Gottfried Wilhelm Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik

Born-reusable scientific knowledge: Concept, implementation, and applications

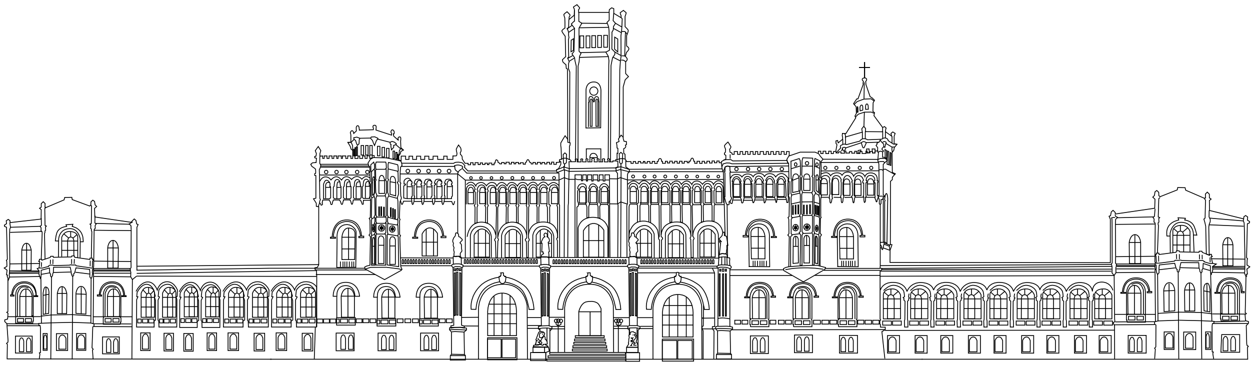
A thesis submitted in fulfillment of the requirements for the degree of
Bachelor of Science in Computer Science

BY

Matthew Anfuso
Matriculation number: 10016428
E-mail: matthew.anfuso@stud.uni-hannover.de

First evaluator: Prof. Dr. Sören Auer
Second evaluator: Dr. Markus Stocker
Supervisor: Dr. Markus Stocker

20 July 2023



Declaration of Authorship

I, Matthew Anfuso, declare that this thesis titled, 'Born-reusable scientific knowledge: Concept, implementation, and applications' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Matthew Anfuso

Signature: ______________________

Date: 20.07.2023

“Everything should be made as simple as possible, but not simpler.”

— Albert Einstein

Acknowledgements

First and foremost, I'd like to thank my supervisor Dr. Markus Stocker for his extraordinary patience, support, and guidance whilst writing this thesis. I'd also like to thank Prof. Dr. Sören Auer for providing me with the opportunity to conduct my thesis in collaboration with the ORKG. Furthermore, I'd like to thank the following: Dr. Lauren Syder for her constant support, particularly in the fields of research synthesis and agroecology; Dr. Ricardo Perez-Alvarez, Dr. Norman Gentsch, and Dr. Daniel Paredes for making this thesis possible by generously granting access to their papers and associated data; Dr. Lars Vogt for his invaluable insights into modeling and ontologies. Last, but certainly not least, the ORKG team for their behind-the-scenes work that keeps the ORKG up and running.

Abstract

The exponentially increasing growth of scientific literature publication presents a significant challenge to effectively read, process, and fully comprehend the wealth of scientific knowledge. The Open Research Knowledge Graph (ORKG) aims to address this challenge by providing infrastructure that aligns with the FAIR principles, to support the creation, curation, and utilization of scientific knowledge. Nevertheless, the current dependence on crowdsourcing and natural language processing (NLP) for post-publication knowledge extraction restricts the scalability and quality of such knowledge bases. In response to these challenges, we present a novel 'born-reusable' approach that seeks to create richly-detailed, machine-reusable descriptions of papers directly within the computing environment where the research was conducted, thus placing the onus on authors to ensure their research findings are FAIR prior to publication. With the help of the ORKG R package, salient scientific knowledge is captured from the paper's associated R source code and serialized to a machine-reusable format (JSON-LD) for harvesting by the ORKG by DOI-lookup. By applying this approach to an unpublished soil science manuscript, we demonstrated how authors are best situated to describe their work in a richly-detailed machine-reusable format. Furthermore, by applying this approach to two published agroecology papers, we demonstrated its relevance to post-publication, thus suggesting that papers which share source code and data sets could be made machine-reusable retrospectively. Finally, a proof-of-concept meta-analysis was conducted to demonstrate how this approach can help facilitate research synthesis by providing FAIR scientific data. We concluded that the 'born-reusable' approach has promising implications for the reusability of scientific knowledge. However, its broad adoption faces several challenges. Therefore, solutions were explored to improve the approach's interoperability with knowledge graphs, assist authors with its implementation into their workflows, and strengthen cooperation with publishers to provide the necessary infrastructure.

Keywords: Knowledge Graph, FAIR, ORKG, Scientific Knowledge

Zusammenfassung

Das exponentiell steigende Wachstum der Veröffentlichung wissenschaftlicher Literatur stellt eine große Herausforderung dar, um das enorme Wissen in der Wissenschaft effektiv zu lesen, zu verarbeiten und vollständig zu verstehen. Der Open Research Knowledge Graph (ORKG) zielt darauf ab, dieser Herausforderung zu begegnen, indem er eine Infrastruktur bereitstellt, die den FAIR-Prinzipien entspricht, um die Erstellung, Kuratierung und Nutzung wissenschaftlichen Wissens zu unterstützen. Dennoch schränkt die derzeitige Abhängigkeit von Crowdsourcing und Natural Language Processing (NLP) zur Wissensextraktion nach der Veröffentlichung, die Skalierbarkeit und Qualität solcher Wissensdatenbanken ein. Als Antwort auf diese Herausforderungen präsentieren wir einen neuartigen 'born-reusable' Ansatz, der darauf abzielt, detailreiche, maschinenverwendbare Beschreibungen von wissenschaftlichen Arbeiten direkt in der Computerumgebung zu erstellen, in der die Forschung durchgeführt wurde. Dies legt die Verantwortung auf die Autoren, um sicherzustellen, dass ihre Forschungsergebnisse vor der Veröffentlichung FAIR sind. Mit Hilfe des ORKG R Package werden wichtige wissenschaftliche Erkenntnisse aus dem zugehörigen R-Quellcode wissenschaftlicher Arbeiten erfasst und in ein maschinenverwendbares Format (JSON-LD) serialisiert, das anschließend vom ORKG mittels DOI-Abfrage gewonnen wird. Durch Anwendung dieses Ansatzes auf ein unveröffentlichtes Manuskript der Bodenwissenschaft haben wir demonstriert, wie Autoren am besten in der Lage sind, ihre Arbeit in einem detaillierten, maschinenverwendbaren Format zu beschreiben. Darüber hinaus haben wir mit dem gleichen Ansatz zwei veröffentlichte Agroökologie-Arbeiten auf seine Relevanz für die Nachveröffentlichung demonstriert, was darauf hindeutet, dass wissenschaftliche Arbeiten, die den Quellcode und Datensätze teilen, rückwirkend maschinenverwendbar gemacht werden könnten. Schließlich wurde eine Proof-of-Concept-Metaanalyse durchgeführt, um zu demonstrieren, wie dieser Ansatz die Forschungssynthese durch Bereitstellung von FAIR-Wissenschaftsdaten erleichtern kann. Wir kamen zu dem Schluss, dass der 'born-reusable' Ansatz vielversprechende Auswirkungen auf die Wiederverwendbarkeit wissenschaftlichen Wissens hat. Die weitreichende Einführung steht jedoch vor mehreren Herausforderungen. Daher wurden Lösungen untersucht, um die Interoperabilität dieses Ansatzes mit Wissensgraphen zu verbessern, Autoren bei der Implementierung in ihre Arbeitsabläufe zu unterstützen und die Zusammenarbeit mit Verlagen zu stärken, um die notwendige Infrastruktur bereitzustellen.
Stichwörte: Knowledge Graph, FAIR, ORKG, wissenschaftliches Wissen

Contents

1 Introduction **1**
1.1 Problem Statement 2
1.2 Research Questions 4
1.3 Structure 4

2 Background **6**
2.1 The FAIR Principles 6
2.2 Open Research Knowledge Graph (ORKG) 7
2.3 ORKG-R Package 8
2.4 Notebook computing environments 9

3 Related Work **10**
3.1 Knowledge bases for scientific information 10
3.1.1 Papers with Code 10
3.1.2 The Cooperation Databank 11
3.1.3 OpenBiodiv 11
3.1.4 The Biology Knowledge Graph 11
3.1.5 Hi Knowledge 12
3.2 Scientific knowledge base production and curation 12
3.2.1 Post-publication 12
3.2.2 At time of publication 15
3.2.3 Pre-publication 16

4 Approach **18**
4.1 Model 19
4.2 Produce 21
4.3 Serialize 24
4.4 Link 26
4.5 Harvest 26

5 Application **30**

5.1 Application to a manuscript . . . 30

5.1.1 Model Creation . . . 32

5.1.2 Template Creation . . . 33

5.1.3 Template Population . . . 34

5.1.4 Harvesting . . . 35

5.2 Application to a proof-of-concept synthesis . . . 38

5.2.1 Model Creation . . . 39

5.2.2 ORKG Contribution Creation . . . 40

5.2.3 Meta analysis . . . 40

6 Discussion **46**

6.1 Implications . . . 46

6.2 Limitations . . . 47

6.3 Future Work . . . 48

7 Conclusions **50**

Bibliography **51**

List of Figures

1.1 The born-reusable approach 3

4.1 Born-reusable process 18
4.2 The modeling process 19
4.3 An example of an ORKG template 20
4.4 Step 1: Harvesting through DOI-lookup using the web interface. 28
4.5 Step 2: Selecting an appropriate research field for the paper. 28
4.6 Step 3. The contributions are generated ready for editing and saving. . . . 29

5.1 Manuscript source materials 31
5.2 Identification and selection of key scientific information. 31
5.3 A conceptual model illustrating the statistical processes. 32
5.4 Resulting ORKG contribution. 35
5.5 The “has implementation” property. 36
5.6 The“has output figure” property. 37
5.7 The “has output dataset” property. 37
5.8 Graph representing the conceptual model for the research contributions. . . 41
5.9 The six contributions from “Contrasting effects of landscape composition on crop yield mediated by specialist herbivores” 43
5.10 The contributions from “Landscape simplification increases Bactrocera oleae abundance in olive groves”. 44
5.11 The results of the meta-analysis inside a Jupyter notebook. 45

List of Tables

5.1 Possible template function parameters. 34
5.2 ORKG Templates inspired by the “Planned Process” conceptual model . . 42

Acronyms

CoDa The Cooperation Databank

CRAN Comprehensive R Archive Network

DOI Digital Object Identifier

FAIR Findable, Acessible, Interoperable, Reusable

JSON-LD JavaScript Object Notation for Linked Data

LMM Linear Mixed Model

ML Machine Learning

NLP Natural Language Processing

ORKG Open Research Knowledge Graph

RDF Resource Description Framework

SciKGTeX Scientific Knowledge Graph TeX

SHACL Shapes Constraint Language

SI Supplementary information

SoTA State-of-the-art

SPARQL SPARQL Protocol and RDF Query Language

URI Universal Resource Identifier

Chapter 1

Introduction

Science is grappling with the exponential growth in scientific literature [1]. It is estimated that in 2018 alone, English-language peer-reviewed journals collectively published over 3 million papers [2]. This publication rate vastly exceeds our capacity to read, process and fully comprehend this flood of new scientific information. This has led to valuable scientific knowledge remaining buried and forgotten within the millions of largely unstructured digital documents such as PDFs. To address this issue, it is critical that scientific data underlying this knowledge is first made reusable. The challenge of scientific data reusability was one of the motivating factors for the introduction of the FAIR principles in 2016. The FAIR principles aim to make scientific data more findable, accessible and interoperable with the help of machines, and ultimately reusable for both machines and humans [3].

Prior to the introduction of the FAIR principles, linked data principles and knowledge graphs were introduced as a method to capture and structure the ever growing volume of factual knowledge online. Examples include DBpedia in 2007 [4] and Google Knowledge Graph in 2012 [5]. In a knowledge graph, information that represents real world knowledge is stored in a directed graph structure where the nodes represent entities and their edges represent the relationships between entities [6]. This form of representation captures the context and the semantic relationships between data.

As a scientific knowledge graph, the Open Research Knowledge Graph (ORKG)[1] seeks to provide infrastructure to support the production, curation, and use of scientific knowledge in alignment with the FAIR principles [7]. It presents scientific

[1] https://www.orkg.org/

contributions in a structured format that is both machine-reusable and user-friendly. It allows data access through a range of options such as SPARQL queries, a RESTful API and libraries for both Python and R. Currently, ORKG content is acquired post-publication of the article, through a combination of human and machine-based approaches such as crowdsourcing and direct data harvesting from third-party knowledge repositories. Researchers at the ORKG have also explored the possibility of using natural language processing (NLP) [8] and hybrid human-machine approaches [9] for the efficient production of high quality structured scientific knowledge.

1.1 Problem Statement

Although knowledge graphs, such as the ORKG, have been shown to be promising platforms for publishing FAIR scientific knowledge, the current reliance on crowdsourcing and NLP to extract scientific information from literature post-publication limits their scalability, richness and overall quality.

Crowdsourcing has successfully been used to construct and curate large factual knowledge graphs such as Wikidata [2] and DBpedia [4]. However, due to the complexity of scientific knowledge, applying crowdsourcing in the research context relies on domain experts in order to ensure rich and accurate results. This restricts the number of potential contributors, increases overall costs and limits scalability.

With the introduction of machine learning (ML) models such as BERT [10], natural language processing (NLP) has made rapid progress in recent years, particularly in the realm of general knowledge graph construction. Nevertheless, research into using NLP for knowledge graph construction from scientific literature is still limited to very specific scientific fields and tasks [11]. Therefore, the use of NLP to construct highly detailed (rich) knowledge graphs from scientific literature still remains impractical [12].

Another limiting factor facing post-publication techniques is the lack of findability and accessibility of associated data such as data sets and source code. Even with the rise of open data and the adoption of data sharing statements, many researchers are not compliant with their statements [13], many fail to use appropriate repositories [14], and many links to data become inaccessible over time [15].

Pre-publication FAIR scientific knowledge production is an alternative to post-publication approaches for scientific knowledge extraction. Pre-publication methods

[2] https://www.WikiData.org/

place the responsibility on authors to make their research findings FAIR prior to the publication of the corresponding paper. Authors have a deep and unique understanding of their own work, which places them in the best position to accurately represent their research in a rich, fine granular, machine-reusable format during the research phase. Authors also have access to the data and program code used to produce the published scientific knowledge, often exclusively.

In this thesis we focus on a pre-publication approach to producing FAIR scientific knowledge - an approach we have titled 'born-reusable'. Specifically, we propose an approach to ensure scientific knowledge is produced machine-reusable during the data analysis phase of the research life cycle, particularly when scientific knowledge is produced in computing environments.

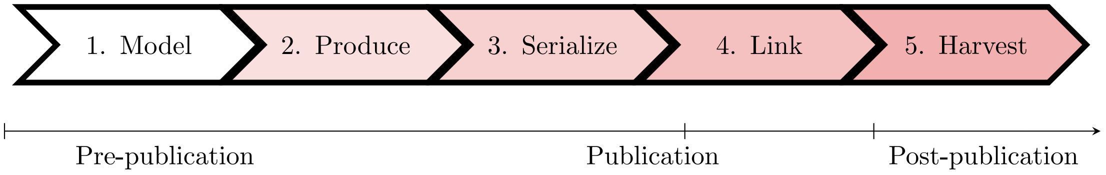


Figure 1.1: The born-reusable approach

The born-reusable approach can be summarized in 5 steps:

1. Model: First, the author identifies the key scientific contributions of their research and determines how the associated data should be modeled in a detailed and semantically rich.

2. Produce: On hand of the model, the scientific knowledge is richly and accurately adapted to a machine-reusable structure from directly within the (statistical) computing environment (e.g., R, Python, MATLAB).

3. Serialize: Subsequently, the structured description of scientific knowledge is then serialized to a machine-reusable format such as JSON-LD and stored within the same programming environment.

4. Link: The serialized machine-reusable descriptions of scientific knowledge and supplementary data are interlinked with the corresponding article using its DOI metadata.

5. Harvest: Finally, machine-reusable data is harvested post publication by scientific knowledge repositories through the use of the publication's DOI.

Although the born-reusable approach is potentially applicable to any programming language and knowledge repository, in this thesis we will focus on the ’born-reusable’ approach as it applies to the R programming language and the ORKG.

1.2 Research Questions

RQ1. How can we produce rich and high-quality machine machine-reusable scientific knowledge?

To address this question, we will apply the born reusable approach to an unpublished manuscript in the field of soil science and two published scientific papers in the field of agroecology. We will employ the pre-existing ORKG R package (see related work) to extract vital statistical and scientific information from the R source code supporting the paper. This information will be structured with the guidance of ORKG templates - predefined structures that help ORKG users to compose their contributions in a standardized way. The extracted information will then be serialized into a machine-reusable format, JSON-LD, stored in a repository, and subsequently linked to within the metadata associated with the paper’s DOI. Finally, the JSON-LD data will be automatically harvested by the ORKG using the paper’s DOI.

RQ2. How can rich and high quality machine-reusable scientific knowledge support research synthesis?

To answer this question, we will conduct a small scale proof-of-concept synthesis, in the form of a meta-analysis, using the two published papers in the field of agroecology that have undergone the born-reusable approach. The papers will then be harvested by the ORKG so that the statistical data required can be directly imported into qn R environment (Jupyter notebook) to help facilitate the meta-analysis.

1.3 Structure

This thesis is structured as follows. Chapter 1 explores the background research that forms the foundation of this thesis. Chapter 2 investigates related work, including the current state of making scientific knowledge reusable through knowledge graphs and the approaches used to extract the scientific knowledge needed for their construction. In Chapter 3, the born-reusable architecture and process are described in detail. In Chapter 4, the born-reusable approach is applied to an unpublished manuscript in the field of soil science, and a proof-of-concept synthesis is conducted to investigate the reusability of born-reusable scientific knowledge in the field of agroecology. Finally, in Chapter 5, the implications and limitations of this approach are discussed as well as potential areas of future research are suggested.

Chapter 2

Background

In this chapter, we explore the research that forms the foundation for this thesis. First, we delve further into the FAIR principles, which are one of the guiding principles for the born-reusable approach. Additionally, we will detail the ORKG knowledge graph and the ORKG R package, which form the technological framework for this thesis.

2.1 The FAIR Principles

The FAIR guiding principles were first outlined in 2016 with the aim of improving the reuse of scientific data [3]. The four foundational principles aim to make digital assets more findable, accessible, interoperable, and reusable.

Findability emphasizes that data and metadata must be identifiable through a unique and persistent identifier such as a DOI and the data must be described using detailed metadata to enable searchability and findability. Additionally, the data and its related metadata should be included in some form of searchable trusted data repository.

Accessibility mandates that metadata and data are retrievable using their identifier over a free and open protocol that provides the option for authentication and authorization, e.g. HTTPS. It is also important that the metadata remains accessible even if the underlying data is not.

Interoperability requires that metadata and data are represented in a machine readable format such as XML, RDF or JSON-LD and that they use standardized vocabularies that also adhere to the FAIR principles. Additionally, the metadata should also include references to any other metadata that helps to fully describe it.

Finally, reusability requires that data and metadata be richly and accurately described. Otherwise, the data provides little practicality for reuse. Also, it is essential to provide data provenance, i.e., where the data comes from, who owns it and who modified it. It is also essential that it is made clear under which usage license the data can be reused.

The born-reusable approach aims to improve reusability, but it should be noted that reusability is highly interconnected with the first three FAIR principles. Data that is not findable is, per definition, not reusable. Once the data is found, it must be technically accessible to enable its reuse. Finally, data, whether text, code or datasets, must be interoperable with other tools and software in order to support its reuse.

2.2 Open Research Knowledge Graph (ORKG)

The ORKG[1]. seeks to provide infrastructure to support the creation, curation and use of FAIR scientific knowledge through means of a knowledge graph [7]. The ORKG knowledge graph allows for fine-grained machine-reusable descriptions of published scientific findings that help promote the reusability of scientific knowledge. Accessibility is maintained through a graphical web interface, RESTful API and SPARQL endpoint[2].

One way in which the ORKG promotes reusability is through enabling researchers to create what are known as comparisons. Comparisons allow researchers to extract relevant information from various research papers and map them into tabular summaries which also allow for filtering and sorting [16]. Additionally, comparisons enable the mapping of properties within contributions to create custom visualizations. Once published, the comparison is assigned a DOI to enhance findability in the global scholarly infrastructure. A practical example of the application of an ORKG comparison is the visualization of published estimated reproductive numbers (R0 value) during the COVID-19 pandemic [17].

The ORKG currently relies on post-publication methods for extracting scientific information from scientific literature. One utilized method, crowdsourcing, is facilitated through the support of curation grants where researchers from various fields

[1] https://www.orkg.org/

[2] https://www.orkg.org/data

contribute ORKG entries over a period of 6 months [18]]. Additionally, the ORKG employs direct data harvesting from other knowledge repositories such as Papers with Code[3].

To help improve scalability, the ORKG has investigated the use of NLP for assisting the construction and curation of the knowledge graph [19]. One example, the Python ORKG-NLP package[4] provides various NLP services tailored to the ORKG, such as named entity recognition (NER) for titles and abstracts in computer science and agriculture publications, as well as ORKG template and research field recommendations based on a paper's abstract [20]. The addition of hybrid human-machine approaches that combine the accuracy of humans with the scalability of NLP has also been explored [21].

2.3 ORKG-R Package

The ORKG R package [22], is software developed for the R language. It works in conjunction with the ORKG API to facilitate the retrieval of ORKG data, such as resources and templates, directly into the R environment. The ORKG R package is highly inspired by and shares a lot of the functionality with the preexisting ORKG Python library [5].

A key component of the ORKG-R package, especially as it relates to the born-reusable approach, is the ability to dynamically create functions that support the creation of structured data based on ORKG templates. It is also capable of serializing this structured data to JavaScript Object Notation for Linked Data (JSON-LD). JSON-LD is a method of encoding Linked Data (i.e RDF) in a standardized serialized format.

The ORKG R package has not been included in the Comprehensive R Archive Network (CRAN), and therefore must be installed either locally or remotely from its GitLab repository[6].

[3] https://www.paperswithcode.com/

[4] https://orkg-nlp-pypi.readthedocs.io/

[5] https://orkg.readthedocs.io/en/latest/

[6] https://gitlab.com/TIBHannover/orkg/orkg-r

2.4 Notebook computing environments

Notebook computing environments are interactive documents that combine code, text, visualizations, and computational outputs, allowing computational processes to be easily shared, understood and replicated [23].

Jupyter [7] is currently one of the most widely adopted notebook computing environments. This open-source software supports various programming languages, including Python and R, and provides interactive computing environments that can be executed locally or remotely via a web-based interface. Utilizing a markup language known as ”Markdown”, formatted text can accompany the code to provide clarity, making it particularly suitable for scientific computing. Notebooks are stored in JSON format with .ipynb extension.

[7] `https://jupyter.org`

Chapter 3

Related Work

In this chapter, we investigate the current state of making scientific knowledge FAIR using knowledge bases, in particular knowledge graphs. Additionally, we examine the various approaches used for extracting scientific information required for their construction.

3.1 Knowledge bases for scientific information

3.1.1 Papers with Code

Papers with Code [1] is an open-source platform supported by Meta AI, that combines abstracts, datasets, code, methods, and evaluation metrics from machine learning (ML) research papers in a machine-reusable format [24]. The specialized focus of Papers with Code allows for detailed, structured machine-reusable metadata that can be accessed using either the web-based GUI or API or through JSON data dumps. The platform also allows for the visual representations of state-of-the-art (SoTa) results in common machine learning tasks. It primarily relies on crowdsourcing for the curation of its knowledge base, assisted by the automated extraction of SoTA results from third-party repositories using their SoTA extractor [2]. This platform has been replicated to create portals focusing on other research fields, such as mathematics, physics and astronomy[3].

[1] `https://paperswithcode.com/`

[2] `https://github.com/paperswithcode/sota-extractor`

[3] `https://portal.paperswithcode.com/`

3.1.2 The Cooperation Databank

The Cooperation Databank[4] (CoDa) is a closed-source knowledge graph focusing on the field of human cooperation within social dilemmas, such as the prisoner's dilemma [25]. The scientific knowledge is extracted from empirical studies and papers, selected based on a systematic search, and annotated by domain experts guided by their own in-house ontology of human cooperation studies. At the time of this thesis's publication, the knowledge graph contained over 1800 papers and over 2600 studies. The use of domain experts and a standard ontology allows for rich machine-reusable descriptions that enable detailed search, visualizations and automated statistical analysis such as meta-regression and meta-analysis through their web interface [5].

3.1.3 OpenBiodiv

OpenBiodiv[6] is a biodiversity knowledge graph containing information automatically extracted biodiversity-related papers that have been semantically annotated [26]. OpenBiodiv provides a web interface, RESTful API and a SPARQL endpoint for access. These papers are semantically annotated by humans either post-publication or at the time of publishing. This information is then mapped to their self-developed OpenBiodiv-O ontology [27]. An ontology is an explicit formal specification of an abstract model [28]. The OpenBiodiv-O ontology includes model biodiversity-related concepts such as treatments and specimen information, allowing for highly detailed SPARQL queries. At the time of this thesis's publication, OpenBiodiv contained over 9600 papers.

3.1.4 The Biology Knowledge Graph

The Biology Knowledge Graph[7] is a commercial closed-source knowledge graph developed by Elsevier [29]. At the time of publishing this thesis, it contained 13.5 million biological relationships grouped into various biological categories such as gene expression and biomarkers. The construction and curation of the knowledge graph are conducted post-publication using NLP techniques. It is regularly updated with PubMed abstracts, full-text papers, clinical trials, and data from third-party

[4] https://cooperationdatabank.org/

[5] https://app.cooperationdatabank.org/

[6] https://openbiodiv.net/

[7] https://www.elsevier.com/solutions/biology-knowledge-graph/

databases such as Drugbank[8] and BioGRID[9]. Human experts consistently review the knowledge graph to ensure accuracy and quality.

3.1.5 Hi Knowledge

Hi Knowledge[10] provides an interactive tool for visualizing 39 hypotheses in the field of invasion biology. The web interface enables users to visualize in a graph-like representation the number of publications which are supporting, questioning, or undecided about specific hypotheses, as well as the relationships between them. The hypothesis network was constructed with the help of a group of domain experts using 39 pre-selected hypotheses [30]. Over 1000 published papers were manually categorized, and their structured information (authors, title, measured disturbance, taxonomic focus, etc.) was extracted [31]. Data accessibility is limited to the export of data in Excel format.

3.2 Scientific knowledge base production and curation

3.2.1 Post-publication

Post-publication approaches, as utilized by ORKG, CoDa, OpenBiodiv and the Biology Knowledge Graph, are currently among the most popular methods for scientific knowledge extraction. However, one limitation of post-publication approaches is the lack of findability and accessibility of the associated data, such as data sets and source code needed to make scientific knowledge reusable. Despite the ever-increasing popularity of open data and the increased adoption of Data Availablity Statements, many researchers are not compliant with their published data sharing statements [13]. Even when compliant, a large percentage of the data referenced within data sharing statements remains embedded in the paper (unstructured text, tables, figures) or is scattered as Supplementary Information (SI) in various data formats [14]. Exacerbating this problem is the fact that many of these shared resources become unavailable over time [15]. Despite these challenges, the use of NLP and

[8] https://go.drugbank.com/

[9] https://thebiogrid.org/

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

crowdsourcing post-publication remain two of the most frequently used techniques for extracting scientific knowledge.

Natural Language Processing (NLP)

In recent years, pre-trained deep learning models like BERT have made rapid progress in autonomously constructing general knowledge graphs as well as completing various essential sub-tasks such as entity recognition (NER), relation extraction (RE), and named entity linking (NEL) [32].

Although NLP has also made promising progress in specific scientific domains [33], using NLP to construct highly detailed knowledge graphs from the scientific literature is not yet feasible [12]. This has led researchers to focus on specialized narrow domain-specific models for scientific literature [34]. One such example in the computer science domain is SCICERO [35].

Its deep learning approach for creating scientific knowledge graphs achieved good precision and recall (F-measure 0.77) when applied to a gold-standard dataset of approximately 6.7 million titles and abstracts from computer science papers. The papers were annotated by human experts according to the "Computer Science Knowledge Graph Ontology" [11]. This ontology contains five entity classes (Tasks, Methods, Materials, Metrics and Other Entities) and 179 properties representing the relationships between entities. However, the authors noted that SCICERO is highly domain-specific, and its application in other scientific fields would require significant modifications. Additionally, they pointed out the high computational costs, specifically memory, as a limitation.

Wider adoption of such NLP models for scientific knowledge graph construction remains challenging for several reasons. Firstly, as exemplified by SCICERO, many pre-trained models, tasks and their associated ontologies are highly domain-specific and require significant modification for application in other scientific domains. Secondly, the substantial human and computational costs of training such models cannot be underestimated. Finally, it remains to be seen whether the scope and detail of the tasks they perform and the accuracy and recall achieved by these models suffice for the construction and curation of real-world scientific knowledge graphs without the need for costly human-supervised quality control.

[11] https://scholkg.kmi.open.ac.uk/cskg/ontology.html

Crowdsourcing

Crowdsourcing is an online activity where an entity (crowdsourcer) delegates a task to a diverse group of individuals (crowd). These individuals contribute their time, work, money, knowledge, or experience, leading to mutual benefit for both the crowdsourcer and crowd [36].

Crowdsourcing has proven to be an effective method for constructing and curating large general knowledge graphs such as Wikidata [12], which contains over 100 million items and has undergone 1,9 billion edits as of July 2023 [37].

Combining crowdsourcing with the support of scientific domain experts allows for highly accurate and standardized knowledge graphs in narrow scientific disciplines as exemplified by The Cooperation Databank (CoDa) [25]. Users can also manually add studies to CoDa along with detailed information such as metadata, treatments, sample and study characteristics, which are later manually reviewed by domain experts.

The ORKG utilizes crowdsourcing through voluntary contributions and the use of paid curation grants where researchers from various fields contribute over a period of 6 months [18]. Organizations can also monetarily sponsor these grants and promote their research field through the use of so-called “Observatories”. Observatories allow organizations to lead curation in specific research fields to gain exposure and to ensure high-quality standards for the ORKG[38].

Nevertheless, the scalability of crowdsourcing remains a challenge. Due to the high complexity of scientific knowledge, curation of scientific knowledge bases remains restricted to scientific domain experts. Also, due to the millions of scientific papers published each year [2] crowdsourcing remains unfeasible for capturing the huge volume of existing and newly created scientific knowledge.

Hybrid human-machine approaches

Hybrid post-publication approaches that combine the scalability of machine learning and the precision of humans have been proposed. One such hybrid approach, PANDA (Platform for Academic kNowledge Discovery and Acquisition), detects and extracts so-called ’knowledge cells’, e.g., tables, figures, and code from scientific papers in PDF format using a hybrid human-machine approach. The framework attempts to detect and extract knowledge cells using heuristics and machine learning. An

[12] https://www.wikidata.org/

algorithm assigns a confidence value for tasks, and any task defined as difficult is passed to a human worker for manual processing [39].

Another proposed hybrid approach, TinyGenius, uses NLP to process unstructured scientific texts and construct a scholarly knowledge graph. Following this, human evaluators manually evaluate the accuracy of the extracted statements via microtasks. The authors envision integrating this approach into the ORKG [40, 9].

The Biology Knowledge Graph represents a hybrid machine-human approach that does not incorporate crowdsourcing [29]. Construction and curation of the knowledge graph rely on machine-based approaches such as NLP, supplemented by expert human reviewers to maintain accuracy and quality.

3.2.2 At time of publication

The widespread adoption of DOIs and their associated metadata within scientific publishing has helped to enhance the findability and accessibility of scientific information. The movement towards open scientific data has also promoted increased sharing of scientific information through the use of supplementary information and data accessibility statements.

Scientific meta-data

Scientific papers use Digital Object Identifiers (DOIs) to provide a unique and persistent identifier to help facilitate their findability and accessibility online. The metadata schema associated with DOIs is usually standardized by the organization that assigns the DOI such as CrossRef [13] or DataCite [14]. This metadata is often harvested by scientific databases to provide information about authors, citations, and research fields. One such example, the Semantic Scholar Academic Graph [15] ingests metadata from various sources to construct a knowledge graph with papers, authors and citation relationships [41]. Connected Papers [16] uses the Semantic Scholar Academic Graph API to provide a visual overview of papers, their interconnections and relative importance within its scientific field [42].

A limiting factor in using DOI metadata for describing scientific knowledge is the limited scope of such metadata, i.e. titles, authors, keywords and other high-level de-

[13] `https://www.crossref.org/`

[14] `https://datacite.org/`

[15] `https://www.semanticscholar.org/product/api`

[16] `https://www.connectedpapers.com/`

scriptors. DOI metadata itself is not intended to describe scientific knowledge richly. However, this thesis explores DOI metadata's potential to facilitate the findability and accessibility of machine-reusable data that richly describes scientific knowledge through its ability to reference supplementary material.

Supplementary Information

The open data movement and the FAIR principles encourage researchers to share their data and code alongside their papers. The increased adoption of data availability statements is making it easier to find, request and access supplementary information provided by the authors [43].

The use of data repositories for supplementary information, such as [17], Zenodo[18] and Figshare[19], enable authors to share data with assigned DOIs, metadata, and usage licenses in accordance with the FAIR principles, thereby significantly improving data reusability.

Nevertheless, compliance with published data availability statements remains a problem among researchers, especially when data is promised based on request [13]. Many authors who provide links to supplementary information neglect to use FAIR repositories, and information remains embedded in the paper (unstructured text, tables, figures) or scattered as supplemental information in various data formats [14]. Although much supplementary information, such as data sets and code, is structured, it requires standardized vocabularies and rich metadata descriptions to facilitate the extraction of scientific knowledge. Exacerbating these problems is the fact that many of the links to supplementary information become unavailable over time, especially when a FAIR repository is not used [15].

3.2.3 Pre-publication

Authors have a unique and deep understanding of their own work and are best situated to accurately represent the nuances and complexities of their research in a machine-reusable format. They can also provide valuable data that may not be included in final publication, such as results directly from source code. Crowdsourcing, i.e. relying on non-authors post-publication, can lead to oversimplification or misrepresentation due to the lack of intimate knowledge of the content.

[17] `https://datadryad.org`

[18] `https://zenodo.org`

[19] `https://figshare.com/`

Having machine-reusable data before publication may also enable faster and more comprehensive reviews by publishers. Machine-based systems could automatically verify the accuracy of statistical data or experimental results, therefore assisting reviewers. Although pre-publication approaches place extra burden on authors, it could reduce dependency on post-publication techniques and increase the accuracy of machine-based approaches such as NLP.

To date, there has been little research into pre-publication approaches. One of the few examples of creating structured scientific information pre-publication is Scientific Knowledge Graph TeX (SciKGTeX) [44]. SciKGTeX allows for the direct annotation of research papers in the LaTeX source code. The contribution data is embedded into the PDF's XMP metadata which can later be automatically ingested by a knowledge graph using a mapping.

Chapter 4

Approach

In this chapter, we outline the born-reusable approach as it is applied in this thesis. Specifically, its implementation using the with the ORKG R package in an R programming environment, DataCite as a DOI provider, and the ORKG as a harvester (Figure 4.1).

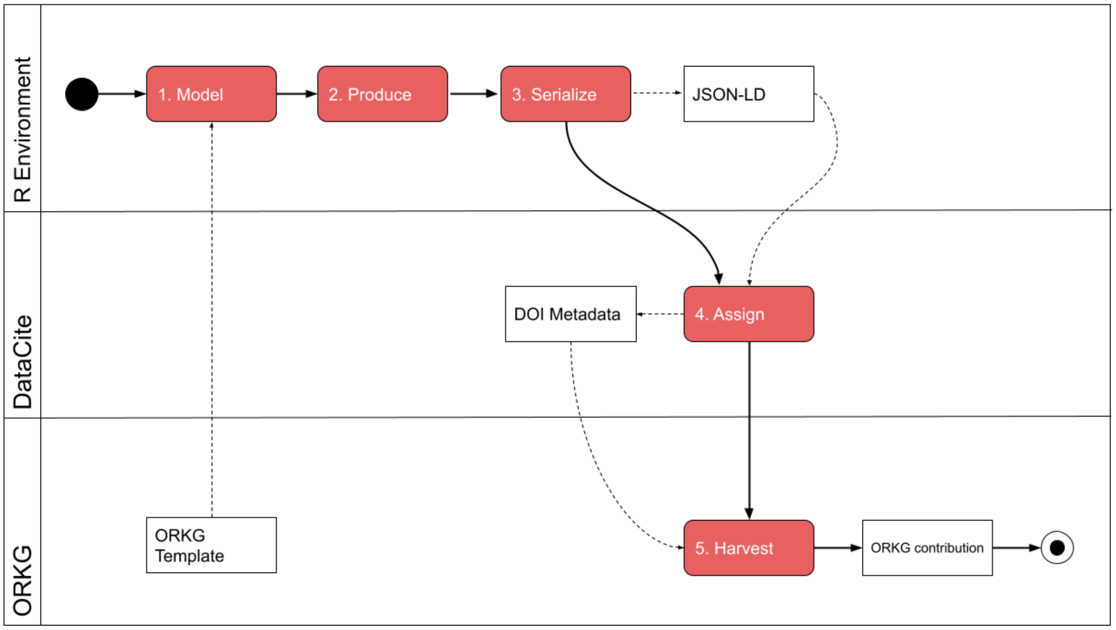


Figure 4.1: The 5 steps of the born-reusable process, demonstrating the overall architecture and the workflow from the perspective of authors and researchers.

4.1 Model

In the first step of the born-reusable approach, the key scientific contributions in the paper and its associated scientific information are identified. With the assistance of a conceptual model or formal ontology, the scientific information is structured into a semantically meaningful structure. An existing ORKG template is selected or newly constructed guided by this model.

ORKG templates serve as structural models for ORKG contributions, aiming to create a standardized representation of research contributions within the ORKG [16]. Templates can be either pre-existing or newly created to suit the specifics of the research contribution with the help of the ORKG web interface.

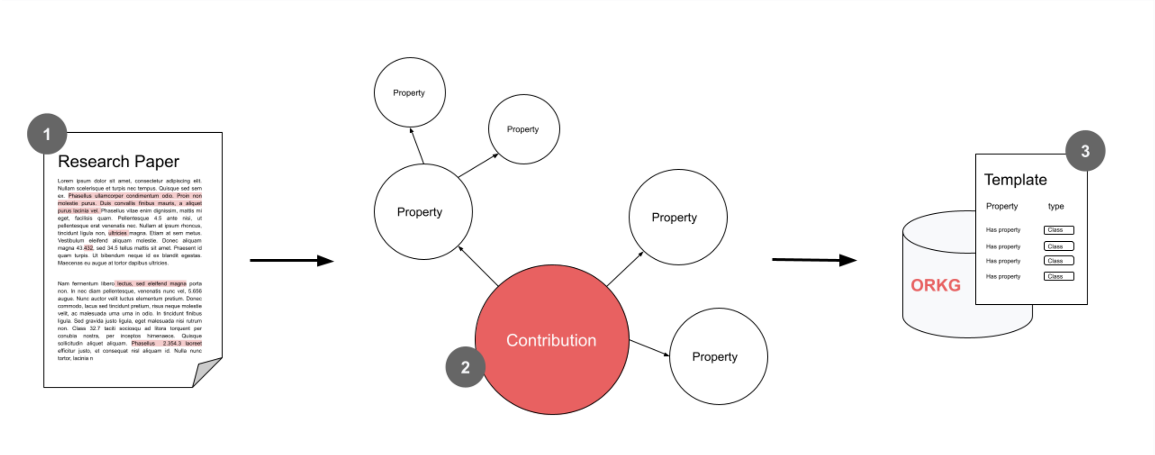
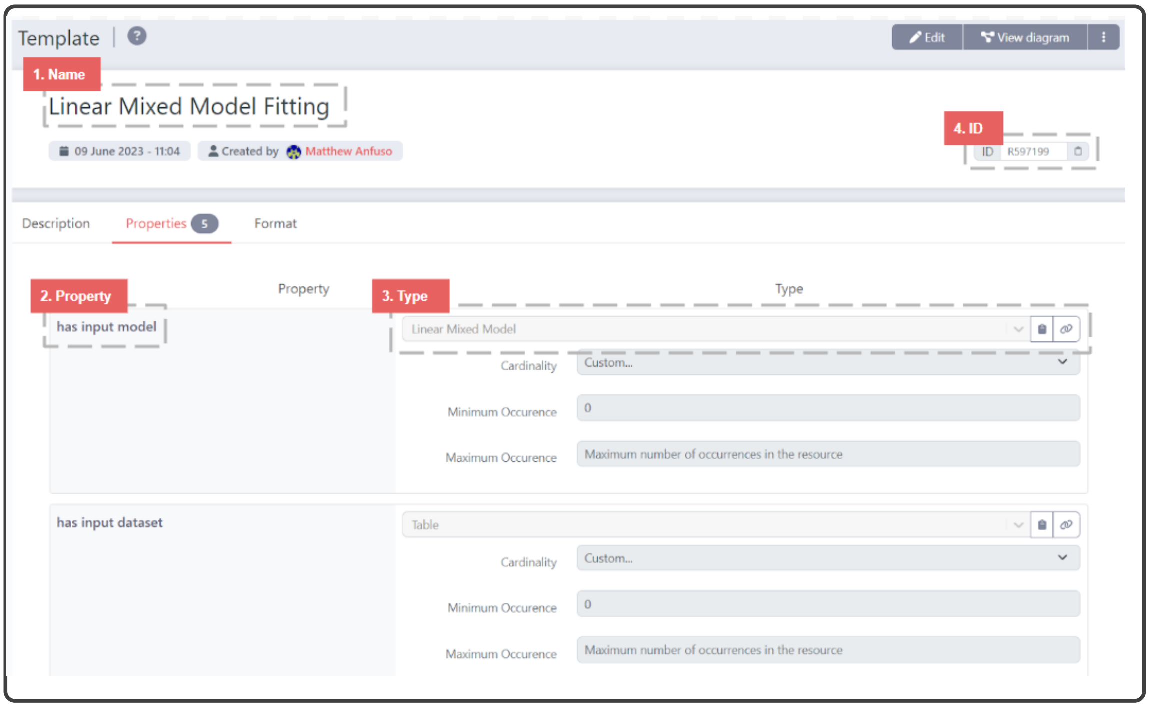


Figure 4.2: The modeling process. 1: Identify key scientific knowledge. 2: Model the scientific knowledge. 3: Translate to an ORGK template.

ORKG Templates

Templates consist of properties and their types (Figure 4.3). Properties represent relationships between entities in the ORKG knowledge graph, whilst types represent an ORKG entity's class. Types are most commonly associated with basic classes such as Text, Decimal, URL, Table, etc., but can also be associated with templates to make nested template structures. Although ORKG templates can be used to mimic ontologies, they currently lack much of the power that an ontology language such as the Web Ontology Language (OWL) provides.



Linear Mixed Model Fitting

09 June 2023 - 11:04

Created by Matthew Anfuso

ID R597199

Property	Type
has input model	Linear Mixed Model Cardinality: Custom... Minimum Occurence: 0 Maximum Occurence: Maximum number of occurrences in the resource
has input dataset	Table Cardinality: Custom... Minimum Occurence: 0 Maximum Occurence: Maximum number of occurrences in the resource

Figure 4.3: An example of an ORKG template. The template *Linear Mixed Model Fitting* (1) has the property *has input model* (2) with a value of type *statistical model* (3). *statistical model* is a class representing a nested template. The resource ID (4) facilitates the import of template information into the ORKG R package.

4.2 Produce

In this stage, the scientific knowledge is produced directly in the R programming environment with the assistance of the ORKG R package and richly and accurately adapted to a machine-reusable structure in accordance with the selected ORKG template's guidelines.

ORKG R Package

The ’born-reusable’ approach outlined in this paper relies on the functionality provided by the ORKG R package. This package facilitates the retrieval of ORKG template specifications into the R environment, allowing the data to be structured within the R environment in line with the template's guidelines, i.e., properties, classes, and types.

Once the data is structured in line with the ORKG template specifications, it can be serialized to a machine-reusable JSON-LD file to facilitate automatic ingestion by the ORKG. Furthermore, the ORKG R package facilitates the automatic import of ORKG resources into the R environment. This feature will be used to help facilitate the proof-of-concept synthesis by importing statistical data directly from born-reusable papers captured by the ORKG (Chapter 5.2).

At the time of writing this theis, the ORKG R package had not been included in the Comprehensive R Archive Network (CRAN). However, two alternative installation methods are available. The first method requires the cloning of the ORKG R package GitLab repository (Listing 4.1). Subsequently, the ORKG R package can be installed using the R console (Listing 4.2). An alternative method to installation is provided by the ”remotes” R package (Listing 4.3). This method allows for the installation of the ORKG R package when *devtools* are restricted, for example in an interactive computing environment such as Jupyter. Further installation instructions are located in the README.md file within the GitLab repository[1].

```
git clone https://gitlab.com/TIBHannover/orkg/orkg-r.git
```

Listing 4.1: Cloning the GitLab repository in terminal

```
devtools::document()
devtools::install()
```

Listing 4.2: Manual installation of ORKG R Package through the R console

[1] https://gitlab.com/TIBHannover/orkg/orkg-r/-/blob/main/README.md

```r
install.packages("remotes")
remotes::install_gitlab("TIBHannover/orkg/orkg-r", force=TRUE)
```

Listing 4.3: Automatic installation of ORKG R Package in the R console using the “remotes” package

Template Materialization

Materializing an ORKG template refers to the process of dynamically creating a corresponding R function that is generated with type-specific arguments in accordance with the specifications of the associated ORKG template [22]. During the materialization process, any nested templates will also be recursively materialized. An example of materializing a template can be seen *materialize_template()* in Listing 4.4.

Once the target template and its nested templates are materialized, a list of all materialized templates can be returned using the *list_templates()* function. Furthermore, documentation for each specific template can be returned by executing a template's function with the parameter *text='doc'* (Listing 4.4).

```
1  library(orkg)
2  orkg <- ORKG(host="https://orkg.org")
3  orkg$templates$materialize_template(template_id = "R597199")
4  tp = orkg$templates$list_templates()
5  keys(tp)
6  # [1] "entity" "linear_mixed_model" "linear_mixed_model_fitting" "
       property" "quantity_value" "qudt_unit"
7  # [7] "variable"
8  tp$linear_mixed_model_fitting(text='doc')
9  # Creates a template of type R597199 (Linear Mixed Model Fitting)
10      # :param label: the label of the resource of type string
11      # :param  has_output_figure : a parameter of type URI (which is
       here considered character )
12      # :param has_input_model: a nested template, use orkg.templates.
       linear_mixed_model
13      # :param  has_input_dataset : a parameter of type Table (which
       is here considered tuple )
14      # :param  has_output_statement : a parameter of type String (
       which is here considered character )
15      # :param  has_output_dataset : a parameter of type Table (which
       is here considered tuple )
16      # :return: a string representing the resource ID of the newly
       created resource
```

Listing 4.4: Materialized templates and their documentation. Comments (blue) represent the resulting R console output. The *keys()* function returns the name of the template and any nested templates within the template. The *'linear_mixed_model_fitting'* is materialised using the *materialize_template()* as its resource ID (3). Subsequently, the template function is executed with the *'text='doc'* parameter, to provide the author information about the template and its parameters.

Template function parameter population

Once the required template functions have been materialized, their parameters can be populated with the relevant scientific information directly from within the R environment (Listing 4.5). Possible parameters include, strings, numbers (e.g., numeric and integer values) other template functions, and tuples containing R data frames. Importantly, this approach allows passing R variables, such as input and computed output data, directly as template function parameter values.

```
1  instance  <- tp$linear_mixed_model_fitting(
2      label="A linear mixed model (LMM) fitting with bactrocera oleae
   abundance (bo) as the response variable and shdi (shannon
   diversity index) as a fixed effect",
3      has_input_dataset= tuple(data, "Raw field data on bactrocera
   oleae abundance"),
4      has_input_model= tp$linear_mixed_model(
5          label="A linear mixed model (LMM) with bactrocera oleae
   abundance (bo) as the response variable and shdi (shannon
   diversity index) as a fixed effect",
6          has_response_variable = var_bactrocera_oleae_abundance,
7          has_fixed_effect_term_i = var_landscape_Shannon_diversity,
8      ),
9      has_output_dataset= tuple(LMMOutput, 'Results of LMM fitting
   with bo as the response variable and shdi as a fixed effect'),
10 )
```

Listing 4.5: The template template function *'linear_mixed_model_fitting'* has been assigned populated parameters in accordance with the template type guidelines. The parameter *'has_input_model'* (5) has type *'nested template'* and references another template function. The parameters *'has_dataset'* and *'has_output_dataset'* represent the ORKG Table class and have values of type R tuple. The first parameter of these tuples references an R data frame.

4.3 Serialize

The next step consists of the serialization of the populated template function parameters into JSON-LD format using the *serialize_to_file()* function. The resulting JSON-LD data facilitates the automatic ingestion of the JSON-LD by the ORKG with the help of two JSON-LD specific keywords [[45]]:

1. `@context` assigns properties IDs to URIs. This provides context to properties and allows for potential interoperability with other knowledge graphs. In the case of the ORKG R package, properties are taken from the materialized ORKG templates and, therefore, limited to URIs within the ORKG.

2. `@type` allows further description of value types through URIs. In the case of the ORKG R package, the @type property assigns ORKG class types to values.

```
{
  "@id": "_:n1",
  "label": "A linear mixed model (LMM) fitting with bactrocera oleae
    abundance (bo) as the response variable and shdi (shannon
   diversity index) as a fixed effect",
  "@type": [
    "https://orkg.org/class/C67001"
  ],
  "P71163": [
    {
      "@id": "_:n2",
      "label": "A linear mixed model (LMM) with bactrocera oleae
   abundance (bo) as the response variable and shdi (shannon
   diversity index) as a fixed effect",
      "@type": [
        "https://orkg.org/class/C67002"
      ],
      "P117004": [
        {
          "@id": "_:n3",
          "label": "Bactrocera oleae abundance in olive groves",
    ...
     "@context": {
        "label": "http://www.w3.org/2000/01/rdf-schema#label",
        "P117004": "https://orkg.org/property/P117004",
        "SAME_AS": "https://orkg.org/property/SAME_AS",
        ...
  }
}
```

Listing 4.6: Excerpt of a JSON-LD file with data describing a linear mixed model fitting

4.4 Link

To ensure that the machine readable JSON-LD data is findable and accessible, the paper's DOI is linked to the JSON-LD files in a Git repository using the DOI's underlying metadata. There are several agencies which provide DOIs on behalf of the non-profit DOI Foundation[2]. In this prototyping phase, we have used a DOI provided by DataCite[3] alongside Datacite's *IsSupplementedBy* and *RelatedIdentifier* metadata properties, which are recommended parameters under version 4.3 of the DataCite metadata schema [46]. This metadata can subsequently be accessed for automatic ingestion by the ORKG using DataCite's API [4]. For production, we will transition to Crossref DOIs and adapt the linking mechanism to its metadata schema. Crossref[5] is widely used by publishers to persistently identify articles using DOI.

```
"relatedIdentifiers":[{"schemeUri":null,"schemeType":null,"
   relationType":"IsSupplementedBy","relatedIdentifier":"https
   ://....article.contribution.1.json","resourceTypeGeneral":"
   Dataset","relatedIdentifierType":"URL","relatedMetadataScheme":
   null}]
```

Listing 4.7: DataCite's DOI metadata with IsSupplementedBy and RelatedIdentifier properties

4.5 Harvest

In the final step, the ORKG fetches the DOI's associated metadata record and detects any JSON-LD files included in the *IsSupplementedBy* property. Each individual JSON file represents an ORKG contribution, a description of a research result together with the employed materials and methods as well as addressed research problem. The contributions are generated on the ORKG in accordance with the JSON-LD structure, with the `@type` and `@context` allowing for automatic detection of ORKG properties and classes. Data can be harvested either with the help of the web interface or with the experimental Python Harvest as part of the ORKG Python

[2] https://www.doi.org/

[3] https://www.datacite.org/

[4] https://support.datacite.org/docs/api

[5] https://www.crossref.org/

library [6].

Harvesting through the web interface is not available in the current production version of ORKG, and requires the installation of the required ORKG frontend version through a Docker[7] container (Listing 4.8).

```
sudo docker run -d -p 3000:80 --env-file .env registry.gitlab.com/
    tibhannover/orkg/orkg-frontend:60
    ef05825bb50e8fee6e230c3e51ecfbbc590b68
```

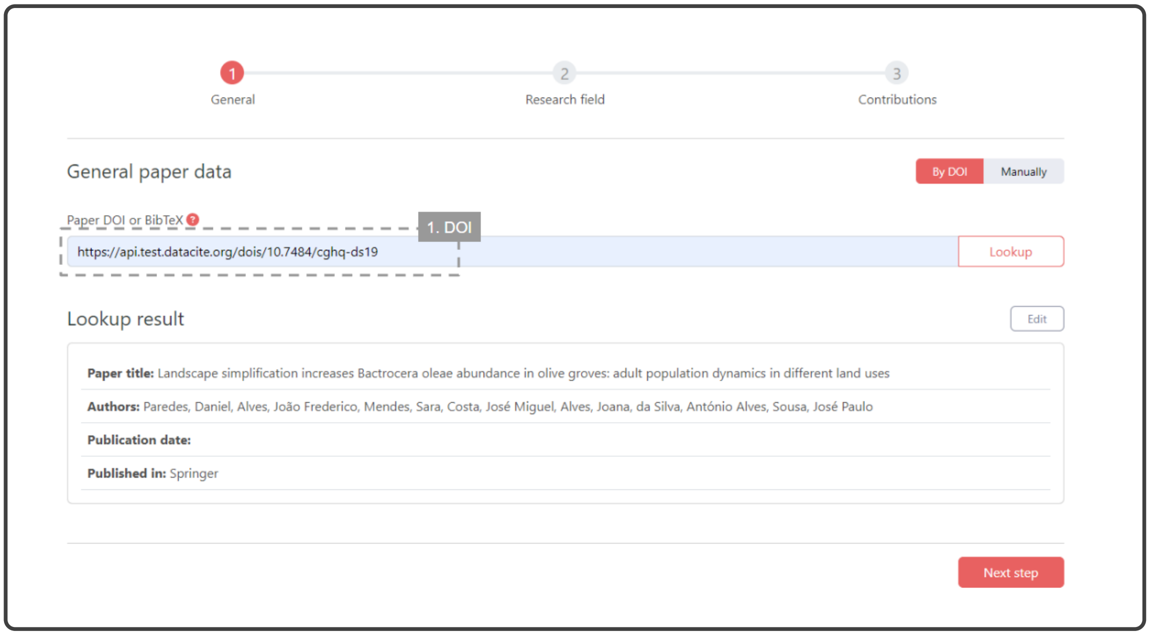
Listing 4.8: Docker installation of ORKG frontend for DOI harvesting

Once the Docker container is started, the web interface can be accessed locally, and a new paper added through the three-step ”Add paper” wizard [8]. In the first step, the paper metadata is fetched by DOI lookup (4.4). In the next step, an appropriate research field is selected for the paper from the list of ORKG research fields (Figure 4.5). In the third step, the research contributions are are generated and displayed for editing through the web interface (Figure 4.6). Finally, after clicking the “Finish” button, the contribution(s) are saved to the ORKG knowledge graph and are made publicly accessible.

[6] https://orkg.readthedocs.io/en/latest/client/harvesters.html

[7] https://docker.com

[8] http://localhost:3000/add-paper



General paper data

Paper DOI or BibTeX: https://api.test.datacite.org/dois/10.7484/cghq-ds19

Lookup result

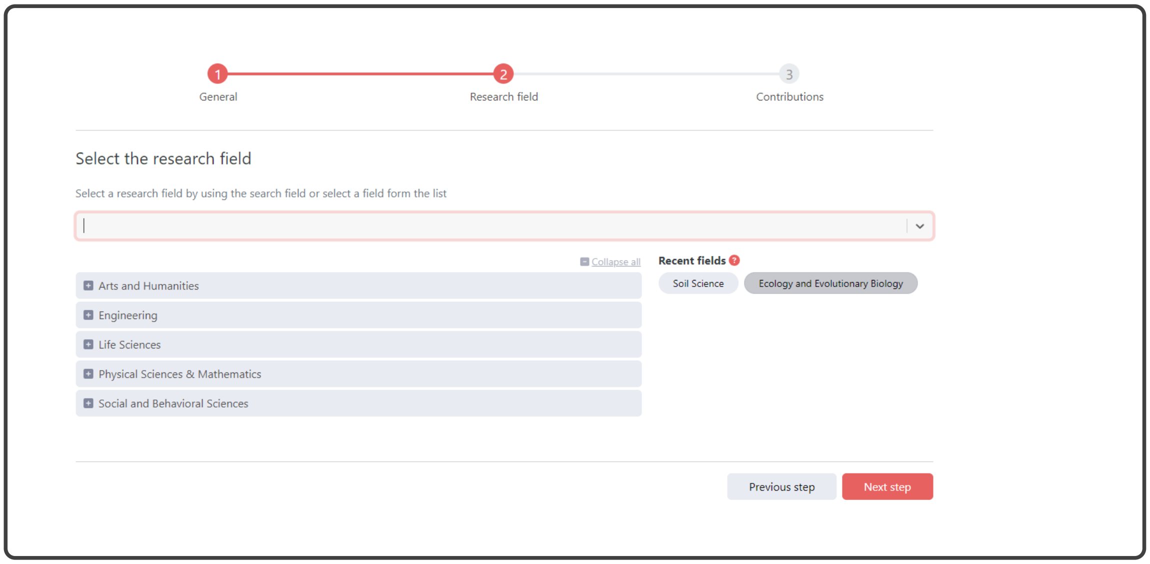
Paper title: Landscape simplification increases Bactrocera oleae abundance in olive groves: adult population dynamics in different land uses

Authors: Paredes, Daniel, Alves, João Frederico, Mendes, Sara, Costa, José Miguel, Alves, Joana, da Silva, António Alves, Sousa, José Paulo

Publication date:

Published in: Springer

Figure 4.4: Harvesting the DOI through the web interface.



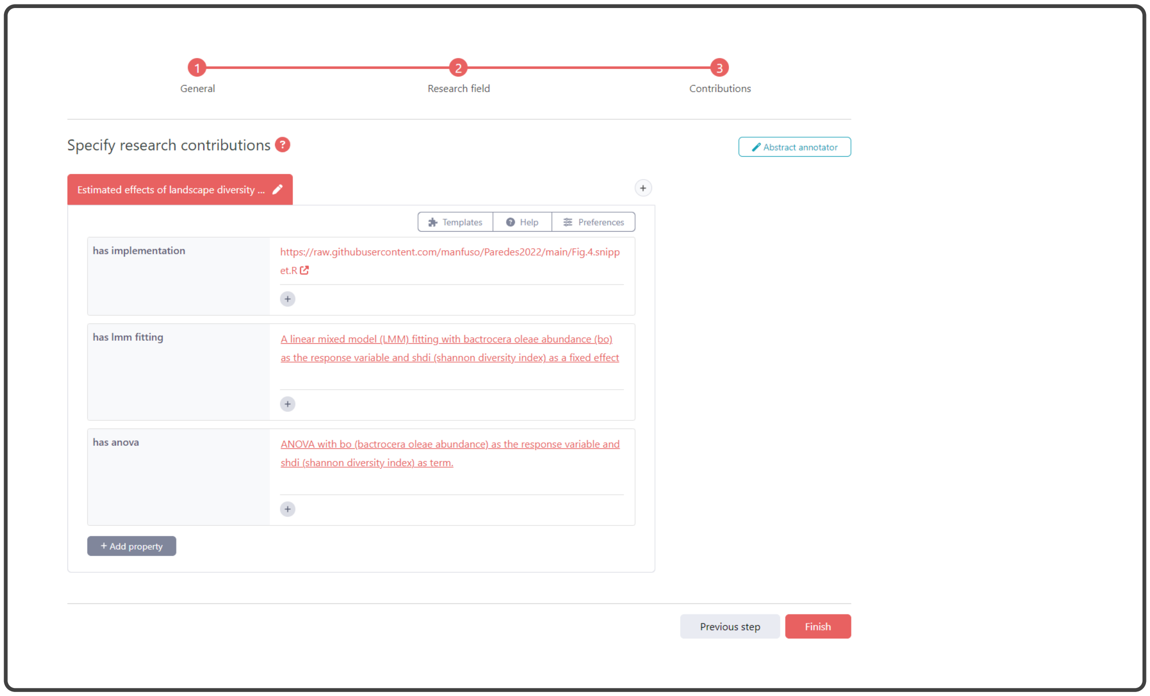
Select the research field

Select a research field by using the search field or select a field form the list

- Arts and Humanities
- Engineering
- Life Sciences
- Physical Sciences & Mathematics
- Social and Behavioral Sciences

Recent fields: Soil Science, Ecology and Evolutionary Biology

Figure 4.5: Step 2: Selecting an appropriate research field for the paper.



Specify research contributions

Estimated effects of landscape diversity ...

has implementation: https://raw.githubusercontent.com/manfuso/Paredes2022/main/Fig.4.snippet.R

has lmm fitting: A linear mixed model (LMM) fitting with bactrocera oleae abundance (bo) as the response variable and shdi (shannon diversity index) as a fixed effect

has anova: ANOVA with bo (bactrocera oleae abundance) as the response variable and shdi (shannon diversity index) as term.

Figure 4.6: Step 3. The contributions are generated ready for editing and saving.

Chapter 5

Application

In this chapter, we apply the born-reusable approach to an manuscript in the field of soil science. Subsequently, we reapply the approach to two published papers and conduct a proof-of-concept meta-analysis to demonstrate how this approach can support research synthesis.

5.1 Application to a manuscript

The manuscript ‘Cover Crops Improve Soil Structure and Change OC Distribution in Aggregate Fractions’, authored by Gentsch et al., was selected to demonstrate the application of the ’born-reusable’ approach to an unpublished manuscript. The co-author, Dr Norman Gentsch, is a researcher at the Institute of Soil Science at the Leibniz University Hannover[1]. This manuscript investigates the role of cover crops in enhancing soil structure and modifying its organic carbon (OC) content. The manuscript is intended for publication in the scientific journal SOIL (Copernicus Publications). The co-author provided the manuscript in PDF format, the supporting statistical code in R, and the associated data set containing raw field sample data in CSV (Figure 5.1). All relevant data is available on GitLab[2].

Salient scientific knowledge, including figures, tables, and their respective captions, were identified in the PDF manuscript, along with the relevant lines of R code used to generate them. In total six essential scientific contributions were identified (Figure 5.2) within the manuscript.

[1] https://www.soil.uni-hannover.de/de/norman-gentsch

[2] https://gitlab.com/TIBHannover/orkg/orkg-papers/-/tree/master/gentsch22cover

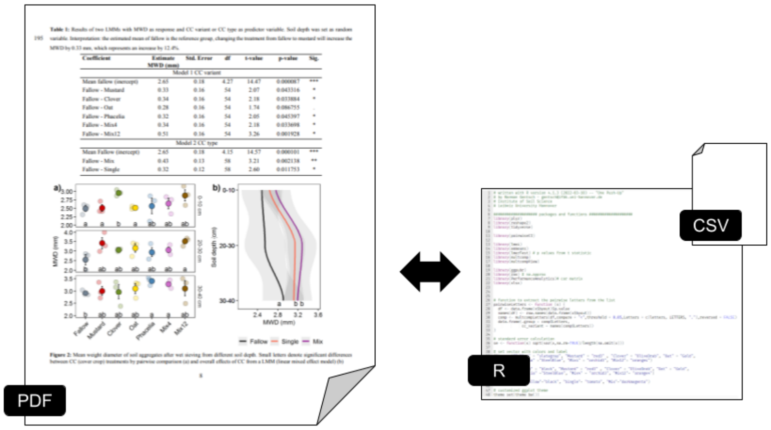
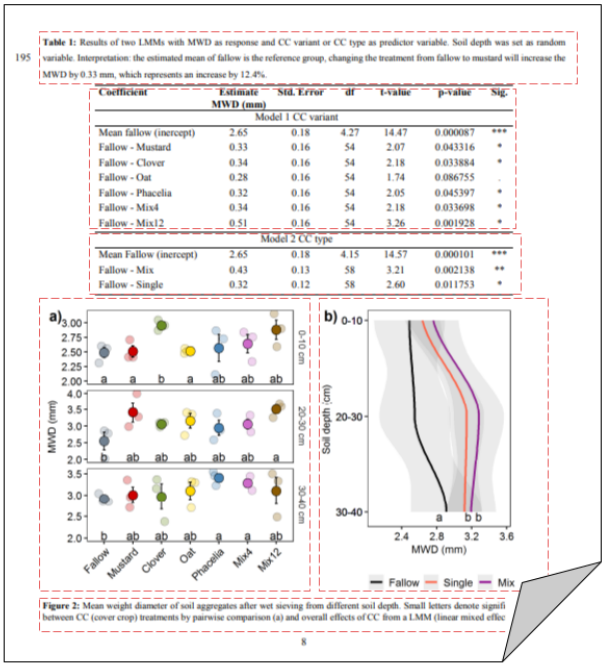


Figure 5.1: Manuscript in PDF format, the supporting statistical code in R, and the associated data set in CSV format containing raw field sample data.



195 **Table 1:** Results of two LMMs with MWD as response and CC variant or CC type as predictor variable. Soil depth was set as random variable. Interpretation: the estimated mean of fallow is the reference group, changing the treatment from fallow to mustard will increase the MWD by 0.33 mm, which represents an increase by 12.4%.

Coefficient	Estimate MWD (mm)	Std. Error	df	t-value	p-value	Sig.
Model 1 CC variant						
Mean fallow (inercept)	2.65	0.18	4.27	14.47	0.000087	***
Fallow - Mustard	0.33	0.16	54	2.07	0.043316	*
Fallow - Clover	0.34	0.16	54	2.18	0.033884	*
Fallow - Oat	0.28	0.16	54	1.74	0.086755	.
Fallow - Phacelia	0.32	0.16	54	2.05	0.045397	*
Fallow - Mix4	0.34	0.16	54	2.18	0.033698	*
Fallow - Mix12	0.51	0.16	54	3.26	0.001928	*
Model 2 CC type						
Mean Fallow (inercept)	2.65	0.18	4.15	14.57	0.000101	***
Fallow - Mix	0.43	0.13	58	3.21	0.002138	**
Fallow - Single	0.32	0.12	58	2.60	0.011753	*

Figure 2: Mean weight diameter of soil aggregates after wet sieving from different soil depth. Small letters denote signifi between CC (cover crop) treatments by pairwise comparison (a) and overall effects of CC from a LMM (linear mixed effec

8

Figure 5.2: The identification and selection (outlined red) of key scientific information inside (page 8) of the manuscript.

5.1.1 Model Creation

All six contributions identified within the manuscript represented some form of a statistical process, either a statistical model fitting, a pairwise t-test or a descriptive statistical calculation. Therefore, a single conceptual model was created to capture the semantics of such statistical processes (Figure 5.3).

The conceptual model included the following entities: An input data set used for the statistical process, an input model representing a statistical model, an output data set representing the results of the process, an output figure as a visual representation of the results, an output statement as a textual representation of the results, and an implementation representing the R code required to generate the contribution.

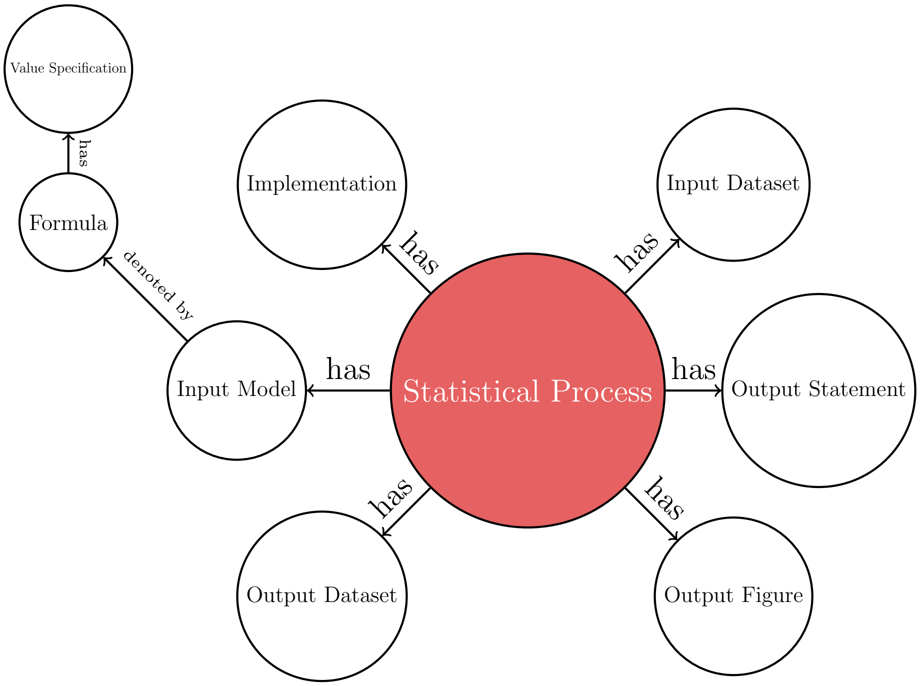


Figure 5.3: A conceptual model illustrating the statistical processes underlying the identified contributions within the manuscript.

5.1.2 Template Creation

The conceptual model was translated into five distinct ORKG templates, representing the various variations of the statistical processes that were implemented in the contributions. These template variations were required, as opposed to a single template, to account for technical limitations in the ORKG R package that prevented the use of optional parameters. For example, all templates represent a subset of the “Model Fitting 3” template (Figure 5.1) which encompasses all properties contained in the conceptual model.

```yaml
’Model Fitting 3’:
  has implementation:
    type: URI
  has input dataset:
    type: Table
  has input output:
    type: Table
  has input model:
    type: Template
    template:
      ’Statistical Model’:
        is denoted by:
          type: Template
          template:
            ’Formula’:
              has value specification:
                type: Template
                template:
                  ’Value Specification’:
                    has specified value:
                      type: String
  has output figure:
    type: URI
  has output statement:
    type: String
```

Listing 5.1: A YAML representation of the “Model Fitting 3” template, including nested templates (blue), ORKG properites (red) and their ORKG types (gray).

5.1.3 Template Population

The materialized template functions' parameters were populated with information directly from the R environment (Listing 5.2) according to the ORKG template type guidelines (Table 5.1).

Parameter	R Type	ORKG Type
label	string	String
has_implementation	string	URI
has_input_dataset	tuple(data frame, string)	Table
has_input_model	function	Template
is_denoted_by	function	Template
has_value_specification	function	Template
has_specified_value	string	String
has_output_dataset	string	URI
has_output_dataset	tuple(data frame, string)	Table
has_output_figure	string	URI
has_output_statement	string	String
has_output_dataset	tuple(data frame, string)	Table

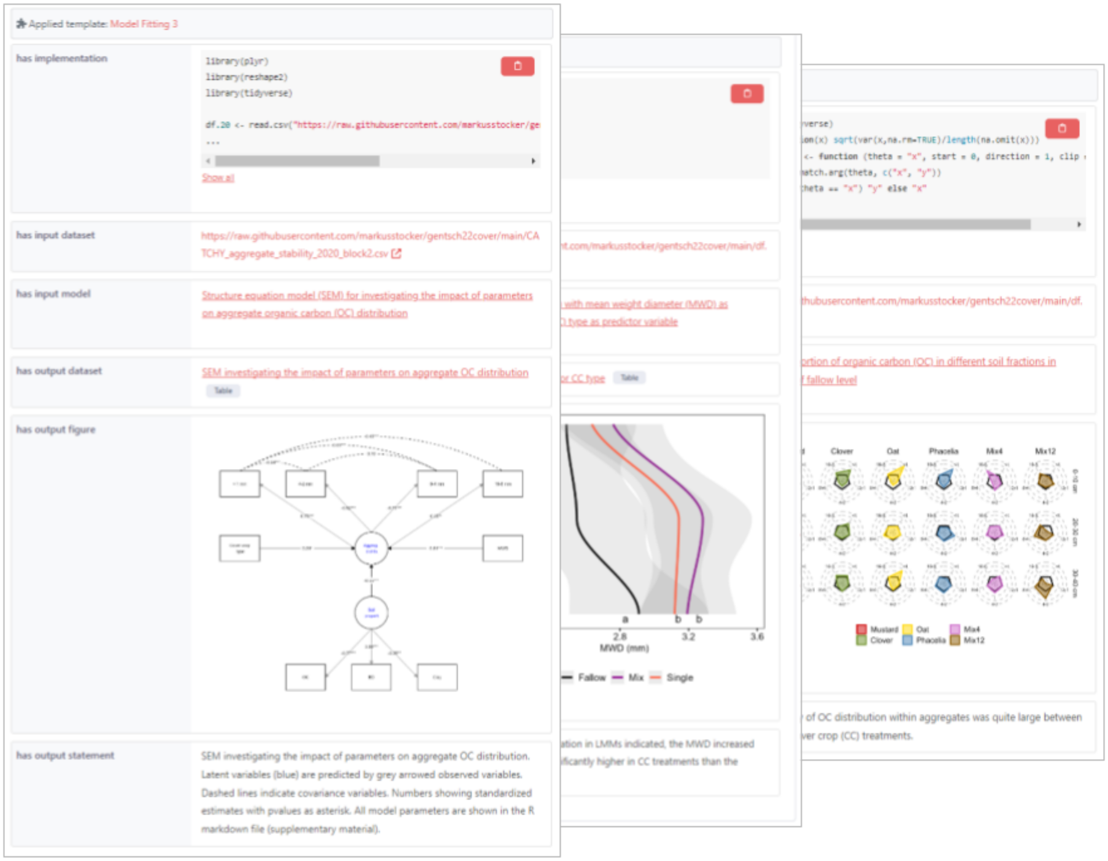
Table 5.1: Possible template function parameters along with their R and ORKG types as mandated by the ORKG templates. Note, ’label’ is a compulsory parameter.

```
instance <- tp$model_fitting_3(
  label="Overall effects of CC from a LMM",
  has_input_dataset="https://.../df.MWD.csv",
  has_input_model=tp$statistical_model(
    label="A linear mixed model (LMM) with mean weight diameter...",
    is_denoted_by=tp$formula(
      label="The formula of the linear mixed model with MWD as...",
      has_value_specification=tp$value_specification(
        label="MWD_cor ~ cc_type + (1|depth)",
        has_specified_value="MWD_cor ~ cc_type + (1|depth)"
      )
    )
  ),
  has_output_dataset= tuple(df.pw.MWD.tot, ’Estimated Marginal...’),
  has_output_figure="https://.../Fig.2b.png",
  has_output_statement= "A comprehensive data evaluation in LMMs",
  has_implementation="https://.../figure2b.snippet.R"
)
```

Listing 5.2: An example (excerpt) of the “Model Fitting 3´´ template population

5.1.4 Harvesting

As the manuscript was unpublished and had not yet been assigned a DOI, we assigned a DataCite test DOI to the manuscript metadata in order to enable the harvesting by DOI lookup. With DOI-lookup based harvesting, the data for the six contributions described in the manuscript were automatically ingested into the ORKG[3] (Figure 5.4) using the process described in Chapter 4.5.



Applied template: Model Fitting 3

has implementation	`library(plyr)`
	`library(reshape2)`
	`library(tidyverse)`
	
	`df.20 <- read.csv("https://raw.githubusercontent.com/markusstocker/ge`
	`...`
	Show all

has input dataset	https://raw.githubusercontent.com/markusstocker/gentsch22cover/main/CATCHY_aggregate_stability_2020_block2.csv
has input model	Structure equation model (SEM) for investigating the impact of parameters on aggregate organic carbon (OC) distribution
has output dataset	SEM investigating the impact of parameters on aggregate OC distribution Table
has output figure	
has output statement	SEM investigating the impact of parameters on aggregate OC distribution. Latent variables (blue) are predicted by grey arrowed observed variables. Dashed lines indicate covariance variables. Numbers showing standardized estimates with pvalues as asterisk. All model parameters are shown in the R markdown file (supplementary material).

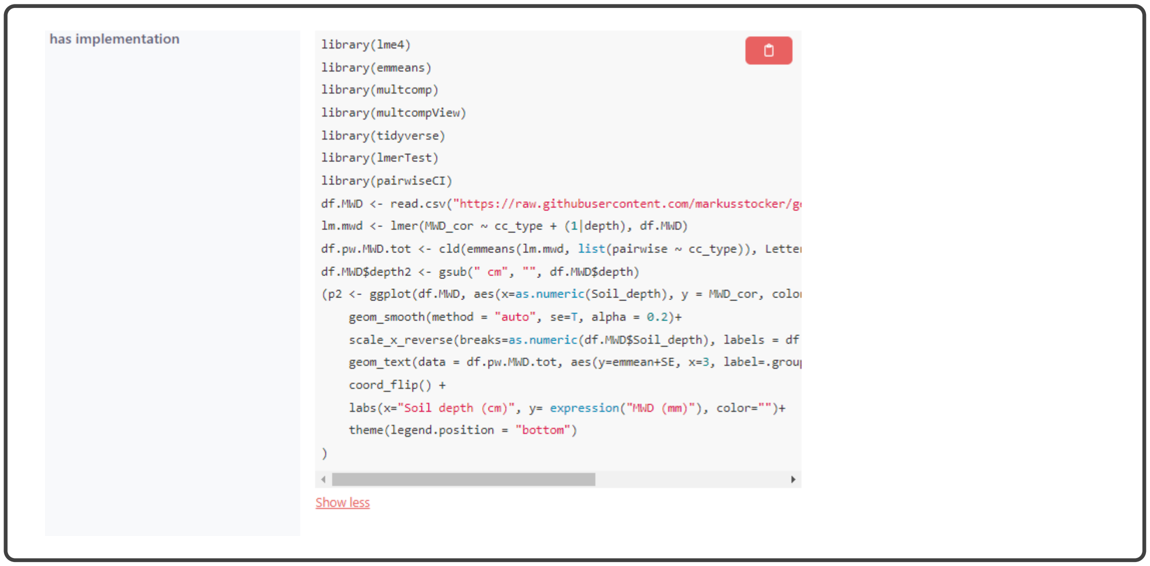
t.com/markusstocker/gentsch22cover/main/df.
with mean weight diameter (MWD) as
) type as predictor variable
or CC type
Table
a
b
b
2.8
3.2
3.6
MWD (mm)
Fallow
Mix
Single
ation in LMMs indicated, the MWD increased
ficantly higher in CC treatments than the
verse)
on(x) sqrt(var(x,na.rm=TRUE)/length(na.omit(x)))
<- function (theta = "x", start = 0, direction = 1, clip
atch.arg(theta, c("x", "y"))
heta == "x") "y" else "x"
thubusercontent.com/markusstocker/gentsch22cover/main/df.
rtion of organic carbon (OC) in different soil fractions in
fallow level
Clover
Oat
Phacelia
Mix4
Mix12
Mustard
Clover
Oat
Phacelia
Mix4
Mix12
of OC distribution within aggregates was quite large between
ver crop (CC) treatments.

Figure 5.4: Resulting ORKG contribution.

The manuscript's title, authors, journal and DOI were automatically extracted from DOI's metadata. Any properties containing Table classes were automatically displayed in a user-friendly table representation, allowing for the option to export the table's content as CSV (Figure 5.7). In addition, strings containing URIs which

[3] https://incubating.orkg.org/paper/R481377

referenced image files were automatically detected and displayed (Figure 5.6). Each contribution included a code snippet representing its implementation ('has implementation'). This code was automatically retrieved from the referenced Git repository and displayed in text form on the ORKG (Figure 5.5). The snippets are designed to be self-contained, requiring only the installation of the necessary R packages, and thus allowing for quick execution in an IDE or an interactive computing environment such as Jupyter.



```
library(lme4)
library(emmeans)
library(multcomp)
library(multcompView)
library(tidyverse)
library(lmerTest)
library(pairwiseCI)
df.MWD <- read.csv("https://raw.githubusercontent.com/markusstocker/g
lm.mwd <- lmer(MWD_cor ~ cc_type + (1|depth), df.MWD)
df.pw.MWD.tot <- cld(emmeans(lm.mwd, list(pairwise ~ cc_type)), Lette
df.MWD$depth2 <- gsub(" cm", "", df.MWD$depth)
(p2 <- ggplot(df.MWD, aes(x=as.numeric(Soil_depth), y = MWD_cor, colo
    geom_smooth(method = "auto", se=T, alpha = 0.2)+
    scale_x_reverse(breaks=as.numeric(df.MWD$Soil_depth), labels = df
    geom_text(data = df.pw.MWD.tot, aes(y=emmean+SE, x=3, label=.grou
    coord_flip() +
    labs(x="Soil depth (cm)", y= expression("MWD (mm)"), color="")+
    theme(legend.position = "bottom")
)
```

Figure 5.5: The "has implementation" property.
The'has implementation" property displaying the associated R code snippet.

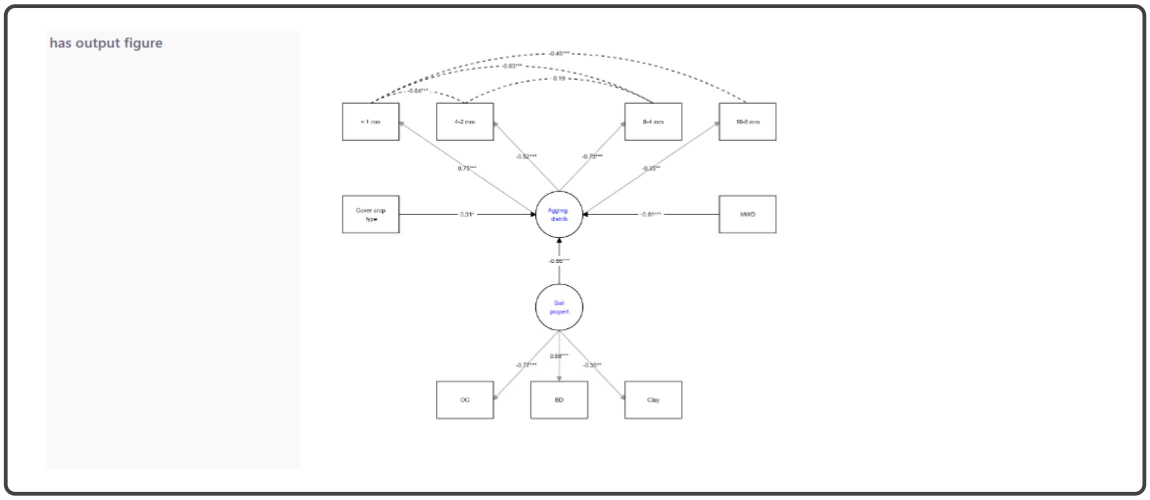
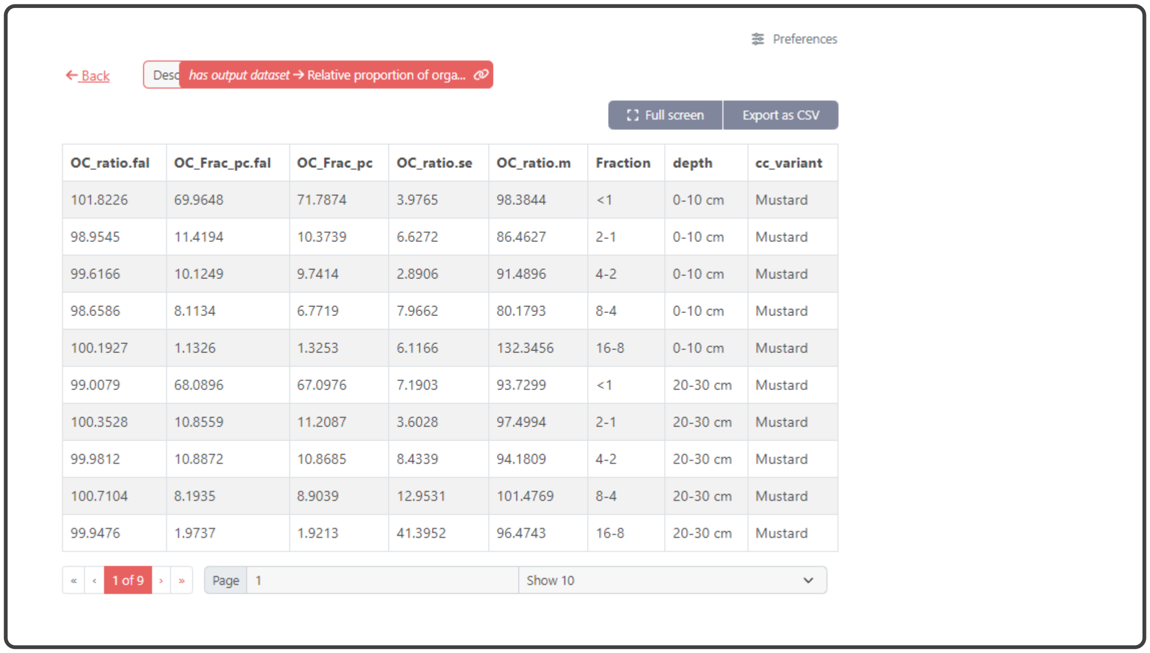


Figure 5.6: The“has output figure” property and the display of the image referenced by its URI.



OC_ratio.fal	OC_Frac_pc.fal	OC_Frac_pc	OC_ratio.se	OC_ratio.m	Fraction	depth	cc_variant
101.8226	69.9648	71.7874	3.9765	98.3844	<1	0-10 cm	Mustard
98.9545	11.4194	10.3739	6.6272	86.4627	2-1	0-10 cm	Mustard
99.6166	10.1249	9.7414	2.8906	91.4896	4-2	0-10 cm	Mustard
98.6586	8.1134	6.7719	7.9662	80.1793	8-4	0-10 cm	Mustard
100.1927	1.1326	1.3253	6.1166	132.3456	16-8	0-10 cm	Mustard
99.0079	68.0896	67.0976	7.1903	93.7299	<1	20-30 cm	Mustard
100.3528	10.8559	11.2087	3.6028	97.4994	2-1	20-30 cm	Mustard
99.9812	10.8872	10.8685	8.4339	94.1809	4-2	20-30 cm	Mustard
100.7104	8.1935	8.9039	12.9531	101.4769	8-4	20-30 cm	Mustard
99.9476	1.9737	1.9213	41.3952	96.4743	16-8	20-30 cm	Mustard

Figure 5.7: The ORKG Table class associated with the “has output dataset” property.

5.2 Application to a proof-of-concept synthesis

To demonstrate how richly-described and high-quality machine-reusable scientific knowledge can support research synthesis, a proof-of-concept synthesis in the form of a meta-anaysis was conducted. The meta-analysis was based on the research question 'how does landscape composition affect the abundance or incidence of agricultural pest species?'. The purpose of this small-scale meta-analysis was not to definitively answer any specific agroecology-related question, but rather simply to demonstrate its applicability in a wider scientific context.

The paper "Contrasting effects of landscape composition on crop yield mediated by specialist herbivores" published by Perez-Alvarez, Nault, and Poveda in 2018 was selected as the foundation for the synthesis [47]. The synthesis was conducted in close collaboration with the original author who indicated the essential research contributions in the paper and provided the associated R source code.

After a small systematic search, the paper "Landscape simplification increases Bactrocera oleae abundance in olive groves: adult population dynamics in different land uses" published in 2023 by Parades was selected [48]. In particular, the research contribution associated with Figure 3 was singled out as a compatible addition to the synthesis. For this purpose, the authors kindly provided the relevant R source code and data set on request. All relevant data is available on GitLab[4,5].

[4] `https://gitlab.com/TIBHannover/orkg/orkg-papers/-/tree/master/Paredes2022`

[5] `https://gitlab.com/TIBHannover/orkg/orkg-papers/-/tree/master/Perez-Alvarez2018`

5.2.1 Model Creation

A conceptual model was created to capture the figures 4a, 4b, 4c, 4d, and 5 in the paper “Contrasting effects of landscape composition on crop yield mediated by specialist herbivores”. These figures depict the relationship between landscape and pest incidence and abundance, and the impact of plant damage on cabbage yield, using linear regression analysis. Of particular interest for the meta-analysis were the statistical processes related to these figures in the associated R source code. These included: the linear mixed model (LMM), the significance of the LMM coefficients, as well as results of an ANOVA.

These statistical processes were conceptualized with the idea of a planned process. A planned process is a process that executes a plan that represents the actualization of a specified plan specification [49]. The planned process was determined to include the following sub-processes: LMM fitting, LMM significance test, LMM prediction, ANOVA, and linear regression, as well as an implementation representing the actualization of the sub-processes. To further help richly describe the processes, notably the variables used in the LMM and linear regression, a conceptual model for variables influenced by the I-ADOPT Framework ontology was integrated. The I-ADOPT framework ontology provides semantic structures designed to model concepts frequently observed in scientific data and make property descriptions more FAIR [50].

The resulting conceptual model (Figure 5.8) was also suitable for capturing the require statistical information associated with Figure 3 from “Landscape simplification increases Bactrocera oleae abundance in olive groves: adult population dynamics in different land uses”, specifically the LMM fitting and the ANOVA results contained in its accompanying R source code.

5.2.2 ORKG Contribution Creation

The conceptual model was translated to an ORKG template structure. Restrictions preventing optional parameters were removed from the ORKG R package, allowing for the creation of a single all encompassing template. Therefore, the “LMM Planned Process”[6] template and its nested templates were able to describe all the contributions in both papers in a machine-reusable form (Table 5.2). Both papers were assigned new test DOIs from DataCite and were harvested by the ORKG using DOI-lookup[7,8] (Figures 5.9 and 5.10).

5.2.3 Meta analysis

The meta-analysis was conducted in R, inside a Jupyter notebook[9] (Figure 5.11). Statistical information was directly retrieved from the required ORKG tables using the *by_id* function in the ORKG R package, which allows for the retrieval of ORKG resources by their ID. The returned data from the *by_id* function was then converted to an R dataframe to facilitate the meta-analysis (Listing 5.3).

```
#retrieve significance testing output dataset for fig.4a
PerezAlvarez2018Fig4aSigTest <- orkg$resources$by_id('R569582')$as_dataframe()

#retrieve significance testing output dataset for fig. 4b (flea beetle abundance)
PerezAlvarez2018Fig4bSigTest <- orkg$resources$by_id('R570502')$as_dataframe()

#retrieve significance testing output dataset for fig. 4c (incidence of lepidopteran
    larvae)
PerezAlvarez2018Fig4cSigTest <- orkg$resources$by_id('R571405')$as_dataframe()

#retrieve ANOVA output dataset for fig. 4 (Bactrocera oleae abundance)
Paredes2022Fig4Anova <- orkg$resources$by_id('R552440')$as_dataframe()
```

Listing 5.3: Retrival of ORKG resources using ORKG R package

[6] https://incubating.orkg.org/template/R492225

[7] https://incubating.orkg.org/paper/R569000

[8] https://incubating.orkg.org/paper/R551746

[9] https://mybinder.org/v2/gl/TIBHannover%2Forkg%2Forkg-notebooks/master?labpath=ecology

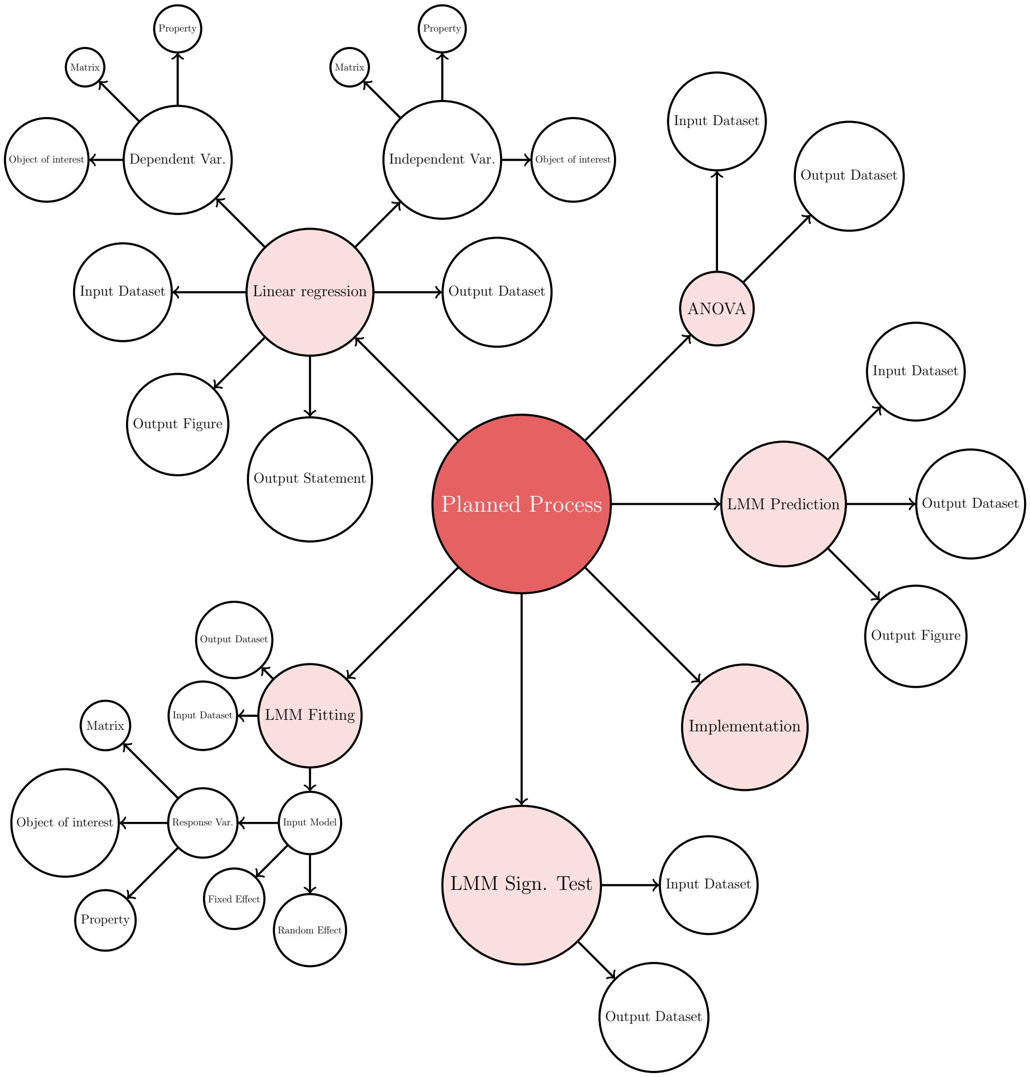


Figure 5.8: Graph representing the conceptual model for the research contributions.

LMM Planned Process	
Property	**Type**
has implementation	URI
has output statement	String
has output figure	URI
has lmm fitting	*Linear Mixed Model Fitting*
has lmm significance testing	*LMM Significance Testing*
has anova	*ANOVA*
has lmm prediction	*LMM Prediction*
has linear regression	*Linear Regression*

Linear Mixed Model Fitting	
Property	**Type**
has input model	*Linear Mixed Model*
has input dataset	Table
has output dataset	Table
has output figure	URI
has output statement	String

Linear Mixed Model	
Property	**Type**
has response variable	*Variable*
has fixed effect term I	*Variable*
has fixed effect term II	*Variable*
has random effect term	*Variable*

LMM Significance Testing	
Property	**Type**
has input dataset	Table
has output dataset	Table

Variable	
Property	**Type**
has object of interest	*Entity*
has property	*Property*
has matrix	*Entity*

ANOVA	
Property	**Type**
has input dataset	Table
has output dataset	Table

Entity	
Property	**Type**
same as	URI
is constrained by	*Quantity Value*

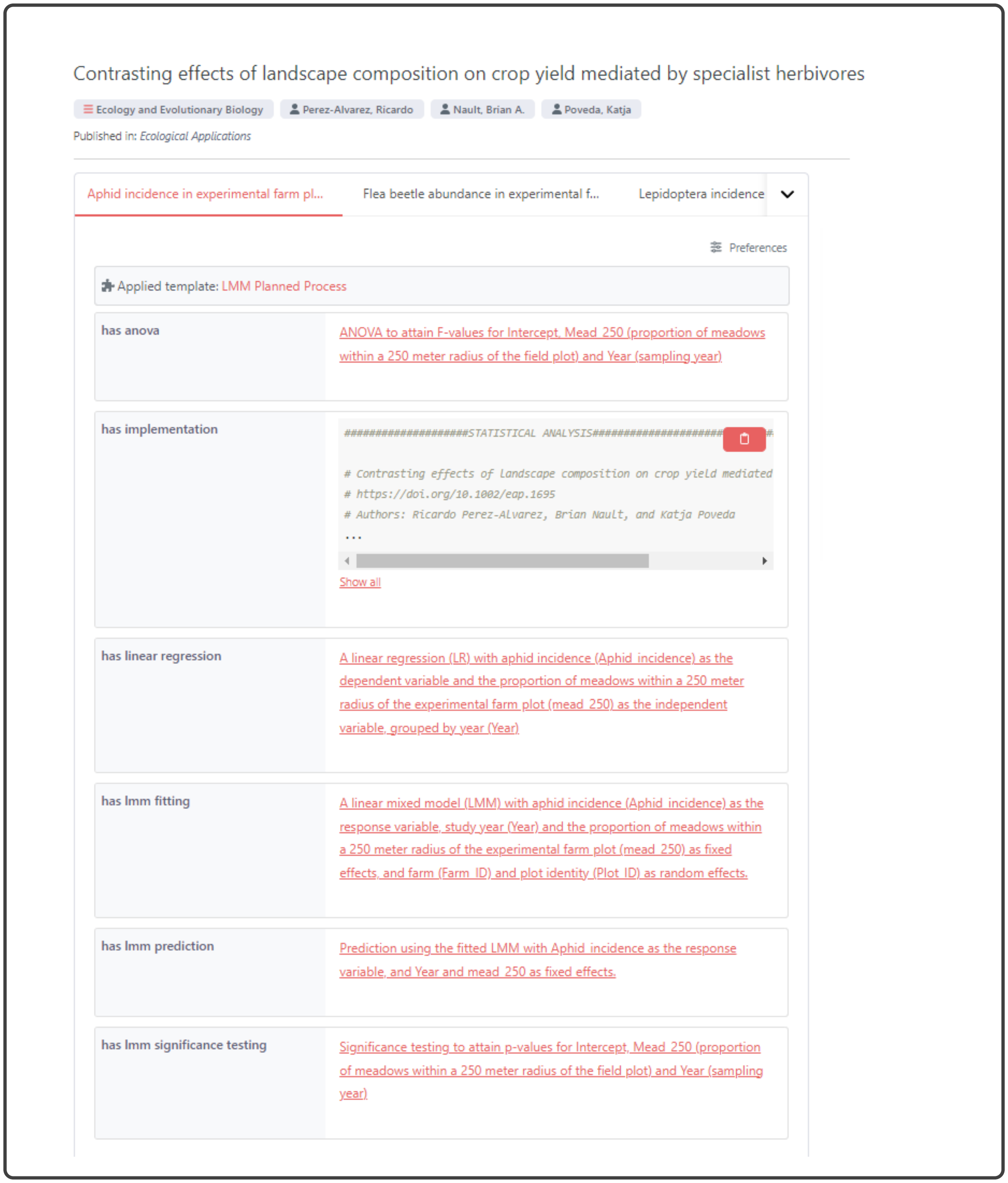
LMM Prediction	
Property	**Type**
has input dataset	Table
has output dataset	Table
has output figure	URI

Property	
Property	**Type**
same as	URI

Linear Regression	
Property	**Type**
has input dataset	Table
Has independent variable	*Variable*
Has dependent variable	*Variable*
has output figure	URI
has output statement	String
has output dataset	Table

Quantity Value	
Property	**Type**
qudt:numericValue	Number
qudt:unit	QUDT Unit

Table 5.2: The resulting ORKG Templates inspired by the “Planned Process” conceptual model.



Contrasting effects of landscape composition on crop yield mediated by specialist herbivores

Ecology and Evolutionary Biology
Perez-Alvarez, Ricardo
Nault, Brian A.
Poveda, Katja

Published in: *Ecological Applications*

Aphid incidence in experimental farm pl...
Flea beetle abundance in experimental f...
Lepidoptera incidence
Preferences

Applied template: LMM Planned Process

has anova

ANOVA to attain F-values for Intercept, Mead 250 (proportion of meadows within a 250 meter radius of the field plot) and Year (sampling year)

has implementation

```
####################STATISTICAL ANALYSIS#####################

# Contrasting effects of landscape composition on crop yield mediated
# https://doi.org/10.1002/eap.1695
# Authors: Ricardo Perez-Alvarez, Brian Nault, and Katja Poveda
...
```

has linear regression

A linear regression (LR) with aphid incidence (Aphid incidence) as the dependent variable and the proportion of meadows within a 250 meter radius of the experimental farm plot (mead 250) as the independent variable, grouped by year (Year)

has lmm fitting

A linear mixed model (LMM) with aphid incidence (Aphid incidence) as the response variable, study year (Year) and the proportion of meadows within a 250 meter radius of the experimental farm plot (mead 250) as fixed effects, and farm (Farm ID) and plot identity (Plot ID) as random effects.

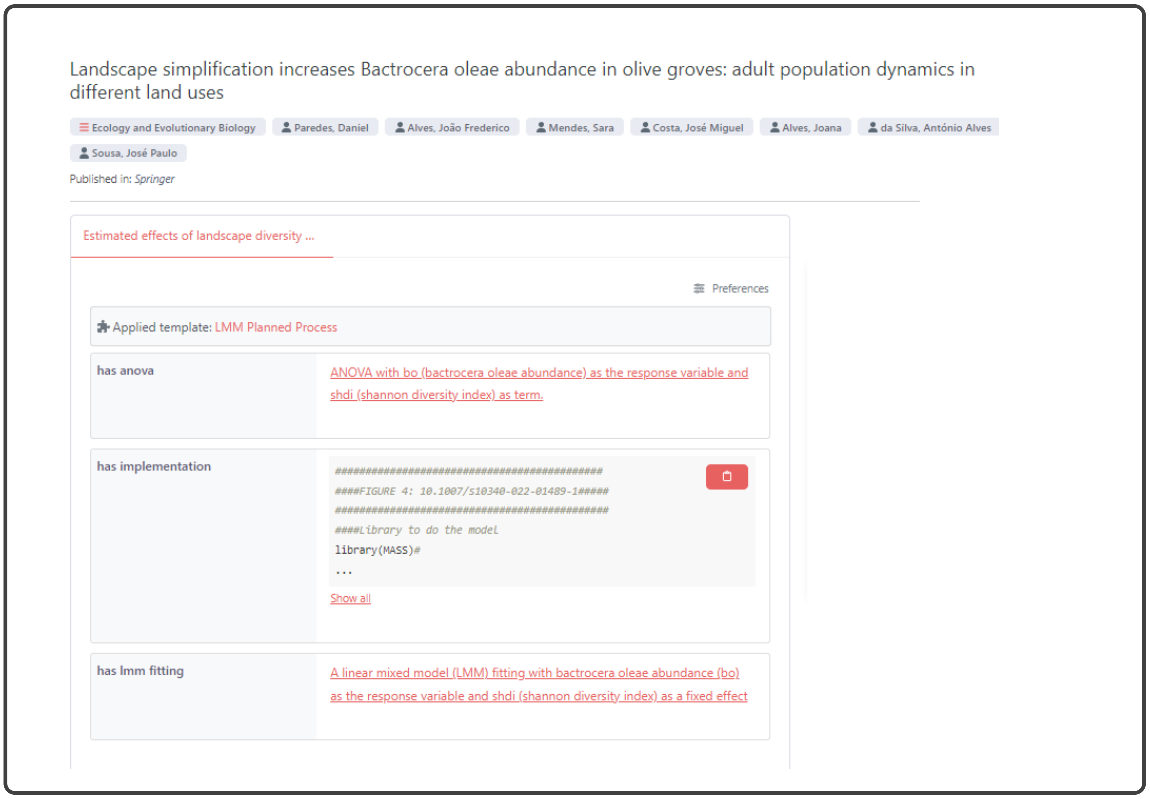
has lmm prediction

Prediction using the fitted LMM with Aphid incidence as the response variable, and Year and mead 250 as fixed effects.

has lmm significance testing

Significance testing to attain p-values for Intercept, Mead 250 (proportion of meadows within a 250 meter radius of the field plot) and Year (sampling year)

Figure 5.9: The six contributions from “Contrasting effects of landscape composition on crop yield mediated by specialist herbivores” in the ORKG.



Landscape simplification increases Bactrocera oleae abundance in olive groves: adult population dynamics in different land uses

Published in: *Springer*

Applied template: LMM Planned Process

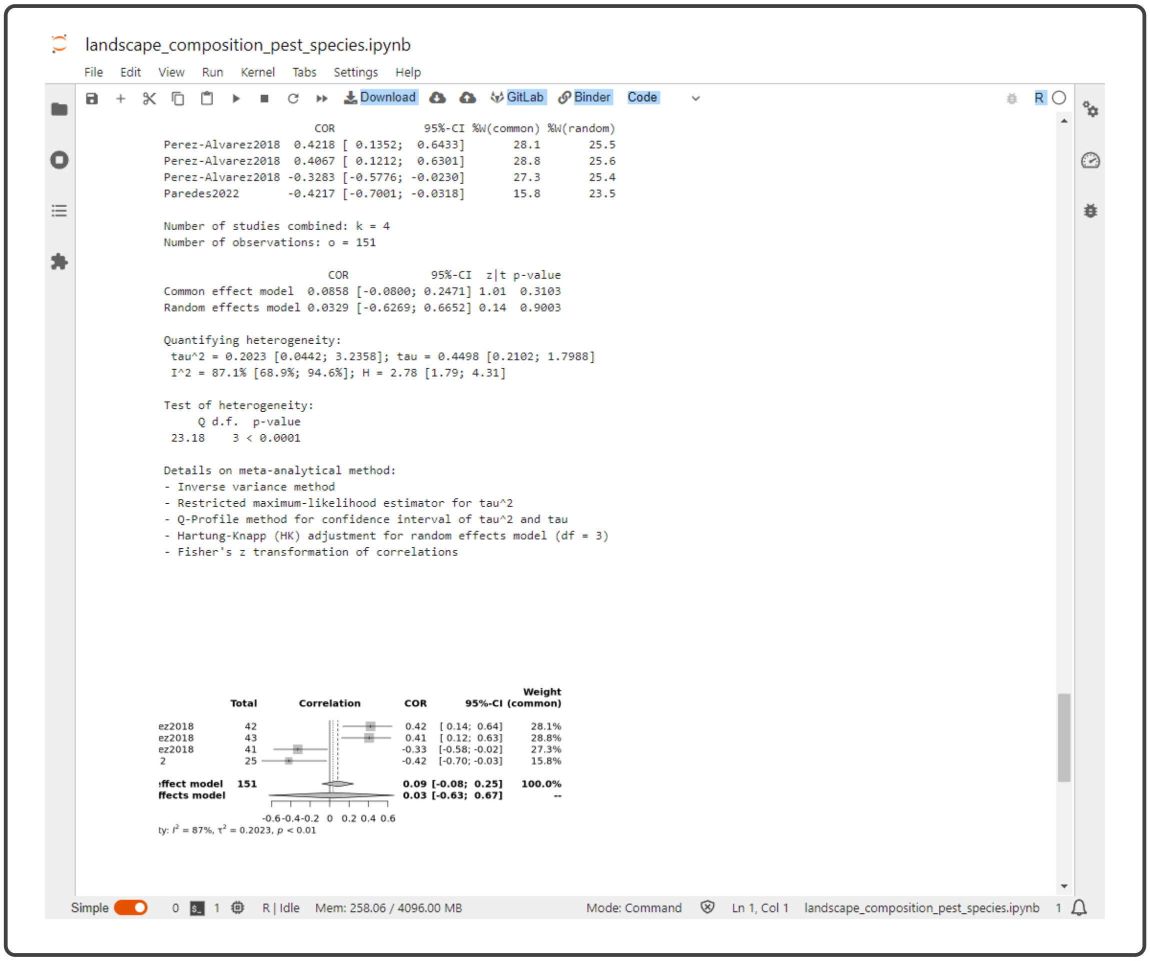
has anova: ANOVA with bo (bactrocera oleae abundance) as the response variable and shdi (shannon diversity index) as term.

has implementation:

```
#############################################
####FIGURE 4: 10.1007/s10340-022-01489-1#####
##############################################
####Library to do the model
library(MASS)#
...
```

has lmm fitting: A linear mixed model (LMM) fitting with bactrocera oleae abundance (bo) as the response variable and shdi (shannon diversity index) as a fixed effect

Figure 5.10: The contribution from "Landscape simplification increases Bactrocera oleae abundance in olive groves: adult population dynamics in different land uses" in the ORKG.



```
                         COR            95%-CI %W(common) %W(random)
Perez-Alvarez2018  0.4218 [ 0.1352;  0.6433]       28.1       25.5
Perez-Alvarez2018  0.4067 [ 0.1212;  0.6301]       28.8       25.6
Perez-Alvarez2018 -0.3283 [-0.5776; -0.0230]       27.3       25.4
Paredes2022       -0.4217 [-0.7001; -0.0318]       15.8       23.5

Number of studies combined: k = 4
Number of observations: o = 151

                           COR            95%-CI  z|t p-value
Common effect model  0.0858 [-0.0800; 0.2471] 1.01  0.3103
Random effects model 0.0329 [-0.6269; 0.6652] 0.14  0.9003

Quantifying heterogeneity:
 tau^2 = 0.2023 [0.0442; 3.2358]; tau = 0.4498 [0.2102; 1.7988]
 I^2 = 87.1% [68.9%; 94.6%]; H = 2.78 [1.79; 4.31]

Test of heterogeneity:
     Q d.f.  p-value
 23.18    3 < 0.0001

Details on meta-analytical method:
- Inverse variance method
- Restricted maximum-likelihood estimator for tau^2
- Q-Profile method for confidence interval of tau^2 and tau
- Hartung-Knapp (HK) adjustment for random effects model (df = 3)
- Fisher's z transformation of correlations
```

Total
Correlation
COR
95%-CI
Weight (common)
ez2018 42 0.42 [0.14; 0.64] 28.1%
ez2018 43 0.41 [0.12; 0.63] 28.8%
ez2018 41 -0.33 [-0.58; -0.02] 27.3%
2 25 -0.42 [-0.70; -0.03] 15.8%
ffect model 151 0.09 [-0.08; 0.25] 100.0%
ffects model 0.03 [-0.63; 0.67] --
-0.6 -0.4 -0.2 0 0.2 0.4 0.6
ty: I^2 = 87%, τ^2 = 0.2023, p < 0.01
Simple
0
1
R | Idle
Mem: 258.06 / 4096.00 MB
Mode: Command
Ln 1, Col 1
landscape_composition_pest_species.ipynb
1

Figure 5.11: The results of the meta-analysis inside a Jupyter notebook.

Chapter 6

Discussion

In the previous chapter, we demonstrated how to produce rich and high-quality machine machine-reusable scientific knowledge (RQ1). Subsequently, we reapplied the approach to two published papers to demonstrate its applicability post-publication. Finally, we conducted a proof-of-concept synthesis on the two published papers after they underwent the born-reusable approach to demonstrate how rich and high quality machine-reusable scientific knowledge can support research synthesis (RQ2).

6.1 Implications

The born-reusable approach has several implications for authors, knowledge repositories, publishers and the scientific community.

We have illustrated the potential benefits of having authors describe their work as machine-reusable data prior to publication. Authors are in the best position to richly describe their papers, as they are the ultimate experts of their work and can provide unique insights that readers may lack. They also have direct access to the raw source code, data sets, and statistical data - as demonstrated in Chapter 5 - much of which is often inaccessible post-publication.

The production of machine-reusable scientific knowledge from within the programming environment used to conduct research may also provide benefits to curators who are tasked with creating richly-described machine-reusable descriptions of research. Firstly, it allows curators to work with authors to gradually create machine-reusable data over the research phase, spreading the workload. Secondly, working with the same research environment is potentially more efficient as it negates the need for external tools that may lead to data processing errors, such as input and conversion errors.

The application of this approach to a published paper in chapter 5 demonstrates its relevance to not only pre-published papers but also published papers. This implies that the knowledge published in papers that share source code and data sets could be made machine-reusable retrospectively.

Research synthesis is of particular interest to ecologists as it is pivotal for advancing ecological knowledge given the immense volume of diverse data [[51], [52]]. By conducting a proof-of-concept meta-analysis on two born-reusable ecology papers (Chapter 5.2), we also demonstrated how this approach could support research synthesis and, ultimately, the reuse of scientific knowledge by making it more FAIR. This approach ensures machine-reusable data is findable through the ORKG knowledge graph using the papers DOI and accessible using either the web interface, SPARQL endpoint or API. Furthermore, the scientific papers have richly described statistical models, statistical results and variables using standard vocabularies, allowing for improved interoperability between various programming languages (i.e. Python, R, Matlab) and statistical packages. Finally, the scientific knowledge is provided for reuse under the ORKG usage license.

The proposed approach also reduces the dependency on post-publication methods such as crowdsourcing and NLP. It could assist humans and improve the scalability of crowdsourcing. Knowledge graph curators could map the JSON-LD data to existing ontologies and subsequently build upon this information. It could also assist current NLP methods with its machine-reusable data. For example, the structured data could be integrated into ML models and combined with current NLP methods to improve the accuracy of knowledge graph construction from unstructured data.

Reviewers and publishers may also benefit from this approach. Reviewers would have machine-reusable data, code snippets and datasets at their disposal prior to publication, thus allowing them to test results directly from the source code. It also facilitates the implementation of machine-assisted reviewing, which may speed up review times, improve review standards and ultimately increase the scientific research's reproducibility.

6.2 Limitations

Although we have outlined a number of promising implications for this approach, it also possesses several limitations that could restrict its wider adoption in the scientific community.

One such limitation, in the context of the implementation in this thesis, is its lack of interoperability. The dependency on ORKG templates limits authors to the use of properties and classes (types) native to the ORKG. This does not allow for the use of external properties, classes, ontologies, or interoperability with other knowledge bases. A move away from the ORKG centric view may potentially allow for wider adoption by authors and other knowledge bases.

Another challenge for this approach is the adoption by authors. It is anticipated that the production of machine-reusable descriptions could result in a significant increase in authors' workloads. This may deter authors who do not understand the reward of richly describing their paper in a machine-reusable format. Also, those from non-computing scientific fields may have little experience with the ideas behind this approach, such as FAIR, knowledge graphs, JSON-LD, and may need additional assistance with integrating this approach into their workflow.

Another potential impediment to the adoption of this approach is the support by publishers. It is highly dependent on the DOIs and their associated metadata to make the JSON-LD data findable and accessible. As this is a novel approach, there are currently no standards in place for handling and storing the associated JSON-LD data. Ultimately, in order to be FAIR, JSON-LD data should be accessible over a DOI. This DOI must be known prior to publication, which is a challenge that needs to be solved. Data provenance is also a crucial factor for these files. It is important that the JSON-LD data can verifiably be linked to the publication and can be tracked for modifications.

6.3 Future Work

The limitations of the proposed approach provide inspiration for several interesting areas of further development.

In order to improve the interoperability of this approach, we propose investigating how to expand beyond the ORKG to support other ontologies and knowledge graphs. One way this could be resolved is by extending support for the importation of OWL ontologies from ontology repositories such as BioPortal [53] to move away from the ORKG centric view. This could also be facilitated through the use of the Shapes Constraint Language (SHACL), which allows for the validation of the JSON-LD data against RDF knowledge graphs [54]. This could possibly be implemented in R using the open-source Jsonld [1] package, which allows for the flexible creation of JSON-LD data within the R environment alongside a SHACL validation engine using an API wrapper. Potential architecture for other program languages, such as Python, could also be considered.

We also envision that the adoption by authors could be supported by machine-assisted automated production of structured data and templates within the programming environment. Potential ML models could identify relevant scientific information from the source code, such as input data sets, statistical methods, statistical results, and relevant figures and automatically structure the data into a relevant template. This data can later be manually verified and modified by authors and researchers.

To promote the support of this approach by publishers, guidelines and standards will need to be introduced. We propose that this can be assisted by the creation of a repository dedicated to storing JSON-LD data. This repository could take influence from popular repositories for data sets such as Dryad [2] and follow FAIR principles. For example, it must store the JSON-LD data indefinitely, assign a DOI, have an open usage license and provide data provenance.

[1] https://cran.r-project.org/web/packages/jsonld/index.html

[2] https://datadryad.org/

Chapter 7

Conclusions

In this thesis, we investigated how the born-reusable approach for the production of machine-reusable scientific knowledge can be applied to an unpublished paper and how the production of richly described scientific information during the research phase can potentially reduce reliance on current methods for the creation and curation of scientific knowledge bases such as crowdsourcing and NLP. Furthermore, we demonstrated that this approach is also applicable to published papers. Finally, we demonstrated how the born-reusable approach could potentially support research synthesis and therefore facilitate the reuse of scientific knowledge.

Despite the potential positive implications for scientific knowledge, the interoperability of this approach, specifically in regards to the production and harvesting of JSON-LD data, must be expanded to support wider adoption by authors and knowledge bases. Furthermore, it is important that authors are supported in the adoption of this approach into their workflows. Finally, improved cooperation with publishers is crucial for expanding the infrastructure (DOIs, metadata) that underpins this approach.

In conclusion, this thesis represents an early demonstration of the production of machine-reusable scientific information prior to publication and presents the potential benefits of producing machine-reusable scientific knowledge directly within programming environments used to conduct research. The insights in this thesis could lay a foundation for further research into improving the reusability of scientific knowledge.

Bibliography

[1] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. "Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases". In: *Humanities and Social Sciences Communications* 8.1 (Oct. 2021). DOI: 10.1057/s41599-021-00903-w. URL: https://doi.org/10.1057/s41599-021-00903-w.

[2] Rob Johnson, Anthony Watkinson, and Michael Mabe. *The STM Report, Fifth Edition.* https://www.stm-assoc.org/2018_10_04_STM_Report_2018.pdf, Accessed: 28-6-2023. Oct. 2018.

[3] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 2016). DOI: 10.1038/sdata.2016.18. URL: https://doi.org/10.1038/sdata.2016.18.

[4] Jens Lehmann et al. "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic Web* 6.2 (2015), pp. 167–195. DOI: 10.3233/sw-140134. URL: http://dx.doi.org/10.3233/SW-140134.

[5] Amit Singhal. *Introducing the Knowledge Graph: things, not strings.* https://blog.google/products/search/introducing-knowledge-graph-things-not/, Accessed: 28-6-2023. May 2012.

[6] Aidan Hogan et al. "Knowledge Graphs". In: *ACM Computing Surveys* 54.4 (July 2021), pp. 1–37. DOI: 10.1145/3447772. URL: https://doi.org/10.1145/3447772.

[7] Markus Stocker et al. "FAIR scientific information with the Open Research Knowledge Graph". In: *FAIR Connect* 1.1 (Jan. 2023). Ed. by Barbara Magagna, pp. 19–21. DOI: 10.3233/fc-221513. URL: https://doi.org/10.3233/fc-221513.

[8] Jennifer D'Souza and Sören Auer. *NLPContributions: An Annotation Scheme for Machine Reading of Scholarly Contributions in Natural Language Processing Literature.* 2020. DOI: 10.48550/ARXIV.2006.12870. URL: https://arxiv.org/abs/2006.12870.

[9] Allard Oelen, Markus Stocker, and Sören Auer. "TinyGenius". In: *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries.* ACM, June 2022. DOI: 10.1145/3529372.3533285. URL: https://doi.org/10.1145/3529372.3533285.

[10] Liang Yao, Chengsheng Mao, and Yuan Luo. *KG-BERT: BERT for Knowledge Graph Completion.* 2019. DOI: 10.48550/ARXIV.1909.03193. URL: https://arxiv.org/abs/1909.03193.

[11] Kun Cao and James Fairbanks. *Unsupervised Construction of Knowledge Graphs From Text and Code*. 2019. DOI: 10.48550/ARXIV.1908.09354. URL: https://arxiv.org/abs/1908.09354.

[12] Jennifer D'Souza and Sören Auer. "Computer Science Named Entity Recognition in the Open Research Knowledge Graph". In: *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries*. Springer International Publishing, 2022, pp. 35–45. DOI: 10.1007/978-3-031-21756-2_3. URL: https://doi.org/10.1007/978-3-031-21756-2_3.

[13] Mirko Gabelica, Ružica Bojčić, and Livia Puljak. "Many researchers were not compliant with their published data sharing statement: a mixed-methods study". In: *Journal of Clinical Epidemiology* 150 (Oct. 2022), pp. 33–41. DOI: 10.1016/j.jclinepi.2022.05.019. URL: https://doi.org/10.1016/j.jclinepi.2022.05.019.

[14] Chenyue Jiao, Kai Li, and Zhichao Fang. "Data sharing practices across knowledge domains: A dynamic examination of data availability statements in iPLOS ONE/i publications". In: *Journal of Information Science* (June 2022), p. 016555152211018. DOI: 10.1177/01655515221101830. URL: https://doi.org/10.1177/01655515221101830.

[15] Lisa M. Federer. "Long-term availability of data associated with articles in PLOS ONE". In: *PLOS ONE* 17.8 (Aug. 2022). Ed. by Jelte M. Wicherts, e0272845. DOI: 10.1371/journal.pone.0272845. URL: https://doi.org/10.1371/journal.pone.0272845.

[16] Sören Auer et al. "Improving Access to Scientific Literature with Knowledge Graphs". In: *Bibliothek Forschung und Praxis* 44.3 (Nov. 2020), pp. 516–529. DOI: 10.1515/bfp-2020-2042. URL: https://doi.org/10.1515/bfp-2020-2042.

[17] Allard Oelen et al. *COVID-19 Reproductive Number Estimates*. en. 2020. DOI: 10.48366/R44930. URL: https://www.orkg.org/orkg/comparison/R44930.

[18] *Curation Grants*. https://orkg.org/about/28/Curation_Grants, Accessed: 1-7-2023.

[19] Marco Anteghini et al. *SciBERT-based Semantification of Bioassays in the Open Research Knowledge Graph*. 2020. DOI: 10.34657/5192. URL: https://oa.tib.eu/renate/handle/123456789/6144.

[20] *ORKG-NLP Services*. https://orkg-nlp-pypi.readthedocs.io/en/latest/services/services.html, Accessed: 14-7-2023.

[21] Allard Oelen. "Leveraging human-computer interaction and crowdsourcing for scholarly knowledge graph creation". en. In: (2022). DOI: 10.15488/13066. URL: https://www.repo.uni-hannover.de/handle/123456789/13171.

[22] Zied Boubakri. "The ORKG R Package and Its Use in Data Science". en. In: (2022). DOI: 10.15488/13072. URL: https://www.repo.uni-hannover.de/handle/123456789/13177.

[23] Thomas Kluyver et al. "Jupyter Notebooks-a publishing format for reproducible computational workflows." In: *Elpub* 2016 (2016), pp. 87–90.

[24] *About Papers With Code*. https://paperswithcode.com/about, Accessed: 9-7-2023.

[25] Giuliana Spadaro et al. "The Cooperation Databank: Machine-Readable Science Accelerates Research Synthesis". In: *Perspectives on Psychological Science* 17.5 (May 2022), pp. 1472–1489. DOI: 10.1177/17456916211053319. URL: https://doi.org/10.1177/17456916211053319.

[26] *OpenBiodiv - About*. https://openbiodiv.net/about, Accessed: 11-7-2023.

[27] Viktor Senderov et al. "OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system". In: *Journal of Biomedical Semantics* 9.1 (Jan. 2018). DOI: 10.1186/s13326-017-0174-5. URL: https://doi.org/10.1186/s13326-017-0174-5.

[28] Rudi Studer, V.Richard Benjamins, and Dieter Fensel. "Knowledge engineering: Principles and methods". In: *Data &amp Knowledge Engineering* 25.1-2 (Mar. 1998), pp. 161–197. DOI: 10.1016/s0169-023x(97)00056-6. URL: https://doi.org/10.1016/s0169-023x(97)00056-6.

[29] *Biology Knowledge Graph — Data structured for insights*. https://www.elsevier.com/solutions/biology-knowledge-graph, Accessed: 9-7-2023.

[30] Martin Enders et al. "A conceptual map of invasion biology: Integrating hypotheses into a consensus network". In: *Global Ecology and Biogeography* 29.6 (Mar. 2020). Ed. by Jonathan Belmaker, pp. 978–991. DOI: 10.1111/geb.13082. URL: https://doi.org/10.1111/geb.13082.

[31] Jonathan M. Jeschke et al. "Towards an open, zoomable atlas for invasion science and beyond". In: *NeoBiota* 68 (Aug. 2021), pp. 5–18. DOI: 10.3897/neobiota.68.66685. URL: https://doi.org/10.3897/neobiota.68.66685.

[32] Liang Yao, Chengsheng Mao, and Yuan Luo. *KG-BERT: BERT for Knowledge Graph Completion*. 2019. DOI: 10.48550/ARXIV.1909.03193. URL: https://arxiv.org/abs/1909.03193.

[33] Kun Cao and James Fairbanks. *Unsupervised Construction of Knowledge Graphs From Text and Code*. 2019. DOI: 10.48550/ARXIV.1908.09354. URL: https://arxiv.org/abs/1908.09354.

[34] Ayoub Harnoune et al. "BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis". In: *Computer Methods and Programs in Biomedicine Update* 1 (2021), p. 100042. DOI: 10.1016/j.cmpbup.2021.100042. URL: https://doi.org/10.1016/j.cmpbup.2021.100042.

[35] Danilo Dessı et al. "SCICERO: A deep learning and NLP approach for generating scientific knowledge graphs in the computer science domain". In: *Knowledge-Based Systems* 258 (Dec. 2022), p. 109945. DOI: 10.1016/j.knosys.2022.109945. URL: https://doi.org/10.1016/j.knosys.2022.109945.

[36] Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara. "Towards an integrated crowdsourcing definition". In: *Journal of Information Science* 38.2 (Mar. 2012), pp. 189–200. DOI: 10.1177/0165551512437638. URL: https://doi.org/10.1177/0165551512437638.

[37] *Wikidata:Statistics*. https://www.wikidata.org/wiki/Wikidata:Statistics, Accessed: 9-7-2023.

[38] *Observatories*. https://orkg.org/about/27/Observatories, Accessed: 1-7-2023.

[39] Zhaoan Dong et al. "Using hybrid algorithmic-crowdsourcing methods for academic knowledge acquisition". In: *Cluster Computing* 20.4 (Sept. 2017), pp. 3629–3641. DOI: 10.1007/s10586-017-1089-8. URL: https://doi.org/10.1007/s10586-017-1089-8.

[40] Allard Oelen. “Leveraging human-computer interaction and crowdsourcing for scholarly knowledge graph creation”. en. In: (2022). DOI: 10.15488/13066. URL: https://www.repo.uni-hannover.de/handle/123456789/13171.

[41] Rodney Kinney et al. *The Semantic Scholar Open Data Platform*. 2023. DOI: 10.48550/ARXIV.2301.10140. URL: https://arxiv.org/abs/2301.10140.

[42] *Connected Papers — Find and explore academic papers*. https://www.connectedpapers.com/about, Accessed: 11-7-2023.

[43] Chris Graf et al. “The Open Data Challenge: An Analysis of 124, 000 Data Availability Statements and an Ironic Lesson about Data Management Plans”. In: *Data Intelligence* 2.4 (Oct. 2020), pp. 554–568. DOI: 10.1162/dint_a_00061. URL: https://doi.org/10.1162/dint_a_00061.

[44] Christof Bless. “SciKGTeX - A LATEX Package to Semantically Annotate Contributions in Scientific Publications”. en. In: (2022). DOI: 10.15488/12462. URL: https://www.repo.uni-hannover.de/handle/123456789/12561.

[45] *JSON-LD 1.1*. https://www.w3.org/TR/json-ld11/, Accessed: 19-7-2023.

[46] DataCite Metadata Working Group. “DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.3”. en. In: (2019). DOI: 10.14454/7XQ3-ZF69. URL: https://schema.datacite.org/meta/kernel-4.3/.

[47] Ricardo Perez-Alvarez, Brian A. Nault, and Katja Poveda. “Contrasting effects of landscape composition on crop yield mediated by specialist herbivores”. In: *Ecological Applications* 28.3 (Apr. 2018), pp. 842–853. DOI: 10.1002/eap.1695. URL: https://doi.org/10.1002/eap.1695.

[48] Daniel Paredes et al. “Landscape simplification increases Bactrocera oleae abundance in olive groves: adult population dynamics in different land uses”. In: *Journal of Pest Science* 96.1 (Mar. 2022), pp. 71–79. DOI: 10.1007/s10340-022-01489-1. URL: https://doi.org/10.1007/s10340-022-01489-1.

[49] *Ontobee*. http://purl.obolibrary.org/obo/OBI_0000011, Accessed: 16-7-2023.

[50] Barbara Magagna et al. “The I-ADOPT Interoperability Framework: a proposal for FAIRer observable property descriptions”. In: (Mar. 2021). DOI: 10.5194/egusphere-egu21-13155. URL: https://doi.org/10.5194/egusphere-egu21-13155.

[51] Benjamin S Halpern et al. “Ecological Synthesis and Its Role in Advancing Knowledge”. In: *BioScience* (Sept. 2020). DOI: 10.1093/biosci/biaa105. URL: https://doi.org/10.1093%2Fbiosci%2Fbiaa105.

[52] Stephanie E Hampton et al. “Big data and the future of ecology”. In: *Frontiers in Ecology and the Environment* 11.3 (Apr. 2013), pp. 156–162. DOI: 10.1890/120103. URL: https://doi.org/10.1890%2F120103.

[53] P. L. Whetzel et al. “BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications”. In: *Nucleic Acids Research* 39.suppl (June 2011), W541–W545. DOI: 10.1093/nar/gkr469. URL: https://doi.org/10.1093/nar/gkr469.

[54] *Shapes Constraint Language (SHACL)*. `https://www.w3.org/TR/shacl/`, Accessed: 9-7-2023.