TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

Towards an Open Research Knowledge Graph

Executive Summary:
The document-oriented workflows in science have reached (or already exceeded) the limits of adequacy as highlighted for example by recent discussions on the increasing proliferation of scientific literature and the reproducibility crisis. Despite an improved and digital access to scientific publications in the last decades, the exchange of scholarly knowledge continues to be primarily document-based: Researchers produce essays and articles that are made available in online and offline publication media as roughly granular text documents. With current developments in areas such as knowledge representation, semantic search, human-machine interaction, natural language processing, and artificial intelligence, it is possible to completely rethink this dominant paradigm of document-centered knowledge exchange and transform it into knowledge-based information flows by representing and expressing knowledge through semantically rich, interlinked knowledge graphs.

The core of the establishment of knowledge-based information flows is the distributed, decentralized, collaborative creation and evolution of information models, vocabularies, ontologies, and knowledge graphs for the establishment of a common understanding of data and information between the various stakeholders as well as the integration of these technologies into the infrastructure and processes of search and knowledge exchange in the research library of the future. By integrating these information models into existing and new research infrastructure services, the information structures that are currently still implicit and deeply hidden in documents can be made explicit and directly usable. This revolutionizes scientific work because information and research results can be seamlessly interlinked with each other and better mapped to complex information needs. As a result, scientific work becomes more effective and efficient, since results become directly comparable and easier to reuse.

In order to realize the vision of knowledge-based information flows in scholarly communication, comprehensive long-term technological infrastructure development and accompanying research are required. To secure information sovereignty, it is also of paramount importance to science – and urgency to science policymakers – that scientific infrastructures establish an open counterweight to emerging commercial developments in this area. The aim of this position paper is to facilitate the discussion on requirements, design decisions and a minimum viable product for an Open Research Knowledge Graph infrastructure. TIB aims to start developing this infrastructure in an open collaboration with interested partner organizations and individuals.

Author

Sören Auer (soeren.auer@tib.eu)

Contributors

Ina Blümel, Ralph Ewerth, Alexandra Garatzogianni, Lambert Heller, Anett Hoppe, Anna Kasprzik, Oliver Koepler, Wolfgang Nejdl, Margret Plank, Irina Sens, Markus Stocker, Marco Tullney, Maria-Esther Vidal, Wilma van Wezenbeek

Publisher

Technische Informationsbibliothek, Welfengarten 1B, 30167 Hannover, www.tib.eu

DOI: 10.5281/zenodo.1157185

From Document- to Knowledge-based Methods for Scholarly Communication

Despite substantial changes in recent decades, the exchange of knowledge continues to be document-based: Researchers produce essays and articles that are made available in online and offline publication media as coarse granular documents. The entire library, technology, service and research landscape is currently geared towards this fundamental approach. This is justified if specific questions can be answered by a single article. If this is not the case, users are hardly supported adequately by the existing infrastructure, but in the best case they will receive a large, disordered amount of more or less relevant documents (information overload) and in the worst case not even that. We are currently observing that the interlinking and processing of information from *various* different publications is becoming increasingly important for scientific work.

With current developments in areas such as knowledge representation, semantic search, human-machine interaction, natural language processing, and artificial intelligence, it is possible to completely rethink this dominant paradigm of document-centered knowledge exchange and transform scholarly communication into knowledge-based information flows by expressing and representing knowledge as structured, interlinked and semantically rich knowledge graphs, which facilitate a whole range of novel exploration, discovery, search and retrieval applications.

Establishing knowledge-based information flows relies on distributed, decentralized, collaborative creation and development of information models, vocabularies, ontologies and knowledge graphs. As communication tools, these will allow to establish a common understanding of data and information between the various stakeholders as well as the integration of technologies in the infrastructures and processes of the research library of the future. By integrating these information models into existing and new research infrastructure services[1], the information structures that are currently still implicit and deeply hidden in documents can be made explicit and directly usable. This revolutionizes scientific work because information and research results can be networked with each other and better linked to complex information needs. As a result, scientific work will become more efficient and effective, results become directly comparable and easier to reuse.

Processes in libraries and memory institutions, in research institutes, universities and educational institutions as well as in research departments of companies and enterprises in general are currently focused on document-based information and knowledge exchange and will have to transform in the coming years. As a central library and information centre for science and technology, the TIB is well positioned to accompany and actively promote this transformation process. Through networking with the institutes of the Leibniz Association, there is a critical mass of application domains and users in order to implement knowledge-based information exchange and to provide corresponding research infrastructures.

Problems of Document-centric information flows

The document-oriented workflows in science have reached (or already exceeded) the limit of suitability. Some examples of the inadequacy of document-based information flows are presented below.

Proliferation of scientific publications. In the last 10 years, the scientific output in the form of published articles has almost doubled[2]. This development is expected to continue as more countries join the international research community (e.g. China, Russia, India, South America). This plethora of

[1] TIB for example is providing or involved in the research infrastructure services TIB bibliography portal, TIB AV-Portal, VIVO and KDF, SlideWiki, ORCID, RADAR, DataCite and others

[2] National Science Foundation: *Science and Engineering Publication Output Trends*. https://www.nsf.gov/statistics/2018/nsf18300/nsf18300.pdf

scientific literature makes it increasingly difficult to keep an overview of the current state of research. As a result, scientists spend a large part of their time reviewing literature, presenting their own research in document form and, in many cases, working independently on very similar research results due to the increasing lack of transparency.

High effort for creating and reading articles. The creation, reading and processing of scientific literature is currently tying up an extremely high cognitive capacity. When scientific publications are created, a lot of redundancy and duplication occurs, since, for example, preliminaries or related works in articles on a topic are repeated over and over again in slightly modified form.

Very limited machine support during processing and searches. Scientific articles are very hard to understand for machines. Although the characters, words, sentences can be indexed and searched, the structure and semantics of text, illustrations, references, symbols, etc. are either currently not accessible to computers or in a very limited way. As a result, modern exploration, retrieval, question answering and visualization interfaces are not applicable to scientific articles, which further hinders the processing of the abundance of daily published scientific literature.

Lack of globally unambiguous identification of concepts in scientific articles. While there are now globally unambiguous systems of persistent identifiers for documents and datasets (DOI) and authors (ORCID), there are no comparable universal identifiers when it comes to terminologies, definitions and concepts for specific subject areas. The referencing is therefore rather granular and is located on the basis of entire publications instead of specific definitions, statements, experiments, etc.

Many frictional losses due to media discontinuities, ambiguity, lack of comparability. Due to the lack of structuring research results, they can often only be compared with much effort. Furthermore, the various artefacts of scientific work (data, publications, software, simulations, models, etc.) are insufficiently linked to each other and are not provided in open and standardized machine-understandable formats, making it difficult (or often impossible) to reproduce them. The FAIR principles[3] are a step in the right direction here, but they are currently too vaguely defined from a technical viewpoint and insufficiently build on existing best practices, such as W3Cs Data on the Web[4].

Existing related initiatives and development

Information exchange becomes increasingly semantic and structured in a number of areas:

Knowledge graphs for encyclopedic knowledge. About a decade ago, DBpedia, Yago and Freebase created the first knowledge graphs for the representation of encyclopedic knowledge, which have evolved considerably in recent years and now provide billions of facts on encyclopedic knowledge in a variety of domains. After DBpedia and Yago showed the value of encyclopedic knowledge graphs, Wikipedia's WikiData now provides a community-curated knowledge graph for Wikipedia, which is used to establish the structured knowledge in Wikipedia in a language-independent and quality-assured way.

Commercial knowledge graphs in companies. In 2007, Freebase was founded as a company to create a community-curated knowledge graph of encyclopedic and common sense knowledge. Freebase was later acquired by Google and became the core of Google's knowledge graph[5]. Google is strategically advancing this knowledge graph by connecting and integrating multiple data sources. Many hundreds of knowledge engineers manually curate Google's knowledge graph. Based on the vocabulary of the schema.org initiative, structured data from billions of web pages are integrated into the knowledge graph. Schema.org is a good example for an industry-wide standard supported by competing search engines, that successfully incentivized the integration of structured, linked data into HTML documents.

[3] https://www.force11.org/group/fairgroup/fairprinciples
[4] https://www.w3.org/TR/dwbp/
[5] https://www.google.com/intl/bn/insidesearch/features/search/knowledge.html

Linking the most diverse data sources and forming the base for the multitude of Google services, Google's knowledge graph now forms the core of the company. In addition to Google, there are now a variety of commercial initiatives to establish corporate knowledge graphs, e.g. through Microsoft/Bing[6], Thomson Reuters[7] or BBC[8].

Starting knowledge graphs in science. Current developments – such as the increasing dissemination of commercial research information systems (e.g. Pure by Elsevier) and social networks (e.g. ResearchGate) as well as non-European initiatives (e.g. Open Knowledge Network[9] of the major US research funders) – demonstrate that the change from document-based to knowledge-based information flows in science and technology is imminent in the next few years. ResearchGate, Elsevier[10] and SpringerNature[11] in particular seem to be actively investing in the development of knowledge graphs. However, if this development is driven solely by non-European or commercial actors, there is a risk that science and technology will become dependent on commercial actors also on knowledge-based information flows, similarly to the dependency and monopolisation of publishers in document-based information flows. To secure information sovereignty, it is therefore of paramount importance to science – and urgency to science policymakers – that scientific infrastructures establish an open counterweight to commercial developments.

Towards an Open Research Knowledge Graph

An Open Research Knowledge Graph (ORKG) should represent original research results semantically, i.e. explicitly and formally, and link existing metadata, data, knowledge and information resources in a comprehensive way (see Figure 1). The graph can be curated collaboratively by research communities. It ensures provenance and represents the scientific discourse and further development. It makes comprehensive and subject-specific concepts unambiguously identifiable and links them semantically (with clearly described relations) to each other and to relevant other artifacts (cf. Figure 2).

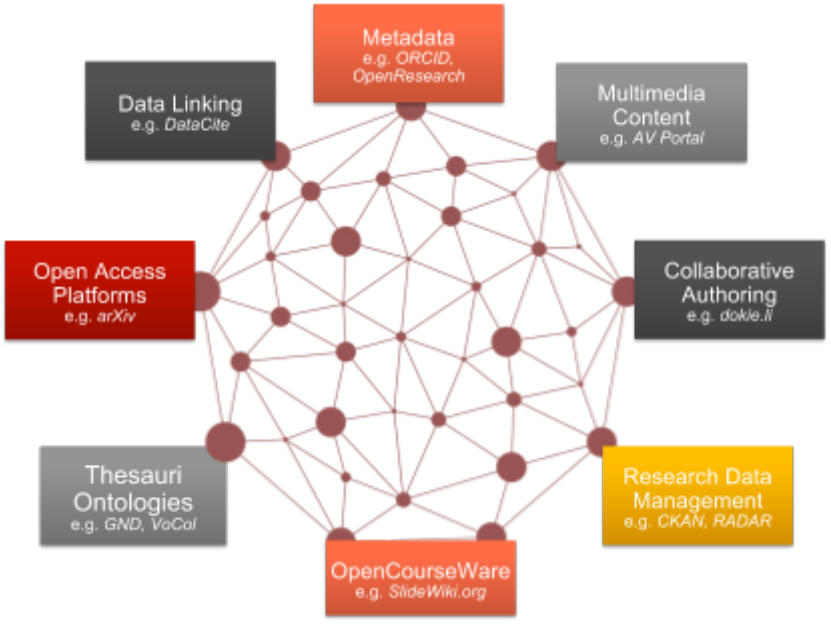

Fig. 1: The planned Research Knowledge Graph as a research infrastructure links an explicit semantic representation of research results with a large number of other information sources and infrastructures.

[6] https://www.bing.com/partners/knowledgegraph
[7] https://www.thomsonreuters.com/en/press-releases/2017/october/thomson-reuters-launches-first-of-its-kind-knowledge-graph-feed.html
[8] https://www.bbc.co.uk/ontologies
[9] https://www.nitrd.gov/nitrdgroups/index.php?title=Open_Knowledge_Network
[10] https://www.slideshare.net/pgroth/knowledge-graphs-at-elsevier
[11] https://www.springernature.com/cn/researchers/scigraph

Synergetic combination of automated and manual procedures. The Open Research Knowledge Graph should be populated and curated from four complementary sources:

1. Existing metadata, data, taxonomies, ontologies, and information models
2. Contributions from scientists who describe their own research supported by intelligent interfaces and automatically generated suggestions
3. Automated methods for knowledge extraction and networking
4. Curating and quality assurance by librarians and information scientists

Only a combination of these different sources and curatorial methods can be successful, since automated procedures alone do not achieve the necessary coverage and accuracy; fully manual curation is too time-consuming; librarians lack the necessary domain-specific expertise; and scientists lack the necessary expertise in knowledge representation. By combining the four strategies in a meaningful way, they can bring their respective strengths to bear and compensate for the weak points.

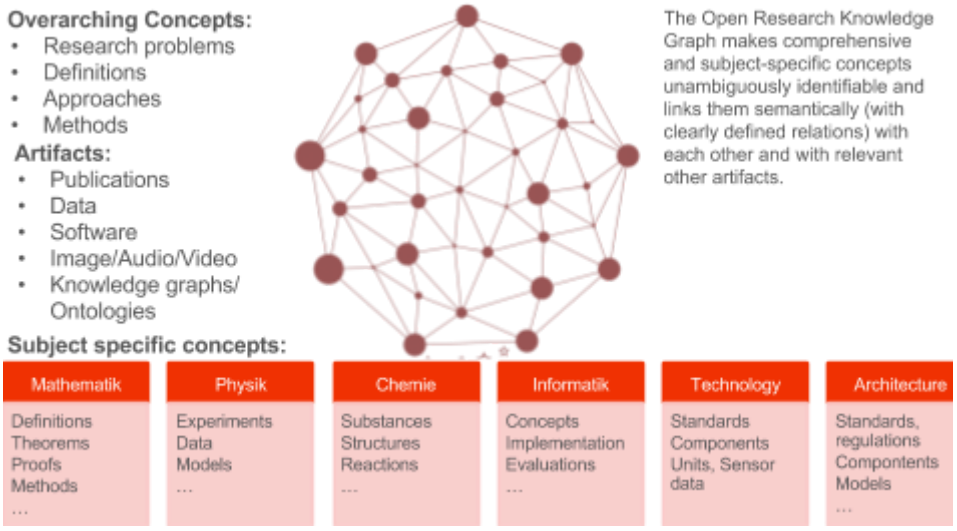

Fig. 2: Interlinking of interdisciplinary and subject-specific concepts and artefacts of scientific work in the different domains (here: TIB subject areas).

The Open Research Knowledge Graph (ORKG) provides interlinking, integration, visualization, exploration, and search functions. It enables scientists to gain a much faster overview of new developments in a specific field and identify relevant research problems. It represents the evolution of the scientific discourse in the individual disciplines and enables scientists to make their work more accessible to colleagues and potential users in industry through semantic description. Figure 3 depicts a research contribution represented in simplified form by a knowledge graph.

Socio-technical ecosystem for knowledge-based science communication. The ORKG service is planned as an open development environment. At its core, it consists of a scalable data management infrastructure with a flexible graph-based data model that can be accessed via lightweight APIs. The service will implement the long-established open standards RDF, RDF-Schema, OWL, Linked Data as well as W3C Data-on-the-Web and FAIR Data Principles to provide maximum interoperability. A central aspect of data storage in the knowledge graph is the preservation of provenance and evolution (similar to wikis), so that changes can be tracked transparently at any time. On the user interface (UI) side, various flexible UI elements should be supported, which can be contributed by advanced users themselves to enable customized domain-specific interactions. In this way it should be possible to represent specific artifacts in the domain (e.g. chemical reactions, mathematical definitions, models and simulations) intuitively and subject-specifically. Figure 4 and 5 show examples from two domains and in [12] we present an initial step for semantically representing research findings.

[12] Said Fathalla, Sahar Vahdati, Sören Auer, Christoph Lange: Towards a Knowledge Graph Representing Research Findings by Semantifying Survey Articles. In TPDL 2017: 315-327

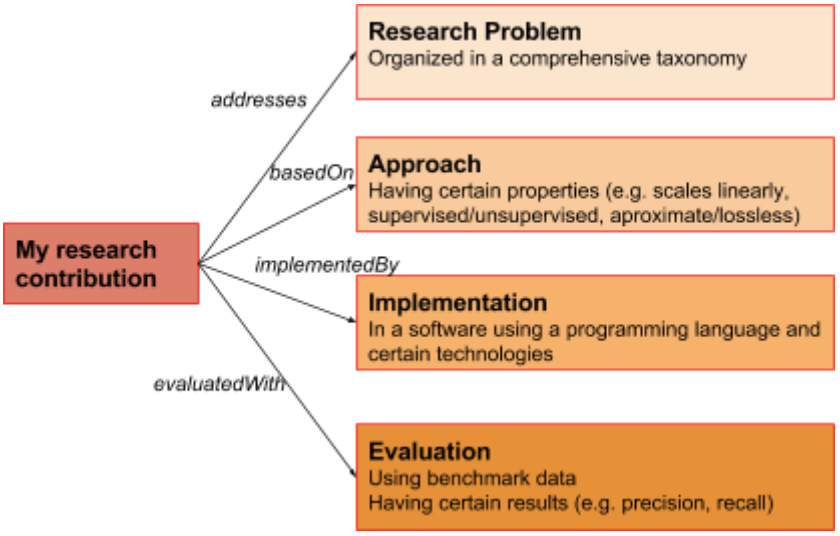

Fig. 3: Possible representation of a research contribution (e.g. in computer science).

Establishment of a network effect. A central challenge in the successful establishment of the ORKG service is to provide researchers with a direct benefit from contributions to the ORKG and thus generate a critical mass of user contributions. We are planning a series of measures to take this into account:

- *Very low-threshold entry and contribution hurdles and crowd-sourcing* – users should be able to submit contributions with just a few clicks.
- *Support through automated procedures*, e.g. for knowledge graph completion or ranking to considerably reduce the effort for contributions.
- *Synergetic combination of automated and manual procedures* through user contributions and curation by information scientists.
- *Integration into publication, peer review, submission systems and repository workflows* through small ORKG widgets, which enable existing publications to be semantically described when submitted/published/reviewed.
- *Generation of direct added value for the contributors*, e.g. by displaying directly related work after a contribution to the ORKG, subscribing to new research results on the topic, and making contributions directly quotable or citable.

Consistent Open Data, Open Science, Open Source Strategy. The ORKG should consistently be implemented as open source software from the outset in order to enable a large number of partners, users and stakeholders to participate in the development. All data and information stored in the ORKG is made available under an open license as open data and open knowledge, so that the community can use this data for integration with other services, new applications or domain-specific analyses.

Transition to Open Access as a basis for the ORKG. In order to make optimal use of published knowledge, it must be as easy as possible to access, i.e. freely accessible, machine-readable and with the possibility of further processing and use (free licenses). The majority of today's publications are currently behind paywalls, i.e. the reader or their institution have to pay to read them. With the transition to Open Access, reading and reuse are free of charge. Only openly licensed and machine-readable publications (including text, video and data) guarantee optimal usability for further and automated processing. Additional major efforts are needed to complete the transition to a sustainable open access landscape, including the comprehensive conversion of subscription contracts with publishers to open-access contracts, the development of new licensing models with publishers/right holders and the establishment of institutional and disciplinary open-access platforms that enable independent and dynamic publishing. We need to actively promote this development and thus support the transition to knowledge-based approaches.

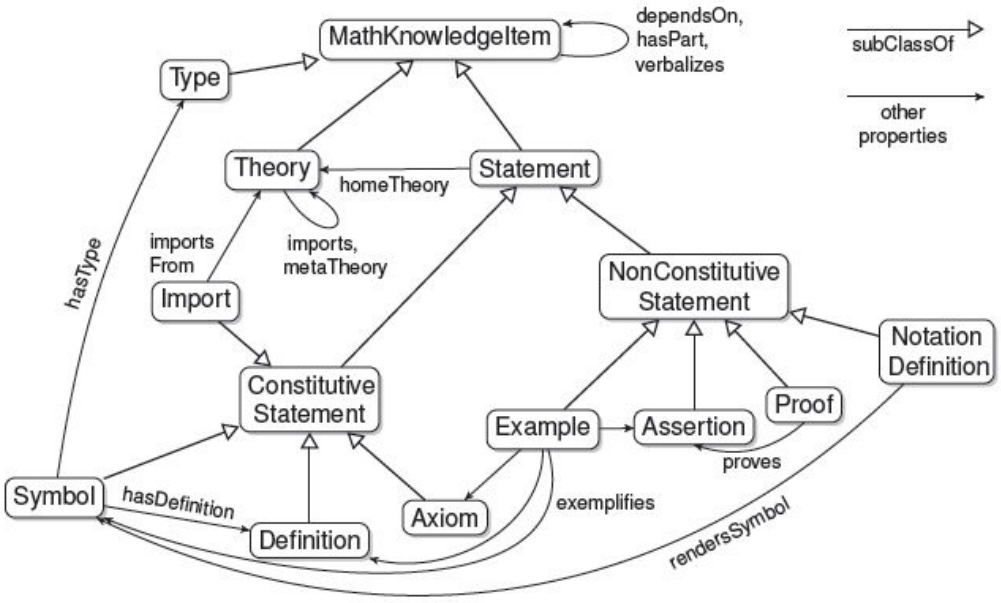

Fig. 4: Ontology for representing mathematical knowledge from [13].

Minimally viable ORKG infrastructure. Together with the community, we aim at realizing a minimally viable ORKG infrastructure. The infrastructure will roughly comprise the following components:

- A data model for semantically representing scholarly communication, which will use RDF and Linked Data as scaffold, but add comprehensive provenance, evolution and discourse information.
- A scalable graph-storage backend infrastructure for storing original ORKG content as well as other key graph assets and exposing a comprehensive API for interacting with the ORKG.
- User interface widgets and components for collaborative authoring and curation of the graph and integration of these widgets into third-party services.
- Semi-automated semantic integration, search, extraction and recommendation services to support the curation of the knowledge graph.

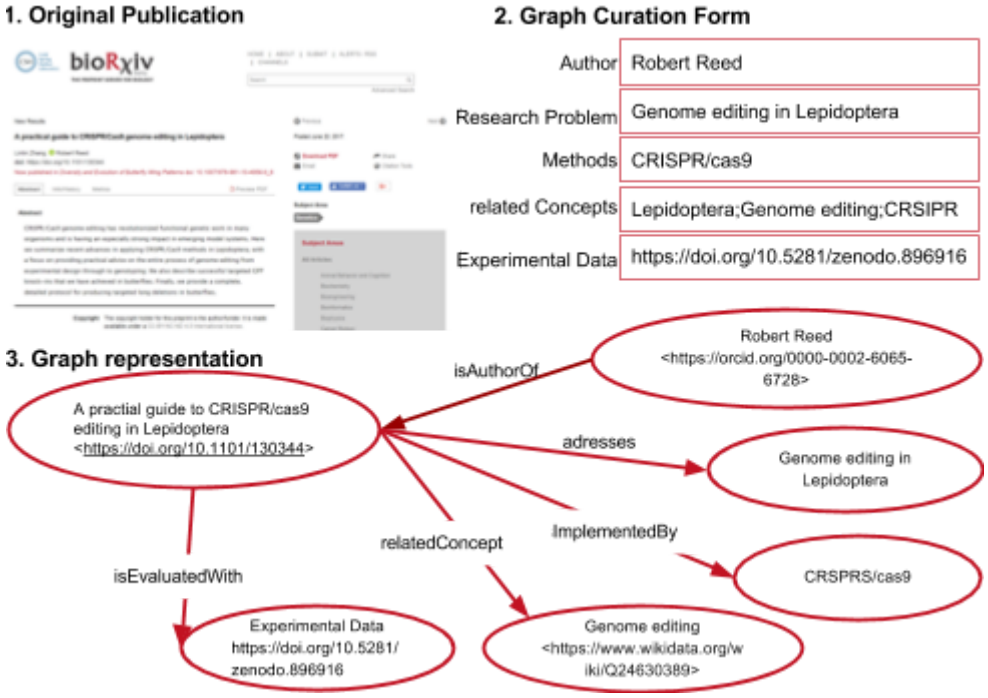

Fig. 5: Acquisition and representation of the CRISPR genome editing method using a knowledge graph.

[13] Christoph Lange: Ontologies and languages for representing mathematical knowledge on the Semantic Web. Semantic Web 4(2): 119-158 (2013)

Exemplary Services based on an ORKG

The ORKG service enables a wide range of applications that offer researchers completely new ways to collaborate faster, more intuitively and more efficiently. Some potential examples of the advantages and possible applications are:

Creation of domain-specific vocabulary and taxonomies (e.g. for research data management). A particularly challenging task in the various scientific fields is the clear definition of the crucial concepts and relationships. At the moment, this is done very costly through a lengthy and implicit crystallization process involving many articles, reviews and surveys. An explicit definition of terms and relationships rarely exists (e.g. in few domain ontologies). The ORKG supports the collaborative creation of domain-specific vocabularies, taxonomies, and ontologies from the outset.

Automatic generation of structured overview visualizations of research areas. Through the structured representation of research results and the use of common vocabularies, it will be possible to filter, organize and present research results according to different criteria. Such overviews are nowadays very resource intensive and time-consumingly compiled by survey or review articles. However, these survey articles do not follow a uniform methodology and have to be created manually for each new criterion or each new selection of works.

Comparison of different research approaches. Due to the terminological inconsistency and heterogeneity of current document-based knowledge communication, comparisons between different research approaches to solve a research problem are very complicated and time-consuming. The structured description of research contributions in the ORKG enables to create comparisons in an automated way.

Survey of opinion formation of scientists. The ORKG might ultimately enable to make the opinions of scientists transparent and thus make possible bias visible. At the moment it is only possible, if at all, through extremely complex analyses.

Visualization of scientific fields. On the basis of the interlinked semantic representation in the ORKG, it is possible to create intuitive visualizations of developments and connections in different scientific fields. This makes it easier to identify and analyse interrelationships.

Proactive notification of new developments in a defined area. Due to the semantic description of the research work, it is possible to realize specific subscriptions to new contributions to the ORKG. At present, researchers are only indirectly exposed to the screening of certain colleagues, journals/conferences or keyword searches for new contributions. This is highly prone to errors and discriminates against young scientists or new conferences and journals. With the help of the ORKG, chemists can be informed about new synthesis possibilities for a certain substance or material scientists can be informed about new experimental properties of a certain material.

Research Analytics. With the help of the ORKG service, specific statistical and qualitative analyses can be carried out e.g. on current research priorities and trends, approaches to challenging research problems, or the impact of research funding in certain areas.

Advantages of graph-based knowledge exchange

An organization of knowledge exchange in research based on the structured, standardized, semantic representation form of a knowledge graph offers numerous advantages. This includes in particular:

- Unique identification of all relevant artifacts, concepts, attributes, relationships, etc.
- More terminological and conceptual precision and sharpness, less ambiguity.

- Better and explicit interrelation of all relevant artifacts and information sources, leading to improved traceability.
- Machine readability of the ORKG enables new search, retrieval, mining, and assistance applications.
- Avoidance of media discontinuities in the various phases of scientific work leading to increased efficiency.
- Use of concepts and relationships across disciplinary boundaries – interdisciplinarity.
- Curbing the proliferation of redundant scientific publications and increasing relevance.
- Facilitating the entry of young, unfamiliar scientists, lay people (Open Science) or researchers with disabilities (e.g. visually impaired).

Research Challenges

Despite the existence of various technologies, standards, and research results, there are a number of research problems and implementation issues that still need to be addressed in order to realize the vision of an Open Research Knowledge Graph. This includes in particular the following:

- Low-threshold integration of users through methods of crowdsourcing, human-machine interaction, and social networks.
- Automated analysis, quality assessment, and completion of the ORKG as well as interlinking with external sources.
- Development of new methods of exploration, retrieval, and visualization of ORKG information.

In the following we briefly discuss some open research questions.

Scholarly Knowledge Representation

Q1: How can we represent Scholarly Communication in various domains in Knowledge Graphs?

While in the meantime it has been demonstrated how to represent general encyclopedic and factual knowledge in knowledge graphs, the question how scientific knowledge from very specialized domains can be represented semantically is still a challenge. We need to devise new methods for a collaborative development of domain models (e.g., vocabularies and ontologies), which can then be used as an underlying semantic structure for ORKG.

Q2: How can we represent discourse, opinion-forming and evolution while maintaining flexibility and simplicity?

A particular characteristic of scholarly communication is that precise conceptual structures only emerge over time and continue to evolve. Thus representing the scholarly discourse needs to accommodate initially fuzzy definitions, diverging opinions, competing conceptualizations and various levels of semantic granularity. When developing a scholarly knowledge representation model for accounting for these requirements, there is the danger that the representation model becomes to complex. Hence, one research challenge is to find a right balance between flexibility and simplicity on the one hand and support for the specifics of scholarly communication on the other.

Scholarly Knowledge Extraction, Completion & Recommendation

Q3: How can we extract Scholarly Knowledge from legacy documents?

It cannot be expected that the current document-oriented scholarly communication workflows can be immediately converted into knowledge-based ones. Hence, we have to support the integration of scientific articles into the knowledge graph. NLP and text mining methods (especially Ontology-based Information Extraction) need to be adapted to leverage the background knowledge already contained in the ORKG (e.g., definitions of existing research problems) to automatically extract such information from legacy scientific publications.

Q4: How can we automatically complete the extracted knowledge graph and generate recommendations?

In order to reduce the manual effort it is of paramount importance to assist users in populating and curating the ORKG. Based on the existing representations in the ORKG we thus need to generate

suggestions and recommendations for completing the graph. By using learning techniques, existing representations in the ORKG can (as training data) help to improve the completion and recommendation of new knowledge.

Knowledge-graph-based Communication & Interaction

Q5: How can we organize collaboration and interaction around the Science Graph?

Automatic methods alone will not reach a satisfactory level of precision and recall in order to populate the ORKG. Consequently, we need to develop methods for integrating a large number of possibly small contributions from researchers and domain experts (who might lack expertise in knowledge representation). The curation of the semantic representations in the ORKG should be supported by knowledge engineers, librarians and information scientists (who in turn lack deep domain expertise).

Q6: How can the incentive systems of document-based scholarly communication be adapted?

Existing scholarly communication incentive measures (e.g. citations, h/i-10 index, impact factor) are document-based. We need to develop novel graph-based incentive measures to assess the impact individual research contributions have. Since the structure and the relationships between different contributions in the graph are much more explicit, it can be expected that the graph-based incentive measures will capture the impact of individual researchers contributions in a more accurate way than coarse-granular document-based measures.

Knowledge Graph Exploration & Question Answering

Q7: How can we facilitate the exploration of the ORKG to fulfil common information needs?

Typical information needs in research are for example finding related work, comparing research approaches addressing a certain research problem, or identifying research contributions to facilitate solving a certain research problem. We should leverage the rich semantic representation of the ORKG for fulfilling such information needs. Also, general semantic information retrieval and faceted search methods should be tailored for search and recommendation in the ScienceGraph.

Q8: How can we answer natural language questions based on the ORKG?

The most intuitive and natural way for researchers to interact with the ORKG is to ask natural language questions. Question answering is already possible for relatively simple factual, encyclopedic knowledge or based on manually created question/query templates (as pursued by commercial QA systems such as Alexa, Google Now or Microsoft Cortana). For more complex information structures such as the ones we anticipate to be comprised in ORKG, reliable and accurate question answering is still a major research challenge.

Conclusion

In this position paper we were outlining the vision of creating an Open Research Knowledge Graph capturing the knowledge generated in science in a semantically structured way. We were discussing some general concepts, advantages and benefits as well as research challenges on the way of its realization. We deem this as a starting point for a discussion in the community ultimately leading to more clearly defined technical requirements, a roadmap and an initiative for realizing the Open Research Knowledge Graph.