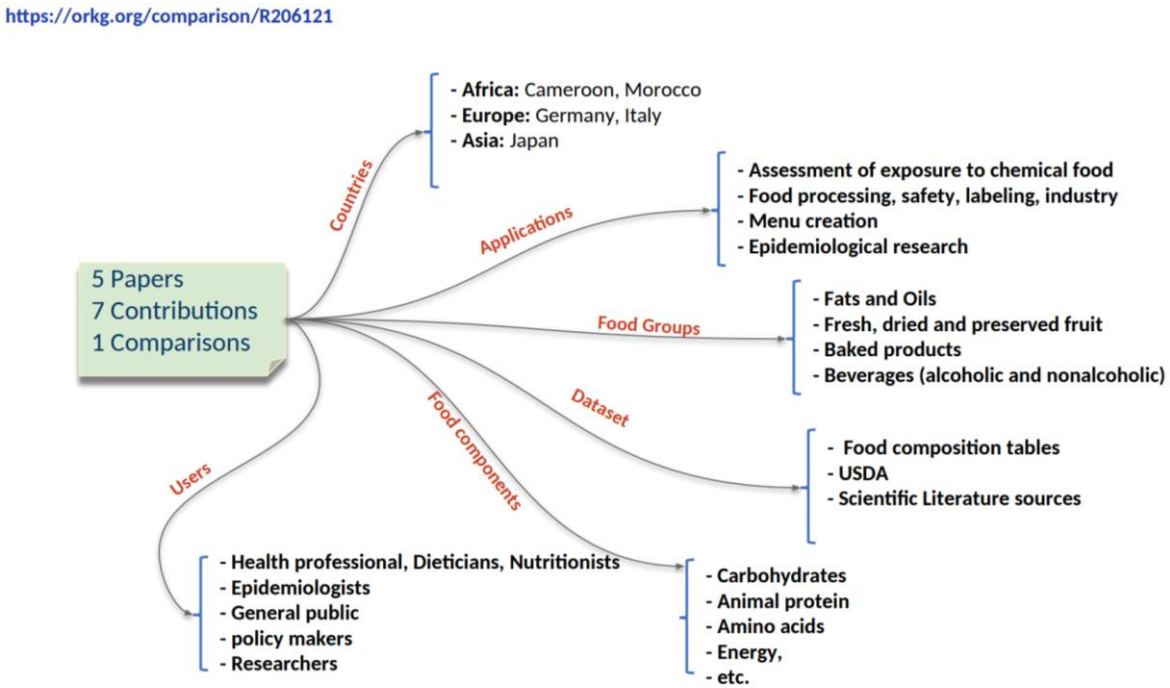

Figure 11.2 Food composition tables

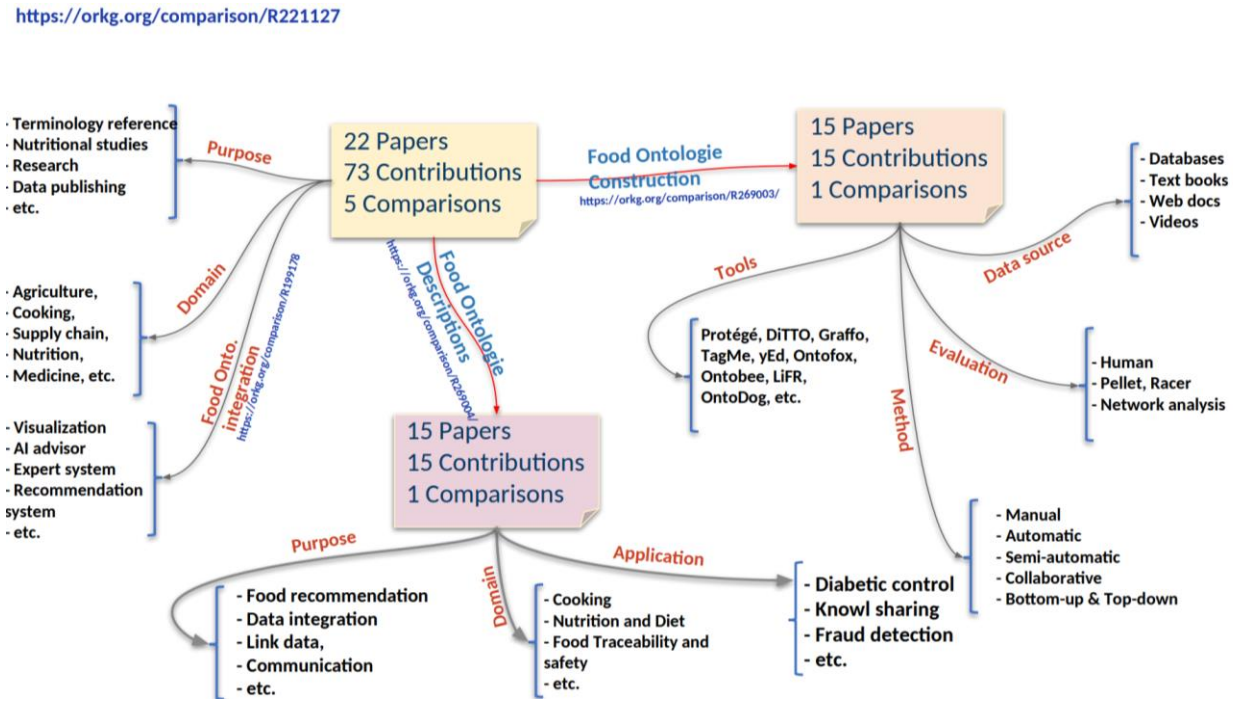

Figure 11.3 Food ontologies

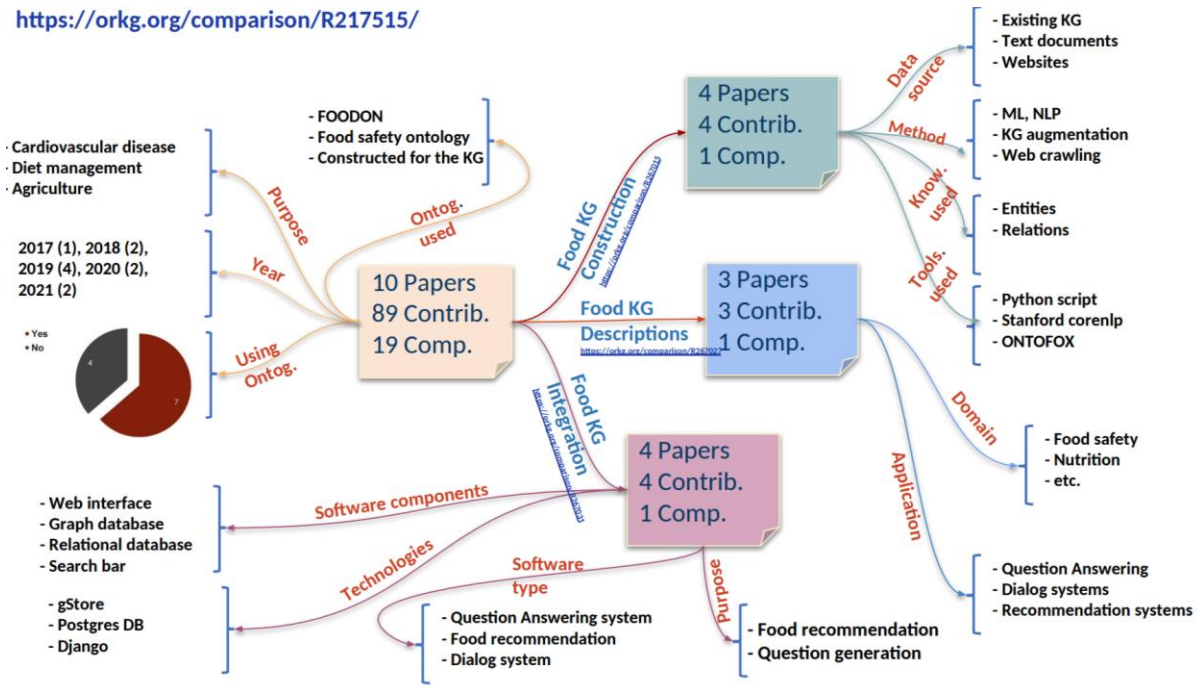

Figure 11.4 Food Knowledge Graph

Comparison tables related to different research problems of food information engineering are also presented in Figure 11.1. Food composition tables are compared according to countries, application, users, datasets used to build, food components and food groups described in the paper are shown in Figure 11.2.

In addition, several comparison tables related to particular topics such as food composition tables (Figure 11.2), food ontologies (Figure 11.3), food knowledge graph (Figure 11.4) and food question answering and dialog systems (Figure 11.5) are also provided. These figures present the different properties used to compare research papers.

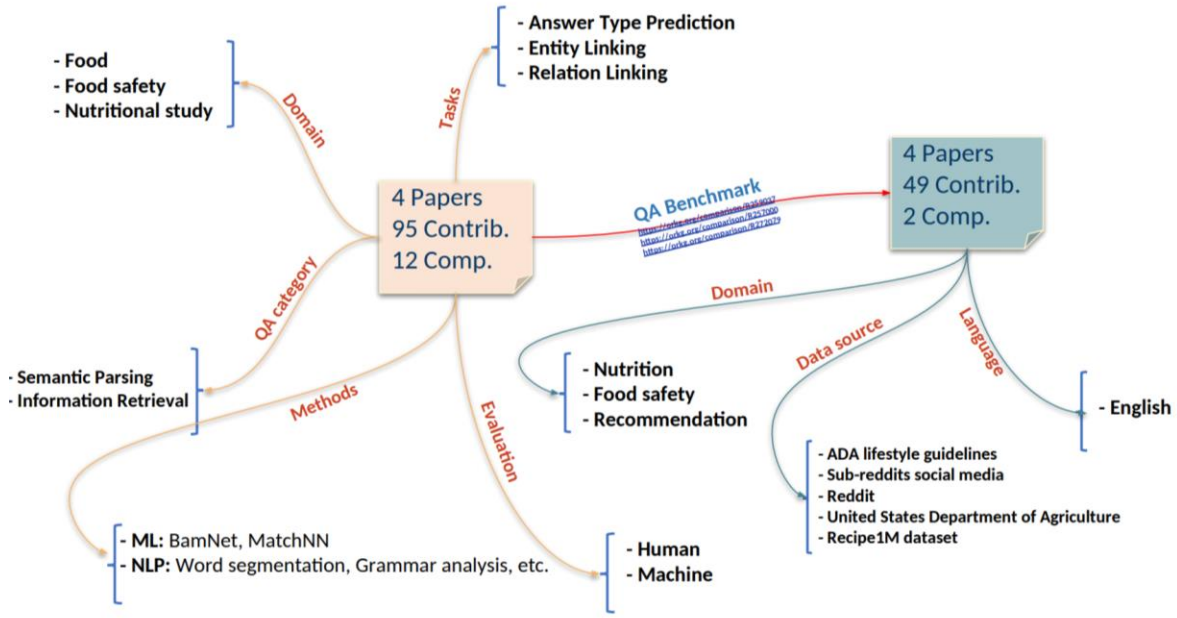

Figure 11.5 Food Question Answering

Finally, smart reviews presenting an overview of the different topics of food information engineering research are provided. Currently, nine smart reviews are included in the observatory.

Food information engineering https://orkg.org/review/R652328 introduces food information engineering, the research methodology being used to curate the observatory and link to collecting, organizing, processing and using food information.

Collecting food information (https://orkg.org/review/R609854) presents an overview of methodologies and tools for collecting food information. Thereafter, it presents the different food datasets.

Food information organization (https://orkg.org/review/R640407) presents different means to organize food information. This review is linked to other reviews describing the organization of food information using food ontologies (https://orkg.org/review/R273392) and food knowledge graph (https://orkg.org/review/R273427).

Food information processing (https://orkg.org/review/R640411) presents methodologies and tools for processing food information

Using food information (https://orkg.org/review/R640411) presents the different stakeholders and how they use food information.

Related work on food information engineering (https://orkg.org/review/R646622) aims to provide a global view of other state of the art research on food information engineering.

11.4 Summary and conclusion

This chapter presents food information engineering and how food information engineering research can be organized. This organization consists of extracting and storing scientific knowledge into the Open Research Knowledge Graph. Currently, around 230 scientific papers are added in the observatory and more papers will be added in the future days.

References

WHO, Malnutrition, 2023. URL: https://www.who.int/news-room/fact-sheets/detail/malnutrition/

A. Jiomekong, Food information engineering: A systematic literature review, volume 37, 2023, pp. 15441–15441. https://doi.org/10.1609/aaai.v37i13.26808 .

M. Kassie, P. Zikhali, K. Manjur, S. Edwards, Adoption of sustainable agriculture practices: Evidence from a semi-arid region of ethiopia, Natural Resources Forum 33 (2009) 189–198. doi:https://doi.org/10.1111/j.1477-8947.2009.01224.x

M. Bortolini, M. Faccio, E. Ferrari, M. Gamberi, F. Pilati, Fresh food sustainable distribution: cost, delivery time and carbon footprint three-objective optimization, Journal of Food Engineering 174 (2016) 56–67. doi:https://doi.org/10.1016/j.jfoodeng.2015. 11.014.

A. Meybeck, V. Gitz, Sustainable diets within sustainable food systems, in: Proceedings of the Nutrition Society, volume 76, 2017, p. 1–11. doi:10.1017/S0029665116000653.

W. Min, S. Jiang, L. Liu, Y. Rui, R. Jain, A survey on food computing, ACM Comput. Surv. 52 (2019). doi:10.1145/3329168.

W. Min, C. Liu, L. Xu, S. Jiang, Applications of knowledge graphs for food science and industry, Patterns 3 (2022) 100484. doi:https://doi.org/10.1016/j.patter.2022. 100484.

S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. Eddine Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving access to scientific literature with knowledge graphs, BIBLIOTHEK – Forschung und Praxis (2020). doi:http://dx.doi.org/10.18452/ 22049.

A. A. Metwally, A. K. Leong, A. Desai, A. Nagarjuna, D. Perelman, M. Snyder, Learning personal food preferences via food logs embedding, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE Computer Society, Los Alamitos, CA, USA, 2021, pp. 2281–2286. doi:10.1109/BIBM52615.2021.9669643.

FAO, Fao/infoods food composition databases, https://www.fao.org/infoods/infoods/tablesand-databases/faoinfoods-databases/en/, 2022. [Online; accessed 2023-09-16].

A. Jiomekong, Food information organization, https://orkg.org/review/R640407, 2023 a. [Online; accessed 2023-09-16].

A. Jiomekong, M. Folefac, H. Tapamo, Food composition knowledge extraction from scientific literature, in: S. Tiwari, F. Ortiz-Rodríguez, S. Mishra, E. Vakaj, K. Kotecha (Eds.), Artificial Intelligence: Towards Sustainable Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 89–103.

A. Jiomekong, Collecting food information, https://orkg.org/review/R609854, 2023 b. doi:10. 48366/r609854, [Online; accessed 2023-09-16].

J. Marín, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, A. Torralba, Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 187–203. URL: https://doi.org/10.1109/TPAMI. 2019.2927476. doi:10.1109/TPAMI.2019.2927476.

A. Jiomekong, M. Uriel, H. Tapamo, G. Camara, Semantic annotation of tsotsatable dataset, in: SemTab@ISWC, 2023 a.

L. Bossard, M. Guillaumin, L. Van Gool, Food-101 – mining discriminative components with random forests, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 446–461.

S. Yagcioglu, A. Erdem, E. Erdem, N. Ikizler-Cinbis, RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1358–1368. doi:10.18653/v1/D18-1166.

A. Jiomekong, Food ontologies, 2022. https://orkg.org/review/R273392.

A. Jiomekong, Food knowledge graphs, 2022 a. https://orkg.org/review/R273427.

A. Jiomekong, Food information processing, https://orkg.org/review/R640411, 2022 b. [Online; accessed 2023-09-16].

A. Parodi, A. Leip, I. De Boer, P. Slegers, F. Ziegler, E. H. Temme, M. Herrero, H. Tuomisto, H. Valin, C. Van Middelaar, et al., The potential of future foods for sustainable and healthy diets, Nature Sustainability 1 (2018) 782–789. doi:10.1038/s41893-018-0189-7.

A. Jiomekong, Using of food information, https://orkg.org/review/R640415, 2022 c. [Online; accessed 2023-09-16].

A. Jiomekong, Food information engineering, https://orkg.org/review/R652328 [Online; accessed 2024-01-31].

A. Jiomekong, Related work on food information engineering, https://orkg.org/review/R642258 , 2023 c. doi:https://doi.org/10.48366/R646622, [Online; accessed 2023-10-30]

Afterword

As we conclude our journey through the pages of this book, which commemorates the fifth anniversary of the ORKG, we stand on the brink of an exciting new era. The chapters you have explored provide a foundational conceptual framework designed to help even non-technical readers grasp the potential and functionalities of the ORKG. Staying true to experimenting emerging technologies, some authors have used ChatGPT in drafting the initial content for some chapters and carefully reviewed and revised the text for accuracy, clarity, and tone, adding references to support the information presented. As with all endeavours at the frontier of knowledge, the journey does not end here.

The next chapter of the ORKG is not just about technology; it is about community. We invite you to join us in this ongoing endeavour. Participate in our future challenges, where you can contribute to testing and refining these new tools. Your involvement will help shape the evolution of the ORKG, ensuring it remains a dynamic resource that continues to meet the needs of its diverse user base.

We encourage you to use the knowledge and insights from this book as a springboard for your own exploration of the ORKG. Whether you are a researcher looking to structure your data more effectively, a scholar eager to discover interconnected research insights, or a curious mind aspiring to contribute to a domain-specific knowledge graph, there is a place for you in the ORKG community.

Glossary

Scholarly Communication: The process by which academics, scientists, and researchers share and publish their findings so that they are available to the wider academic community and beyond.

Open Data: Data that is freely available to everyone to use and republish as they wish, without restrictions from copyright or other mechanisms of control.

Semantic Web: A set of standards promoted by the World Wide Web Consortium (W3C) that enable users to create data stores on the Web, build vocabularies, and write rules for handling data.

FAIR Principles: Guidelines that aim to enhance the ability of machines to automatically find and use data, and support its reuse by individuals. Stands for Findable, Accessible, Interoperable, and Reusable.

Knowledge Graph (KG): A network of entities (nodes) and their interrelationships (edges), structured as a graph, used to model complex sets of data and their interactions.

Triple: A triple is the fundamental data structure in semantic web technologies and knowledge graphs. It consists of three components: a subject, a predicate, and an object. The subject is the resource being described, the predicate is the property that defines the relationship or attribute, and the object can be another resource or a literal. Triples are used to make assertions about resources and their relationships, effectively building the graph's structure.

Metadata: Data that provides information about other data, used to help understand, use, and manage the data.

Linked Data: A method of publishing structured data that allows data to be interconnected and become more useful through semantic queries.

API (Application Programming Interface): A set of rules and protocols for building and interacting with software applications, which allows different software programs to communicate with each other.

JSON-LD (JavaScript Object Notation for Linked Data): A method of encoding linked data using JSON, facilitating the easy interchange of data on the Web.

SPARQL (SPARQL Protocol and RDF Query Language): A query language and protocol used for querying and managing data stored in Resource Description Framework (RDF) format.

Interoperability: The ability of different systems, platforms, or organizations to work together and share data seamlessly.

Ontology: In the context of knowledge management, an ontology represents a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain.

Data Curation: The process of organizing, integrating, and managing data collected from various sources. It includes annotation, publication, and presentation of the data to ensure that it is maintained over time and remains available for reuse and preservation.

DOI (Digital Object Identifier): A unique alphanumeric string assigned to identify a digital object, such as an electronic document, and provide a persistent link to its location on the Internet.

Resource: In the context of knowledge graphs, a resource refers to any identifiable entity or concept that can be described within the graph. Resources are typically represented as nodes in the graph and can include things like people, places, concepts, or any other objects relevant to the domain of the knowledge graph.

Property: A property in a knowledge graph defines the attributes or relationships of resources. It acts as an edge connecting two nodes in the graph or as an attribute that describes a specific characteristic of a node. For example, in a knowledge graph about books, a property might connect authors to their books or define attributes like the genre or publication year of a book.

Classes: Classes are the categories or types into which resources are grouped in a knowledge graph. They represent the general concepts under which resources are classified, such as 'Person', 'Organization', 'Event', etc. Classes help in structuring the knowledge graph by defining common characteristics shared by resources within the same class.

Literals: Literals are specific values or constants used to define the properties of resources in a knowledge graph. They are basic, non-decomposable values such as strings, numbers, or dates. For example, the birthdate of a person or the name of a city would be represented as literals in a knowledge graph.

Machine Learning: A branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention.

LLMs (Large Language Models): Advanced artificial intelligence models trained on extensive datasets to understand, generate, and manipulate natural language text. LLMs can interpret complex queries, provide information, and assist in generating human-like text based on patterns learned from their training data.