ORKG: Facilitating the Transfer of Research Results with the Open Research Knowledge Graph

Sören Auer‡, Markus Stocker‡, Lars Vogt‡, Grischa Fraumann‡, Alexandra Garatzogianni‡

‡ TIB Leibniz Information Center for Science and Technology, Hanover, Germany

Corresponding author: Grischa Fraumann (gfr@hum.ku.dk)

Abstract

This document is an edited version of the original funding proposal entitled 'ORKG: Facilitating the Transfer of Research Results with the Open Research Knowledge Graph' that was submitted to the European Research Council (ERC) Proof of Concept (PoC) Grant in September 2020 (https://erc.europa.eu/funding/proof-concept). The proposal was evaluated by five reviewers and has been placed after the evaluations on the reserve list. The main document of the original proposal did not contain an abstract.

Keywords

knowledge graph, technology transfer, open science, publishing industry

Section 1: The idea – Excellence in Innovation potential

1a. Brief description of the idea to be taken to proof of concept

1a.1 The problem

According to the UNESCO Institute of Statistics, we are spending almost US$1.7 trillion per year worldwide for acquiring new knowledge through research (UNESCO Institute of Statistics 2020). Currently, however, this is not a good investment and, each year, an ever-increasing share of this investment is wasted. The reason for this is that, for representing and sharing research findings, we use antique methods, which were developed many centuries ago. Since the beginning of modern science – with the publishing of the first scientific journals – the Journal des Sçavans and the Transactions of the Royal Philosophical Society in 1665 (Mack 2015, Spinak and Packer 2015) – we use the same methods for representing and sharing scholarly knowledge: scientific articles. At the time of the polymath Gottfried Wilhelm Leibniz in the 17th and 18th centuries, a single researcher could still read the entire published scientific literature.

Today, each year, 2.5 million new research articles are produced. Even in a relatively narrow scientific field, it is impossible to read, comprehend and make sense of all scientific articles. For example, publications from 1980 to 2012 show an exponential growth rate of 3% annually (Bornmann and Mutz 2015).

For the genome editing method CRISPR/Cas9, for example, the research search engine Google Scholar lists a quarter-million publications available as PDF articles. If a researcher is interested in how good the method is compared to other genome editing methods, what specifics it has when applied to insects and who has applied it to butterflies, a researcher needs either years of experience or is very likely not to find what he or she is looking for. Imagine that, to order a new iPhone, you had to compare prices by checking dozens of mail order catalogues published as PDF or, to navigate to a hotel, you would need to look at a PDF scan of a street map. This is exactly how the exchange of research findings works today – the previously analogue articles from scientific journals are now made available and distributed as PDF documents.

The new methods of the digital world, such as filtering large amounts of data and information, integrating information from different sources or involving users via crowdsourcing to review and help to organise the information, are non-existent in scholarly communication. Researchers are drowning in a flood of millions of pseudo-digitalised PDF publications. As a result, some research is seriously flawed: many research results cannot be reproduced by other researchers, peer-review struggles to cope with volume, speed and quality and we have more and more redundancy. Major social challenges, such as handling the COVID-19 pandemic and infodemic (WHO 2020) or implementing climate neutrality, require interdisciplinarity and putting bits and pieces from different disciplines together, which is currently extremely cumbersome and resource-intensive.

1a.2 The solution

In the ERC ScienceGRAPH project, we are researching and devising foundational concepts for organising scholarly communication in a knowledge-based way, leveraging a new formal model – cognitive knowledge graphs. According to this model, research contributions are represented in a human and machine-readable manner – the knowledge graph. As a result, completely new ways of machine assistance, such as semi-automatic generation of state-of-the-art overviews, visualisations or even question-answering applications become possible. To prepare the demonstration of the ScienceGRAPH results, the ERC project partner TIB Leibniz Information Center for Science and Technology (also directed by the ERC grant holder Sören Auer) started to develop the Open Research Knowledge Graph service, available at https://orkg.org. As an example, Fig. 1 shows a state-of-the-art comparison of different studies targeting the research question about the R_0 base infection rate of COVID-19.

Based on such a structured semantic and machine-readable representation, various other exploration and assistance tools are also possible, for example, a chart visualisation,

aggregating the results from the various studies. This example illustrates the solution to problems for various stakeholders:

- **Researchers** in the field (here epidemiologists and virologists) can get a quick overview of the state of the scholarly discourse related to a particular research question and determine gaps or how they can devise their approach to make their contributions stronger.
- **Peer-reviewers** can quickly assess the merits of a particular approach and view it in comparison to the current state-of-the-art.
- **Publishers** have a tool for assisting their editors, editorial managers, reviewers and authors to make contributions stronger and better positioned in the scientific discourse. In addition, publishers using such semantic descriptions and comparisons will dramatically increase the attraction of their journals.
- **Equipment and instrumentation manufacturers** can ensure that important configurations of materials used in research are documented and the use of their devices is properly acknowledged and visible.
- **Industrial and societal stakeholders** get faster and better access to the state-of-the-art and can, thus, more efficiently and effectively realise research-based products and services.

While some user groups will not pay directly for this solution (e.g. researchers and peer-reviewers) and the ORKG will be an open infrastructure in general, we see the potential for various value-added services for publishers, equipment and instrumentation manufacturers and other industrial and societal stakeholders.

To realise the potential, with this ERC PoC project, we aim to demonstrate some key results attained in the first two years within the ORKG.org proof-of-concept:

- Integrate the crowd- and expert-sourcing authoring and curation model for cognitive knowledge graphs, based on the knowledge graph cells concept (Vogt et al. 2020).
- Integrate persistent identifiers for scientific sensors and instruments to support the provenance and reproducibility of research results from experiment to publication.
- Develop approaches for generating comprehensive state-of-the-art overviews for a specific research question from the semantic knowledge graph representations of corresponding contributions.

1b. Demonstration of Innovation Potential

The ORKG is completely unique in its idea to describe scientific contributions in a knowledge graph. There are several other knowledge graph projects for scholarly communication also from commercial players, such as SciGraph from Springer Nature or the Microsoft Academic Graph. However, these initiatives solely focus on bibliographic

information and do not comprise a rich structured representation of the actual content of the publications. Other related initiatives are text-mining projects, such as SemanticScholar, which generate some relatively shallow semantic descriptions automatically. However, due to the low precision and recall of text mining methods (in particular for relation extraction), this does not go beyond relatively simple classifications, annotations and summarisation of the content and, thus, does not suffice for creating a comprehensive knowledge graph representation and exploration services, such as comparisons, visualisations, question answering, etc.

Section 2: The Expected Impact

2a. Identification and description of any effect or benefit to the economy, society, culture, public policy/services.

The results of this ERC PoC project can have a dramatic impact on the effectiveness and efficiency of research and how research results are transferred into applications. We expect that research will be at least 10-15% more efficient with corresponding positive effects on the effectiveness of the annual research spending of almost US$1.7 trillion worldwide. Especially the scholarly publishing industry, with an annual US$10 billion market (Research and Markets 2020), would significantly benefit from the results of this project. In the following, we describe the impact on the research instrumentation industry in more detail.

Sensors and scientific instruments are important in the research cycle for several academic disciplines. Sensors, for example, are used for permanent measurements in agriculture and scientific instruments are used in laboratories to carry out scientific measuring. There is a need to develop persistent identifiers (PIDs) for sensors and scientific instruments and several initiatives are working towards that goal. The Digital Object Identifier (DOI) is a common example of a PID widely used for publications and research datasets and further identifiers are, for example, handles. Sensor platforms in agriculture have assigned PIDs and there is widespread use in the scientific community, but scientific instruments are usually not citable in publications. The proposed ORKG PoC will generate several benefits for the economy. There is a need to introduce the project outcomes of the ERC-funded ScienceGRAPH in the market of sensors and scientific instruments. Manufacturers for scientific instruments operate in a global market and the 20 top companies in 2018, according to the value of instrument sales, are based in the US (8), Europe (7) and Japan (5). The top five companies in 2018 included Thermo Fisher Scientific, Shimadzu, Roche Diagnostics, Agilent Technologies and Danaher (Chemical & Engineering News (c&en) 2019). The global market for scientific instruments is estimated at US$60 billion in 2020 and is expected to grow to US$79.9 billion by 2025 (Markets & Markets 2020). The aim is to develop use cases as part of the PoC and initiate an innovation process which focuses on close collaboration with industry partners. The R&D team at TIB has close ties to several important players in the market, such as LI-COR Biosciences, Zeiss and Leica. LI-

COR Biosciences mainly focuses on sensors in the agriculture market and instruments for research purposes are part of the product portfolio of the company.

While persistent identifiers for research datasets are increasingly used in research and are citable in academic publications (Robinson-Garcia et al. 2017), PIDs for scientific instruments are a more recent development (Stocker et al. 2020). The citation of instruments in publications that were used to carry out the research (e.g. measuring) would contribute to more transparent communication of research results. Some exceptions are already mentioned in publications, such as electron microscopes or particle accelerators. Instrument citation could be achieved by extending the DataCite schema that is currently being used for research data, amongst others. This extension could include, for example, the model number of instruments, date of purchase, use in a research project, maintenance of instruments and the calibration of instruments. The business office of DataCite is located at TIB and the R&D team has already held discussions on this topic. If the DOI suffix of a publication is extended by mentioning the related scientific instrument, this would provide several advantages. Scientific instruments could be initially registered by the manufacturer, which would require a new membership to register DOIs via DataCite.

The PoC would build on the basic research that is being carried out as part of the ERC-funded ScienceGRAPH project, but would provide an automatic connection to the Open Research Knowledge Graph (ORKG) that is also operated at TIB and focuses on applied R&D. Furthermore, we will prepare a use case in the Integrated Carbon Observation System (ICOS) research infrastructure in collaboration with LI-COR Biosciences. Further use cases would include more academic disciplines, such as engineering at Leibniz University Hanover (LUH) and life sciences at Hanover Medical School (MHH). There is already a well-established collaboration with Collaborative Research Centres (SFBs funded by the German Research Foundation or DFG), such as the SFB “Tailored Forming” at the Hanover Centre for Production Technology as part of LUH.

Structured machine-readable data will provide a competitive advantage for our industry partners since instruments, registered with a PID, will have an advantage over those from other companies. Apart from manufacturers of sensors and scientific instruments, the PoC will generate benefits for academic publishers, researchers and research infrastructures. Academic knowledge is generated at different points in time and not only while publications are being written by researchers. As such, saved metadata from instruments would make these efforts visible. Potential reuse in further follow-up projects might be applied in laboratory information management systems (LIMS). This could be done, for example, in collaboration with the Julius Kühn Institute, a federal research centre for cultivated plants in Germany, which already collaborates closely with the R&D team at TIB through other projects. Furthermore, TIB established contacts with the software engineering company Limsophy LIMS. The outcome of the PoC will be a prototype with TRL 7 that can be further developed by the industry partner in collaboration with researchers.

Apart from economic benefits, the ORKG PoC will also generate benefits for society. The coronavirus pandemic demonstrated once again that there is a need for transparent measurements of scientific results. The proposed project will enable FAIR (findable,

accessible, interoperable and reusable) research information and research data for several stakeholders. The reproducibility crisis fuels an ongoing debate in research and research policy (Fanelli 2018). Furthermore, this also relates to issues with replicability and several projects try to tackle this challenge (Whole Tale 2020). The project outcomes will reduce challenges of reproducibility and replicability in certain academic disciplines. What is more, sensors are strongly promoted in public policy and services, for example, with regard to digitising European industry and advancing the Internet of Things (IoT). As such, they also contribute to building a Digital Single Market, one of the key priorities of the European Commission (European Commission 2018).

2b. Outline of the value creation process

To maximise the societal benefit from the results of the ERC and this PoC project, the core ORKG service will be an open infrastructure following the Open Science, Open Access and Open Source principles. This also enables rigorous and large-scale testing and evaluation of the outcomes of the project with real user communities. TIB is prepared to sponsor and further develop, maintain and operate the ORKG service in the long term. In addition to the open strategy, we envision various commercialisation opportunities including:

- **Providing value-added services tailored for commercial scientific publishers**, such as Springer Nature, Wiley and IEEE Publishing.
- **Providing commercial data, analytics and question answering services** for speeding up the spread and transfer of research results in industrial applications.
- **Partnering with industrial stakeholders, in particular scientific instrument manufacturers,** regarding sponsoring of the ORKG and integration of their instrument descriptions.

TIB has long-term established R&D collaborations and customer relationships with small and large industrial stakeholders. TIB already provides commercial literature access services to > 100 customers and aims to expand this to the research analytics services offered on the ORKG service infrastructure.

Section 3: The proof of concept plan

3a. Project-management plan including risk and contingency measures

3a.1 Organisational structure and decision-making process

Since the ORKG project is relatively focused, we envision a lean organisational structure depicted in Fig. 2.

In addition to the PI and the ORKG development lead, the organisational structure will involve leads of the three ORKG work packages, an industrial advisory board as well as an ORKG community board.

The industrial advisory board will advise the project team in matters related to the commercialisation of the results, such as product features, product and service offerings, IPR, pricing, as well as legal matters. We will organise quarterly meetings of the board. We have been in touch with several industry representatives about joining the board.

The ORKG Community board will advise the development team with regard to community requirements and comprise experts from various research fields, research data infrastructures and open-access publishers. We plan to organise quarterly webinars or workshops with the community advisory board (possibly in conjunction with larger scholarly communication events).

Decision-making and development methodology. The size of the project allows it to follow a lean focused decision-making process, where most of the decisions are made in the regular weekly ORKG project meeting by involving the whole team. For all developments, we follow the agile KANBAN-inspired development methodology aiming at establishing a constant active development process by optimising the issue burn rate and establishing a proactive communication culture.

3a.2 Plan for the identification and acceptance or off-setting of possible risks

We aim at identifying, evaluating and eliminating or minimising potential risks that may jeopardise the success of the project. While some relevant project risks and how to address them are already identified, risk management will be conducted throughout the project. It is a continuous process in which known risks will be regularly reviewed and new risks will need to be recognised to handle and control them adequately. Their assessment will lead to the formulation of appropriate mitigation measures that should help to prevent and overcome a risk or reduce its effects to an acceptable level. The process behind risk management can be broken down as follows:

1. Risk identification (i.e. recognise and describe risks).
2. Risk analysis (i.e. analyse likelihood and consequences of risks).
3. Risk assessment (i.e. determine magnitude/acceptability of risks for the project).
4. Risk response planning (i.e. create and execute an action plan to prevent or minimise risks).
5. Risk control (i.e. monitor, track and review risks and mitigation actions).

3a.3 Plan for unforeseen non-scientific events

Table 1 contains some examples of risks and corresponding mitigation strategies that we already identified.

3b. Description of the team

3b.1 Team, achievements and experience

The team is led by ScienceGRAPH PI Prof. Dr. Sören Auer. He is supported by the ORKG project head Dr. Markus Stocker, who has been leading related research and development activities for almost two years. In addition, a seasoned team is already established, including experienced PostDoc researchers (e.g. Dr. Jennifer D'Souza and Dr. Lars Vogt), more than five PhD students, software developers (Manuel Prinz and Kheir Eddine Farfar) and business and technology transfer experts (especially in the TIB departments), which can be dynamically involved in the project as required.

Sören Auer. Following positions at the Universities of Dresden, Ekaterinburg, Leipzig, Pennsylvania, Bonn and the Fraunhofer Society, Prof. Auer was appointed Professor of Data Science and Digital Libraries at Leibniz Universität Hanover and Director of the TIB in 2017. Prof. Auer has made important contributions to semantic technologies, knowledge engineering and information systems. He is the author (resp. co-author) of over 200 peer-reviewed scientific publications. He has received several awards, including an ERC Consolidator Grant from the European Research Council, a SWSA ten-year award, the ESWC 7-year Best Paper Award and the OpenCourseware Innovation Award. He has led several large collaborative research projects, such as the EU H2020 flagship project BigDataEurope. He is co-founder of high potential research and community projects, such as the Wikipedia semantification project DBpedia, the OpenCourseWare authoring platform SlideWiki.org and the innovative technology start-up, eccenca.com (now employing more than 40 people). Prof. Auer was founding director of the Big Data Value Association and led the semantic data representation in the International Data Space.

Dr. Markus Stocker is head of the Knowledge Infrastructures research group at TIB. Markus holds a PhD in Environmental Informatics from the University of Eastern Finland, an M.Sc. in Environmental Science from the University of Eastern Finland and a Diploma (M.Sc.) in Informatics from the University of Zurich, Switzerland. He is author of 40 peer-reviewed journal and conference proceedings papers, with more than 1000 citations. He has managed partner contributions and been involved in various H2020 projects, including THOR, ENVRIplus, OpenAIRE, FREYA, ENVRI-FAIR, as well as nationally-funded projects in Finland and Germany. Prior to TIB, Markus held a postdoctoral research associate position at PANGAEA, the Data Publisher for Earth & Environmental Science, at the MARUM Center for Marine Environmental Sciences, University of Bremen, Germany. As a member of the Research Data Alliance (RDA), Markus is involved in various groups, in particular the WG Persistent Identification of Instruments and the IG From Observational

Data to Information. He has several years of professional experience in software development and semantic technologies, with positions at Hewlett Packard Labs, Bristol, UK and Clark & Parsia, Washington DC, USA.

Alexandra Garatzogianni is the Head of the Knowledge and Technology Transfer Department at TIB, leading a diverse and inclusive team, which, besides providing impulses for innovation and developing future technologies, offers comprehensive consulting, research and support in order to enable sustainable access to the market for research output. She is the Coordinator of the H2020 project TRUSTS Trusted Secure Data Sharing Space, of the H2020 MediaFutures, Data-driven innovation hub for the media value chain and WP leader for the H2020 project PLATOON (Digital PLAtform and analytics TOOls for eNergy). She leads the project management of the Leibniz Joint Lab Data Science & Open Knowledge amongst TIB, the Leibniz University of Hanover (LUH) and the L3S Research Center, which serves as a nucleus for further initiatives in the field of research and innovation. She co-founded the IDSA Competence Center at the Leibniz Joint Lab Data Science & Open Knowledge in June 2019 and received the BDVA iSpaces award on behalf of the L3S Research Center (November 2019), which signifies that L3S is a Trusted Data Incubator targeted to accelerate take-up of data-driven innovation in commercial sectors.

3b.2 Roles of the team and main strengths and weaknesses

The role of the PI Prof. Dr. Sören Auer is to develop and communicate the strategic vision of the project and to devise the key development milestones and priorities. He will advise and mentor the PhD students and PostDocs on the project and work closely with the ORKG development lead Dr. Markus Stocker. A further focus of the PI is to build strategic partnerships, attract further funding, sponsoring or investments. The ORKG project development head, Dr. Markus Stocker, will lead the day-to-day operations and developments of the project. He will lead the regular KANBAN sessions together with the development deputy Manuel Prinz and guide the research and development along with the community and advisory board defined requirements and strategic priorities. Alexandra Garatzogianni will lead the business development strategy and contribute to building and maintaining sustainable sponsor, partner and customer relationships for the ORKG service ecosystem throughout and beyond the project's duration. The main strengths and weaknesses of the team include the following:

Key strengths

- Successful track record of translating research excellence into large scale applications including successful commercialisation in a spin-off.
- A long history of industrial collaborations.
- ORKG innovation concept with an enormous value potential.

- A seasoned team including a variety of backgrounds and skills: experienced PostDoc researchers, PhD students, software developers and business experts, who can be dynamically involved in the project.

Weaknesses

- Limited resources compared to commercial entities (e.g. commercial publishers).
- Community and industrial buy-in just starting to develop.
- More advocacy, policy backing for the transition/digitisation in scholarly communication required.
- Initially limited possibilities for automation using AI and machine learning due to the lack of training data.

3c. Plan of the Proof of Concept – Action description

Objectives:

The overall objectives of the ORKG project are:

- Mature the existing ORKG service prototype, establish interoperability with publishing platforms, prototype services for research result exploitation and devise possible business models.
- Integrate support for persistent identifiers and semantic descriptions for scientific sensors and instruments and evaluate the integration with concrete research infrastructures and vendors.
- Enable FAIR semantic descriptions and the generation of SOTA Surveys for automatically generating survey and review publications from the ORKG infrastructure.

Description of work:

Table 2 summarises the tasks and corresponding resources planned in the three work packages.

Allocation of resources: The lump sum will be primarily used to fund the personal resources of the team. There are some further minor cost items, such as travel or minor equipment expenses, which will also be financed from TIB directly.

WP1 ORKG Service Maturation and Business Model Development

The goal of this work package is to mature the ORKG service by integrating two functions particularly important for the exploitation of the results. This includes: 1) the establishment of interoperability interfaces with traditional journal and proceedings publishing platforms of commercial publishers and 2) the prototyping of services for research exploitation and

transfer analytics, based on the current ORKG knowledge graph infrastructure. Finally, we will work on the business development by outlining commercial offering options with the corresponding market and pricing analysis.

T1.1 Interoperability with traditional scholarly publishing platforms

Traditional commercial scientific publishing platforms organise the submission, peer-review and publication process of scientific articles (e.g. in platforms, such as Clarivate's ScholarOne Manuscripts). Each of these three steps is highly relevant regarding integration with the ORKG:

1. In the submission process, authors can be encouraged to create an ORKG representation of their key contributions, thus facilitating the comparability of the state-of-the-art.
2. Peer-reviewers can subsequently use such comparisons, visualisations and further aggregated views to assess the merits of the scientific contribution.
3. After publishing an article, the semantic representation in the ORKG along with additional comparisons, explorations and visualisations will provide further context and insights to the readers of the published article. We will provide a REST API integration interface, where small user interface widgets can be directly integrated with minimal efforts into the respective publishing management systems.

Result: Integration interface for embedding UI widgets directly into publishing management systems.

T1.2 Services for research exploitation and transfer analytics

Based on the structured semantic representations in the ORKG, completely new analytical services for the exploitation of research results become possible. In this task, we will prototype such services, which can be a key pillar for commercial exploitation via an attractive service for the research, innovation and product development departments in enterprises. For example, for a particular research problem, the most promising approaches addressing this problem with regard to certain framework conditions can be identified. In addition, the impact and consequences of following particular approaches can be compared and analysed.

Result: Prototypical research exploitation and transfer analytics services.

T1.3 Business Model Development

In this task, we will develop a portfolio of possible business models, based on the ORKG services developed in this PoC project. For each of the possible service offerings, we will analyse the competition, market, competitive advantage, customer profiles, pricing options along the business model canvas paradigm. We will also compile a list of possible options for further funding and investment to advance the ORKG service to the next commercialisation and exploitation level. Aspects, such as impact assessment, exploitation, sustainability roadmap and implementation, will be appropriately researched

and implemented, ensuring thus the successful and sustainable uptake of the project's output.

Result: Prioritised list of business model options organised along the business model canvas paradigm.

WP 2 Persistent Identifiers for Scientific Sensors and Instruments

Instruments play an essential role in creating research data. Given the importance of instruments and associated metadata for the assessment of data quality and data reuse, globally unique, persistent and resolvable identification of instruments is crucial. The Research Data Alliance Working Group Persistent Identification of Instruments (PIDINST), chaired by Dr. Markus Stocker, developed a community-driven solution for persistent identification of instruments (Stocker et al. 2020). Based on an analysis of 10 use cases, PIDINST developed a metadata schema and prototyped schema implementation with DataCite and ePIC as representative persistent identifier infrastructures and with Helmholtz-Zentrum Berlin für Materialien und Energie (HZB) and British Oceanographic Data Centre (BODC) as representative institutional instrument providers.

In this work package, we plan to implement and integrate the concept for persistent identification and semantic description of sensors and instruments into the ORKG service infrastructure project, thus greatly facilitating reproducibility and reusability of research results.

T2.1 Integration of persistent identification and description of sensors and instruments into the ORKG

In this task, we will integrate key functionality for the persistent identification and semantic description of scientific instruments into the ORKG infrastructure. This will involve the integration of the PIDINST metadata schema, the creation and alignment of identifiers, the management of revisions, provenance tracking and the integration of interfaces for automatic import and alignment with vendor-supplied instrument and equipment descriptions. For the latter, we envision a JSON-LD REST interface, which will enable vendors to directly represent and upload their descriptions according to the PIDINST schema.

Result: Comprehensive representation and integration of scientific instrumentation in the ORKG.

T2.2 Evaluation with concrete research infrastructure providers and equipment vendors

In this task, we will work with concrete research infrastructure providers and equipment vendors on testing and evaluating the integration developed in T2.1 and creating demonstrations and showcases for attracting further research infrastructure providers and scientific instrumentation equipment vendors. We already identified a shortlist of infrastructures, such as ICOS, Leibniz DSMZ or the virology labs at TWINCORE and Hanover Medical School (MHH). Concerning instrument vendors, we have close ties to

important players in the market, such as LI-COR Biosciences, Zeiss and Leica. In addition, we plan to outreach to Thermo Fisher Scientific, Shimadzu, Roche Diagnostics, Agilent Technologies and Danaher to scale the number of showcases and integrations.

Result: Comprehensive portfolio of research infrastructure and scientific instrument showcase integrations.

WP 3 FAIR Semantic Descriptions of Research Quests, Contributions and SOTA Surveys

The goal of this WP is to organise scholarly communication in a structured knowledge graph-based manner. We will, thus, go beyond static PDF publications and make research problems, approaches, algorithms, implementations and evaluations FAIR and first-class citizens of the scholarly discourse.

Science typically involves the definition of research problems or questions and corresponding research approaches contributing to solving these problems or questions. Examples of research problems or questions are Named Entity Recognition, Question Answering, Machine Translation, Image Recognition or Data Clustering. Contributions addressing these problems are typically following a particular approach and are evaluated using some benchmark dataset. Currently, all this information is deeply hidden in unstructured articles, often published as PDFs. In this measure, we will make research problems, questions, contributions and their description first-class citizens of the scholarly Data Science communication. We will build on the already established Open Research Knowledge Graph (ORKG) platform (https://www.orkg.org) and expand it in three yearly iterations with crucial functionality for data science and AI research. Subsequently, we will further evaluate, broaden the applications and scale the use of the platform.

Task 3.1 Development of templates for semantic descriptions of science contributions

In this task, we will develop a comprehensive library of semantic templates for research question and contribution descriptions. The templates will be represented in a formal way (e.g. according to the W3C SHACL standard) and, thus, facilitate interoperability between various services. In particular, we will demonstrate the applicability of the templates with the Open Research Knowledge Graph, which provides an environment for authoring, organising and curating semantic research question and contribution descriptions. We will also integrate techniques to automatically extract and represent information from articles according to the templates.

Result: Library of semantic templates for research question and contribution descriptions.

Task 3.2 SOTA Comparisons and Leaderboards

We will use the semantic descriptions of data science and AI approaches and publications to generate comparative overviews and leaderboards on the approaches addressing a particular research question or problem. The approach for generating such SOTA overviews will be highly automated, but enabled to be configured and fine-tuned by users.

We will integrate functionality to publish (using DOIs), integrate and link such comparative overviews directly from traditional publications (e.g. via LaTeX/BibTeX or Word export). Leaderboards will give a comprehensive overview on the evolution of the SOTA over time with regard to concrete performance indicators (e.g. precision/recall) attained on community-defined benchmarks.

Result: Automatic comparison and leaderboard generation with a focus on the SOTA evolution.

Task 3.3 Authoring environment for cognitive knowledge-graph-based surveys and reviews

In this task, we will integrate the service elements and functionalities developed in other tasks of this measure into a comprehensive environment for creating structured SOTA survey articles for specific Data Science and AI research questions. The structured elements will comprise a motivation of the research problem, its definition, a classification taxonomy and qualitative (functional) and quantitative approach characterisations, as well as problem-specific visualisations and leaderboards. The survey article will be compiled automatically and directly from the structured semantic knowledge graph representations, but represented as a self-contained article publishable as a Web resource (or PDF). We will assign DOIs and enable the publication of these surveys in traditional publications outlets, such as journals and OA repositories.

Result: Publishing environment for structured surveys and reviews with integration with traditional publishing outlets.

Acknowledgements

We would like to thank Gino Erkeling for supporting this grant proposal.

Funding program

This European Research Council (ERC) Proof of Concept (PoC) Grant proposal is based on the ERC Consolidator Grant ScienceGRAPH (Grant agreement ID: 819536) and co-funded by TIB Leibniz Information Centre for Science and Technology.

Grant title

ORKG: Facilitating the Transfer of Research Results with the Open Research Knowledge Graph

Hosting institution

TIB Leibniz Information Centre for Science and Technology

Ethics and security

Not applicable.

Author contributions

The authors contributed equally to this grant proposal.

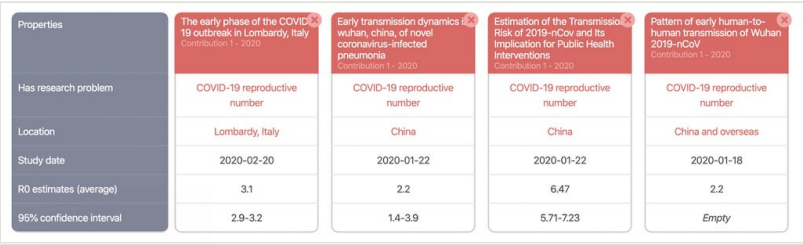
Conflicts of interest

The authors declare no conflict of interest.

References

- Bornmann L, Mutz R (2015) Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology *66* (11): 2215-2222. https://doi.org/10.1002/asi.23329
- Chemical & Engineering News (c&en) (2019) *Top instrument firms of 2018*. https://cen.acs.org/business/instrumentation/Top-Instrument-Firms-2018/97/i9. Accessed on: 2020-9-14.
- European Commission (2018) Digitising European Industry. https://ec.europa.eu/digital-single-market/en/policies/digitising-european-industry. Accessed on: 2020-9-14.
- Fanelli D (2018) Opinion: Is science really facing a reproducibility crisis, and do we need it to? Proceedings of the National Academy of Sciences 115 (11): 2628-2631. https://doi.org/10.1073/pnas.1708272114
- Mack C (2015) 350 Years of Scientific Journals. Journal of Micro/Nanolithography, MEMS, and MOEMS 14 (1). https://doi.org/10.1117/1.JMM.14.1.010101
- Markets & Markets (2020) Life science instrumentation market by technology. https://www.marketsandmarkets.com/Market-Reports/life-science-chemical-biotech-instrumentation-market-38.html. Accessed on: 2020-9-14.
- Research and Markets (2020) Global $10B Scientific & Technical Publishing Industry Report, 2019-2023. https://www.globenewswire.com/news-release/2020/01/29/1976933/0/en/Global-10B-Scientific-Technical-Publishing-Industry-Report-2019-2023.html. Accessed on: 2020-9-14.
- Robinson-Garcia N, Mongeon P, Jeng W, Costas R (2017) DataCite as a novel bibliometric source: Coverage, strengths and limitations. Journal of Informetrics 11 (3): 841-854. https://doi.org/10.1016/j.joi.2017.07.003
- Spinak E, Packer AL (2015) 350 years of scientific publication: from the “Journal des Sçavans” and Philosophical Transactions to SciELO. https://blog.scielo.org/en/2015/03/05/350-years-of-scientific-publication-from-the-journal-des-scavans-and-philosophical-transactions-to-scielo/#.YH1A8T_RY2w. Accessed on: 2021-4-21.

- Stocker M, Darroch L, Krahl R, Habermann T, Devaraju A, Schwardmann U, D'Onofrio C, Häggström I (2020) Persistent Identification of Instruments. Data Science Journal 19 (18). https://doi.org/10.5334/dsj-2020-018
- UNESCO Institute of Statistics (2020) How much does your country invest in R&D? http://uis.unesco.org/apps/visualisations/research-and-development-spending/. Accessed on: 2020-9-14.
- Vogt L, D'Souza J, Stocker M, Auer S (2020) Toward Representing Research Contributions in Scholarly Knowledge Graphs Using Knowledge Graph Cells. In: Huang R, Wu D, Marchionini G (Eds) Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual event, China, August 2020. https://doi.org/10.1145/3383583.3398530
- WHO (2020) Infodemic Management - Infodemiology. https://www.who.int/teams/risk-communication/infodemic-management. Accessed on: 2020-9-14.
- Whole Tale (2020) What is Whole Tale? https://wholetale.org/index.html#what-wt. Accessed on: 2020-9-14.



Properties	The early phase of the COVID-19 outbreak in Lombardy, Italy Contribution 1 - 2020	Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia Contribution 1 - 2020	Estimation of the Transmission Risk of 2019-nCov and Its Implication for Public Health Interventions Contribution 1 - 2020	Pattern of early human-to-human transmission of Wuhan 2019-nCoV Contribution 1 - 2020
Has research problem	COVID-19 reproductive number	COVID-19 reproductive number	COVID-19 reproductive number	COVID-19 reproductive number
Location	Lombardy, Italy	China	China	China and overseas
Study date	2020-02-20	2020-01-22	2020-01-22	2020-01-18
R0 estimates (average)	3.1	2.2	6.47	2.2
95% confidence interval	2.9-3.2	1.4-3.9	5.71-7.23	*Empty*

Figure 1.

State-of-the-art comparison of different studies targeting the research question about the R_0 base infection rate of COVID-19.

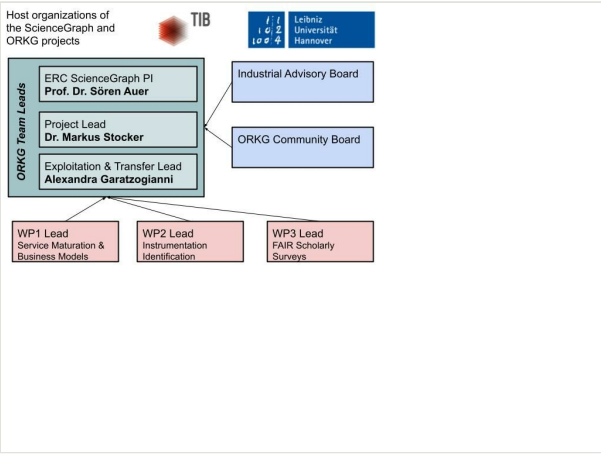


Figure 2.

Organisational structure and decision-making process.

Table 1.

Plan for unforeseen non-scientific events.

Description of the risk	Proposed risk-mitigation measures
Entrance of new competitors	We aim to gain as much competitive advantage as possible and to increase user/customer fidelity by open science infrastructure. In addition, we aim to build an open interoperable ORKG service ecosystem.
Lack of qualified personnel	As a research institute with a close connection to a university department, we have direct access to skilled master graduates. In addition, we have built an international reputation making us an attractive target for qualified international candidates.
Lack of user and customer adoption	We align the development process as closely as possible with user/customer requirements and, thus, aim to maximise adoption success. In addition, we follow an iterative development process with regular intermediate evaluations and community building.
Leaving of a key person	Already now, the ScienceGRAPH/ORKG team divides the work on several individuals, thus reducing the dependency on a single person. In addition, the skills to perform key activities are aimed to be made available by at least two people.
Lack of funding and investors	The ORKG Service is of strategic interest to TIB and even in the absence of further external funding, TIB is committed to sponsoring ORKG. In addition, we will actively work on attracting further sponsors, create awareness in politics for the open infrastructure and build a sustainable business model on top of the ORKG, based on value-added services.

Table 2.

Description of work.

Work Packages / Tasks	**Resources**
WP1 ORKG Service Maturation and Business Model Development	*8 PM*
T1.1 Interoperability with traditional publishing platforms	2 PM
T1.2 Services for research exploitation and transfer analytics	4 PM
T1.3 Business Model Development	2 PM
WP2 Persistent Identifiers for Scientific Sensors and Instruments	*8 PM*
T2.1 Integration of persistent identification and semantic description of sensors and instruments into the ORKG	4 PM
T2.2 Evaluation with concrete research infrastructures and equipment vendors	4 PM
WP3 FAIR Semantic Descriptions of Research Quests, Contributions and SOTA Surveys	*9 PM*
T3.1 Development of templates for semantic descriptions of science contributions	2 PM
T3.2 SOTA Comparisons and Leaderboards	3 PM
T3.3 Authoring environment for cognitive knowledge-graph-based surveys and reviews	4 PM
SUM	**25 PM**