Research Ideas and Outcomes 8: e83789
doi: 10.3897/rio.8.e83789

Grant Proposal

SKG4EOSC - Scholarly Knowledge Graphs for EOSC: Establishing a backbone of knowledge graphs for FAIR Scholarly Information in EOSC

Markus Stocker[‡,§], Tina Heger[|,¶], Artur M. Schweidtmann[#], Hanna Ćwiek-Kupczyńska[¤], Lyubomir Penev[«], Milan Dojchinovski[»,^], Egon Willighagen[˅], Maria-Esther Vidal[§,‡], Houcemeddine A. Turki[¦], Daniel Balliet[ˀ], Ilaria Tiddi[ˀ], Tobias Kuhn[ˀ], Daniel Mietchen[|,ˁ], Oliver Karras[‡], Lars Vogt[‡], Sebastian Hellmann[»], Jonathan M. Jeschke[|,ˁ], Paweł Krajewski[¤], Sören Auer[‡,§]

‡ TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany
§ Leibniz University Hannover, Hannover, Germany
| Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany
¶ Technical University of Munich, Germany; TUM School of Life Sciences, Freising, Germany
# Department of Chemical Engineering, Delft University of Technology, Delft, Netherlands
¤ Institute of Plant Genetics, Polish Academy of Sciences, Poznań, Poland
« Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
» Institut Für Angewandte Informatik e.V., Leipzig, Germany
^ Faculty of Information Technology, Czech Technical University in Prague, Prague, Czech Republic
˅ Maastricht University, Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht, Netherlands
¦ Data Engineering and Semantics Research Unit, Faculty of Sciences of Sfax, University of Sfax, Sfax, Tunisia
ˀ Vrije Universiteit Amsterdam, Amsterdam, Netherlands
ˁ Institute of Biology, Freie Universität Berlin, Berlin, Germany

Corresponding author: Markus Stocker (markus.stocker@tib.eu)

Reviewable v 1

Received: 14 Mar 2022 | Published: 15 Mar 2022

Citation: Stocker M, Heger T, Schweidtmann AM, Ćwiek-Kupczyńska H, Penev L, Dojchinovski M, Willighagen E, Vidal M-E, Turki HA, Balliet D, Tiddi I, Kuhn T, Mietchen D, Karras O, Vogt L, Hellmann S, Jeschke JM, Krajewski P, Auer S (2022) SKG4EOSC - Scholarly Knowledge Graphs for EOSC: Establishing a backbone of knowledge graphs for FAIR Scholarly Information in EOSC. Research Ideas and Outcomes 8: e83789. https://doi.org/10.3897/rio.8.e83789

Abstract

In the age of advanced information systems powering fast-paced knowledge economies that face global societal challenges, it is no longer adequate to express scholarly information - an essential resource for modern economies - primarily as article narratives in document form. Despite being a well-established tradition in scholarly communication, PDF-based text publishing is hindering scientific progress as it buries scholarly information

into non-machine-readable formats. The key objective of SKG4EOSC is to improve science productivity through development and implementation of services for text and data conversion, and production, curation, and re-use of FAIR scholarly information. This will be achieved by (1) establishing the Open Research Knowledge Graph (ORKG, orkg.org), a service operated by the SKG4EOSC coordinator, as a Hub for access to FAIR scholarly information in the EOSC; (2) lifting to EOSC of numerous and heterogeneous domain-specific research infrastructures through the ORKG Hub's harmonized access facilities; and (3) leverage the Hub to support cross-disciplinary research and policy decisions addressing societal challenges. SKG4EOSC will pilot the devised approaches and technologies in four research domains: biodiversity crisis, precision oncology, circular processes, and human cooperation. With the aim to improve machine-based scholarly information use, SKG4EOSC addresses an important current and future need of researchers. It extends the application of the FAIR data principles to scholarly communication practices, hence a more comprehensive coverage of the entire research lifecycle. Through explicit, machine actionable provenance links between FAIR scholarly information, primary data and contextual entities, it will substantially contribute to reproducibility, validation and trust in science. The resulting advanced machine support will catalyse new discoveries in basic research and solutions in key application areas.

Keywords

Scholarly literature, Scholarly information, Scholarly communication, Knowledge Graphs, FAIR, Information extraction

1. Excellence

1.1 Objectives and ambition

Expressing scholarly information primarily in narrative documents is outdated and hindering scientific progress. The use of printed articles (and their now pseudo-digitalized static PDFs) is a relic of historic developments dating back to the very beginning of science (see Fig. 1). In the age of advanced information systems powering fast-paced knowledge economies that face numerous global societal challenges, it is no longer adequate to express scholarly information - an essential resource for modern economies - primarily as article narratives in document form. Text and data mining tries to overcome these issues, with only limited results, unless the extracted information and data are liberated, quality checked and FAIRified to allow for a wider and efficient reuse.

The human effort required to comprehend information expressed in such form can no longer keep pace with the overall speed of science and subsequent demands on all research lifecycle phases, including information production and review. The urgency of results, seen for instance in vaccine research among other global societal challenges, or the relentless growth of scholarly information, makes it increasingly hard to gain or maintain an overview of the state of the art (Jeschke et al. 2019). The current lack **of**

FAIRness of scholarly information is a serious impediment to science productivity, and the more efficient use of scholarly information is an important current need of the research community at large and, more broadly, the global society. We argue that the need to more efficiently use scholarly information will only grow stronger in the future, yet will be hindered by the growing number of documents added to the total.Hence, the European Open Science Cloud (EOSC) and its infrastructures urgently need to increase their service offering and capabilities to address the FAIRness of scholarly information. Indeed, the 2016 first report and recommendations of the Commission High Level Expert Group on EOSC (European Commission 2016) had already identified "*New modes of scholarly communication (with emphasis on machine actionability) need to be implemented*" as a key factor for the effective development of the EOSC as part of Open Science. In 2018, the European Commission expert group on FAIR data provided a follow-up report (European Commission 2018) including an action plan with recommendations for stakeholders to take towards a FAIR ecosystem, calling for discipline-aware implementations of EOSC-related infrastructure centred around FAIR Digital Objects, including scholarly information. Similarly, the 2020 EOSC Secretariat Workshop (Akgun et al. 2020) on "*Co-creating the EOSC: Needs and requirements for future research environments*" underscored the need for services for machine actionable scholarly information sharing. Below, we illustrate this need with an example user story and associated scenario to highlight the problems of the current document-centric information flows and how our proposal SKG4EOSC innovatively addresses them.

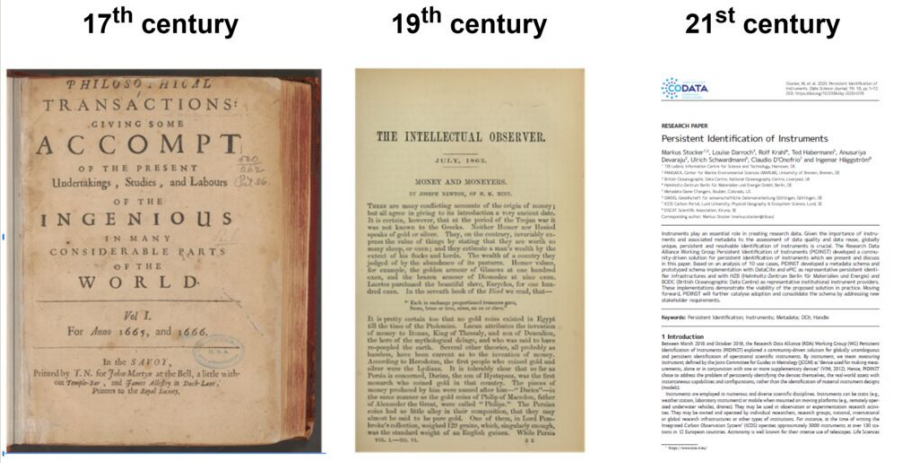

Figure 1. doi

While other information-intense domains were completely disrupted (e.g. mail order catalogues, street maps), scholarly information continues to be published as static documents, a root cause of major challenges for the scientific method overall, including proliferation of publications and reproducibility crisis. (Sources: 17th century, https://doi.org/10.1098/rstl.1665.0001; 19th century, https://www.biodiversitylibrary.org/page/39489576; and 21st century, https://doi.org/10.5334/dsj-2020-018.)

User Story

As a researcher, I want to discover relevant work in a research area to get an overview of the state of the art.

Scenario

Catherine is beginning her doctoral studies with a focus on circular economy. As a newcomer to this research area, she uses digital libraries of major publishers to discover relevant work. In this way, she learns which approaches are currently leading and how they are evaluated. Catherine will skim through hundreds of papers, most of which turn out to be irrelevant for her research. Dozens of articles will need to be read in detail and the essential information Catherine needs for her research will be manually organized.

Problems

- Initially, Catherine does not know the right keywords and mostly finds irrelevant literature.
- Search results include articles rather than the information Catherine needs.
- Because the information in documents is not machine actionable, Catherine spends substantial time manually extracting, organizing, and processing the required information.
- Since Catherine cannot easily share her evolving literature review, Alex - another PhD student - facing the same problems as Catherine cannot build on her (cognitive) work.
- Being a brilliant early career researcher, Catherine will advance the state-of-the-art, but can only communicate her findings with another article.

How SKG4EOSC innovates

- Services for FAIR scholarly information production, curation, and use in EOSC will enable exploiting scholarly information in a fine-grained manner, not merely at the level of articles.
- Researchers are presented with the state-of-the-art information known about a research problem.
- Next-generation semantic publishing tools ensure machine actionability of content at the time of publication.
- Integration of machine actionable content and data pre- and post-publication into knowledge graphs will bring together the legacy and future of our scientific knowledge.

1.1.1 Overall objectives

The main objective of SKG4EOSC is to **improve science productivity with services for the production, curation, and use of FAIR scholarly information**. Scholarly information is information*[1] expressed as scholarly literature or as databases with information extracted from the literature (see Fig. 2 for representative concepts from various domains).

Hence, the project applies the FAIR data principles (Wilkinson et al. 2016) to scholarly information and the technical infrastructure to support it. Specifically, SKG4EOSC will establish the *Open Research Knowledge Graph*. (Jaradeh et al. 2019) (ORKG*[2], orkg.org) as a Hub for access to FAIR scholarly information in the EOSC and leverage the Hub to advance innovative and customizable EOSC services for the production, curation, and use of FAIR scholarly information as summarized in Table 1. With the aim to improve machine-based and efficient scholarly information use, SKG4EOSC addresses an important current and future need of researchers. Furthermore, it extends the application of the FAIR data principles from the research data lifecycle to the scholarly communication lifecycle, therefore more comprehensively covering the entire research lifecycle. The objectives are thus pertinent to the work programme topic and destination. Given the substantial related work by SKG4EOSC partners, these objectives are realistically achievable.

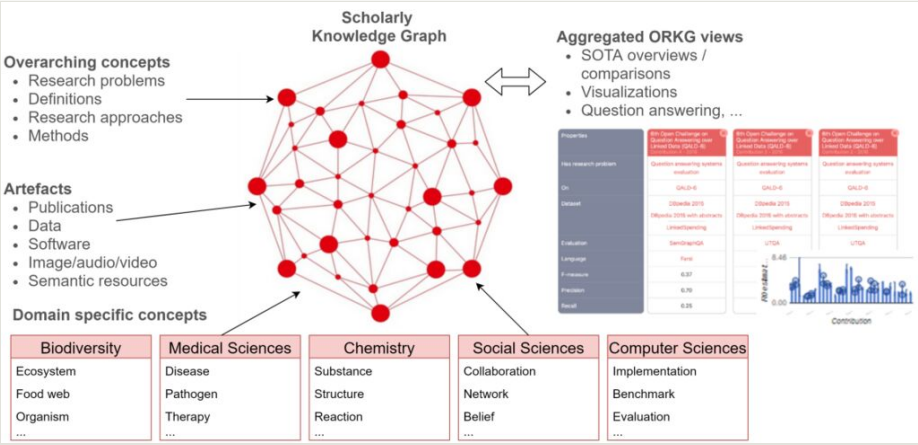

Figure 2. doi

Vision of semantically organizing and interlinking overarching, generic concepts and artefacts as well as domain-specific concepts of the research lifecycle in a knowledge graph.

1.1.2 Progress beyond the state of the art

There exist numerous commercial and non-commercial services that publish structured scholarly metadata following FAIR and/or Linked Data Principles, most of which also provide programmatic access to content via Web-based APIs. Predictably, most of the large and global scale services publish metadata about artefacts, in particular scholarly articles, datasets, and software, or metadata about other entities, e.g. people and organizations. Metadata standardization has a long history, and as a result, there are numerous widely-used schemas (e.g. DataCite Metadata Schema, Dublin Core, W3C PROV, etc.) and established curation workflows that ensure metadata about scholarly artefacts and their contextual entities are FAIR and facilitate finding and accessing the described artefacts. With such standardization, it became possible to build infrastructures with excellent global coverage, thus enabling finding and accessing millions of artefacts

and information describing them. **However, metadata is only of very limited direct value to researchers aiming to answer research questions.**

Table 1.

Overview of the challenges the project addresses, as well as the expected results and their impact.

Challenge	Objective	Approach	Result	Impact
Scholarly information is inefficient to use	Make scholarly information FAIR and a first-class citizen in the EOSC	Leverage the established *ORKG* to federate community initiatives and enable the development of common services	*ORKG* as a Hub for large-scale FAIR scholarly information covering millions of research findings in the EOSC	More efficient discovery, analysis, and reuse of scholarly information in the EOSC
The vast majority of scholarly information is buried and dispersed in legacy documents and is thus machine inactionable	Efficient, scalable, granular, and accurate extraction of scholarly information from the literature in the EOSC	Realize FAIR scholarly information extraction using hybrid AI strategies (e.g., NLP, ML) incorporating the wisdom of the crowd (i.e. crowdsourcing).	Novel EOSC service for extracting FAIR scholarly information from literature and related assets such as workflows	Scalable, maximally automated post-publication FAIRification of scholarly information in the literature
Structure & semantics of scholarly information produced in research lifecycles is not adequately preserved (cf. reproducibility crisis)	Scholarly information produced and published FAIR in the research lifecycle	Extend digital tools (e.g., for data analysis and scholarly communication) used by researchers with features that ensure scholarly information is produced and published FAIR	Transferrable approaches and tools that embed scholarly information FAIRification into the research lifecycle	Scalable, maximally automated pre-publication FAIR-at-Birth scholarly information, contributing to the provenance and reproducibility of results
Limited opportunities to create advanced services (e.g. discovery, analysis, visualization) for static research articles	Advanced infrastructures for scholarly information exploitation	Leverage the *ORKG* Hub as the harmonized single-point-of-entry for FAIR scholarly information in the EOSC for complex digital research activities through service composition	Advanced information infrastructure services for scholarly information discovery, state-of-the-art comparison, systematic reviews, data science, scholarly communication	Increased efficiency of disciplinary and interdisciplinary research
Numerous disjoint and uncoordinated disciplinary efforts to FAIRify scholarly information	Synergistic harmonization and integration of disciplinary efforts to enable FAIR scholarly information use in interdisciplinary research	Elevate disciplinary approaches to the *ORKG* Hub for FAIR scholarly information in the EOSC, adapt the Hub-enabled services for production, curation and use in (inter-) disciplinary research	Adaptable services for the production, curation, and use of FAIR scholarly knowledge in disciplinary and interdisciplinary research	Direct and measurable added value of FAIR scholarly information and services in EOSC for research and research communities

As a consequence, for the actual data/content within scholarly articles, i.e. the scholarly information, there are disciplinary efforts and infrastructures being developed, but none has succeeded at the large scale as seen for bibliographic metadata. The *ORKG* is a FAIR-driven infrastructure for scholarly information developed at TIB since 2018. The ORKG implements digital library services that support acquiring, curating, publishing, and processing FAIR scholarly information in a variety of disciplines. In SKG4EOSC, the *ORKG* is the central component providing unified access to the heterogeneous scholarly information published by disciplinary infrastructures.

The EOSC has a strong focus on the application of the FAIR principles to research data in a classical sense, i.e. primary (e.g. sensor or experimental) or derivative data products in tabular text or binary forms. **SKG4EOSC will make the first steps in applying the FAIR principles to the content of scholarly articles in the EOSC.**

Table 2 lists some well-known (global) services for publishing metadata about articles, datasets, people, and organizations. In contrast to all these services, SKG4EOSC will go beyond the state of the art by enabling the access to FAIR scholarly information, i.e. not merely metadata but also the article contents (scientific assertions and claims), as well as the linking of data and (bibliographic) metadata. Table 3 summarizes how SKG4EOSC will go beyond some leading infrastructures for publishing structured scholarly information, many of which are involved in SKG4EOSC. Aspects shared with all these infrastructures for how SKG4EOSC advances the state of the art include interdisciplinary access and integration of FAIR scholarly information and the leveraging of generic services and tools for information visualization and processing.

Table 2.

Knowledge graphs and databases publishing bibliographic metadata or metadata about other artefacts such as datasets or entities such as people and organizations.

Existing service	Description
SciGraph (CC BY 4.0)	SpringerNature service, providing access to linked metadata about SpringerNature publications.
Crossref (CC BY 4.0)	Non-profit organization supporting the persistent identification of scholarly artefacts and publishing of metadata about them.
DataCite (CC0)	Non-profit organization that provides persistent identifiers for research data and other research outputs.
Open Citations (CC0)	Non-profit organization providing bibliographic and citation metadata for scholarly publications.
ResearchGraph (CC BY 4.0)	A non-profit metadata organization initiative closely aligned with the Research Data Alliance.
Semantic Scholar (ODC-BY)	A search engine that uses NLP methods to improve publication searching.
WikiCite/Scholia (CC0)	A Wikimedia initiative for organizing bibliographic information and visualizing it as scholarly and topic profiles for Wikipedia/Wikidata.
OpenAIRE Research Graph (CC BY 4.0)	A knowledge graph that enables integrated metadata search on funders, organizations, researchers, research communities, and publishers.

Existing service	Description
ORCID (CC0)	A non-profit organization that provides a persistent identifier for researchers and enables linking researchers with research.
PID Graph (CC0)	A DataCite service that uses a *GraphQL* interface to enable integrated metadata searches on entities, especially data, publications, and people.
CultureGraph (CC BY-NC)	A service that links metadata of the library networks of Germany and Austria, as well as the German National Library.
Open Knowledge Maps (CC BY 4.0)	Visualization frontend for searching scholarly literature indexed in the Bielefeld Academic Search Engine (BASE)
Connected Papers (ODC-BY)	A service that visualizes connected papers as a graph to explore academic fields, for example, to discover the most relevant prior and derivative works.

Table 3.

Disciplinary databases and services publishing domain-specific structured scholarly information.

Existing service	Description	How SKG4EOSC will go beyond
Products and services planned to be integrated in SKG4EOSC		
ORKG (CC BY-SA)	A FAIR-driven infrastructure for scholarly information that includes digital library services to support acquiring, curating, publishing, and processing FAIR descriptions of research contributions in a variety of disciplines.	Establish the *ORKG* as the central Hub for providing unified access to heterogeneous disciplinary infrastructures.
Hi Knowledge (CC0)	An interactive visualization tool that structures scholarly knowledge on invasion biology from more than 1.100 publications into a network of 12 invasion hypotheses. The data are published as static Excel files for download.	Programmatic access to FAIR scholarly information by means of Web APIs.
Cooperation Databank (CoDa)(CC BY-SA)	A machine-readable history of cooperation research to search and select studies for on-demand meta-analysis. The data are accessible via a SPARQL endpoint.	Automated access to FAIR scholarly information through new Web APIs. Improvement of the Interoperability principle by means of new vocabularies and mappings to external datasets.
The CLARIFY Knowledge Graph (CC BY-NC-SA)	Integrates structured EHRs of lung, breast, and lymphoma cancer patients with biomedical data extracted (e.g. drugs and side effects) from open scientific databases (e.g. DrugBank).	The *CLARIFY Knowledge Graph* will be integrated to FAIR scholarly information to support traceability, reproducibility, and explainability of the outcomes.
Linear Mixed Models (LMM) KG (CC BY-SA)	Allows processing and storing results of linear model fitting (parameter estimates, hypothesis testing results) in structured RDF objects. Used mainly for plant experimental data.	SKG4EOSC will extend the tool to other statistical models and application areas and integrate the objects into more general KGs and publications.
OpenBiodiv (CC-BY)	An RDF-based Biodiversity Knowledge Graph, encompassing data extracted from full-text article XMLs, integrated in a graph database following the OpenBiodiv-O (Senderov et al. 2018) ontology and RDF version of the GBIF taxonomic backbone. The data is available through a SPARQL endpoint.	Development of apps and Nanopublication modules which will turn *OpenBiodiv* into a key, LOD-based reference tool for data about biological species to be used also by other domains through federation and Nanopublications.

Existing service	Description	How SKG4EOSC will go beyond
Products and services not (yet) planned to be integrated in SKG4EOSC		
Metadataset (Unknown)	Collection of open data from scientific publications about the management of agricultural and natural resources.	SKG4EOSC will cover a much broader range of research areas and domains, while providing similar or more advanced functionalities
Energy in Buildings and Communities (IEA-EBC)(CC0)	Structured data relevant to Occupant-Centric Building Design and Operation extracted from the literature. The Excel data are accessible via the Open Science Framework (OSF).	Programmatic access to FAIR scholarly information by means of Web APIs.
Papers-with-Code (PwC) (CC BY-SA)	Facebook AI service that collects research contributions, especially algorithms, in a structured way according to the scheme task-metric-benchmark and creates rankings of them.	Transfer and application of PwC features such as leader boards to other disciplines.
EuropePMC (Mix)	European database of article metadata and full texts of OA articles, enriched with text mining results and links to external databases.	We will provide SKG4EOSC as a knowledge source to be disseminated by EuropePMC LabLinks.

Table 4.

Overview of the SKG4EOSC components, their current content, user base, current and resulting TRL. SKG4EOSC components are italicised throughout the proposal text.

SKG4EOSCComponent	Content coverage	User base	Current TRL	Resulting TRL
Open Research Knowledge Graph	~500 state-of-the-art comparisons; ~10,000 contributions in 450 fields	500 users and 20 organizations	7	9
Hi Knowledge	>1,100 extracted scientific paper contributions, mapped to 12 hypotheses	~1000 visits per month	7	9
Cooperation Databank	SKG of 2,641 studies on human cooperation (1958-2017) conducted in 78 countries involving 356,680 participants. Expert annotatations with 312 variables, incl. quantitative results (13,959 effect sizes).	~400 users from psychology, economics, sociology, and political science.	7	9
OpenBiodiv	RDF & GraphDB KB of ~700M triples extracted from more than 30,000 article XMLs.	~300 users, mostly biodiversity scientists and informaticians.	7	9
Nanopublications	General purpose schema and technology stack for packaging and publishing Linked Data independent of content type.	125 Nanobench users; 23 code bases using nanopub-java library	7	9

SKG4EOSCComponent	Content coverage	User base	Current TRL	Resulting TRL
GraphQL	General purpose graph database independent of content type.	~100 major businesses (e.g. KLM)	9	9
Wikidata / Wikibase / WikiCite / Scholia	13 billion statements about 37 million scholarly publications as well as 284 million scientific citations. Overall, 95 million items covering specialized knowledge in many research fields.	~23,000 active contributors per month, 61 administrators, 324 bots, >5M registered users	8 / 9 / 7 / 7	9 / 9 / 8 / 8
Medical-KG	Portfolio of data-driven tools to transform and integrate heterogeneous data sources into an RDF knowledge graph.	12 EU projects, >2 Billion RDF triples	6	9
Linear Mixed Models (LMM) KG	~ 6 million statements, 52 models	users in plant research community	4	7
DBpedia Snapshot KG	13 billion multilingual statements (with 850 mln ENG), 62 million linking statements to LOD	100,000 users (technology domain), 12M daily queries	9	9
DBpedia Stack	A catalogue of KG services for quality control, hosting, text analysis, search that can be automatically deployed with the *Databus*	>100,000 downloads	4-9	7-9
DBpedia Databus	Open Source platform for distributed file sharing and collaborative quality curation	>80 million hits, several interdisciplinary deployments	6	9

1.1.3 R&I maturity

Despite its highly innovative nature, SKG4EOSC builds on a number of mature technology components (Table 4) with sizable established user bases. The SKG4EOSC service ecosystem results from the integration and advancement of these components to support the production, curation, and use of FAIR scholarly information in the EOSC. As a whole, the SKG4EOSC service ecosystem is, however, highly innovative and as a socio-technical system it is currently situated at a lower TRL than the leveraged components. *ORKG* has demonstrated a system prototype in an operational environment. SKG4EOSC will lift this prototype and numerous disciplinary scholarly information infrastructures into the EOSC. Some technology required for such lifting, e.g. for the integration of ORKG with disciplinary infrastructures, have been validated in relevant environments (e.g. with *Hi Knowledge*) and, thus, begin at TRL 5. For other required technology, e.g. for deploying FAIR scholarly information production, curation, and use in the EOSC we have currently only observed basic principles and, thus, begin at TRL 1-2. Given the substantial groundwork with ORKG

and disciplinary scholarly information infrastructures, SKG4EOSC will be able to demonstrate a system prototype for FAIR scholarly information production, curation, and use in the EOSC as the operational environment during the project's lifetime.

1.2 Methodology

1.2.1 Overall Methodology

In the proposed architecture (see Fig. 3 for a schematic overview and Table 5 for the key enabling technologies), the *Open Research Knowledge Graph (ORKG)* implements the Hub for harmonized access to FAIR scholarly information in the EOSC. The Hub's role is twofold: (1) abstract from the technological heterogeneity of disciplinary scholarly information infrastructures, and (2) enable the efficient development of EOSC services for the production, curation, and use of FAIR scholarly information in EOSC stakeholder communities and beyond. In the hourglass metaphor, *ORKG* acts as the narrow neck, harmonizing access and catalysing services.

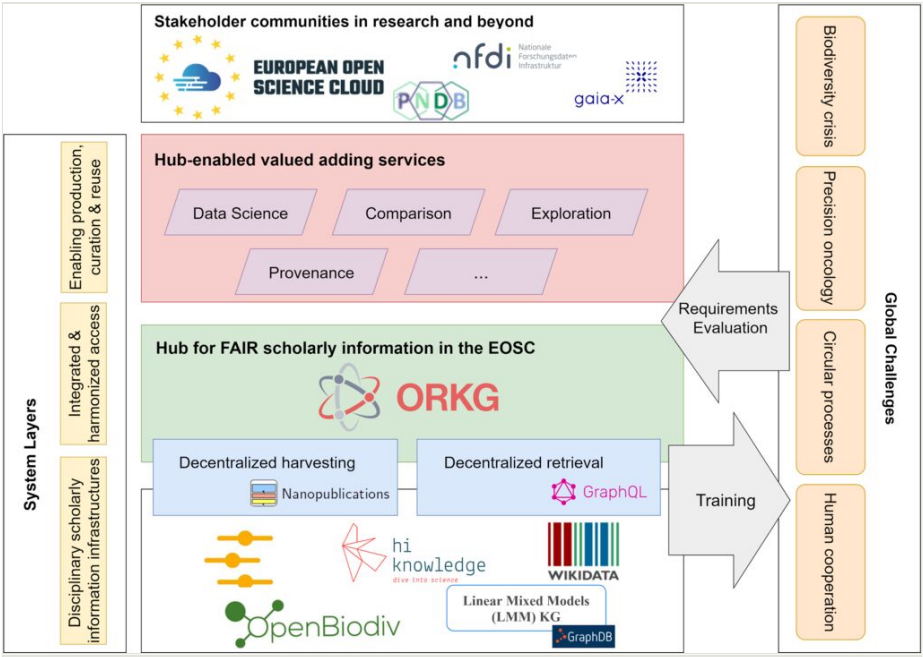

Figure 3. doi

Schematic overview of the SKG4EOSC architecture.

As part of the lower glass bulb, SKG4EOSC involves numerous disciplinary scholarly information infrastructures. These infrastructures are widely used in the respective research communities. Many of these infrastructures adhere to the FAIR data principles and, thus, individually publish machine-based reusable content. However, technological heterogeneity hinders their integration as an ecosystem in the EOSC. These heterogeneity

issues also complicate both the development of generic services for FAIR scholarly information production, curation, and use in the EOSC, as well as the transfer of services and approaches developed in one community to other communities.

Table 5.
Key enabling technologies.

Technology	Enables
Graph databases and query languages (e.g. Neo4j, GraphDB)	FAIR scholarly information management and retrieval
Nanopublications	FAIR scholarly information representation and data exchange
Semantic technologies / knowledge representation	Formal (machine actionable) representation of data semantics
Semantic resources (e.g. terminologies, vocabularies, ontologies)	Reusability of scholarly information meeting community standards
MNatural language processing	Scholarly information extraction from the literature and other assets
Computational environments (e.g. Jupyter)	FAIR scholarly information production in data analysis
Provenance modelling (e.g. PROV-O)	Associate FAIR scholarly information with detailed provenance data

Hence, **SKG4EOSC builds on and extends two approaches that will provide technological harmonization** of access (in terms of data formats and exchange protocols) to FAIR scholarly information serviced by the involved disciplinary scholarly information infrastructures:

1. ***Nanopublications*-based decentralised harvesting** (Groth et al. 2010). The core idea here is that disciplinary scholarly information infrastructures publish (e.g. using the Nanopub server network and the Nanobench client) their contents as *Nanopublications*. *Nanopublications* are FAIR Digital Objects in line with the EOSC Interoperability Framework and are consumed by ORKG as well as potentially by other systems. As an example of this approach, the linear mixed model computation and the resulting findings published by Gentsch et al. (2020) in their Figure 1 can be described using the Statistical Methods Ontology (STATO). Differences between soil treatments that influence carbon flux are presented in a box plot depicting data distribution for different experimental factor levels, classified within statistically homogeneous groups. Such visual information should be published with its machine-actionable counterpart, i.e. the numerical values and the semantics of the depicted statistics and their provenance. A disciplinary scholarly information infrastructure can publish this machine-actionable counterpart as a *Nanopublication*. Using the *Linear Mixed Model KG*, we can demonstrate this by means of a SPARQL CONSTRUCT query that constructs a *Nanopublication* that can be directly harvested and ingested by *ORKG* (Fig. 4).

2. **GraphQL-based decentralized retrieval**. The core idea here is that disciplinary scholarly information infrastructures publish their contents in a heterogeneous manner (using arbitrary protocols, including *GraphQL, SPARQL, REST*) as is currently the case and harmonization occurs within a *GraphQL* endpoint implemented by *ORKG*. TIB has prototyped this approach*[3] with a *GraphQL*-based integration of *ORKG* and PID Graph that enables cross-walking metadata about articles, datasets, people, or organizations and the data published in the scholarly literature (Haris et al. 2021) (Fig. 5).

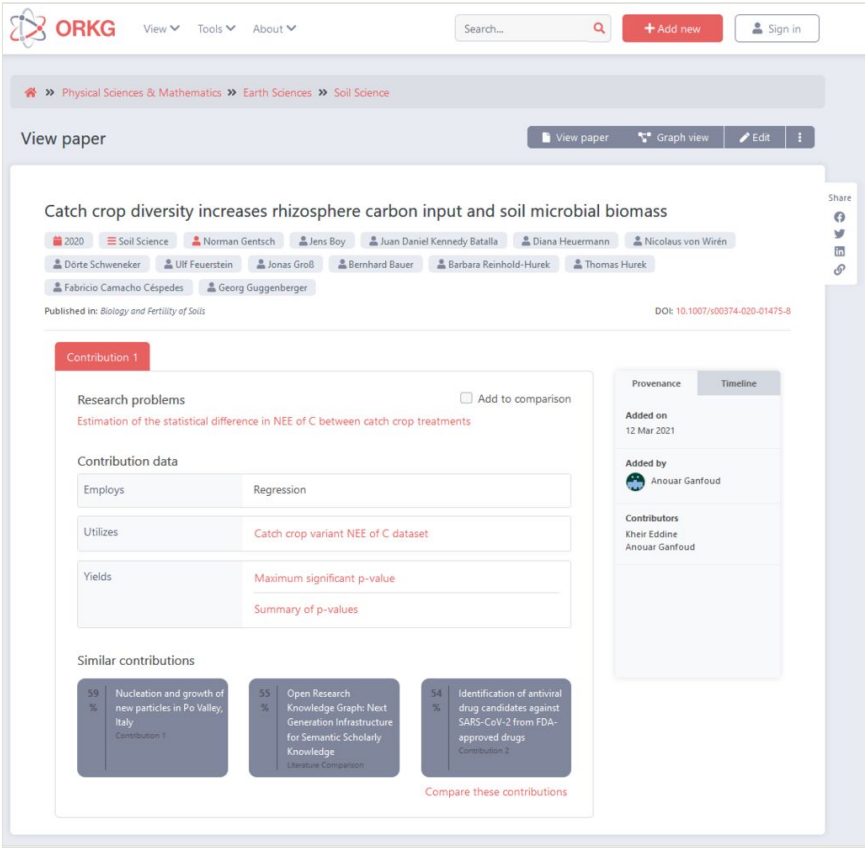

Figure 4. doi

Result of Nanopublication-based decentralized harvesting as an approach towards harmonised access to FAIR scholarly information in the EOSC, here exemplified with content published by a Linear Mixed Model KG as Nanopublication that can be harvested and automatically ingested by ORKG, shown on results by Gentsch et al. (2020).

Both approaches **harmonise the syntax and protocols of exchanged data. In order to harmonise the semantics of exchange data**, SKG4EOSC will leverage existing semantic resources (e.g. EBI-OLS, BioPortal) to ensure that the same information published by different infrastructures (e.g. a statistical hypothesis test) is described the same way (i.e. using the same terminologies). The main difference between the approaches is that in the

case of *Nanopublications*-based decentralised harvesting, the (relevant) content of disciplinary scholarly information infrastructures is cached in *ORKG*. In contrast, in *GraphQL*-based decentralised retrieval, the content of disciplinary infrastructures is not centrally cached and retrieval is, thus, truly decentralised. SKG4EOSC will explore both approaches in order to determine their individual advantages and disadvantages and decide whether only one or both approaches have their merits in that they enable different use cases and services.

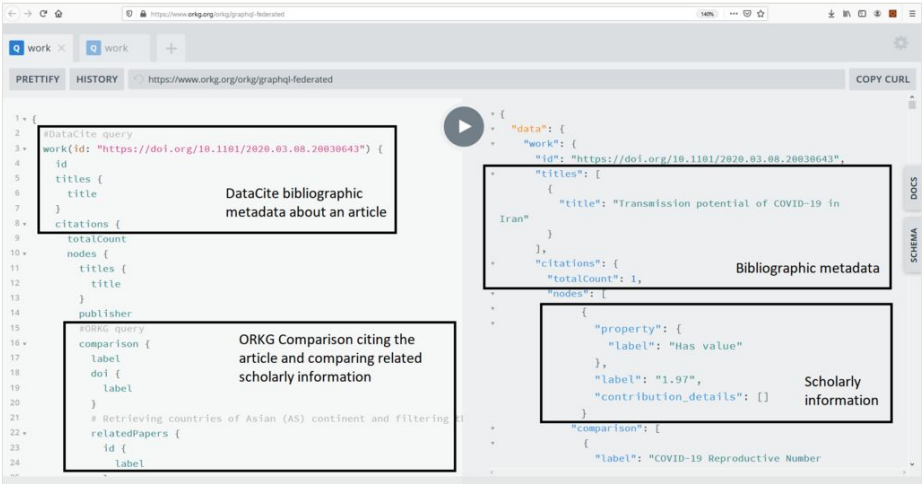

Figure 5. doi

GraphQL-based decentralised retrieval as an additional approach towards harmonised access to FAIR scholarly information in the EOSC, here exemplified with a GraphQL endpoint integrating the PID Graph with DataCite bibliographic metadata and FAIR scholarly information published by ORKG.

As part of the upper bulb of the hourglass, SKG4EOSC will provide numerous Hub-enabled value adding services for the production, curation, and use of FAIR scholarly information in the EOSC, by both humans and machines. Services include, among others:

- *Scholarly information comparison*, e.g. precision and recall of deep learning algorithms;
- *Exploration and visualization*, e.g. visualisation of hypothesis networks in invasion biology;
- *Integrating FAIR scholarly information in data science*, e.g. as data sources in systematic reviews; or
- *Provenance services* ensuring FAIR scholarly information relates to the primary data.

Some of these services will be powered directly by *ORKG*. Others will be standalone services, part of the overall ecosystem of Hub-enabled value-adding services that enable the production, curation, and use of FAIR scholarly information in the EOSC.

The EOSC and the served research communities are primary SKG4EOSC stakeholder communities. All services involved and newly developed in **SKG4EOSC will be discoverable in the EOSC** (EOSC Portal Marketplace). SKG4EOSC also ensures that the ecosystem of services is composable, i.e. researchers will be able to select multiple EOSC services (SKG4EOSC services and others) needed to accomplish a particular task knowing that the services will interoperate. For instance, a researcher in biodiversity may use EGI Notebooks to execute a data analysis task and the SKG4EOSC *Linear Mixed Models* (LMM) KG to store FAIR scholarly information resulting in data analysis by using the newly developed SKG4EOSC Python and R libraries in research software.

We pilot the SKG4EOSC methodology for global societal challenges in four disciplines: biodiversity crisis, precision oncology, circular processes, and human cooperation (WP5). With an iterative and inclusive development in close cooperation/co-design with the relevant research communities, these pilots contribute with requirements and to the agile development and evaluation of SKG4EOSC services.

An essential aspect is that the **SKG4EOSC methodology relies on the three complementary approaches: post-publication enrichment, FAIR-at-birth, and crowdsourcing for FAIR scholarly information production and curation**. These three approaches cover different phases of the research lifecycle, i.e. information production. The classical approach is to extract information post-publication using natural language processing and text mining (WP2). In addition, SKG4EOSC develops approaches to ensure scholarly information is produced FAIR at birth (FAIR-by-design). Rather than burying information into unstructured text, the aim of WP3 is to develop approaches and services that can be embedded in data analysis in order to ensure that the produced information is FAIR. Finally, SKG4EOSC leverages Crowdsourcing, which is central to *ORKG*, as a third approach for both FAIR scholarly information production and curation.

Another essential aspect is that the **SKG4EOSC methodology integrates data**, i.e. the scholarly information expressed in articles, **and metadata**, i.e. the bibliographic information about articles. Bibliographic metadata is a valuable resource that describes a scholarly publication with information about its output, layout and authorship (Turki et al. 2021). In some applications, the title and abstract of a research paper can hold sufficient information about the findings and outcomes of the research publication, making full text analysis optional (Gu et al. 2016). Moreover, bibliographic metadata driven (co-)citation network analysis is a scientometric instrument widely used to gain insight about trends, clusters, bias, etc., in the scholarly record. Furthermore, keywords, particularly, the controlled ones like the MeSH Keywords in PubMed, can be leveraged to identify the topics of scholarly publications (Valderrama-Zurián et al. 2021). The analysis of keyword co-occurrences using a variety of techniques can be useful to extract information about the findings of a research publication (Li et al. 2016). By integrating scholarly information (data) and bibliographic information (metadata), SKG4EOSC will enable entirely novel analyses of the scholarly record. These possibilities will be explored and demonstrated in WP5 pilots.

1.2.2 Building on national or international research and innovation activities

SKG4EOSC will collaborate with several global, largely international and national infrastructures, networks and projects to ensure a wider access, inclusivity and uptake of the tools, services and data developed in the project (Table 6).

Table 6.

Infrastructures, projects, networks and initiatives SKG4EOSC will collaborate with to ensure a wide uptake of its products.

Activity	Description	How SKG4EOSC will leverage
ScienceGraph ERC	The ERC Consolidator Grant (Prof. Sören Auer) Knowledge Graph based Representation, Augmentation and Exploration of Scholarly Communication explores the research foundations of machine actionability in scholarly communication.	SKG4EOSC leverages the basic research performed in the context of this ERC CoG and the resulting *ORKG* service, and extends this work by advancing the *ORKG* as a Hub for FAIR scholarly information in the EOSC that harmonizes access to disciplinary scholarly information. The link will be established through the involvement of TIB in both SKG4EOSC and ScienceGraph.
EOSC	The European Open Science Cloud is an initiative aiming at developing an infrastructure providing its users with services promoting open science practices.	SKG4EOSC leverages the ongoing EOSC developments, including services such as the EOSC Portal Marketplace. SKG4EOSC will also leverage EOSC services that can be discovered in the Marketplace, such as EGI Notebooks. Finally, SKG4EOSC will be a provider of new services to EOSC. The link will be established through membership in the EOSC Association, active involvement in EOSC Working Groups as well as other INFRAEOSC projects with direct or indirect involvement of SKG4EOSC partners.
GAIA-X	GAIA-X is a project for the development of the next generation of federated, efficient, competitive, secure, and trustworthy European data infrastructure which is supported by representatives of business, science and administration from Germany and France, together with other European partners.	SKG4EOSC leverages the ongoing development of GAIA-X, since it is another important European initiative that provides open interfaces and standards to link data and make it available to a wide audience to create different types of innovation platforms. The link will be established through the involvement of TIB in the German national project FAIR Data Spaces that builds a common cloud-based data space for industry and research by connecting the National Research Data Infrastructure (NFDI) and GAIA-X by following the FAIR data principles.
NFDI	The German National Research Data Infrastructure is a research community driven national initiative, ultimately involving roughly 30 disciplinary consortia, aiming at the implementation of a research data infrastructure following the FAIR data principles.	SKG4EOSC leverages the ongoing development of NFDI since it will also be linked to international initiatives such as EOSC and GAIA-X to participate in their development. The link will be established through involvements of SKG4EOSC partners in different disciplinary consortia of the NFDI, including, NFDI4Chem, NFDI4Ing, NFDI4Culture, and NFDI4DataScience.

Activity	Description	How SKG4EOSC will leverage
DBpedia	A large-scale multidisciplinary multilingual knowledge graph extracted from the semi-structured information provided by the language editions of Wikipedia such as infobox data, wikilinks and Wikipedia Categories. That being said, DBpedia only supports the entities that meet the Wikipedia notability guidelines.	Proven as an efficient resource to drive semantic technologies like semantic similarity measures and word embeddings (Lastra-Díaz et al. 2019), the DBpedia information can be useful to enhance the semantic similarity between extracted data about the findings of scholarly publications, allowing a better representation of scientific information from the perspective of data accuracy and granularity. This is made possible thanks to the availability of a SPARQL endpoint, of an API and of RDF dumps for the DBpedia database. The link will be established through SKE4EOSC partner INFAI who represent DBpedia.
Wikidata	A large-scale multidisciplinary open, FAIR and collaborative multilingual knowledge graph hosted by the Wikimedia Foundation and driven by *Wikibase*. It represents various types of domain knowledge including cultural heritage data, biomedical information and scholarly metadata thanks to crowdsourcing and user contributions. *Wikidata*'s notability criteria are a good fit for scholarly reference data intended for reuse.	Represented in RDF format, *Wikidata* can be processed by machines thanks to the MediaWiki API, to the *Wikidata* SPARQL endpoint and to the Python library Wikibase Integrator. These tools allow users to extract information from *Wikidata* and modify it when needed. SKG4EOSC will use bibliographic metadata from *Wikidata* created within the framework of the *WikiCite* Project to drive the knowledge graph enrichment and refinement of *ORKG*. SKG4EOSC will also use the specialized structured semantic knowledge of *Wikidata* to guide the extraction of research findings from scholarly publications for the support of *ORKG* data. The link will be established through partner USFAX, who represents the Wikimedia Community.
Project funded by the call HORIZON-INFRA-2021-EOSC-01-06 - FAIR and open data sharing in support of cancer research	The aim of this call is to fund the development of the EOSC federated framework to store, share, access, analyze, and process cancer-related research data following the FAIR principles.	SKG4EOSC will cooperate with this call-funded project. This synergy will pave the way for developing services to reproduce oncological clinical outcomes observed on clinical data and reported in scientific literature. The link will be established through SKG4EOSC partners LUH, SERMAS and UPM involved in Pilot T5.2.
GO FAIR	Bottom-up, stakeholder-driven and self-governed initiative aiming to implement FAIR principles through Implementation Networks (INs).	SKG4EOSC partners participate in GO FAIR Implementation Networks (IN), such as the Chemistry IN (Coles et al. 2020).
EU H2020 CLARIFY	CLARIFY aims at paving the way for the support of cancer patients' profiling for improving quality of life after treatment.	SKG4EOSC partners SERMAS is the coordinator of the project and UPM and TIB lead the tasks of data integration and analytics, and knowledge graph creation.
RDA	The Research Data Alliance (RDA) builds the social and technical bridges to enable the open sharing and re-use of data.	SKG4EOSC partner TIB will continue to be actively involved in the Open Science Graphs for FAIR Data IG.

1.2.3 Interdisciplinary approach

SKG4EOSC involves science pilots and respective research communities for four global societal challenges (WP5): Biodiversity crisis (biodiversity), precision oncology (life sciences), circular processes (chemical engineering) and human cooperation (social sciences), to showcase the interdisciplinary approach as follows:

- Involved research communities support and evaluate the developments in WPs 1-4 in Agile requirements analysis and implementation processes (co-design).
- Evaluate and further develop the scholarly information production/curation/use methods and approaches, scholarly information types and tools for information production (e.g., for data analysis such as Python/R in Jupyter, SPSS, MAXQDA, Stata, etc.) used for various purposes in diverse contexts.
- Leverage the expertise and adopt existing methods developed in one research community (WPs 2 and 3), e.g. for scholarly information extraction using compact identifiers, generalize these approaches (WP4) and transfer them to other communities, where applicable.
- Not just vertically implement the pilots, but also identify an interdisciplinary pilot showcasing how FAIR scholarly information from multiple disciplines can be used to conduct research (T5.5).
- Evaluate the developed approaches in 1-2 additional open call pilots (e.g. on digital technologies for teaching and learning with papers published by CEUR-WS) during the project's lifetime.

1.2.4 Integration of social sciences and humanities

SKG4EOSC will work towards an integration of the social sciences and humanities in SKGs. A core use case in the social sciences for this project is the *Cooperation Databank*. This is a knowledge graph of social science studies about human cooperation, including both experimental manipulations and correlations. These studies represent research done within psychology, economics, sociology, and political science. This is an excellent basis to begin to expand the knowledge graph to include other topics within the social sciences, and this will be done with a focus on research about how human beliefs, attitudes and behaviours affect climate change. One goal would be to include a knowledge graph of this research on human behaviour, co-operation, and climate change, which scientists can use with an application to produce queries that output on-demand meta-analyses of research on these topics.

SKG4EOSC will also link knowledge graphs from the social sciences with existing knowledge graphs in the humanities. Doing so can provide the immediate benefit of using information about variation across societies, such as history, institutions, and economies, that can be linked to variation in human behaviour observed across studies. For example, the knowledge graph of the humanities could include information that can be linked to the outcome of the social science studies. This could help inform policy makers about how behavioural interventions for pro-environmental behaviours could be tailored to the societal and cultural context in which the interventions have been proven to be most effective.

1.2.5 Gender dimension

Diversity has many dimensions, and they interact with the project in several ways: Direct involvement in the design and implementation of the project by way of one or more project partners; indirect involvement through community-facing activities; and passive involvement in terms of the scholarly information represented in the literature and databases that the project interacts with. There are clear biases in terms of what has been published, what has been published about, what is/was considered notable for inclusion in databases, what has been digitized or whose contributions have been recorded. There are also biases inherent to some approaches, e.g. in NLP (Lu et al. 2020). This affects the biases in what scholarly information is FAIRifed, both with pre-publication as well as with post-publication approaches. SKG4EOSC acknowledges that such biases exist and is going to take them into account when planning, refining and executing its activities. While planning the project and defining its concept, priorities and approaches, several diversity dimensions have been taken into account, e.g. disciplinary, geographic, linguistic, career stage and gender. The networking activities will include promotion of gender equality and other dimensions of diversity in outreach and training events. Particular attention will be given to diversity measures among speaker invitations, discussions and event contributions of any kind as well as to diversity aspects of the use cases and content we handle. To this end, we will collaborate with existing diversity-related initiatives*[4] and consider adopting or adapting some existing workflows*[5] around highlighting priority areas for engaging with diversity and around visualizing progress in this regard. To ensure that the gender balance but also other inclusivity and equality issues are properly supervised through the project lifetime, we will appoint a Diversity and Equality Champion from the project partners.

1.2.6 Open Science Practices

The project as a whole is designed to provide public benefit, and as such defaults to openness of both its activities and their outcomes. Activities will be publicly documented and open for participation. Outcomes, in particular data, software, reports, articles, will be made available in formats and under non- or less-restrictive licenses that maximize reuse. This is reflected in the nature of the tasks and deliverables and their relationships. For each activity undertaken within a task, we will consider and document the potential benefits or harms of sharing or not sharing the process, the outcomes or any other aspects of the activity, and what an appropriate timing for sharing would be.

Developing a project with this level of openness entails open cooperative work and systematic sharing of knowledge and tools as early and widely as possible in the process. The consortium has extensive experience with sharing processes and outcomes at each stage of the research cycle. As a testimony to this consortium's distinctive Open Science practices, half of the project partners have in the past published grant proposals*[6] in the journals *Research Ideas and Outcomes* and *NeoBiota*, both published by PENSOFT.

Scholarly output will be published Open Access with CC BY license in top-ranked, peer-reviewed, renown and international conferences or journals. Furthermore, SKG4EOSC

research contributions will be described in *ORKG* and cited in the corresponding article to ensure

1. that SKG4EOSC scholarly output reuse is maximized and
2. to actively use and demonstrate the effectiveness of SKG4EOSC services.

Research data will be published following the FAIR data principles using a trusted data repository (listed in re3data), licensed as open as possible and closed as necessary (in the context of the WP5 Precision Oncology pilot, some data may not be published openly). Research software will be published Open Source with a suitable license within the framework of the Open Source Initiative (MIT or similar). All software will be managed from the beginning in a Git repository, openly in the cloud (GitHub, GitLab or similar). At project end, all software will be deposited in their final versions on Zenodo.

As the key measure to ensure reproducibility of research outputs, SKG4EOSC will leverage own services to not only describe SKG4EOSC research contributions, but also in SKG4EOSC research activities, e.g. use the *ORKG* Comparison service in literature reviews. The created artefacts will be accordingly cited in SKG4EOSC articles. As such, SKG4EOSC will practice openness also at the level of research activities and ensure that important assets generated during the research lifecycle are also accessible and reusable. As an additional measure, research data and research software will be managed following the FAIR data principles, including describing these assets with metadata following community standards, depositing the assets in repositories that support the persistent identification, and the linking of assets by persistent identifiers in metadata. Research software will be developed openly, thus allowing for the involvement of and the reuse by relevant knowledge actors, including the public at large.

2. Impact

2.1 Project's pathways towards impact

Europe spends ~2.18% of its GDP amounting ~300 Billion Euro annually for R&D*[7]. In addition, with climate change, de-carbonization, clean mobility, social inclusion and responsibility, supply chain resilience, etc., we face significant societal challenges, which can only be mastered by research and development. With SKG4EOSC we aim to **make European research at least 30% more efficient and effective**, since the currently extremely cumbersome processes of literature work (search, exploration, and ingestion) will be greatly improved through the SKG4EOSC state-of-the-art overview, querying, analysis, and visualization services. In addition, SKG4EOSC services will help to reduce duplications in research, guide researchers towards important research goals and significantly improve reproducibility and peer-review. Last but not least, the transfer from research to industrial and societal applications will be greatly improved, since industrial and societal stakeholders can easily obtain current overviews of the state-of-the-art about their innovation challenges.

Table 7 explains how SKG4EOSC applies the FAIR data principles. Table 8 summarizes the steps towards the achievement of the expected impacts of the project over time, including beyond the duration of the project. Table 9 summarizes the SKG4EOSC's unique contributions towards the outcomes specified in the topic of this call and the wider impacts specified in the respective destination.

Table 7.
How the FAIR data principles will be applied to the three SKG4EOSC research data categories.

	Scholarly information	**SKG4EOSC research data**	**SKG4EOSC research software**
Findability	DataCite DOIs for the persistent identification of scholarly information in *ORKG*. As an EOSC service sustainably operated by TIB, *ORKG* is a trusted infrastructure for FAIR scholarly information.	Research data will be described and deposited in a trusted repository that implements DOI-based identification (e.g. Zenodo or similar).	Research software will be managed on GitHub and deposited on Zenodo.
Accessibility	*ORKG* employs HTTP-based open protocols for access to content. The content is licensed CC BY-SA and is also available as dumps. TIB ensures the long-term storage and preservation of this content.	The trusted repository used for SKG4EOSC research data will employ HTTP based open protocols for access to content. SKG4EOSC research data will be open access.	GitHub and Zenodo employ HTTP based open protocols for access to content. SKG4EOSC research software will be open source.
Interoperability	*ORKG* uses state-of-the-art graph database and semantic technologies for knowledge representation and reasoning. Content is represented using standard formats (e.g. JSON-LD). SKG4EOSC employs vocabularies to describe data semantics to ensure machine actionability of its content.	SKG4EOSC research data will be formatted following standards that are appropriate for the respective data (e.g. model evaluations) and described using corresponding vocabularies.	SKG4EOSC research software will be written in major computer languages (e.g., Python, Java) and described using suitable vocabulary.
Reusability	In SKG4EOSC, scholarly information will be richly described using community standards, including rich provenance (supported by *Nanopublications*).	SKG4EOSC research data will be described to maximize reuse, using relevant community standards (e.g. for the description of models) and required provenance, in line with the metadata collected by the selected trusted data repository.	SKG4EOSC research software will be documented following open source community standards and richly described with metadata, including provenance, in order to maximise reuse.

Efficient, machine supported, use of scholarly information has been, is and will be a need for research. Standing on the shoulders of giants relies on literature reviews; each community periodically conducts systematic reviews; synthesis of published results is performed to increase the statistical power of claims. All these research activities rely on information extraction and organization from literature, processes that are currently manual and inefficient. **SKG4EOSC is the first EOSC project that tackles this problem** head on. In the first phase, SKG4EOSC develops its service offering for FAIR scholarly information production, curation and use in the EOSC for four communities directly involved in the

project. In a second phase, 1-2 communities will be additionally onboarded during the project as further pilots not directly involved in the project developments. In parallel, *ORKG* has been onboarding diverse communities through *ORKG* Observatories. Overall, SKG4EOSC thus reaches tens of thousands researchers during its project lifetime, a reach that is a key pathway to ensuring the project's impact.

By leveraging approaches and enabling technologies for FAIR scholarly information production in both pre-publication and post-publication phases, SKG4EOSC not only develops services that extract information from millions of existing legacy documents and the millions that are going to be written in the coming years but, importantly, also **develops a pathway for a future in which scholarly information is produced FAIR** during the research lifecycle. While this approach comes with its own challenges (e.g. considerable upgrade of the research infrastructures and tools currently in use) the pre-publication approach (WP2) has the potential to fundamentally transform the production of FAIR scholarly information.

Table 8.

Steps towards the achievement of the expected impacts of the project over time, including beyond the duration of the project.

Time	Project result	Measurable KPI
M3	*First research community workshops.* Elicit the requirements for SKG4EOSC pilots (WP5).	Conducted workshops with SKG4EOSC research communities involving >200 representatives. Comprehensive set of user stories, requirements and feature descriptions
M6	*Alpha SKG4EOSC service offering.* Covers key features to demonstrate the SKG4EOSC platform to relevant communities	Three initial demos showcasing advanced FAIR scholarly information production, curation and use Open-Source software platform and collaboration infrastructure
M9	*Comprehensive set of dissemination and engagement materials and initial set of stakeholder events*	>5 tutorial video screencasts >3 demonstration showcases recorded as video >10 engagement events with >500 participants Established stakeholder advisory group with >30 members from various scientific fields
M12	*Federated SKG4EOSC service*	>10 research domains comprehensively covered >100,000 scholarly information items
M18	*Significantly expanded set of dissemination and engagement material and ramp-up of stakeholder events*	>20 video tutorial video screencasts >10 demonstration showcases in various domains >25 engagement events with >2,000 participants >1,000 regular users

Time	Project result	Measurable KPI
M24	*Beta SKG4EOSC service offering.* Covers pre- and post-publication FAIR scholarly information production, curation and use services	>50 research domains comprehensively covered >1,000,000 scholarly information items >10 research communities utilize SKG4EOSC service offering to conduct research >5 industry stakeholders (e.g. journalists or publishers) utilize SKG4EOSC services
M30	*Fully established dissemination and engagement*	Sustainable international governance structure for the SKG4EOSC services is established >50 tutorial video screencasts tailored for various research fields >25 comprehensive demonstration showcases for various research domains >100 engagement events with >10,000 participants >5,000 regular users
M36	*Final SKG4EOSC service offering release*	>100 research domains comprehensively covered >5,000,000 scholarly information items
After the project end		
Y1	*The use of scholarly knowledge graphs in research workflows becomes a disruptive network effect*	>10% of European researchers regularly engage with SKG4EOSC services >5% of publications are accompanied by SKG4EOSC content >20% of research domains are comprehensively covered >5% efficiency increase of research through SKG4EOSC services
Y3	*Majority of research involves representing artefacts and outputs as well as interacting with scholarly knowledge graphs*	>25% of European researchers regularly engage with SKG4EOSC services >15% of publications are accompanied by SKG4EOSC content >50% of scientific fields is comprehensively covered >10% efficiency increase of research through SKG4EOSC services
Y5	*Scholarly knowledge graphs are the workhorse of almost all research workflows*	>50% of European researchers regularly engage with SKG4EOSC services >30% of publications are accompanied by SKG4EOSC content >90% of research domains are comprehensively covered >20% efficiency increase of research through SKG4EOSC services

Beyond the project's lifetime, the SKG4EOSC service offering will reach further communities. This is ensured by sustained operations of the services, which is guaranteed by our infrastructure partners. *ORKG* as the proposed Hub and the Hub-enabled services are and will be sustainably operated by TIB. Disciplinary scholarly information infrastructures operated by partners (e.g. IGB, Wikimedia, etc.) have their own sustainability plans. In the event that a service is retired, we will leverage the SKG4EOSC

approach for *Nanopublication*-based decentralized harvesting to ensure the respective content continues to be available through *ORKG*. TIB will thus actively coordinate with partnering infrastructures to ensure content and service availability beyond the project's lifetime. Through these measures, **SKG4EOSC and its service offering has the potential to reach a majority of researchers and their respective research communities** to ultimately fundamentally transform scholarly communication from being purely human actionable to being also machine actionable, eventually reaching the envisioned 30% efficiency gains in annual R&D expenditures in Europe, and globally. Hence, SKG4EOSC lays the foundations to fundamentally transform the way researchers create, share and exploit scholarly information, as well as the way the public and private sectors can exploit scholarly information. Additionally, SKG4EOSC will sustainably address an important gap in the EOSC, namely the seamless access to and management of increasing volumes of scholarly literature following the FAIR principles. Finally, with improved machine actionability in scholarly communication and consequent service offering for machine supported processing of FAIR scholarly information, including the reliable tracking of its provenance and the structured description with formal semantics of materials and methods, SKG4EOSC will substantially contribute to improving the reproducibility of and the trust in science.

Table 9.

Unique contributions of SKG4EOSC project results towards the outcomes specified in this topic and the wider impacts specified in the respective destination.

#	Work program objectives	How addressed
Topic		
1	Increase service offer and capabilities beyond the present landscape in addressing the current and anticipated needs of the research community at large.	By tackling scholarly information FAIRness in the EOSC, SKG4EOSC closes an important gap in the EOSC that addresses current as well as future needs of the research community at large.
2	Increase availability of (pre)operational services that can be customized and integrated in the existing workflows of researchers across different disciplines.	SKG4EOSC increases, by 2-3 steps, the TRL of numerous disciplinary infrastructures for structured scholarly information by lifting them into the EOSC, ensuring their interoperability and composability to enable their integration in research lifecycles.
3	Facilitate cross-disciplinary collaboration, reducing the time to results and increasing productivity.	By harmonizing access to and ensuring interoperability of disciplinary infrastructures, SKG4EOSC facilitates cross-disciplinary scholarly information integration. Through FAIRification of scholarly information, SKG4EOSC will substantially reduce the time to results and increase research productivity (estimated 30%).
4	Provide researchers with a set of highly innovative new services that would exploit, in a structural way, cloud-based EOSC technologies and European compute and data management capacities.	SKG4EOSC provides researchers an ecosystem of highly innovative new services that exploit EOSC technologies (e.g. *ORKG*). In the FAIRification of scholarly information resulting in the data analysis phase of the research lifecycle, SKG4EOSC leverages EGI Notebooks, and therefore, European compute capacities.

#	Work program objectives	How addressed
5	Development and improvement of existing pre-operational software, tools and open source services, aiming to be integrated to the service-based architecture offered through the EOSC.	By lifting a dozen open source and open data services and tools in four disciplines into the EOSC, SKG4EOSC improves their TRL and ensures their findability, access and composability in the EOSC (EOSC Portal Marketplace).
6	Iterative and inclusive development in close cooperation/ co-design with the relevant user communities	By adopting the Agile methodology, SKG4EOSC developments are iterative, guided, informed and evaluated, by the researchers of represented communities (Biodiversity, Life Sciences, Chemical Engineering, Social Sciences and Humanities).
7	Use of open source software and tools for wide availability and uptake.	Software and data used and produced by SKG4EOSC are open source and open data (as open as possible, as closed as necessary, e.g. in Life Sciences).
8	Wide service application range for data intensive science.	SKG4EOSC services cover a wide application range, including extraction of information from scientific literature; pre-publication FAIRification of scholarly information; repeatability and reproducibility in science; advanced machine assisted discovery and reuse of scholarly information.
9	Make use of various enabling technologies.	SKG4EOSC makes use of the following enabling technologies: Artificial intelligence and machine learning; natural language processing and text mining; knowledge representation and reasoning.
10	Developments should be tested against 2-3 real life use cases from a variety of scientific domains.	SKG4EOSC developments are tested against use cases in biodiversity and ecosystem crisis, precision oncology, circular processes, and human cooperation.
11	Cooperate with other relevant and related projects and e-Infrastructures and large user communities. Joint use cases and testing across individual project boundaries are encouraged.	Through its partners, SKG4EOSC will cooperate in national (e.g., German NFDI, GAIA-X) and international (e.g. related EU projects, EOSC, the emerging GOSC) projects as well as in international initiatives such as GO FAIR and RDA (e.g. the RDA Open Science Graphs for FAIR Data IG).
12	The services should be integrated in the EOSC core service platform.	SKG4EOSC services (existing integrated and novel) will be integrated in the EOSC core service platform (EOSC Portal Marketplace).
13	Proposals should include sufficient provisions to address service integration, including, appropriate IPR and licence agreements.	Disciplinary scholarly infrastructures will be integrated through distributed approaches by harmonizing protocols and access. These infrastructures are standalone. Integration occurs at the level of the proposed Hub (*ORKG*) and TIB will ensure that all license agreements are satisfied with integrated parties, during the project and beyond.
14	Clearly identify the resources that the services will offer.	FAIR scholarly information is the resource offered by SKG4EOSC services, which will support the production, curation and use of FAIR scholarly information.
15	Sustainability model for the long-term availability of services can rely on EOSC.	As an EOSC Provider, TIB guarantees the long-term availability of the services and well as the served FAIR scholarly information.
16	Participation of industry players, including SMEs.	SKG4EOSC partner PENSOFT is a publisher and SME directly involved in the development and further exploitation of the project results.
17	Technologies aiming to reach TRL7 or higher by the end of the project.	See Table 4.
Destination		

#	Work program objectives	How addressed
18	Enable and enhance seamless access to and reliable re-use of FAIR research outputs (i.e. data and other digital objects) covering the whole research data life cycle.	SKG4EOSC is the INFRAEOSC project that will ensure that the FAIR data principles are applied to the scholarly information output of the scholarly communication lifecycle, thus covering an important phase of the research lifecycle. Currently, the FAIR principles are limited to the research data lifecycle.
19	Transform the way researchers as well as the public and private sectors create, share and exploit research outputs.	SKG4EOSC lays the foundations to fundamentally transform the way researchers create, share and exploit scholarly information, as well as the way the public and private sectors can exploit scholarly information, leading to better quality, validation, more innovation and higher productivity of research.
20	Facilitate scientific multi-disciplinary cooperation.	Through advanced machine support, FAIR scholarly information with standardized syntax as well as formal semantics will bring multi-disciplinary cooperation to the next level, and has the potential to support discoveries in basic research and solutions in key application areas.
21	Seamless access to and management of increasing volumes of research data following FAIR principles (that are open as possible) and other research outputs.	SKG4EOSC will enable machine based access to and management of the increasing volumes of scholarly literature and information communicated therein, following the FAIR principles. Such advanced knowledge-based systems will stimulate the development and uptake of a wide range of new innovative and value-added services from public and commercial providers.
22	Improve trust in science through increased FAIRness, openness and quality of scientific research in Europe.	Through explicit, machine actionable provenance links between scholarly information and the primary data from which information is derived, as well as contextual agents and activities, SKG4EOSC will substantially contribute to reproducibility, validation and trust in science. The novel services leveraging FAIR scholarly information will support more meaningful monitoring, including peer-review, and advanced, machine-based re-use of research results. FAIR scholarly information is furthermore an opportunity to innovate communication of science to the public.
23	Developed software should be published open source under an open source licence.	Developed software will be published Open Source and licensed MIT or similar.
24	Projects are expected to participate in concertation activities in the framework of the EOSC Partnership.	Through its partners, SKG4EOSC will be involved in various EOSC Partnership activities, in particular partnership in the EOSC Association, involvement in EOSC WGs, GO FAIR Implementation Networks.

The desired global impact of the SKG4EOSC service offering relies on a concerted effort to lift FAIR scholarly information to a first class citizen status in the ecosystem of research objects. As the primary artefact in scholarly communication, articles have held this status for centuries. Required is the same for a corresponding machine actionable expression of scholarly information. The SKG4EOSC project will pave the way, but more investment and regulatory actions will be needed for a global transition. As it can be automated only to a certain extent, the production of FAIR scholarly information relies on researchers and will need to be incentivised, primarily through excellent services that directly add value to researchers.

The desired global impact can face a number of potential barriers. First, change may be very slow, primarily because research and research practices are ingrained activities with

established methods and tools that are difficult to advance and steer in new directions. Secondarily, change may be actively resisted by actors that perceive FAIR and open scholarly information as a threat, e.g. to their business models. Furthermore, technology may not mature as fast as needed for the problem at hand. Of particular concern are text and data mining as well as natural language processing, and thus our inability to efficiently extract granular information from documents. This area of research has a decade old history and has not yet achieved the performance needed for scholarly information. Hence, the technological maturity in this area may be a potential barrier, especially in scaling the SKG4EOSC service offering to the massive corpus of legacy articles. An additional potential barrier may be that crowdsourcing does not perform in the scholarly context for scholarly information as well as in other contexts, e.g. for encyclopedic or geospatial information. Successful crowdsourcing typically relies on the 90-9-1 rule, whereby 90% of users only consume content, 9% of users curate existing content and only 1% of users create new content. It is unclear whether in the scholarly context, we can rely on a mere 1% of researchers to produce FAIR scholarly information. Moreover, advancing the existing research infrastructure so that scholarly information is produced FAIR may also prove to be a mammoth endeavour. Hence, the pre-publication approach proposed by SKG4EOSC will also come with significant barriers. Finally, the regulatory framework may not give enough emphasis on machine actionability of scholarly information in the context of Open Science, the EOSC and equivalent international initiatives.

Naturally, progress will occur, in technical as well as in human infrastructures. Computer science will make further progress, especially also on information extraction, meaning that our ability to extract granular scholarly information from the literature is likely to further improve. Modifying the existing and future research infrastructures, services and tools, both open source and commercial so that scholarly information is produced FAIR at birth, is less of a technical challenge. Indeed, the technologies required to do so exist and are mature enough to be adopted in production environments. Along this dimension, the evolution needed is in social infrastructures, especially the willingness of researchers to adopt more advanced services and tools as well as the willingness of commercial toolmakers to advance their systems. As our understanding of how the use of FAIR scholarly information benefits stakeholders will evolve, it will become clearer how these aspects can be accordingly incentivised.

Given that all approaches - post-publication text and data mining, pre-publication FAIRification, and crowdsourcing - have their potential barriers, SKG4EOSC builds on all three approaches to FAIR scholarly information production and curation in order to mitigate the respective barriers. SKG4EOSC argues that the challenge relies on the effective integration of automated and manual approaches.

2.2 Measures to maximise impact - Dissemination, exploitation and communication

Dissemination. SKG4EOSC will tailor various uni- and bi-directional dissemination channels to the needs of each defined potential user group and audience, eliciting

expertise, knowledge and perceptions from stakeholders as part of the project's co-design engagement activities. The preliminary mapping of dissemination channels, target groups, related impacts and relevant KPIs for which they will be applied can be seen in Table 11. All SKG4EOSC partners will be actively engaged in the dissemination process by:

- Providing content to the Communication and Dissemination work package;
- Using own personal and/or institutional networks, social media and websites to promote the project;
- Using relevant conferences to present the project results and distribute dissemination materials;
- Publishing research and data papers in reputable international scientific journals, in line with their academic and institutional policies;
- Participating in campaigns and events (conferences, expert round tables, webinars, and workshops) specifically designed to raise visibility of the new community and increase engagement from relevant actors beyond the project consortium.

These efforts will be streamlined in the project **Plan for Exploitation, Dissemination and Communication (PEDCOM)**, which will be a regularly updated 'living' document, serving as a management tool for dissemination actions, available to all partners from M6 and updated in M24. These updates will include any necessary modification and adapt appropriately to project progress and new circumstances, including feedback from stakeholders and end-users.

Exploitation. To maximise the exposure of project results and their potential for exploitation, the project will take advantage of the EC's Horizon Results Platform and appoint Pensoft as a lead partner for these activities. This platform will serve as a bridge towards policy-makers and researchers, giving access to the project's main and prioritised results with a high potential value (Key Exploitable Results, Table 10). In addition, SKG4EOSC will consider the Horizon Results Booster for dissemination and exploitation of results so that the added value of the KER is amplified. The exploitation and sustainability of the SKG4EOSC results and products assumes two levels of responsibilities: (1) products and services developed at the base of either project partners or RI will be a responsibility of the respective partner or RI; (2) the key synthetic product of SKG4EOSC, namely the ORKG Hub, will be hosted and run after the project end by TIB at the first level and a consortium of projects partners and RIs who provide services through the Hub at the second level.

Table 10.

List of key globally exploitable results (KER).

KER	Main novelty	Stakeholders	Potential outcomes & impact
ORKG as an EOSC service	EOSC Hub and single-point-of-entry for FAIR scholarly information	Researchers in all disciplines as well as other public and private sector stakeholders	Fundamentally transforming machine-based discovery and use of scholarly information in Europe and beyond

KER	Main novelty	Stakeholders	Potential outcomes & impact
Blueprint to onboard disciplinary infrastructures in the *ORKG* Hub	Approach to efficiently scale the FAIR scholarly information accessible in EOSC	Research infrastructures, researchers	SKG4EOSC approaches will ultimately scale to virtually all of research
Technology for post-publication literature FAIRification	Lifting disciplinary approaches for scholarly information extraction into *ORKG* and interdisciplinary transfer.	Researchers, research infrastructures	Exploitation of tested disciplinary tools across research fields, scaling the production of FAIR scholarly information
Technology for pre-publication scholarly information FAIRification	Ensuring scholarly information is FAIR-by-design, upon production	Researchers, research infrastructures	Embedding scholarly information FAIRification in the research lifecycle, avoiding post-publication extraction, scaling the production of FAIR scholarly information
ARPHA semantic authoring tool	Use of ontologies embedded in the pre-publication authoring process; generation of machine-readable structured content	Researchers and data scientists; publishers	Changing the way scholarly literature is published and re-used through authoring of structured, machine-actionable and ontology-related content; saving costs of post-publication data liberation
Technology for visualization, exploration and use of FAIR scholarly information	Fundamentally transformative approaches for machine-based interaction with scholarly information	Researchers, data scientists, public and private sector stakeholders	Efficiency gains in scholarly information use, increased reproducibility and trust in science
Disciplinary scholarly knowledge graphs	Establishing knowledge graphs as infrastructures for the curation and publishing of FAIR scholarly information	Biodiversity and life-science specialists, bioinformaticians, conservationists and practitioners, chemical engineers, social scientists among other disciplines	Trusted disciplinary infrastructures for FAIR scholarly information with their content discoverable and accessible in the EOSC through the *ORKG* Hub
Integration of *Wikidata* into EOSC	Multilingual representation of FAIR scholarly resources across domains	Researchers, educators, journalists and others, including automated tools, citizen scientists and the public at large	FAIR-first approach to multilingual curation and exploration of scholarly resources, integrated with the Wikipedia ecosystem

The sustainability will be enforced by the uptake of the products and services by the starting community through actions and measures described in the PEDCOM (D6.2). An essential element of the project sustainability is the adherence to the long term data preservation and accessibility via the repositories and RIs involved in compliance with the EOSC long term sustainability plans supported by the Member States and infrastructures. To ensure also the long term commitment to Open and FAIR data, SKG4EOSC will adopt whenever relevant the RDA FAIR Maturity KPIs to check the Fairness of the data infrastructures involved*[8].

Communication. In order to achieve maximum exposure and impact, we will prioritise our communication channels based on the ones that are actively used by our target audiences.

With our professional audiences our focus will be on establishing two-way communication, whereas for the general public we will adopt the “getting our message out there” mantra (mainly through collaboration with associations and networks, press releases and social and mass media).

Internal communication. The SKG4EOSC website platform will consist of a password protected internal communication platform (ICP). The ICP will have the following main features: internal repository where all registered users can upload files and all internal documents related to the activities of the project will be stored; a user section containing the profiles of all project members that are granted access to the ICP; upload options files with restricted access, intended only for the consortium members; option to upload news and events; dissemination report forms (symposia & meetings, general dissemination, scientific publications and open research data); living documents containing a view-only copy of important project forms and tables (including contact lists and dissemination reports); a comprehensible step-wise ICP user manual. Additionally, a business communication GDPR-compliant platform will be implemented as a central channel for internal communication. The platform allows for an easy exchange of messages and calls, hence avoiding the need of unnecessary email exchanges.

External communication. SKG4EOSC external communication strategies will be bi-directional, i.e. not only disseminating project outputs to targeted actor groups and the public at large, but also eliciting expertise, knowledge and perceptions as part of the project’s engagement activities. A short guidebook with standard processes and best communication practices (a Handbook of Communication, MS6.2) will be created, together with the communication strategy document. This document will include issues like:

- tips and tricks on how to create social media posts, news articles, press releases and policy briefs,
- instructions on how to shoot and create engaging videos,
- guidelines on how to acknowledge EU funding etc.

The different Communication & dissemination tools and targets, measures to maximise impact and KPIs are outlined in Table 11.

Table 11.

Communication & dissemination tools and targets, measures to maximise impact and KPIs.

Tool: Target	**Measures to maximise impact**	**Key Performance Indicators**
General project website and MS Teams: Project partners, Policy, Practice, General public	Inform and discuss specific topics of common interest; engage interested parties through information to the project’s main outcomes. The project website will act as a hub for all our online communication efforts, and as a repository for all public information on the project.	Number of visits >50,000/project duration; average session duration >120 sec; returning visitors >30%; 20% average increase in web traffic per year.

Tool: Target	Measures to maximise impact	Key Performance Indicators
Presentations at meetings, webinars, conferences, events and workshops: Practice, Policy	Direct raising of awareness among stakeholders; interaction with key persons and direct conversations with a number of relevant public-sector bodies (within the EU and outside), industry bodies, consumers, waste managers, decision makers from cities, and other organisations.	Participants feedback forms >80% satisfaction rate
Data sets: Project partners, Policy, Practice, General public	SKG4EOSC data will be openly shared through automated workflows with relevant repositories, including but not limited to *ORKG*, *OpenBiodiv*, etc. All data and models, both generated as part of SKG4EOSC and obtained from other sources, will be annotated, using internationally recognised keywords and meta-tags. Output from SKG4EOSC will be organised in an easily accessible and interpretable format. The necessary tools, standards and protocols for making SKG4EOSC data accessible, findable, exchangeable and secured in the long term will be made available to all SKG4EOSC partners and users.	Small to medium-sized data sets collected and generated (incl. quality control) >500,000; 8-10 major sources integrated into SKG4EOSC
Project-relevant mailing lists and networks: Practice, Policy	Dissemination and discussion of specific topics of interest; facilitate collaboration/uptake	Subscribers of mailing lists and networks >1,000
Training video series: Practice, General public	Project results and developments translated in an easy to digest format for practitioners and private persons. The website will host various training materials (video tutorials and slide presentations), which will provide clear guidance on the processes involved in using the tools developed by SKG4EOSC	Number of views on YouTube >1,000/video, comments on social media
Social media (Twitter, Facebook, YouTube, etc.): Policy, Practice, General public	Create communities and inform members about project developments, results and recommendations	+200 new followers/ Twitter, Facebook in the first three months; number of impressions on Twitter >100,000/ project duration; 25% increase of followers per year
Posters: Practice, Policy	Promotion and raising awareness of the project	Number of downloads of electronic copies >1,000, number of distributed printed copies >500
Leaflets: Practice, Policy, General public	Increase awareness about the topics dealt with by the project	Number of downloads of electronic copies >1,000, number of distributed printed copies >500
SKG4EOSC e-Newsletter: Policy and decision makers, Industry, SMEs, Practice, General public	Disseminate ongoing activities, results, other project relevant news and events	Number of clicks and links opened >30%

Tool: Target	Measures to maximise impact	Key Performance Indicators
Scientific publications :Policy, Researchers	Presentation of project research in open research journals including Open Research Europe	Number of papers in open access journals >15; >10,000 reads, downloads, citations
Policy briefs: Practice, Policy	Knowledge transfer from the project to policy-makers for key issues	Number of visits >1,000, downloads >200
Guidelines (e.g., in the form of the FAIR Cookbook[*9]*)*: Practice, Policy	Transferring key results directly to SKG4EOSC end-users to ensure maximum uptake and use.	Number of visits >500, downloads >200
Factsheet, infographics & visuals: Project partners, Policy, Practice, General public	Enhance communication of project outputs to facilitate knowledge transfer	Number of visits >500, downloads >200
Press releases: Journalists, media	Announcement of significant project results	Number of press releases issued >2/ year; visits >1,000/ press release
Publications in specialist and popular media: Practice	Raising public awareness	List of publications or broadcasts
A collection of results in the Research Ideas and Outcomes (RIO) journal: Project partners, Researchers, Practice	A collection of results in the Research Ideas and Outcomes (RIO) journal will host an open science compendium (Mietchen et al. 2021) of data, factsheets, policy briefs, project reports and infographics published with a permanent DOI to ensure SKG4EOSC collective knowledge is available, citable and reusable beyond the project lifetime.	Number of publications: >20
Partner's existing communication structures: Policy, Practice, General public	We will leverage the existing communication structures - such as partner websites, newsletters, social media, events and online communities - to disseminate project developments which are relevant to the field of activity of each partner. This will greatly extend our reach at minimal additional effort.	Number of channels >10, frequency of action >1/month

Open data and open science strategy. The Horizon Europe work program highlights the need to have research data and software tools openly used, by maximizing open science practices, access and re-use of all research cycle outcomes. To coordinate the research data management within the project, SKG4EOSC will develop a guiding Data Management Plan (DMP) (D7.2). The DMP will specifically cover: handling of research data during and after the project; data collection and processing; methodologies and standards; data sharing and open access; curation and preservation. The DMP will also provide the dataset metadata specification that will be used in the data registry, following an appropriate relevant standard. It will specify the recommended licensing schemes, preferably using the Creative Commons Public Domain (CC0) and Attribution (CC BY) licenses as suggested by Horizon Europe. In the cases where the datasets cannot be publicly shared, the reasons will be mentioned in its metadata description (e.g. ethical, rules of personal data,

intellectual property, commercial, privacy-related, security-related). Below is a preliminary description of all major points to be covered in detail within the project DMP:

- **What types of data will the project generate/collect**? Numerous and varied data sets will be collected or generated by SKG4EOSC project partners, including specific data types, e.g. data extracted from literature. The project will not only openly share data, but will provide a unique new level of linking open data between different science domains through advanced Linked Open Data technologies (LOD).
- **What standards will be used**? To ensure interoperability, the SKG4EOSC project aims to collect and document the data in standardized formats (i.e. RDF or tabular data) to ensure that the datasets can be understood, interpreted and shared with accompanying metadata and documentation and relevant supporting material. Metadata standards will depend on the discipline and/or the methodology that was used to produce the data. SKG4EOSC partners will use both discipline-specific repositories and common/standard metadata requirements and ontologies (example for biodiversity: Darwin CoreTaxPub, OpenBiodiv-O, Ecological Metadata Language (EML), etc.), including generic ISO-90155 compliant metadata libraries dependent on discipline-specific or institutional repositories.
- **How will this data be exploited and/or shared/made accessible for verification and re-use**? SKG4EOSC data will be openly shared through automated workflows with relevant repositories, both generic (Table 2) or domain-specific (Table 3).
- **How will this data be curated and preserved**? All data and models, both generated as part of SKG4EOSC and obtained from other sources, will be annotated, using ontology-aligned keywords and meta-tags. Output from SKG4EOSC will be organised in an easily accessible and interpretable format. The necessary tools, standards and protocols for making SKG4EOSC data accessible, findable, exchangeable and secured on the long term will be made available to all SKG4EOSC partners and users.
- **Management of internal knowledge in SKG4EOSC.** The terms of Intellectual Property Rights (IPR) management will be specified in detail in the DESCA Consortium Agreement to be signed at the beginning of the project after discussing and encountering the specific IPR policies and legitimate interests of all partners. SKG4EOSC partners will work on a cooperative basis without commercial interest. However, for future maintenance of software, models and data mutual agreements on ownership and access conditions are essential to build trust and to respect interests relevant for durable cooperation. Issues about ownership, access rights and use conditions will be described transparently in the Consortium Agreement to ensure optimal cooperation among the SKG4EOSC partners. To that end, the Consortium Agreement will define use conditions. User groups, already foreseen in this project, will be asked to agree to these conditions using partnership agreements.
- **Management of external knowledge in SKG4EOSC.** Non-confidential results will be disseminated on the project website and through open access, i.e. free online access, applying the ‘gold’ open access model. Each WP leader has responsibility

to manage external knowledge available to the general public according to the dissemination plan mentioned above. Fundamental scientific results will be freely disseminated through appropriate channels including scientific publications, presentations at international conferences and workshops. The publication venues will be primary scientific high-impact open access journals, especially the Open Research Europe platform. SKG4EOSC will follow the guidelines on open access to scientific publication and research data stated in Horizon Europe. Some budget for supporting publication in open access form will be dedicated and managed by the Executive Committee in accordance with the dissemination plan of the project. Deliverables and other important project outputs will be published in a dedicated open access collection in the open science Research Ideas and Outcomes (RIO) Journal.

- **Open source policies.** The software tools or plugins produced within the SKG4EOSC will be available as open source code under an OSI-approved license and published in the open science RIO Journal or other appropriate journals to ensure findability and reusability of all open source resources. The aim of SKG4EOSC's open source approach is to ensure that a framework for new contributions is established that allows them to continue to be developed as open source in order to facilitate the further adoption and update of the technologies by all stakeholders.

2.3 Summary

Specific needs

What are the specific needs that triggered this project?

Scientific needs. Scholarly information is a vital resource for modern societies. By burying scholarly information into text and documents it is, however, not prepared for modern information processing. An enormous amount of time is required and wasted in manually discovering, extracting, processing, and interpreting scholarly information published in the literature. There is an urgent need to ensure the published scholarly information is re-used in a cost-efficient and effortless way for generation of new knowledge.

Societal challenges. The pace of global changes require rapid analyses and prognoses based on both knowledge of the past (big data liberated from the legacy literature published over centuries of scientific work) and knowledge of the future (newly generated data and FAIR scholarly information).

Policy scope. The decisions of policy makers and governments addressing societal challenges should be supported by quick access to large machine-actionable corpora of knowledge and FAIR, cross-domain, interoperable data.

D & E & C MEASURES

What dissemination, exploitation and communication measures will you apply to the results?

Communication. Online activities will use the website as a hub and a repository for all public information on the project, supported by the use of social media channels (Twitter, LinkedIn and YouTube).

Dissemination. We will actively disseminate key project developments by leveraging existing communication structures (partner websites, newsletters, social media, events and online communities, especially those in the Wikipedia ecosystem). Consortium partners will represent the project at international, national or regional events, and directly engage with interested actors. We will reach interested professionals through scientific publications, mobilise research journalists and send press releases to newspapers and online media platforms, as well as primary scientific publication.

Exploitation. The project will take advantage of the European Commission's Horizon Results Platform, which will serve as a bridge towards policy-makers and researchers, giving access to the project's main KPIs. A dedicated collection in the RIO journal will hold all SKG4EOSC outputs, making them available beyond the project's lifetime. The Horizon Results Booster will be considered for dissemination and exploitation of results so that the added value of the KER is amplified.

Expected results

What do you expect to generate by the end of the project?

Primary product. User-friendly, openly available, *Open Research Knowledge Graph* (*ORKG*) as a Hub and single-point-of-entry for FAIR scholarly information in the EOSC. Through *ORKG*, the project will lift disciplinary scholarly information infrastructures in four disciplines into EOSC and provide EOSC an ecosystem of interoperable services for the production, curation and use of FAIR scholarly information.

Integrated capability. Integrating scholarly knowledge graphs from various domains through a systems approach will reduce uncertainty and increase efficiency of use of FAIR scholarly information for producing new scientific hypotheses and knowledge, predictive modelling of processes of the future and decision support system for policy evaluation.

Data and IT. SKG4EOSC will develop radically new methods and tools for handling cross-domain FAIR data through innovative solutions such as *Nanopublications*, scholarly graph integration, and semantic, ontology-aligned publishing.

Target groups

Who will use or further uptake the results of the project? Who will benefit from the results of the project?

Primary end-users are researchers in all disciplines. They will benefit with fundamentally new possibilities in machine-supported discovery, processing, and analysis of scholarly information.

Tier II target group includes private and public sector actors, e.g. innovation hubs and funds, patent advisors and agencies, journalists who will benefit with new possibilities to exploit scholarly information. SKG4EOSC will devise an efficient and straightforward mechanism for other disciplinary scholarly information infrastructures to join the Hub and make their services and content available to EOSC.

Tier III target groups are industries dealing with post-publication (TDM, NLP, AI) processing of scholarly information and publishers seeking to implement semantic tools and workflows to produce FAIR-at-birth scholarly information, thus supporting the primary and secondary end-users.

Future development. The project will provide a basis for future development and expansion of scientific networks dealing with extraction and production of FAIR scholarly information in other domains and inclusion of an ever growing number of SKGs in the *ORKG* Hub. The Open Science and Open Source approaches will facilitate this process.

Outcomes

What change do you expect to see after successful dissemination and exploitation of project results to the target group(s)?

SKG4EOSC pilot communities of researchers will actively use *ORKG*, the disciplinary scholarly information infrastructures, and the ecosystem of services for the production, curation and use of FAIR scholarly information in their research. For instance, we expect researchers to leverage FAIR scholarly information and the devised services in conducting literature or systematic reviews. Moreover, we expect that additional 1-2 communities of different disciplines currently not involved as pilots (e.g. CEUR Workshop Proceedings in Computer Science) will join the Hub with their own infrastructure and content still during the project's lifetime.

Furthermore, we expect **secondary target groups**, e.g. journalists to start exploiting the project's results during its lifetime.

Finally, we expect to see integrations with **relevant e-Infrastructures** such as OpenAIRE as well as **Research Infrastructures** such as the Integrated Carbon Observation System, and further adoption among researchers and research communities beyond the project's lifetime.

Ultimately, we expect that SKG4EOSC will be the INFRAEOSC project that catalyses the **transformation to the advanced scholarly information systems of the future**.

Impacts

What are the expected wider scientific, economic and societal effects of the project contributing to the expected impacts outlined in the respective destination in the work programme?

SKG4EOSC will drive the application of the FAIR data principles to the information expressed in the scholarly literature. It will thus extend the application range of the principles from the research data lifecycle to scholarly communication and contribute to ensuring a broader coverage of the entire research lifecycle. As such, SKG4EOSC will **ultimately fundamentally transform the way researchers as well as the public and private sectors create, share and exploit scholarly information**, which is the ultimate research output. With advanced machine-based information processing, machine actionable scholarly information will substantially contribute to making the exploitation of the increasing volumes of scholarly information more efficient; it will further improve multi-disciplinary research; and through advanced provenance tracing of scholarly information it will ultimately improve the reproducibility of and trust in science.

3. Quality and efficiency of the implementation

3.1 Work plan and resources

The SKG4EOSC work programme is structured in seven WPs (Fig. 6). **WP1** establishes the *ORKG* as a Hub for FAIR scholarly information in the EOSC. **WP2** devises innovative approaches for post-publication scholarly information extraction from the literature and related assets, such as figures and tables. **WP3** devises innovative approaches for pre-publication production of FAIR scholarly information, especially in data analysis and scholarly communication phases of the research lifecycle. Thus, WP2 and WP3 develop the approaches required in disciplinary scholarly infomation infrastructures to produce the FAIR scholarly information that will be made accessible in a harmonized manner through the Hub. **WP4** builds on the Hub and develops Hub-enabled services, as well as their composition with other relevant services in the EOSC. These services build the technical foundations in support of **WP5** pilots that leverage FAIR scholarly information and services to support the science underpinning global societal challenges. **WP6** ensures to maximize the impact of SKG4EOSC project results through communication, engagement, dissemination, and exploitation instruments and activities. Finally, **WP7** is concerned with project management. Each WP has its own roadmap and objectives, and the consortium members will work towards these throughout the project. However, as the high-quality outputs generated by the individual WPs are essential to the success of the overall project, all WPs must strive for success in the overall work plan, requiring close integration and alignment between the executed activities. The time frame of the work plan tasks and milestones is shown in Fig. 7 as a Gantt chart.

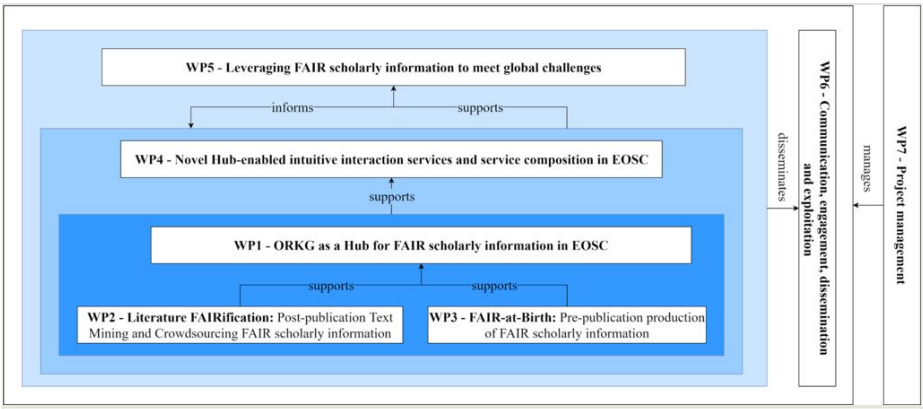

Figure 6. doi

Work package structure and relations.

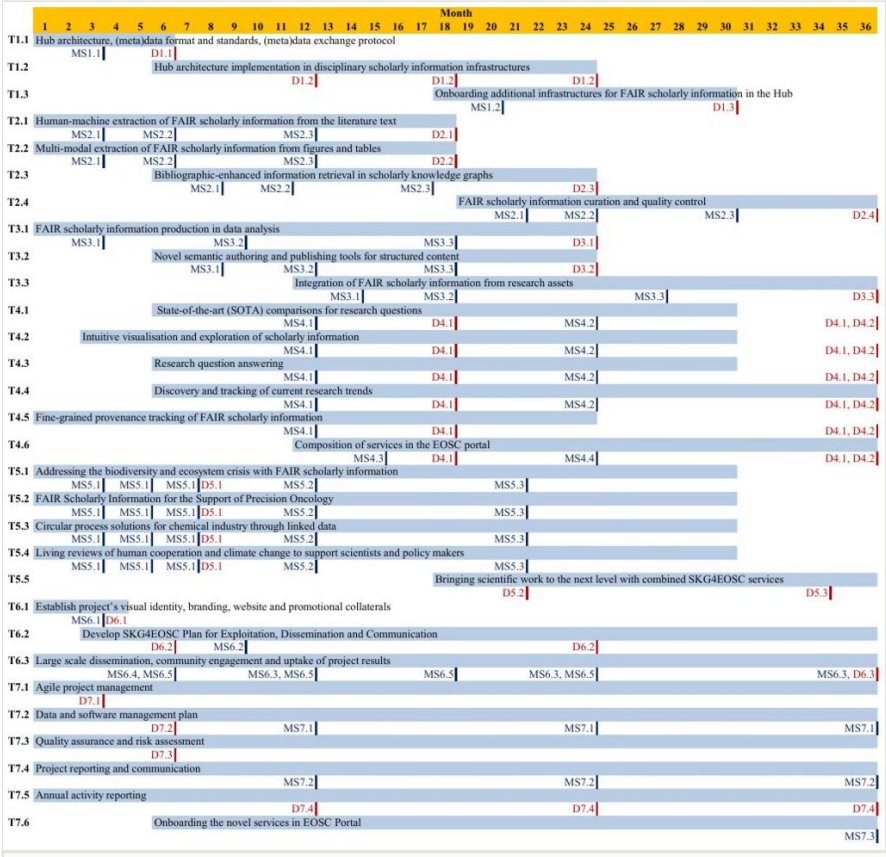

Figure 7. doi

SKG4EOSC Gantt chart showing Task durations, Deliverable due dates, and Milestones.

3.2 Capacity of participants and consortium as a whole

The project's objective is to establish an EOSC service ecosystem for FAIR scholarly information production, curation and use by lifting heterogeneous disciplinary scholarly information infrastructures into EOSC and devising innovative services. To match this core objective, the SKG4EOSC consortium involves top leaders in inter-disciplinary technology, infrastructure and service research and innovation and top leaders in disciplinary scholarly knowledge organization. With the *ORKG*, TIB has established its leadership in the area of FAIR scholarly information with knowledge graph technologies. As SKG4EOSC, TIB brings into the consortium vision and multi-year experience in developing services for FAIR scholarly information, as well as a world-class record in research and innovation on relevant technologies. This expertise is complemented by partner INFAI, which brings into the consortium strong technology knowledge to lead the WP4 developments on ORKG enabled services for FAIR scholarly information use in the EOSC. Technology expertise is further complemented by partner VUA, which has been leading the research and development on *Nanopublications*, a technology component that will be fundamental to one approach for distributed harvesting pursued within SKG4EOSC. As a representative of Wikimedia and its ecosystem of services including *Wikidata*, *Wikibase*, *WikiCite*, SKG4EOSC partner USFAX brings into the consortium strong technology knowledge, especially in regard to scalable Crowdsourcing. With USFAX, SKG4EOSC will lift the Wikimedia service ecosystem into the EOSC by making services such as *Wikidata* and *Wikibase* compatible with the proposed approaches to distributed content harvesting and retrieval. *Wikidata* and *Wikibase* will thus be additional service offerings that disciplinary initiatives can employ to manage FAIR scholarly information.

SKG4EOSC builds on numerous and heterogeneous disciplinary scholarly information infrastructures in four disciplines (biodiversity, life sciences, chemical engineering, and social sciences) with their respective partners. With the services *Hi Knowledge*, *Linear Mixed Model KG*, and *OpenBiodiv*, the partners IGB, IPG PAS and PENSOFT have developed state-of-the-art infrastructure and services for FAIR scholarly information production, curation, publishing and use in biodiversity. In SKG4EOSC, these partners will demonstrate the production, curation and use of FAIR scholarly information in a pilot on the biodiversity and ecosystem crisis (T5.1). With the Linear Mixed Model KG, IPG PAS is one of few institutions that have applied the FAIR data principles to scholarly information during the data analysis phase of the research lifecycle. In addition to supporting the biodiversity pilot (T5.1), with its leading experiences in analysis of biological data, IPG PAS will also lead the WP3 developments on pre-publication FAIRification of scholarly information. With Hi Knowledge, IGB has demonstrated disciplinary leadership in information extraction for biodiversity literature, the organization of such information and the development of visualization services using the organized information. With their long-standing experiences in this area and as a strong disciplinary partner, IGB will lead WP5 activities on the development of pilots that will demonstrate the use of FAIR scholarly information in research, building the knowledge base to tackle global societal challenges. With *OpenBiodiv* and as a world leader in Open Access publishing tools, infrastructure and

services, PENSOFT further complements the biodiversity pilot and also brings into the consortium strong technology expertise.

With partners, LUH, SERMAS, UPM, and IDIPHIM, SKG4EOSC will demonstrate the production, curation, and use of FAIR scholarly information in a pilot on precision oncology (T5.2). These institutions have long-standing collaborations in numerous international projects. They merge world-class technology expertise in knowledge graph applications for knowledge organization in the sciences with disciplinary knowledge in life sciences. Partner LUH brings expertise in knowledge graph technologies into the consortium and further complements SKG4EOSC with technical knowledge. UPM are experts in AI technologies for processing clinical data and forecasting patterns for patient stratification. Lastly, SERMAS and IDIPHIM are leading healthcare institutions in cancer treatment and share strong knowledge on cutting-edge clinical methods for cancer patient profiling.

Partners TU Delft and UM are internationally recognized leaders in digital technologies and services for chemical engineering and cheminformatics and will thus demonstrate the production, curation and use of FAIR scholarly information in a corresponding pilot on circular process solutions for the chemical industry (T5.3). As WP2 lead, TU Delft brings into the consortium world class technology expertise in information extraction from literature, text, images and other information assets. Similarly, UM is a recognized leader in developing tools, infrastructure and services for cheminformatics, promoting Open Science, for example in the Journal of Cheminformatics.With the development of Cooperation Databank, partner VUA has demonstrated world class leadership in applying knowledge graph technologies for the production, curation and use of FAIR scholarly information in social sciences. In SKG4EOSC, VUA will leverage their technology expertise and disciplinary knowledge in the development of a pilot on human cooperation and climate change (T5.4).

As such, the SKG4EOSC consortium is a perfectly balanced set of partners, bringing together the necessary disciplinary and inter-disciplinary knowledge to address the project's objective. The consortium includes expertise in social sciences and humanities and develops a corresponding pilot. As underscored in Section 1.2.6, all partners have been leading Open Data, Software, Knowledge and Science champions for decades.

With PENSOFT, SKG4EOSC also involves an SME in the scholarly publishing sector that is well-known among academics worldwide with its technologically advanced peer-reviewed Open Access journals publishing in the domain of biodiversity as well as the development of advanced digital services in scholarly communication such as the ARPHA-XML publishing workflow and the *OpenBiodiv* knowledge graph participating in this project. PENSOFT’s project department consists of a motivated team of active scientists, project managers and science communicators offering long-standing expertise in delivering the full set of science communication services. PENSOFT has been involved in the science communication of over 30 projects, which guarantees the company’s experience and competence. The company is actively developing new tools, workflows and methods for text- and data publishing, dissemination of scientific information and technologies for semantic enrichment of an articles’ content. PENSOFT will lead WP6 and support

communication and dissemination activities for the entire project. PENSOFT will be appointed a lead partner to coordinate the exploitation of results in accordance with the PEDCOM (D6.2, see also Section 2.2). For example, the open source tools developed in the project will be supported by the relevant documentation and made available for use for any interested organisation, mostly industrial entities but also libraries and non-commercial organisations.

Table 12 provides an overview of partner expertise in relevant areas.

Table 12.

Competence/expertise vs. Partner Matrix.

	TIB	LUH	VUA	PENSOFT	IPG PAS	TU Delft	UM	IGB	USFAX	SERMAS	UPM	IDIPHIM	INFAI
Partner #	1	2	3	4	5	6	7	8	9	10	11	12	13
Technologies													
Semantic Technologies	X	X	X	X	X		X	X	X				X
User interfaces	X	X	X	X	X								X
Information extraction	X	X	X	X		X	X	X	X		X		X
Graph federation	X	X	X	X			X						X
Query Processing		X		X	X			X					X
Integrity Validation		X		X	X			X					X
NLP		X				X			X		X		X
Linked Open Data	X	X	X	X			X	X	X				X
Horizontal aspects													
Digital services	X	X	X	X	X	X	X	X	X				X
User engagement	X			X			X	X					X
Agile development	X	X					X		X				X
Data science	X	X	X	X	X	X	X	X	X	X	X	X	X
Application domains													
Computer Science	X	X	X				X		X		X		
Biodiversity				X	X			X					
Life Sciences		X			X		X			X	X	X	

	TIB	LUH	VUA	PENSOFT	IPG PAS	TU Delft	UM	IGB	USFAX	SERMAS	UPM	IDIPHIM	INFAI
Information Science	X	X					X	X	X				X
Chemistry						X	X						

4. Workpackage description

WP1 - *ORKG* as a Hub for FAIR scholarly information in EOSC

The objective of this WP is to harmonize access to FAIR scholarly information in the EOSC by standardising data format and exchange protocols of disciplinary scholarly information infrastructures (directly involved in SKG4EOSC, additionally onboarded during the project, or joining the federation after the project ended). Harmonized access is to enable the efficient development of a wide range of generic and specialized services for the production, curation, and use of FAIR scholarly information in the EOSC.

T1.1 - Hub architecture, (meta)data format and standards, (meta)data exchange protocol

The objective of T1.1 is to evaluate two approaches for harmonizing access to FAIR scholarly information in the EOSC and determine the optimal approach or whether both should be pursued to power WP4 services and WP5 applications. The first approach is decentralized harvesting and leverages *Nanopublications*. Here, disciplinary infrastructures publish scholarly information as *Nanopublications* and *ORKG* harvests the published assertions. The task will evaluate the use of *nanopub-servers**[10] (*Kuhn et al. 2016*) for the publishing of *Nanopublications* by disciplinary scholarly information infrastructures, thus leveraging existing services. The second approach is decentralized retrieval and leverages *GraphQL*. Here, disciplinary scholarly information infrastructures implement an API of their choice (*GraphQL*, REST, SPARQL, etc.). Instead of harvesting data, *ORKG* implements a *GraphQL* endpoint that enables distributed access to disciplinary scholarly information infrastructures and their content. Both approaches enable harmonized access to FAIR scholarly information and efficient use by Hub-enabled services (WP4).

T1.2 - Hub architecture implementation in disciplinary scholarly information infrastructures

Building on T1.1, the objective of T1.2 is to implement the approaches for harmonizing access to FAIR scholarly information in all disciplinary scholarly information infrastructures. In a first stage, T1.2 will prototypically implement the architectures for the two devised approaches as a minimal viable product. In a second stage, the implementation of the approach that is primarily required by Hub-enabled services (WP4) is then further refined by all disciplinary scholarly information infrastructures for deployment in production environments.

T1.3 - Onboarding additional infrastructures for FAIR scholarly information in the Hub

The objective of T1.3 is to develop the technical specification which current and future infrastructures for FAIR scholarly information are required to implement in order to integrate with the Hub. T1.3 will evaluate the practical viability of the developed specification in onboarding additional (1-2) infrastructures for FAIR scholarly information during the project's lifetime (e.g. CEUR-WS). The specification will be published Open Access with CC BY license to support future adoption by other infrastructures and thus catalyse the production, curation, and use of FAIR scholarly information in the EOSC.

WP2 - Literature FAIRification: Post-publication Text Mining and Crowdsourcing FAIR scholarly information

The objective of this WP is to design and implement open source, web-based interactive tools for the extraction of information from published scholarly articles. This includes the use of natural language processing to extract information from text, computer vision to extract information from images, and the analysis of bibliometric metadata. Moreover, the extracted information is curated to ensure high quality data.

T2.1 - Human-machine extraction of FAIR scholarly information from the literature text

The goal of T2.1 is the design and implementation of an interactive web-based information extraction tool for scholarly literature. This human-centred approach will leverage the joint power of human experts and machine intelligence for understanding and structuring scholarly information at large scale. In the first step, the topics of scientific publications are identified through a (dynamic) topic model that is trained on the abstracts and relevant metadata. In the web-tool, a user selects or uploads a publication within his or her field and marks relevant text passages for information extraction (c.f. building on the *ORKG* PDF annotator tool*[11]). The selected text is automatically imported into our tool and a number of pre-processing tasks are executed. This includes the recognition of technical vocabulary from taxonomies. Also, we train and apply a pre-trained BERT transformer model (Devlin et al. 2019) for named entity recognition and relation extraction (Xue et al. 2019). Notably, there exist pre-trained BERT models for multi-domain scientific publications (Beltagy et al. 2019). The results of the named entity recognition are suggested to the user who can confirm or change the model predictions. Similarly, an active learning approach is used to map the entity to a template from a pre-defined ontology. This interactive procedure increases data quality and allows for an active learning approach, which will improve the model performance over time. Moreover, the metadata of the publication, text passage, user ID, and model version will be saved in the KG to improve trustworthiness. This task will initially focus on a specific scientific domain related to the use cases in WP5 with established ontologies (e.g. extraction of chemical compounds and properties from literature). Then, the developed framework will be extended to other domains.

T2.2 - Multi-modal extraction of FAIR scholarly information from figures and tables

Figures and tables contain densely-packed semi-structured information in scientific literature. In T2.2, we implement a service deployable in *ORKG* or disciplinary infrastructures that extracts information from figures and tables. First, figures and tables will be extracted from the scholarly documents through established tools. We will compare existing tools, integrate the best tool as a service in the *ORKG*, and improve its performance through an active learning setup. This includes PDF processing toolkits (e.g. PyMuPDF) and computer vision tools (DeepFigures, PDFFigures 2.0, DeepPDF). Second, we aim to extract information from identified figures and tables. While domain-independent methods for information extraction from tables exist (academic (Pinto et al. 2003) and commercial*[12]), the extraction of information from figures is often domain-specific. We aim to select and integrate an established algorithm for information extraction from tables. For the classification of figures, we will develop a semi-supervised approach that suggests labels based on captions (using named entity recognition from T2.1). A user will check the labels. Then, we will train an image classification algorithm for the labelling of figures. In a subsequent step, domain-specific information extraction approaches will be developed, and existing algorithms will be incorporated. This includes the extraction of information from chemical block flow diagrams using object detection (Kang et al. 2019) and the extraction of chemical structure descriptions using transformer models (c.f. Rajan et al. (2021)). A subset of the extracted figures and associated FAIR data will be uploaded to Wikimedia Commons with *structured data* annotation for reuse (e.g. by using existing infrastructures*[13]).

T2.3 - Bibliographic-enhanced information retrieval in scholarly knowledge graphs

There are currently several open bibliographic databases (e.g. *DBLP*, *OpenCitations* and *Wikidata*) that can be used to support the information retrieval of research findings from scholarly publications, including recognition of citation types. Leveraging these resources with rich and interlinked metadata describing the context of research (e.g. publications, datasets, people, organizations, etc.) can bring an added value to the algorithms aiming to extract scholarly outputs from full texts. The objective of T2.3 is to couple these databases with full text analysis (T2.1) and processing using semantic similarity measures, co-word analysis, semantic annotation and advanced machine learning techniques including graph and word embeddings, deep learning and machine learning. This will enable the generation, validation, adjustment and the addition of reference support to FAIR scholarly information published in *ORKG*.

T2.4 - FAIR scholarly information curation and quality control

A technical challenge in SKG4EOSC is the orchestration of intertwined, iterative data and software lifecycles with the final objective to reach sufficient quality to achieve satisfaction of researcher requirements (end user). Harvested payload (content/data), metadata describing the payload, integration metadata (conversion, links, mappings), crowd- and domain-expert annotations, user-assisting tools, trained and retrained models for these tools are continuously iterated in SKG4EOSC and form a hyper-dimensional Petri net

(distributed, discrete event dynamic system) of dependencies that requires efficient, automated and effective quality control. T2.4 will leverage INFAI's *Databus* to track and version artefacts, software, metadata in a Persistent ID Graph, that already implements all FAIR principles and will be extended in SKG4EOSC to additionally

1. adhere to TRUST principles (Lin et al. 2020),
2. provide ingrained continuous integration (CI) data quality management processes,
3. measure quality of interconnectedness (links) and
4. deploy pre-packaged "software with data" developed in WP2 to the EOSC cloud as services.

Building on previous work in the context of DBpedia and on the transparency of *Databus*, we will implement delegation of data issues in the opposite direction of the data flow from end user over intermediate nodes to source and notify consumers, curators and creators of usage, updates and prioritized issues. INFAI's role is to develop the *Databus* platform and collaboratively integrate work by other partners, in particular Trav-SHACL*[14] (an SHACL engine by LUH) and other WP2 tasks.

WP3 - FAIR-at-Birth: Pre-publication production of FAIR scholarly information

The objective of this WP is to ensure scholarly information is produced FAIR by retaining its semantics and exposing it as FAIR digital objects to scholarly communication (data services and publications). In particular:

- Implement novel technologies and workflows for next-generation semantic publishing of structured content;
- Pre-publication tools for conversion and embedding both human-readable and machine-actionable observational and statistical results into manuscripts and "living articles";
- Align the use of identifiers and bi-directional linking of FAIRified data between literature and research infrastructures to enable distributed systems to interface with one another;
- Improve and extend automated workflows for seamless post-publication interoperation between knowledge, information and data through RDF-conversion of published content into LOD resources, including *Nanopublications* and formalization papers.

T3.1 - FAIR scholarly information production in data analysis

The objective of T3.1 is to ensure that scholarly information produced in data analysis is produced FAIR. We will develop methodology to preserve (e.g. statistical) information directly from the computational environments in which analytical results are being generated, and make them machine-actionable for storage and further use in data services and publishing. In addition to disciplinary scholarly information infrastructures, we will explore using the DBpedia *Databus* platform as a storage for FAIR scholarly information produced in data analysis. Based on initial user requirement analysis, for selected

computational environments and data analysis types we will provide tools (e.g. libraries in Python or R) that allow data scientists to document and expose the key results of data processing (i.e., meaningful statistics, such as model parameter estimates or research hypotheses tests (Ćwiek-Kupczyńska et al. 2020)) as FAIR digital objects. The information will be modelled and annotated according to adequate general and domain-specific ontologies (e.g. STATO, QB), integrate persistent identifiers to link contextual entities (e.g. InChI for chemicals, Gene Ontology for biological processes and functions), and made available in EOSC through the *ORKG* Hub by means of the project-proposed approaches (e.g. *Nanopublications*). The role of VUA will be to assist with the nanopublication modelling and their integration into data analysis environments.

T3.2 - Novel semantic authoring and publishing tools for structured content

The objective of T3.2 is to develop the ARPHA Writing Tool (AWT) into an independent, innovative, standard-aligned, collaborative authoring and editing environment with extensions and plugins for import of structured data and metadata into texts and semantic enrichment of the narrative. It will integrate narrative and data in an efficient and highly automated way to produce FAIR scholarly information expressed in machine-readable formats (e.g. as *ORKG* templates or key findings expressed in *Nanopublications*) before the manuscript is submitted to journal publishing workflows, including conversion to Linked Open Data. We will thus take a definitive step towards resolving the “PDF impediment” to knowledge sharing through facilitating data-driven and semantically enriched publishing, based on both generic and domain-specific standards and ontologies. The AWT will generate JATS XML documents and include persistent identifiers (e.g. DOI, compact identifiers, etc.). These augmented documents can be used for:

1. submission to journal publishing workflows, for example the TIB Open Publishing or WikiJournals, or
2. computational workflows for efficient harvesting and further re-use of semantically enriched content.

SKG4EOSC will demonstrate the workflow through publication of its entire research cycle in a project-branded collection in RIO (T6.3). The role of VUA will be to assist with the nanopublication modelling and their integration into the writing tool and the final publications.

T3.3 - Integration of FAIR scholarly information from research assets

Research assets are structured and information-rich data files that are critical for the reusability, reproducibility, and thus credibility of research results. Common research assets that are published together with scholarly publications include computer code, models, simulation files, electronic lab notebooks, data management plans, supplementary media files, and workflows. The automated integration of information from those assets with the publication itself is highly desirable. The objective of T3.3 is to develop information extraction services for research assets. This includes the use of specialized NLP tools (e.g. FALCON) for named entity recognition of text-based files (e.g. code). Domain-specific tools

for information extraction from research assets are also used and further advanced. This includes information extraction from chemical process simulation files, where adoption of interoperability standards will be coordinated with the GO FAIR Chemistry IN. Extracted scholarly information is represented as *Nanopublications* using suitable templates. The role of VUA will be to assist the nanopublication modelling and to create the needed semantic templates.

WP4 - Novel Hub-enabled intuitive interaction services and service composition in EOSC

The ultimate objective of this WP is to develop novel services for FAIR scholarly information and enable use of the services in the EOSC platform. The developed services will facilitate the work in WP5 in relation to the identified global challenges. WP4 will ensure that the developed approaches in WP2 and WP3 are generic enough to be utilized in novel, yet unknown scenarios. Also, the goal is to ensure WP2 and WP3 approaches are composable services.

T4.1 - State-of-the-art (SOTA) comparisons for research questions

The objective of T4.1 is to enable retrieval of state-of-the-art information for concrete research questions. This will be implemented as a service where users can explicitly state (a) research question(s) and retrieve a state-of-the-art comparison for the provided question(s). The service will be built on top of the foundations of the *ORKG* comparison capabilities. A critical, non-obvious objective of T4.1 is data quality (DQ) of created SOTA objects in particular regarding breadth and coverage to achieve complete, unbiased (and potentially inter-disciplinary) SOTA comparisons. We will leverage T2.1-2.3 and implement use of ontologies for research questions (IGB).

T4.2 - Intuitive visualisation and exploration of scholarly information

The objective of T4.2 is to develop customizable services for visualisation and exploration of FAIR scholarly information published by *ORKG*. This will make the access and exploitation of the offered information more efficient. The service will provide faceted search capabilities and personalized retrieval of scholarly information. The service will support scientists to get an overview and latest advancements in a particular field via an interactive UI. A particular focus will be put on the temporal aspects, e.g. how cooperation research changed over time (Balliet et al. 2020). The service will provide predefined visualisations, which will be designed in close collaboration with the domain experts in WP5. Visualisations will be provided in a format which will be easy to use and integrate in scientific papers, with features such as citable snapshots to enhance reproducibility.

T4.3 - Research question answering

The objective of T4.3 is to develop different information retrieval services utilizing question answering (QA) methods. Building on the extracted data resulting from T2.1 and T2.2, ontology-based machine learning models and natural language processing techniques are

employed to expose a new natural language interface for users. The service will enable users to explore research contributions and scientific data via posing queries formulated in natural text and getting precise answers (such as resources, papers, and comparisons). Moreover, with the availability of structured information in the form of a knowledge graph, the services' QA module will be able to find relations between entities and better aggregate information and deliver answers (Jaradeh et al. 2020). Methods such as named entity recognition and disambiguation, and relation extraction are base techniques for the QA system to build on in order to comprehend a posed question and try to find the answer. For instance, an example question would be "What is the most common machine learning method used by state-of-the-art papers addressing entity linking?". To answer this question, the schema of the knowledge graph needs to be understood by the system and entities to be recognized. Though extracted research data is stored in a KG, different data objects have different representations in the graph and require different traversals and comprehension techniques. As such, task **T4.3** requires different types of QA systems to address different data forms and information representations (e.g. QA on tables, QA on figures, QA on datasets). Such systems will have user interfaces for user consumption and API access for other types of clients.

T4.4 - Discovery and tracking of current research trends

The objective of T4.4 is to develop a service that enables users to discover and track current research trends. The service will exploit machine learning techniques for identification of topics. The service will reuse the models developed in T2.1 for topic identification. Researchers and other interested parties can use the service to track the popularity of particular research topics over time. The service will also enable users to get insight information about user specified topics, but also get information about the currently "hot", most attractive topics in the recent period. In addition to topic tracking, the service will enable identification and tracking of other scholarly information such as datasets, code and other related materials. The service will be accompanied by an automated notification system, which will in an automated manner announce updates and trends to its subscribers (e.g., RSS feed, Twitter bot, mailing list).

T4.5 - Fine-grained provenance tracking of FAIR scholarly information

The objective of T4.5 is to increase the trust in science by improving transparency and reproducibility of results. Research lifecycles generate considerable amounts of provenance information in each step, and this task will develop an EOSC service for fine-grained provenance tracking of FAIR scholarly information covering the whole lifecycle. The service will relate FAIR scholarly information to primary data, capture and expose contextual information, i.e. the activities and the agents involved in these activities by building on PROV-O. Research results will be made more sustainable via an archiving model that enables to load and re-run archived experiments and analyses on EOSC via the *Databus* model to guarantee sustainable reproducibility and reusability via composition (T4.6) beyond the lifetime of projects, a big technical challenge in science. A custom graphical provenance browser will support user groups of WP5 in traversing and tracing provenance information.

T4.6 - Composition of services in the EOSC portal

The objective of T4.6 is to enable the composition of the developed and integrated services in the EOSC portal. A common API and protocol will be designed so that the services are easily integrated and composed in novel workflows. The services will rely on common data, which will enable effortless creation of service compositions. For example, using the federated graph visualization and exploration service, a researcher could explore the research knowledge graph and identify a topic associated with a particular research article. Next, the researcher could query the research topic trends service and retrieve the popularity information for this particular topic. The “composability” capability will be exploited in WP5 in domain specific and generic, cross-domain use cases.

WP5 - Leveraging FAIR scholarly information to meet global challenges

This work package involves experts from four domains (biodiversity, biomedicine, chemistry, social sciences) and is a testbed for the functionality and fitness of the services developed in WP1-4. In each domain, a current lack of FAIR information is impeding usage of scholarly knowledge for meeting global challenges. The domain experts will closely interact with their communities as well as with the other SKG4EOSC team members in defining user requirements, in adapting the existing domain-specific services for integration in EOSC, and in testing the services developed in WP1-4. Cross-disciplinary use cases will be constructed to exemplify the operability of the services across domains.

T5.1 - Addressing the biodiversity and ecosystem crisis with FAIR scholarly information

Given that 25% of animals and plant species are threatened with extinction and ecosystems are deteriorating worldwide (IPBES 2019), there is an urgent need for science-based applied solutions. For example, efficient management of invasive species relies on knowledge about their introduction pathways, current occurrences, potential for further spread and impacts. Efficient restoration of ecosystems, as a second example, requires knowledge about methods for facilitating the recovery of soils, re-establishment of species and triggering of ecosystem functions. In the respective research areas, specifically invasion biology and restoration ecology, publication workflows are impeding progress, since much of the locally produced scholarly knowledge remains hidden in PDFs that are hard to discover, not machine readable, and often hidden behind paywalls (Jeschke et al. 2021). Databases exist for providing some of the much needed information, but these focus on particular groups of organisms (e.g. GAVIA on birds (Dyer et al. 2017), GloNAF on plants (van Kleunen et al. 2018)) or on certain aspects of invasive species (e.g. their introduction pathways (Saul et al. 2016) or first records in different countries (Seebens et al. 2017)). For every species management plan or specific restoration project, experts will have to visit the disparate resources bit by bit to retrieve the information they need. A tool that has been developed to address some of these challenges is the hierarchical network of invasion hypotheses at *Hi Knowledge*. In a current project[73], it is being developed into an evolving knowledge resource, structuring and visualizing curated information on research questions, hypotheses and their empirical basis in the domain of invasion biology.

OpenBiodiv is another useful research infrastructure providing information on biological species. However, these services are not interoperable, and are known and used only in subgroups of the biodiversity research community. In T5.1, experts in the biodiversity domain will therefore define user requirements and use cases and test which of the two pathways is most useful for enhancing the production, curation and regular use of FAIR scientific data in the field: (a) *Nanopublications*, (b) feeding the local knowledge into KGs that are then connected via the SKG4EOSC hub. Either way, the services that were previously restricted to a local user group because of limited scope and prominence will become available to the European research community at large and could thus become a central source for scholarly information for the biodiversity domain, providing an overview on current research, easing discoverage of relevant information and allowing close involvement of the research community. All this will bring a steep increase in the usability of scholarly information and thus a much stronger basis for actions against the biodiversity and ecosystem crisis.

T5.2 - FAIR scholarly information for the support of precision oncology

Cancer is a leading cause of death worldwide; it generates a tremendous psychological, financial, and physical burden. Advances in oncological treatments enable the potential control of the disease. However, because of the yearly increase in the medical literature, physicians require enormous hours to maintain track of new medical research, hindering, thus, novel treatments' reproducibility. In this task, FAIR Scholarly Information (FSI) will be created from fine-grained descriptions of oncological literature, clinical trials, and oncological treatment response and disease prognosis of patient populations (e.g. lung and breast cancer patients). FSI will be applied in other populations to evaluate reproducibility. So far, systematic literature review tools (Scott et al. 2021), pre-trained embeddings (Rasmy et al. 2021), and literature mining (Zhao et al. 2020) are devised for solving specific tasks (e.g., disease prediction). Still, considerable manual work is needed to extract clinical variables required for tracing reproducibility of patients' outcomes and novel treatments. In the context of the EU H2020 project CLARIFY, an oncology KG has been built that comprises biomedical entities and relations extracted from scientific articles and databases (e.g., PubMed and DrugBank), and a fine-grained description of clinical notes for lung and breast cancer patients from SERMAS. Clinical data includes wearables, physical examination, oncological treatments, long-term toxicities, and tests. The *CLARIFY KG* is used to identify patterns to understand long-term toxicities generated by oncological treatments, and relapse and progression of the cancer. Analytical services against the *CLARIFY KG* enable the detection of patterns in treatment responses, disease prognosis, and toxicities. However, the lack of FSI hinders reproducibility validation of uncovered patterns and medical outcomes reported in the literature. Based on a specification of user requirements and formulation of use cases, in T5.2 biomedical entities and relations (e.g. genes, proteins, metabolites from biological pathways, and interactions) extracted from the literature will be integrated into an *SKG* with oncology articles and their comparisons included in *ORKG*. The *SKG* will be linked to the *CLARIFY KG* and existing KGs (e.g. *Wikidata*, *DBpedia*, *Bio2RDF*, and *ELIXIR Core Resources*). Analytical methods and services for reasoning and decision support will enable real-time systematic reviews,

epidemiological studies, and clinical trials based on patient information and related results represented in FAIR scholarly information.

T5.3 - Circular process solutions for chemical industry through linked data

The European Union aims at achieving climate neutrality by 2050. To achieve this goal, the chemical industry requires a transformation towards more sustainable and circular practices, eliminating its dependencies on fossil fuels and limiting its impact on the environment (Clark 2017, Kätelhön et al. 2019). This has further implications for many other fields, like medicine and food production. However, sustainability problems are wicked problems that cover almost every aspect of society and often include unforeseen implications (Norton 2015). Thus, they cannot be solved through traditional disciplinary approaches. One of the main problems is the lack of linked and FAIR multidisciplinary scholarly information that is needed for model building and (optimal) decision making (Weber et al. 2021, Patel et al. 2012). In T5.3, we will extract and link information for the support of holistic (Social-)Life Cycle Assessment of processes and products. This includes a multitude of aspects such as process information from chemical engineering, toxicity information from biology, nutrients in food production, synthesis routes from chemistry, and various regional data (e.g., water availability, human rights, and labour rights). For this, we will identify the key facts in the literature to monitor progress in knowledge around the selected processes and products in the four domains according to specific user requirements and focal use cases that will be developed. For each key fact, essential concepts and identifiers will be defined that will allow information extracted from literature to be analysed and linked to other databases. This will be done in collaboration with European initiatives such as EU NanoSafety Cluster (e.g. H2020 NanoCommons, RiskGONE, SbD4Nano) (Karcher et al. 2018), VHP4Safety, FNSCloud, who are stakeholders interested in this knowledge. Thus, our tool will provide researchers, companies, and policy-makers the linked information for the evidence-based multi-criteria decision-making. The tool supports extraction of product composition, chemical structure and chemical properties or nutritional value data from literature, information about chemical processes, product content, and identification of waste and potential pollution. Extracted facts will be accessible through *ORKG* (WP1). Finally, molecular property models (e.g., graph neural networks (Schweidtmann et al. 2020)) will be trained on the linked FAIR information and will be used to support the (optimal) identification of circular process solutions.

T5.4 - Living reviews of human cooperation and climate change to support scientists and policy makers

Human behaviour is causing climate change and producing an existential crisis on a global scale. Scientific research about how human behaviour is affecting the environment and climate is moving at a fast pace, including institutional and behavioural interventions (Alló and Loureiro 2014, Nisa et al. 2019, Hornsey et al. 2016). We need to continually update our understanding of this scientific information to build and implement policies that enable humans to cooperate to solve the problem of climate change. Yet, scientific publications are made available in PDFs and datasets and are not easily integrated and meta-analysed.

The *Cooperation Databank (CoDa)* (Spadaro et al. 2020) is a knowledge graph of scientific studies about human cooperation, which can be used to produce on-demand meta-analyses and living meta-analytic reviews. Living meta-analytic reviews are summaries of scientific research that can be automatically updated with the emergence of new scientific findings. The *CoDa* knowledge graph has a rich description of the provenance of scientific results and has an application that allows users to aggregate scientific results and then analyse how the results vary across different contexts. *CoDa*, however, currently has no links to environmental or other datasets that allow for living reviews about the interconnection of human cooperation with ongoing environmental changes. Also, *CoDa* now does not represent different theories and hypotheses, and in T5.4 we will do this by using automated analyses over the text of the papers. Identified theories and predictions about cooperation and climate change could then be tested with on-demand meta-analytic analyses and within living meta-analytic reviews. *CoDa* is Findable, as it contains rich metadata descriptions, and it is Accessible, as its data is represented in standard format (RDF), and it is Reproducible as it contains clear data usage and access license. That said, *CoDa* is not yet Interoperable, as it is not fully linked to other datasets. *CoDa* does not include references to other datasets and is limited in the reuse of vocabularies (both points related to the “Interoperability” principle). Interoperability is needed in order to create living meta-reviews on climate change. In T5.4, we will use the *CoDa* knowledge graph to build an ontology and knowledge graph of studies about human attitudes, beliefs and behaviours that affect climate change, and the various forms of interventions that have been studied to affect behavioural change. It will then be possible to create queries across the knowledge graph that can create living reviews for scientists and policy makers to keep them up to date with the literature.

T5.5 - Bringing scientific work to the next level with combined SKG4EOSC services

With T5.5, partners from several research domains will demonstrate how the combined application of services developed in WP1-4 can be leveraged to solve challenges scientists are facing across domains. They will also explore ethical aspects of the challenges and solutions, particularly those that apply across domains, such as data sovereignty, knowledge equity or the environmental footprint of knowledge graphs. Recurring challenges are, for instance, the connection of single study results to existing hypotheses and theories (Heger et al. 2020), determination of the level of empirical support for research hypotheses by meta-analyses, identification of research gaps and biases and analysis of the robustness of results (‘reproducibility crisis’)*[15] (Spadaro et al. 2020). Respective work on these issues is sparse, disconnected, and occurring in different disciplines. The researchers collaborating on this proposal have made pioneering contributions to these issues, such as the *Cooperation Databank* providing living reviews of research gaps and biases*[16]. However, results of such efforts often stay hidden in local research environments or are published in the usual scientific outlets with the known shortcomings. The full benefits of the rich portfolio of services that will be developed in WP1-4 will become exploitable once a combined access via the EOSC is made possible (T4.6). In the second half of the funding period, domain experts will develop use cases to

demonstrate how the SKG4EOSC services can be integrated in scientific workflows in order to:

- regularly check for gaps and biases to direct research efforts,
- maximize robustness of results by pointing to areas in demand of repeated analyses, and
- establish procedures facilitating close connection of empirical results with theory and affected communities.

WP6 - Communication, dissemination, engagement, and exploitation

WP6 aims to maximise the impact of SKG4EOSC by communication and engagement with the best practices and knowledge generated by the project with relevant user groups. Using state-of-the-art methods for community engagement, the project has the specific objectives to:

1. raise awareness of the project outputs through a recognisable project brand and active dissemination and promotion;
2. maximise the impact and outreach of SKG4EOSC results to user groups;
3. engage participating research infrastructures in sharing open standards and practices;
4. build a strong scientific community for uptake of results, and
5. maximise access to and re-use of research data generated by SKG4EOSC.

WP6 will be focusing on facilitating access to the knowledge graphs and their backing research infrastructures by modern and technically advanced active dissemination and communication methods, and ensure the sustainability of the project results through engagement with the *ORKG* Hub developed in WP1. This includes the engagement in EOSC through the deployment of most up-to-date technologies. WP6 will achieve its objectives via an ambitious and structured approach, relying on both best practices in the field of science communication, and continuous monitoring and upgrade of communication and dissemination tools. To reach out to all relevant target groups and share project progress and results, WP6 will implement custom communication and dissemination methods for each group, such as training webinars, hackathons, videos and more.

T6.1 - Establish project's visual identity, branding, website and promotional collaterals

Within the first months of the project, a recognisable project visual identity will be developed alongside an initial marketing pack (a project logo, brand manual, brochure, poster, stickers, letterhead and presentation templates). These will ensure that SKG4EOSC is communicated effectively and professionally with the objective to raise awareness and build a community from the very start. A modern and user-friendly public website will provide an easy-to-navigate, continuously updated platform, allowing for fast access to general information about SKG4EOSC and its activities. The website will be secured for rapid uptake and long-term persistence by bringing together and multiplying

the strong brands of the participating research infrastructures, their extensive experience and existing communities.

T6.2 - Develop SKG4EOSC Plan for Exploitation, Dissemination and Communication (PEDCOM)

A comprehensive Plan for Exploitation, Dissemination and Communication of project results will set the rules and Key Performance Indicators (KPIs) to guide and measure communication activities during the project's duration. The Plan will guarantee outreach to all target groups and will ensure the uptake of results during and after the project duration by defining:

1. key dissemination actors and targets;
2. key messages that the project wants to deliver;
3. multi-modal mix of communication channels, e.g. social media, events, press releases etc.;
4. bilateral communication approaches, and
5. KPIs for each outreach activity.

The Plan will give special emphasis to engage with stakeholders and to attract new communities of users along the entire data life cycle (see T6.3). A revision of the Plan (M24) will provide a midterm evaluation and upgrade according to the performance of communication and dissemination tools.

T6.3 - Large scale dissemination, community engagement and uptake of project results

These activities will be operating on several levels: internal communication system, based on a professional content management software hosted on EU servers (e.g. Teamwork), open notebook science, open access publication of all important project outcome, blog and public relations interface, social media profiles in Twitter and Facebook, and others, to ensure effective integration, prioritization, cost-effectiveness and sustainability of the community's communication interface, networking activities and operations during the project lifetime and especially beyond it through the *ORKG* Hub (WP1). Following a mix of traditional and innovative approaches and best practices in science communication and state-of-the-art tools, and based on the project's **PEDCOM** (T6.2), a series of FAIR-by-design dissemination and training tools and events tailored to the needs of the different stakeholder groups will support the knowledge transfer and capacity building (e.g. tutorial video screencasts, demonstration showcases recorded in video, webinars, and other engagement events, see Table 8). In close relationship with the virtual services and use cases developed under WP4 and WP5, a set of 4 expert round tables for tackling user requirements, technology of access, interoperability standards and contributing to WP1 activities will be organized to collate expert advice from various stakeholders, including industrial actors and to foster cross-disciplinary fertilisation (M6.5). The *ORKG* Hub will further advance the SKG4EOSC approach with virtual workshops, curriculum, best practices and models for collaboration at all scales (Tables 8, 11). To reach a wide, global

and cross-disciplinary audience beyond the scientific realm, the project will contribute to the Wikipedia ecosystem, in particular *Wikidata* and Wikimedia Commons. The most relevant and impactful results will be published in authoritative open access journals and gathered together in a SKG4EOSC-branded collection of articles in the Research Ideas and Outcomes (RIO) journal, together with other community-related documents produced along the entire SKG4EOSC research cycle (grant proposal, methods, data management plan, workshop and project reports and other most important deliverable, standards, guidelines, policy briefs, training programme, etc), thus ensuring the FAIR and Open Science spirit and practice at all instances of the project lifetime.

WP7 - Project management

The aim of this WP is to ensure the high quality level of achievement of the project's results via the continuous monitoring of the implementation and completion of the project tasks, activities, milestones and deliverables, safeguarding their proper and timely development according to the DoW and the project's work plan, while ensuring the successful collaboration among the partners.

T7.1 - Agile project management

This task ensures the high quality, efficient and timely administrative coordination of the project. It incorporates Administration Management activities, including procedures and guidelines for activity planning and monitoring, cost and time management, submission of periodic progress reports and cost statements, preparation of annual review reports, review presentations, and timely submission of high quality deliverables to the Commission.

T7.2 - Data and software management plan

Aiming to improve and maximise access to and re-use of research data and software generated by Horizon Europe projects, SKG4EOSC will develop a Data and Software Management Plan within the first six months of the project, publish it and keep it up to date during the project's lifetime (annual updates). The main aim of the Plan is to adhere to the FAIR data management criteria and thus to leverage openness of the project's design and results. In particular, the Plan will describe what data/software types, licences, formats, access and archiving will be used within the project. To add value to the Plan, an additional one-pager with Data and Software Management Guidelines will be produced and shared with partners, in order to acquaint them with the recommendations valid for the project and serve as a guiding tool when generating/developing, collecting or using research data or software.

T7.3 - Quality assurance and risk assessment

The task focuses on defining and specifying the appropriate mechanisms and processes that will be established in order to maintain a high quality level in the whole project structure and outcomes. Additionally, T7.3 deals with the identification of potential project management risks and the respective monitoring of each risk profile as well as with the definition and timely application of contingency plans.

T7.4 - Project reporting and communication

The project coordinator will act as the point of contact for partners in communications with the Commission. The coordinator will ensure that the annual reporting to the EC, semi-annual technical internal reporting, milestone review, midterm review will be implemented.

T7.5 - Annual activity reporting

On an annual basis, project reports will be drafted and released focusing on the progress and intermediate results, and updated plans for the following period.

T7.6 - Onboarding the novel services in EOSC Portal

The objective of this task is to ensure that by project end, all adopted and newly developed SKG4EOSC services are discoverable in the EOSC Portal Marketplace (https://marketplace.eosc-portal.eu).

Funding program

This proposal was submitted to Horizon Europe Framework Programme (HORIZON). It did not get funded.

Grant title

Innovative and customizable services for EOSC

Hosting institution

TIB - Leibniz Information Centre for Science and Technology

Ethics and security

SKG4EOSC uses and develops methods in Artificial Intelligence, in particular machine learning, natual language processing, andknowledge representation and reasoning. These methods are used for the following purposes: Information extraction from literature,representation of machine actionable scholarly information, and processing of machine actionable scholarly information. Such use of Artificial Intelligence does not raise ethical concerns related to human rights and values.

Author contributions

Stocker, Heger, Schweidtmann, Ćwiek-Kupczyńska, Penev, Willighagen, Mietchen, Jeschke, and Auer contributed to ideation, writing, and review. Dojchinovski, Vidal, Turki,

Balliet, Tiddi, Kuhn, Karras, Vogt, Hellmann, and Krajewski contributed to writing and review.

Conflicts of interest

The authors declare no conflicts of interest.

References

- Akgun OC, Bazilinskyy P, Budroni P, Demoulin B, Dias F, Flicker K, Gazzola V, Giroletti J, Grossi V, Günes M, Leidel S, Leppert L, Mangeney J, Nogues Gonzalez I, O'Neill G, Ramadan H, Rauber A, Rezaei M, Rychnovská D, Sanchéz Solís B, Saurugger B, Steiner U, Stocker M, Susi T, Verde L (2020) Co-creating the EOSC: Needs and requirements for future research environments. https://doi.org/10.5281/ZENODO.3701194
- Alló M, Loureiro M (2014) The role of social norms on preferences towards climate change policies: A meta-analysis. Energy Policy 73: 563-574. https://doi.org/10.1016/j.enpol.2014.04.042
- Balliet D, Spadaro G, Markovitch B, Beek W (2020) How did cooperation research change over time? Cooperation Databank. URL: cooperationdatabank.org/data-stories/how-did-cooperation-research-change-over-time/
- Beltagy I, Lo K, Cohan A (2019) SciBERT: A Pretrained Language Model for Scientific Text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) https://doi.org/10.18653/v1/d19-1371
- Clark J (2017) From waste to wealth using green chemistry: The way to long term stability. Current Opinion in Green and Sustainable Chemistry 8: 10-13. https://doi.org/10.1016/j.cogsc.2017.07.008
- Coles S, Frey J, Willighagen E, Chalk S (2020) Taking FAIR on the ChIN: The Chemistry Implementation Network. Data Intelligence 2: 131-138. https://doi.org/10.1162/dint_a_00035
- Ćwiek-Kupczyńska H, Filipiak K, Markiewicz A, Rocca-Serra P, Gonzalez-Beltran A, Sansone S, Millet E, van Eeuwijk F, Ławrynowicz A, Krajewski P (2020) Semantic concept schema of the linear mixed model of experimental observations. Scientific Data 7 (1). https://doi.org/10.1038/s41597-020-0409-7
- Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. URL: arxiv.org/abs/1810.04805
- Dyer E, Redding D, Blackburn T (2017) The global avian invasions atlas, a database of alien bird distributions worldwide. Scientific Data 4 (1). https://doi.org/10.1038/sdata.2017.41
- European Commission (2016) Realising the European Open Science Cloud. https://doi.org/10.2777/940154
- European Commission (2018) Turning FAIR into reality. https://doi.org/10.2777/1524
- Floridi L (2013) The Philosophy of Information. Oxford University Press, 405 pp. [In English]. [ISBN 978-0199232390]

- Gentsch N, Boy J, Batalla JDK, Heuermann D, von Wirén N, Schweneker D, Feuerstein U, Groß J, Bauer B, Reinhold-Hurek B, Hurek T, Céspedes FC, Guggenberger G (2020) Catch crop diversity increases rhizosphere carbon input and soil microbial biomass. Biology and Fertility of Soils 56 (7): 943-957. https://doi.org/10.1007/s00374-020-01475-8
- Groth P, Gibson A, Velterop J (2010) The anatomy of a nanopublication. Information Services & Use 30: 51-56. https://doi.org/10.3233/isu-2010-0613
- Gu J, Qian L, Zhou G (2016) Chemical-induced disease relation extraction with various linguistic features. Database 2016 https://doi.org/10.1093/database/baw042
- Haris M, Farfar KE, Stocker M, Auer S (2021) Federating Scholarly Infrastructures with GraphQL. URL: arxiv.org/abs/2109.05857
- Heger T, Aguilar-Trigueros CA, Bartram I, Braga RR, Dietl GP, Enders M, Gibson DJ, Gómez-Aparicio L, Gras P, Jax K, Lokatis S, Lortie CJ, Mupepele A, Schindler S, Starrfelt J, Synodinos AD, Jeschke JM (2020) The Hierarchy-of-Hypotheses Approach: A Synthesis Method for Enhancing Theory Development in Ecology and Evolution. BioScience 71 (4): 337-349. https://doi.org/10.1093/biosci/biaa130
- Hornsey M, Harris E, Bain P, Fielding K (2016) Meta-analyses of the determinants and outcomes of belief in climate change. Nature Climate Change 6 (6): 622-626. https://doi.org/10.1038/nclimate2943
- IPBES (2019) Summary for policy makers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.
- Jaradeh MY, Oelen A, Farfar KE, Prinz M, D'Souza J, Kismihók G, Stocker M, Auer S (2019) Open Research Knowledge Graph. Proceedings of the 10th International Conference on Knowledge Capture https://doi.org/10.1145/3360901.3364435
- Jaradeh MY, Stocker M, Auer S (2020) Question Answering on Scholarly Knowledge Graphs. Digital Libraries for Open Knowledge 19-32. https://doi.org/10.1007/978-3-030-54956-5_2
- Jeschke JM, Lokatis S, Bartram I, Tockner K (2019) Knowledge in the dark: scientific challenges and ways forward. FACETS 4 (1): 423-441. https://doi.org/10.1139/facets-2019-0007
- Jeschke JM, Heger T, Kraker P, Schramm M, Kittel C, Mietchen D (2021) Towards an open, zoomable atlas for invasion science and beyond. NeoBiota 68: 5-18. https://doi.org/10.3897/neobiota.68.66685
- Kang S, Lee E, Baek H (2019) A Digitization and Conversion Tool for Imaged Drawings to Intelligent Piping and Instrumentation Diagrams (P&ID). Energies 12 (13). https://doi.org/10.3390/en12132593
- Karcher S, Willighagen E, Rumble J, Ehrhart F, Evelo C, Fritts M, Gaheen S, Harper S, Hoover M, Jeliazkova N, Lewinski N, Marchese Robinson R, Mills K, Mustad A, Thomas D, Tsiliki G, Hendren CO (2018) Integration among databases and data sets to support productive nanotechnology: Challenges and recommendations. NanoImpact 9: 85-101. https://doi.org/10.1016/j.impact.2017.11.002
- Kätelhön A, Meys R, Deutz S, Suh S, Bardow A (2019) Climate change mitigation potential of carbon capture and utilization in the chemical industry. Proceedings of the National Academy of Sciences 116 (23): 11187-11194. https://doi.org/10.1073/pnas.1821029116

- Kuhn T, Chichester C, Krauthammer M, Queralt-Rosinach N, Verborgh R, Giannakopoulos G, Ngonga Ngomo A, Viglianti R, Dumontier M (2016) Decentralized provenance-aware publishing with nanopublications. PeerJ Computer Science 2 https://doi.org/10.7717/peerj-cs.78
- Lastra-Díaz J, Goikoetxea J, Hadj Taieb MA, García-Serrano A, Ben Aouicha M, Agirre E (2019) A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art. Engineering Applications of Artificial Intelligence 85: 645-665. https://doi.org/10.1016/j.engappai.2019.07.010
- Li H, An H, Wang Y, Huang J, Gao X (2016) Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: Based on two-mode affiliation network. Physica A: Statistical Mechanics and its Applications 450: 657-669. https://doi.org/10.1016/j.physa.2016.01.017
- Lin D, Crabtree J, Dillo I, Downs R, Edmunds R, Giaretta D, De Giusti M, L'Hours H, Hugo W, Jenkyns R, Khodiyar V, Martone M, Mokrane M, Navale V, Petters J, Sierman B, Sokolova D, Stockhause M, Westbrook J (2020) The TRUST Principles for digital repositories. Scientific Data 7 (1). https://doi.org/10.1038/s41597-020-0486-7
- Lu K, Mardziel P, Wu F, Amancharla P, Datta A (2020) Gender Bias in Neural Natural Language Processing. Logic, Language, and Security189-202. https://doi.org/10.1007/978-3-030-62077-6_14
- Mietchen D, Penev L, Georgiev T, Ovcharova B, Kostadinova I (2021) Open science in practice: 300 published research ideas and outcomes illustrate how RIO Journal facilitates engagement with the research process. Research Ideas and Outcomes 7 https://doi.org/10.3897/rio.7.e68595
- Nisa C, Bélanger J, Schumpe B, Faller D (2019) Meta-analysis of randomised controlled trials testing behavioural interventions to promote household action on climate change. Nature Communications 10 (1). https://doi.org/10.1038/s41467-019-12457-2
- Norton B (2015) A Brief Philosophy of Adaptive Ecosystem Management. Sustainable Values, Sustainable Change. https://doi.org/10.7208/9780226197593-006
- Patel A, Meesters K, den Uil H, de Jong E, Blok K, Patel M (2012) Sustainability assessment of novel chemical processes at early stage: application to biobased processes. Energy & Environmental Science 5 (9). https://doi.org/10.1039/c2ee21581k
- Pinto D, McCallum A, Wei X, Croft WB (2003) Table extraction using conditional random fields. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03 https://doi.org/10.1145/860435.860479
- Rajan K, Zielesny A, Steinbeck C (2021) DECIMER 1.0: deep learning for chemical image recognition using transformers. Journal of Cheminformatics 13 (1). https://doi.org/10.1186/s13321-021-00538-8
- Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D (2021) Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digital Medicine 4 (1). https://doi.org/10.1038/s41746-021-00455-y
- Saul W, Roy H, Booy O, Carnevali L, Chen H, Genovesi P, Harrower C, Hulme P, Pagad S, Pergl J, Jeschke JM (2016) Assessing patterns in introduction pathways of alien species by linking major invasion data bases. Journal of Applied Ecology 54 (2): 657-669. https://doi.org/10.1111/1365-2664.12819

- Schweidtmann A, Rittig J, König A, Grohe M, Mitsos A, Dahmen M (2020) Graph Neural Networks for Prediction of Fuel Ignition Quality. Energy & Fuels 34 (9): 11395-11407. https://doi.org/10.1021/acs.energyfuels.0c01533
- Scott AM, Forbes C, Clark J, Carter M, Glasziou P, Munn Z (2021) Systematic review automation tool use by systematic reviewers, health technology assessors and clinical guideline developers: tools used, abandoned, and desired. medRxiv https://doi.org/10.1101/2021.04.26.21255833
- Seebens H, Blackburn T, Dyer E, Genovesi P, Hulme P, Jeschke J, Pagad S, Pyšek P, Winter M, Arianoutsou M, Bacher S, Blasius B, Brundu G, Capinha C, Celesti-Grapow L, Dawson W, Dullinger S, Fuentes N, Jäger H, Kartesz J, Kenis M, Kreft H, Kühn I, Lenzner B, Liebhold A, Mosena A, Moser D, Nishino M, Pearman D, Pergl J, Rabitsch W, Rojas-Sandoval J, Roques A, Rorke S, Rossinelli S, Roy H, Scalera R, Schindler S, Štajerová K, Tokarska-Guzik B, van Kleunen M, Walker K, Weigelt P, Yamanaka T, Essl F (2017) No saturation in the accumulation of alien species worldwide. Nature Communications 8 (1). https://doi.org/10.1038/ncomms14435
- Senderov V, Simov K, Franz N, Stoev P, Catapano T, Agosti D, Sautter G, Morris R, Penev L (2018) OpenBiodiv-O: ontology of the OpenBiodiv knowledge management system. Journal of Biomedical Semantics 9 (1). https://doi.org/10.1186/s13326-017-0174-5
- Spadaro G, Tiddi I, Columbus S, Jin S, Teije At, Balliet D (2020) The Cooperation Databank: Machine-Readable Science Accelerates Research Synthesis. Perspectives on Psychological Science https://doi.org/10.31234/osf.io/rveh3
- Turki H, Hadj Taieb MA, Ben Aouicha M, Fraumann G, Hauschke C, Heller L (2021) Enhancing Knowledge Graph Extraction and Validation From Scholarly Publications Using Bibliographic Metadata. Frontiers in Research Metrics and Analytics 6 https://doi.org/10.3389/frma.2021.694307
- Valderrama-Zurián JC, García-Zorita C, Marugán-Lázaro S, Sanz-Casado E (2021) Comparison of MeSH terms and KeyWords Plus terms for more accurate classification in medical research fields. A case study in cannabis research. Information Processing & Management 58 (5). https://doi.org/10.1016/j.ipm.2021.102658
- van Kleunen M, Pyšek P, Dawson W, Essl F, Kreft H, Pergl J, Weigelt P, Stein A, Dullinger S, König C, Lenzner B, Maurel N, Moser D, Seebens H, Kartesz J, Nishino M, Aleksanyan A, Ansong M, Antonova L, Barcelona J, Breckle S, Brundu G, Cabezas F, Cárdenas D, Cárdenas-Toro J, Castaño N, Chacón E, Chatelain C, Conn B, Sá Dechoum M, Dufour-Dror J, Ebel A, Figueiredo E, Fragman-Sapir O, Fuentes N, Groom Q, Henderson L, Inderjit, Jogan N, Krestov P, Kupriyanov A, Masciadri S, Meerman J, Morozova O, Nickrent D, Nowak A, Patzelt A, Pelser P, Shu W, Thomas J, Uludag A, Velayos M, Verkhosina A, Villaseñor J, Weber E, Wieringa J, Yazlık A, Zeddam A, Zykova E, Winter M (2018) The Global Naturalized Alien Flora (GloNAF) database. Ecology 100 (1). https://doi.org/10.1002/ecy.2542
- Weber J, Guo Z, Zhang C, Schweidtmann A, Lapkin A (2021) Chemical data intelligence for sustainable chemistry. Chemical Society Reviews 50 (21): 12013-12036. https://doi.org/10.1039/d1cs00477h
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S,

Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

- Xue K, Zhou Y, Ma Z, Ruan T, Zhang H, He P (2019) Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) https://doi.org/10.1109/bibm47256.2019.8983370
- Zhao S, Su C, Lu Z, Wang F (2020) Recent advances in biomedical literature mining. Briefings in Bioinformatics 22 (3). https://doi.org/10.1093/bib/bbaa057

Endnotes

*1 Semantic information, i.e. truthful, meaningful, well-formed data (see Floridi (2013))

*2 SKG4EOSC components are *italicised*.

*3 *ORKG* GraphQL endpoint integrating the DataCite PID Graph endpoint and *ORKG* orkg.org/orkg/graphql-federated

*4 e.g. www.wikidata.org/wiki/Wikidata:WikiProject_LGBT

*5 e.g. whgi.wmflabs.org/

*6 e.g. 10.3897/rio.7.e68513 10.3897/neobiota.68.66685, 10.3897/rio.7.e73858, 10.3897/rio.6.e52052, 10.3897/rio.6.e53921

*7 ec.europa.eu/eurostat/statistics-explained/index.php?title=R_%26_D_expenditure&oldid=503835

*8 www.rd-alliance.org/groups/fair-data-maturity-model-wg

*9 fairplus.github.io/the-fair-cookbook/content/home.html

*10 github.com/tkuhn/nanopub-server

*11 orkg.org/orkg/pdf-text-annotation

*12 nanonets.com/blog/table-extraction-deep-learning/

*13 commons.wikimedia.org/wiki/User:Open_Access_Media_Importer_Bot

*14 github.com/SDM-TIB/Trav-SHACL

*15 cooperationdatabank.org/data-stories/whats-in-the-databank/

*16 cfpub.epa.gov/si/si_public_record_Report.cfm?Lab=NHEERL&dirEntryID=113250