Preprints der Zeitschrift *BIBLIOTHEK – Forschung und Praxis*, 2020, AR 3377 **Auer et al.**

AR 3377

Sören Auer, Allard Oelen, Muhammad Haris, Markus Stocker, Jennifer D'Souza, Kheir Eddine Farfar, Lars Vogt, Manuel Prinz, Vitalis Wiens and Mohamad Yaser Jaradeh

Improving Access to Scientific Literature with Knowledge Graphs

Abstract: The transfer of knowledge has not changed fundamentally for many hundreds of years: It is usually document-based - formerly printed on paper as a classic essay and nowadays as PDF. With around 2.5 million new research contributions every year, researchers drown in a flood of pseudo-digitized PDF publications. As a result research is seriously weakened. In this article, we argue for representing scholarly contributions in a structured and semantic way as a knowledge graph. The advantage is that information represented in a knowledge graph is readable by machines and humans. As an example, we give an overview on the Open Research Knowledge Graph (ORKG), a service implementing this approach. For creating the knowledge graph representation, we rely on a mixture of manual (crowd/expert sourcing) and (semi-)automated techniques. Only with such a combination of human and machine intelligence, we can achieve the required quality of the representation to allow for novel exploration and assistance services for researchers. As a result, a scholarly knowledge graph such as the ORKG can be used to give a condensed overview on the state-of-the-art addressing a particular research quest, for example as a tabular comparison of contributions according to various characteristics of the approaches. Further possible intuitive access interfaces to such scholarly knowledge graphs include domain-specific (chart) visualizations or answering of natural language questions.

Keywords: Subject classification, Knowledge Graph, Semantic Web, Crowdsourcing, Text Mining

Verbesserter Zugang zu wissenschaftlicher Literatur mit Wissensgraphen

Zusammenfassung: Der Verbreitung wissenschaftlicher Erkenntnisse hat sich seit vielen hundert Jahren nicht grundlegend verändert: Er erfolgt in der Regel dokumentenbasiert - früher als klassischer Aufsatz auf Papier gedruckt und heute online als PDF. Mit rund 2,5 Millionen neuen Forschungsbeiträgen pro Jahr ertrinken Forscher in einer Flut von pseudo-digitalisierten PDF-Publikationen. Als Folge davon wird die Forschung stark geschwächt. In diesem Artikel plädieren wir dafür, wissenschaftliche Beiträge in strukturierter und semantischer Form als Wissensgraph zu repräsentieren. Der Vorteil ist, dass die in einem Wissensgraph dargestellten Informationen für Maschinen und Menschen lesbar sind. Als Beispiel geben wir einen Überblick über den Open Research Knowledge Graph (ORKG), einen Dienst, der diesen Ansatz umsetzt. Für die Erstellung des Wissensgraph

setzen wir eine Mischung aus manuellen (crowd/expert sourcing) und (halb-)automatisierten Techniken ein. Nur mit einer solchen Kombination aus menschlicher und maschineller Intelligenz können wir die erforderliche Qualität der Darstellung erreichen, um neuartige Explorations- und Unterstützungsdienste für Forscher zu ermöglichen. Im Ergebnis kann ein Wissensgraph wie der ORKG verwendet werden, um einen komprimierten Überblick über den Stand der Technik in Bezug auf eine bestimmte Forschungsaufgabe zu geben, z.B. als tabellarischer Vergleich der Beiträge nach verschiedenen Merkmalen der Ansätze. Weitere mögliche intuitive Nutzungsschnittstellen zu solchen wissenschaftlichen Wissensgraphen sind domänenspezifische Visualisierungen oder die Beantwortung natürlichsprachlicher Fragen mittels Question Answering.

Schlüsselwörter: Sacherschließung, Wissensgraph, Semantic Web, Crowdsourcing, Text Mining

1 Introduction

Scientific libraries must adapt to the changing requirements of science. The digitization of scientific working methods, processes and forms of publication is a central challenge. The methods of scholarly communication have been more or less static text articles for centuries. Although these can now be reproduced electronically as PDF or HTML and quickly accessed via the Internet, the basic representation as unstructured, static articles has not changed fundamentally. On the other hand, other information domains have changed fundamentally and developed completely new digital forms of representation. The only remaining encyclopedia, for example, is Wikipedia, which is not simply a digital PDF copy of an analog encyclopedia, but has realized completely new forms of processing, representation and organization for encyclopedic content, thus enabling, for example, the realization of encyclopedia versions for hundreds of languages and a wide variety of target groups in a completely new depth and breadth. Further examples of entirely digital information services include

- digital map applications (such as Google Maps or OpenStreetMaps), which have now almost completely replaced physical street maps
- Online stores and e-commerce applications with completely new search, evaluation and data networking functions instead of the classic mail order catalogs,
- digital communication applications, which have made telephone books obsolete, for example.

All these examples illustrate that analog forms of representation (books and documents) have not simply been "pseudo-digitized" as PDF, but have been realized completely new as digital-born applications. Such "digital-first" applications are based on a fundamentally new, structured and data-oriented information organization and thus enable completely new support through intelligent search and filter functions, the integration of diverse additional information and services, crowdsourcing etc.

The currently still only pseudo-digitized scientific exchange of information is roughly comparable to a situation in which we would have to pick out products from PDF catalogs sent by an e-mail or find our way to our vacations on a PDF road map, which causes great problems for scientific work:

- We are confronted with a constantly growing number of scientific publications, which of course can be produced faster with digital tools. In the field of technology and natural sciences, for example, the number of publications per year has almost doubled within a decade (NSF NCSES 2018).
- Due to the dramatic growth in the number of publications, the quality of peer reviews is often insufficient: On the one hand, there are too few qualified reviewers with sufficient time, and on the other hand, it is increasingly difficult to determine the review, i.e. the contribution of the research
- The majority of scientific publications cannot be reproduced by researchers (and often even by the authors themselves) (Baker 2016). A main reason for this is the unstructured presentation in static PDF articles, where important information may be missing.
- Different research approaches can hardly be compared due to the unstructured presentation, which makes it extremely difficult to determine the state of research, especially for younger or interdisciplinary researchers.
- The unstructured presentation of research results does not allow for any or only very insufficient machine support. Research contributions cannot be effectively searched, filtered or visualized. Assistance systems such as those already available for everyday situations with Google Now or Alexa are currently unthinkable for coping with the flood of scientific information.

Of course, there are a number of initiatives that have set themselves the goal of addressing these problems. However, it seems that often the symptoms are being worked on rather than addressing the fundamental problems. In part, such solutions seem to be also based on false assumptions.

One misconception, for example, is that text and data mining can solve the problem of indexing and exploring scientific articles. Fully automated text mining and natural language processing methods alone will not provide sufficient accuracy for the extracted information for real use. Such methods often achieve only medium precision and recall for the recognition of entities (named entity recognition). The actual performance highly depends on the domain - while standardized terms (e.g. genes, datasets or countries) can be discovered more reliably, the recognition of other more vaguely defined entities (e.g. materials, processes) is very error-prone (Brack 2020). For relation extraction, which is essential for improved machine support, the correct results are often hardly one third of extracted relations, which are not sufficient to realize reasonable applications in most cases.

Another misconception is that the scientific information can only be organized with fully automated procedures and possibly machine learning. Machine learning methods only work where sufficient training data is available. This is not the case with the structured extraction of scientific results from unstructured articles and will not be possible for the foreseeable future. With crowd-sourcing, or rather expert sourcing, we could, however, master the structured organization of scientific information - possibly supported by machine learning methods. Initially, neither a complete processing of the scientific literature nor the involvement of a large percentage of scientists is necessary. It would initially be sufficient if only a relatively small percentage of scientists were involved in the curation and organization and if, under certain circumstances, only a few research problems and special fields were covered. This would be sufficient for potential applications in these areas and could lead to the establishment of a network effect later on, thus covering broader scientific fields. A good example in this regard is

OpenStreetMaps, a crowdsourcing application with which a few thousand collaborators have created an open world map that is much more detailed than commercial offerings in many areas. Thanks to its innovative data organization, it can be used for a wide range of applications from disaster control and navigation to mobility for the disabled and bicycle maps. Applied to scholarly communication, we therefore need a new form of representation of scientific knowledge that is highly semantically structured and allows for large-scale collaboration between specialist scientists, knowledge engineers, information scientists, librarians and users, while at the same time enabling an evolutionary transition from previous scientific publication and incentive systems. The Open Research Knowledge Graph developed by the TIB and partners, which we present in this article, aims precisely to incorporate these requirements.

2 Overview of the Open Research Knowledge Graph

The following we give a brief overview of the most important features of the Open Research Knowledge Graph (ORKG).

Structured description of the research contributions. The ORKG allows to describe the research contributions traditionally described in scientific articles in a structured and semantic way. For this purpose, articles are added to the ORKG by retrieving (or manually adding) key metadata of the article via DOI from CrossRef and then describing the content of the research articles using specialized input fields. Such structured content descriptions of scientific contributions should describe the addressed research problem, the materials and methods used, and the results achieved in such a way that the contribution becomes comparable with other articles addressing the same research problem. The semantic description follows the RDF subject-predicate-object paradigm and can be flexibly extended by users to include their own additional predicates (properties or attributes) at any time. A suggestion function makes it easy to find and reuse existing predicates and entities. Figure 1 illustrates the structured input of a research article on the effect of ozone on plant growth.

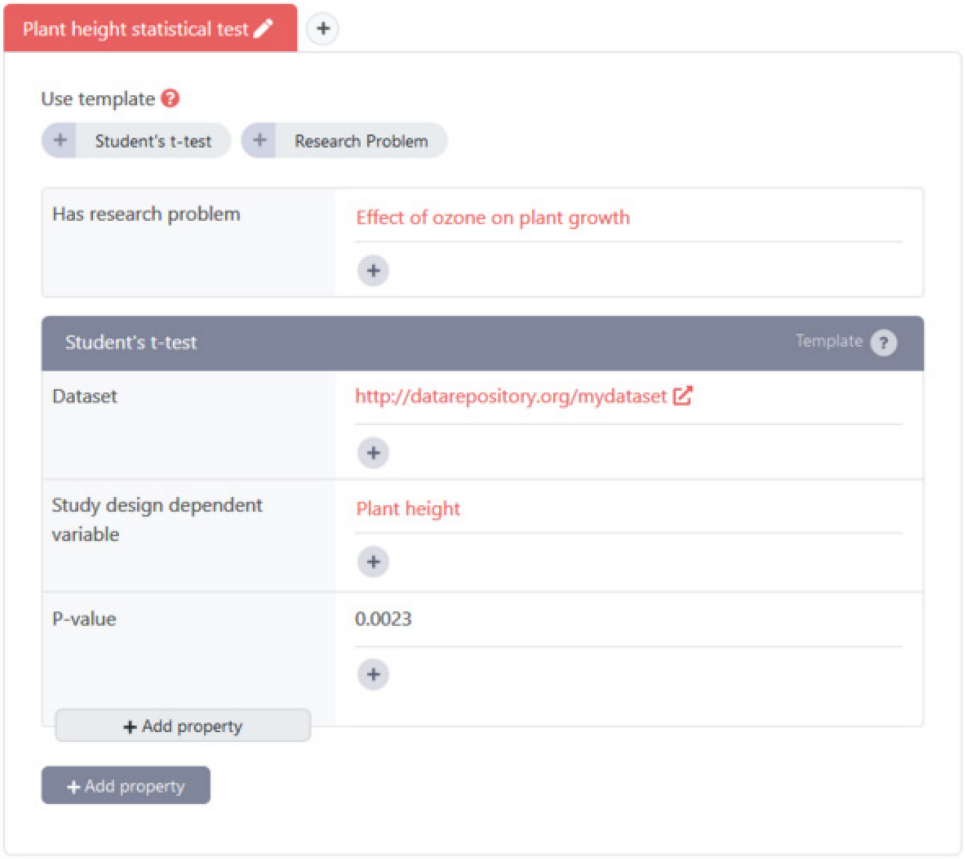


Figure 1: Structured input of a research contribution in ORKG.

Templates. The structured description of research contributions is often not an easy task, because the description of scientific findings is complex and based on expert knowledge. On the one hand, it must be decided in which granularity a research contribution should be described. On the other hand, research contributions dealing with the same problem should be comparable. For this reason, the ORKG supports the possibility to create templates that specify the structure of scientific information. Templates can then be reused in the description of research contributions to facilitate data entry and ensure comparability.

SOTA Comparisons. The ORKG enables the automated comparison of research contributions that deal with a specific problem. Comparisons support users in obtaining a state-of-the-art overview. A classic example in computer science is a comparison of the best/worst-case performance of sorting algorithms or the precision and recall of algorithms for vehicle recognition in images. For researchers in virology and epidemiology it is interesting to be able to compare the reproduction numbers of different viruses. Such comparisons provide an overview of key information on a research problem over dozens or hundreds of papers and are thus a valuable tool for obtaining an overview of the state of the art in a field.

Graph Visualization. Since the ORKG is a knowledge graph, research contributions can also be visualized as graphs. The graph visualization is a sophisticated user interface for visual exploration of a scientific contribution and is therefore a way to interact with ORKG content. The graph is automatically arranged optimally on the

screen. Nodes can be easily expanded, collapsed or removed. Users can also search for information directly in the graph.

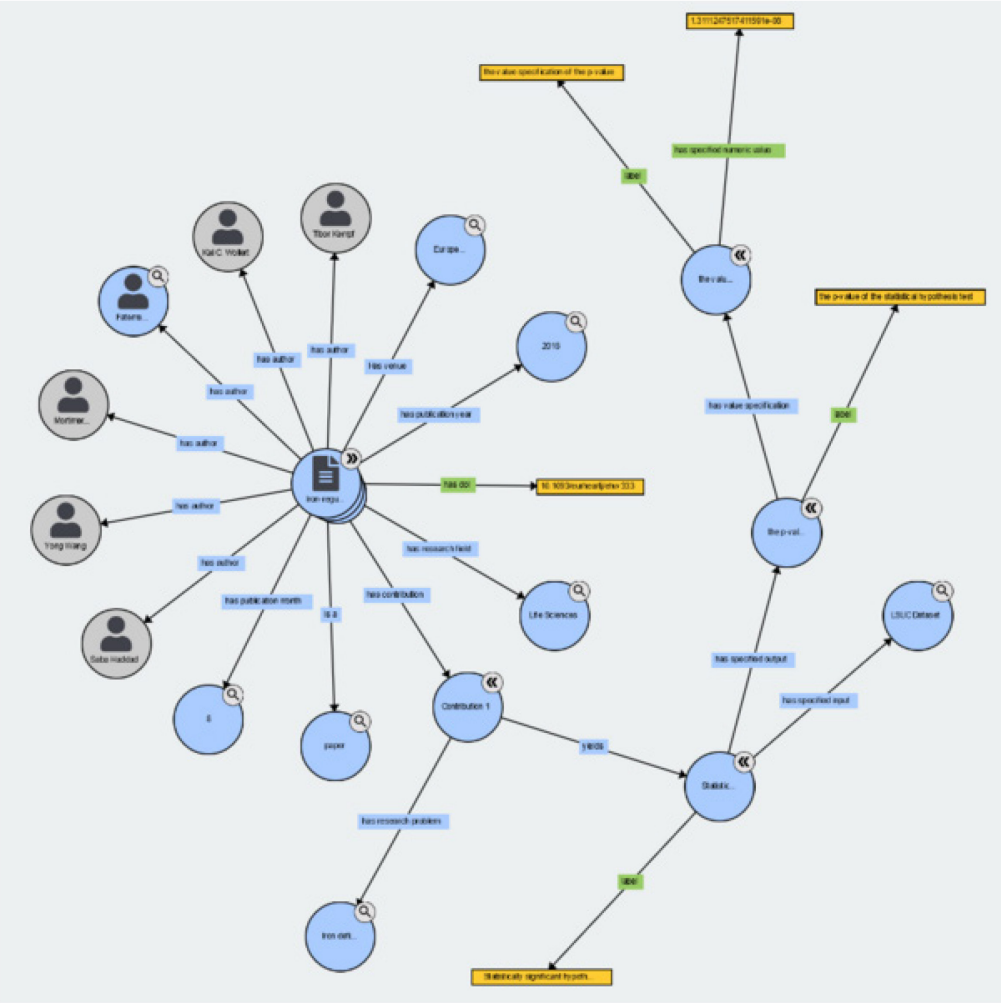


Figure 2: Dynamic graph visualization in ORKG.

Observatories. The ORKG relies heavily on expert content curation and knowledge organization. In order to pool disciplinary expertise, we developed the ORKG Observatories. Observatories bring together experts from different institutions that curate and organize ORKG content for a specific discipline or research problem. Observatories and their experts can contribute in many ways. In addition to adding and describing contributions or curating existing contributions, observatories play a crucial role in the organization of knowledge in a research area. Observatories can, for example, create templates relevant to a specific field. In this way, observatories help to ensure the creation of high-quality and comparable structured scientific knowledge for their field. Since knowledge curation and organization is resource-intensive, ORKG acknowledges the contributions of experts and corresponding observatories and institutions. are made prominently visible in the ORKG. The provenance information in Figure 3 shows how the research contributions of an observatory in the ORKG has been acknowledged.

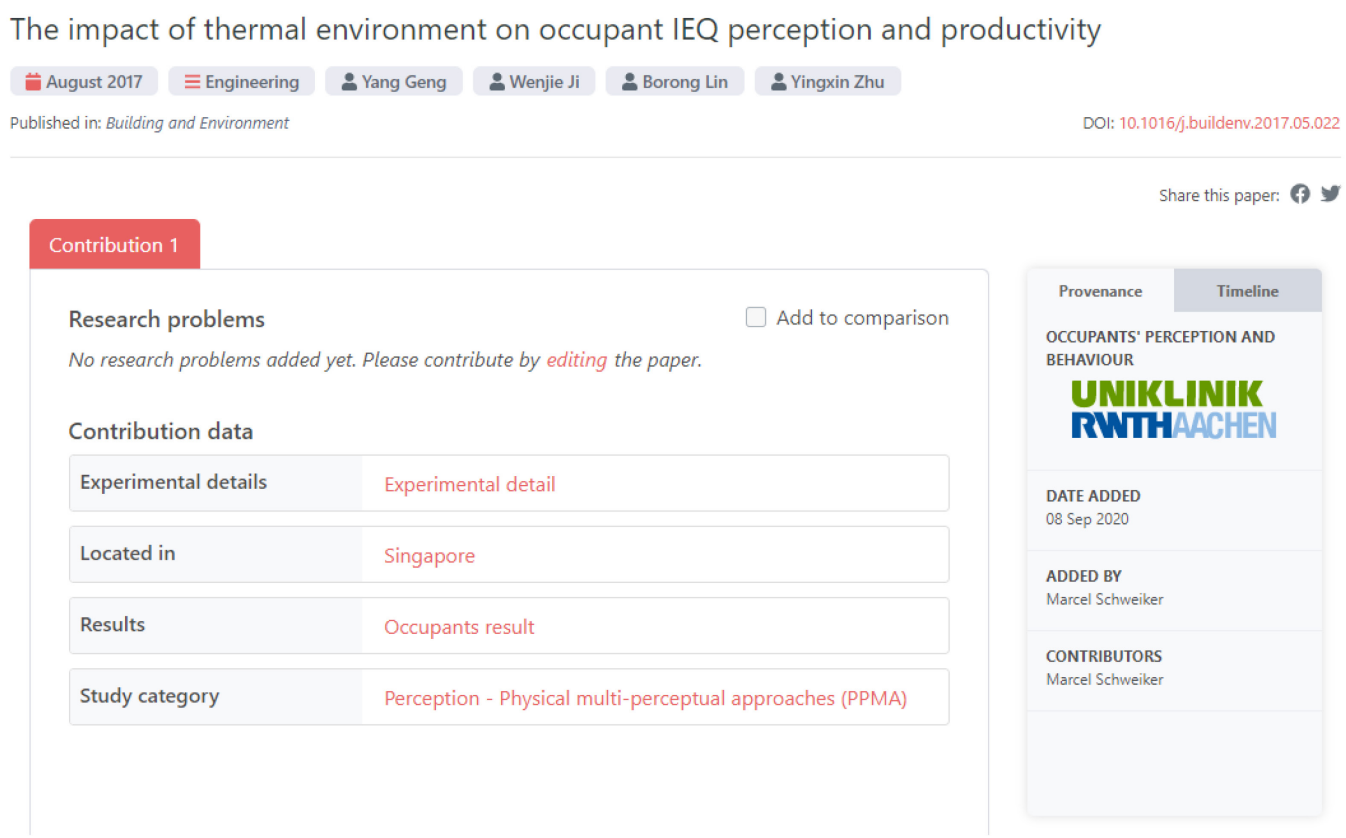


Figure 3: Provenance information of a research contribution in ORKG (box on the right).

Abstract Annotator. An abstract annotator tool has been developed to automatically extract key information from the abstracts of scholarly articles. Different natural language processing and machine learning techniques have been employed to annotate the abstract in an efficient way. While adding the paper, users can use the abstract annotator which could help them to extract the important information such as research problems, methods and materials used to solve the problem. Once these annotations are extracted these can be added in ORKG for a particular paper. Moreover, we plan to automate the content curation as much as possible by implementing the text summarization techniques. Text summarization would summarize the content such as methodology while ensuring to preserve the key information elements.

Question Answering. Answering scientific questions with text is an important part of any research lifecycle. The acquisition of knowledge and appropriate answers is hardly possible due to the following main reasons: machine inactionable, ambiguous and unstructured content in publications. We have also developed a question answering system in ORKG which maps the natural language queries to the graph and finds the answer for that query (Jaradeh 2020). Furthermore, analyzing and searching data from tables is a difficult aspect so a question answering system for ORKG comparisons has also been introduced. Users or authors can ask questions from comparisons to find the relevant information.

Knowledge integration. The structured and semantic description of the knowledge enables a simplified knowledge re-use. The comparisons described above are only one type of knowledge re-use. In fact, in science, knowledge in literature is re-used in countless ways. For example, to support this diversity, ORKG implements a web-based interface (REST API), which can be used with the Python programming language. This allows to load ORKG content (individual article descriptions and comparisons) into a data analysis environment such as Jupyter notebooks, to process it and to create domain-specific applications and visualizations. This allows to easily create data visualizations, but also to implement complex data-enabled activities, which integrate ORKG-data with other data integrators, data interpreters, models, etc.

2.1 Extraction from Survey and Review Articles

A method to populate ORKG is to leverage already existing scholarly knowledge from survey articles. Survey articles, also called review articles, present an overview of the state-of-the-art for a specific research area. The overviews are generally manually curated and of high quality. Additionally, surveys indicate what the current trends are and are therefore providing relevant information for researchers (Gall et al., 1996). Within survey articles, the overviews are often presented in (semi-)structured tabular format. Compared to generating a knowledge graph from natural text, building a graph from a table is more straightforward because of the already existing structure.

To import survey articles within ORKG, we employ a human-in-the-loop approach. The full approach and evaluation are described in Oelen et al. (2020b). While most of the steps are conducted automatically, the human curator is responsible for fixing potential extraction errors and adding additional metadata. All steps required to import a table are integrated in a single User Interface (UI). We now discuss the individual steps required to import a survey article and indicate where human labor is required. Firstly, the curator uploads a PDF version of the survey article. Afterwards, the region of the table is selected (as depicted in Figure 4). Only tables that list related work (i.e., tables with references) should be selected. The selected table is extracted using Tabula, a state-of-the-art PDF table extraction tool (Corrêa et al., 2017). The extracted table is presented in a spreadsheet editor, as displayed in Figure 5. The curator is responsible for fixing extraction errors and for formatting the table in such a way that it is suitable for import. This means that each row should represent a single paper, in some cases a table has to be normalized to accomplish this. Once the table is formatted by the curator, the references are extracted using GROBID. GROBID is a state-of-the-art bibliographic data extraction tool that supports parsing references in PDF articles (Lopez, 2009). In this step, the curator is responsible for validating the extracted references, and if needed, adding missing references. Multiple columns are appended to the table to include the parsed reference data (among others, the title, authors and publication date).

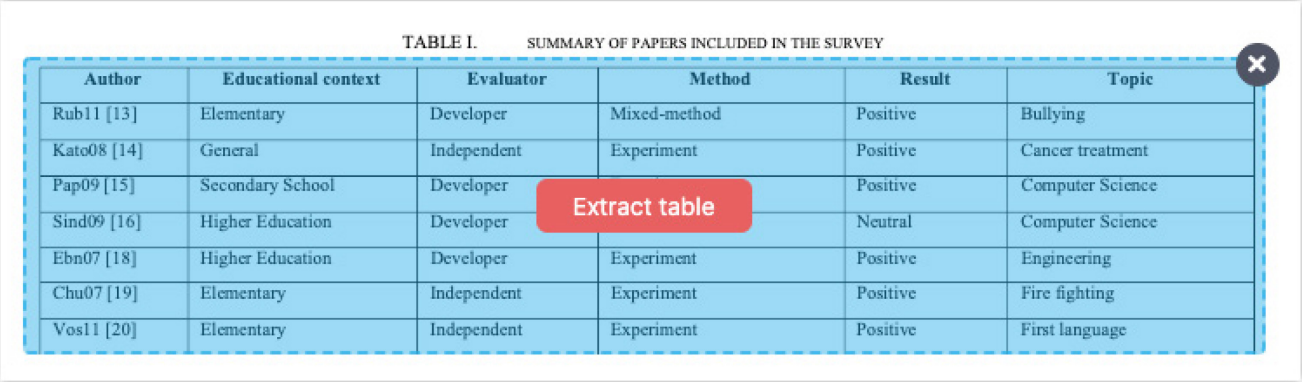


Figure 4: Select the table region within the PDF article to perform the table extraction.

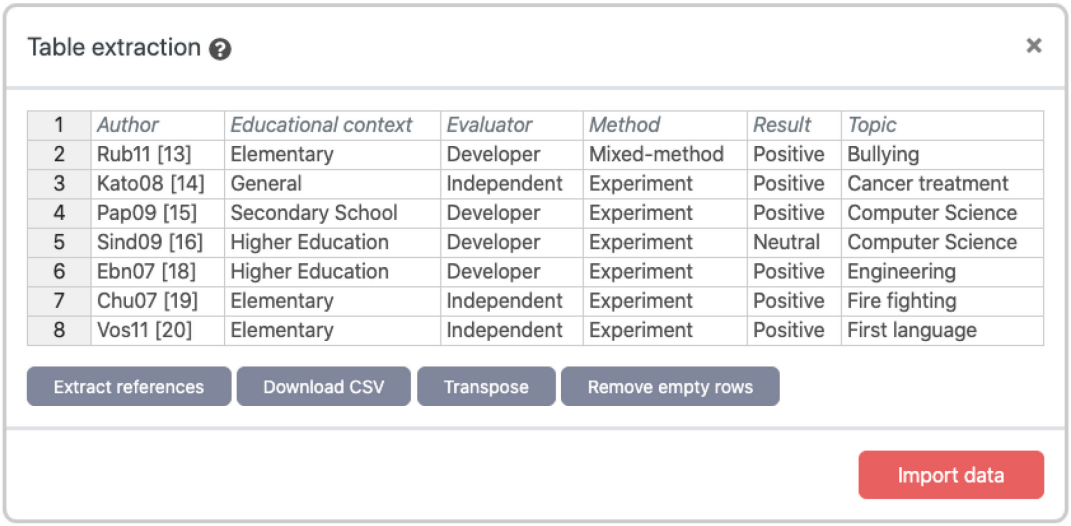


Figure 5: Spreadsheet editor to fix extraction errors, perform table formatting and to extract references.

The resulting knowledge graph for a single paper is depicted in Figure 6. The figure depicts an illustrative example table with three rows. For each row, a separate paper subgraph is generated. Based on the first column of the table, the metadata of the cited paper is extracted (the resulting data is colored blue in the figure). The remaining tabular data is displayed in orange colored nodes. The predicates of the orange nodes are defined by the table header. Finally, the white nodes are ORKG specific concepts. This includes the *Research Contribution* and the *Research Problem*.

Author	Educational context	Evaluator	Method	Result	Topic
Rub11 [13]	Elementary	Developer	Mixed-method	Positive	Bullying
Kato08 [14]	General	Independent	Experiment	Positive	Cancer treatment
Pap09 [15]	Secondary School	Developer	Experiment	Positive	Computer Science

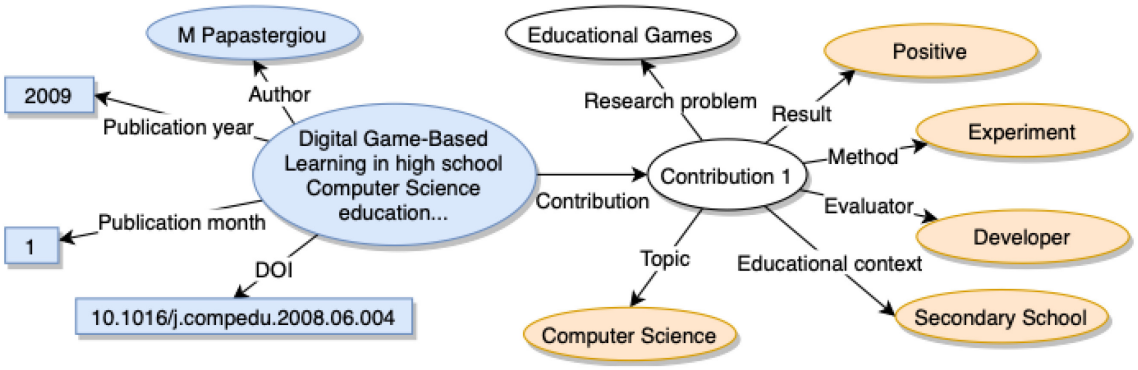


Figure 6: The resulting graph for a single paper comprising extracted metadata (blue), the tabular data (orange) and ORKG specific concepts (white)

The method of extraction survey tables provides a means of relatively quickly populating a knowledge graph without the need of domain experts. To evaluate our approach, we used the previously described method to import 160 survey tables (Oelen et al. 2020b). This resulted in a total amount of 2,626 papers that were ingested in ORKG. In the same work, we estimated that approximately 20% of recently published survey articles use tables to represent reviewed articles. Importing a table in ORKG is not the final step, on the contrary, it provides a first step to create more comprehensive and dynamic overviews of the literature. One of the weaknesses of the current review method is that published reviews are static (Oelen et al., 2020a). When survey articles are published, they are rarely updated and do therefore not always accurately represent the current state-of-the-art. Once a survey is imported in ORKG, it enables researchers to add their results to the literature comparison (by means of adding an additional row). Additionally, a survey could be extended by including more comparison criteria (by adding extra columns). This highlights another benefit of this approach, namely the ability to collaboratively work on literature overviews. Not only the viewpoint of a few authors are represented in a comparison, but of the whole community.

2.2 Further functions

In addition to the examples presented, the ORKG concept includes a number of other functions already implemented or planned for the near future:

- Tabular comparisons of the state of research on a particular research problem can be published independently with DOI and exported in various formats.
- Overview tables on the state of research can be imported semi-automatically from PDF documents into ORKG.
- All ORKG contributions are versioned so that all changes can be discussed by the professional community.
- Automatic extraction functions support the filling of the knowledge graph.

- Research problems can be described independently, provided with relevant sources and assigned to a taxonomy of research areas.
- Specialist scientists can independently realize domain-specific visualizations based on the ORKG.

3 Knowledge Graph Use Cases

In this section we will show how knowledge graphs and in particular the ORKG can be used to structure information for three different areas. For each domain an ORKG comparison is generated to give an overview of the state of the art for this particular example.

3.1 Computer Science

In computer science, a number of recurring characteristics of scientific contributions are of interest to the ORKG. These properties include evaluation results (F-measure, precision, recall), data sets, benchmarks, model properties and implementation approaches. These properties are applicable to a considerable number of computer science articles and are of interest to many researchers. In this example we show how these recurring properties are used in the field of Question Answering (QA). The goal of the QA task is to automatically provide answers to questions in natural language. To organize QA-related information in ORKG, methodological approaches and evaluation results of different papers and the implemented approaches are represented and compared. An interesting aspect of these problems is that the compared QA systems and the evaluation results of these systems are not originally published in the same articles or by the same authors. However, due to the dynamic aspects of the ORKG it is possible to establish links between the systems and evaluations afterwards. Figure 7 shows a comparison of different QA systems and their evaluation results in terms of precision, recall and F-measure in the QALD benchmark. Each QA system is presented in a separate article published in different journals but can be compared directly with other approaches. It is therefore possible to get a quick overview of the state of the art and the performance of the systems. Furthermore, the comparison of these evaluation results provide information about the progress for a specific research area.

Properties	LIMSI participation at QALD-5@CLEF Contribution 1 - 2015	Cross-Lingual Question Answering Using Common Semantic Space Contribution 1 - 2016	CASIA@ V2: a MLN-based question answering system over linked data Contribution 1 - 2014
Has research problem	Question answering systems	Question answering systems	Question answering systems
Implementation	SemGraphQA	UTQA	CASIA
Disambiguation task	Local disambiguation	Local disambiguation	Local disambiguation MLN
Query construction task	Using info. from the QA	*Empty*	Using machine learning
Question analysis task	Dependency parser NE n-gram strategy	POS learned	Dependency parser NE n-gram strategy POS learned

On	QALD-6	QALD-6	QALD-6
Dataset	DBpedia 2015 DBpedia 2015 with abstracts LinkedSpending	DBpedia 2015 DBpedia 2015 with abstracts LinkedSpending	DBpedia 2015 DBpedia 2015 with abstracts LinkedSpending
Evaluation	SemGraphQA	UTQA	UTQA
Language	Farsi	English	Spanish
F-measure	0.37	0.65	0.68
Precision	0.70	0.70	0.76
Recall	0.25	0.61	0.62

Figure 7: Comparison of question-answer systems and their evaluation results in the lower part

3.2 Epidemiology and COVID-19

The COVID-19 pandemic also leads to an abundance of daily new articles in research. To support COVID-19 research, many publishers have decided to publish open access articles related to COVID-19. While access is crucial, organizing the information published in articles is essential for effective research. Unfortunately, however, this is extremely time-consuming and thus a major obstacle. An example of a fundamental research finding that is spread over numerous published (preprint) articles is the COVID-19 baseline reproduction number R0, the respective 95% confidence interval, the location of the population under study and the observation period. In contrast to the conventional document-based publication of such information in natural language texts, tables or figures, ORKG enables to publish such information in a structured, semantic form. Information can thus be evaluated automatically and across publications. Machine readable scientific knowledge opens a number of interesting possibilities. As already mentioned, the structured semantic representation of scientific knowledge in the knowledge graph makes it possible to automatically create literature comparisons. Figure 8 illustrates the comparison for our use case of the COVID-19 basic reproduction number. In addition, ORKG comparisons can develop further in contrast to static PDF overview articles (surveys or reviews). If new literature on the basic reproduction number is published, it is easy to continuously expand such a comparison, which thus continues to reflect the current state of knowledge in a comparable way.

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 8: Automatic comparison of basic reproduction figures published in the literature

However, the true strength of knowledge graphs become apparent when the graph is used for further comprehensive data analyses. We demonstrate this by connecting Jupyter notebooks to ORKG to take advantage of the flexibility of data science environments and programming languages such as Python to visualize COVID-19 comparison data. Figure 9 shows a possible visualization of the R0 values and the 95% confidence interval over time. Of course, data from other sources can be used for such analyses, for example statistical data on the death rate in this case.

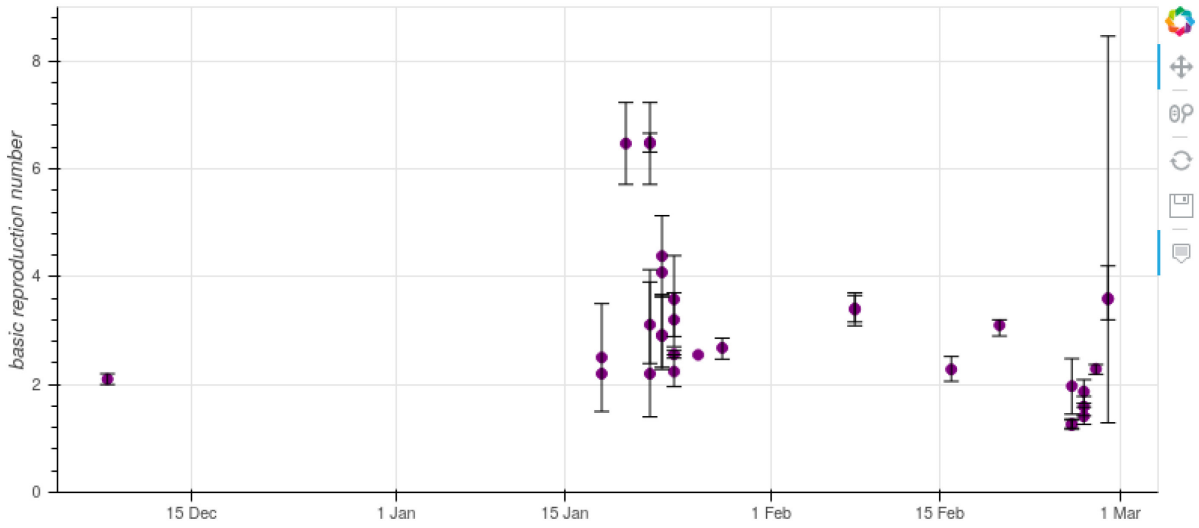
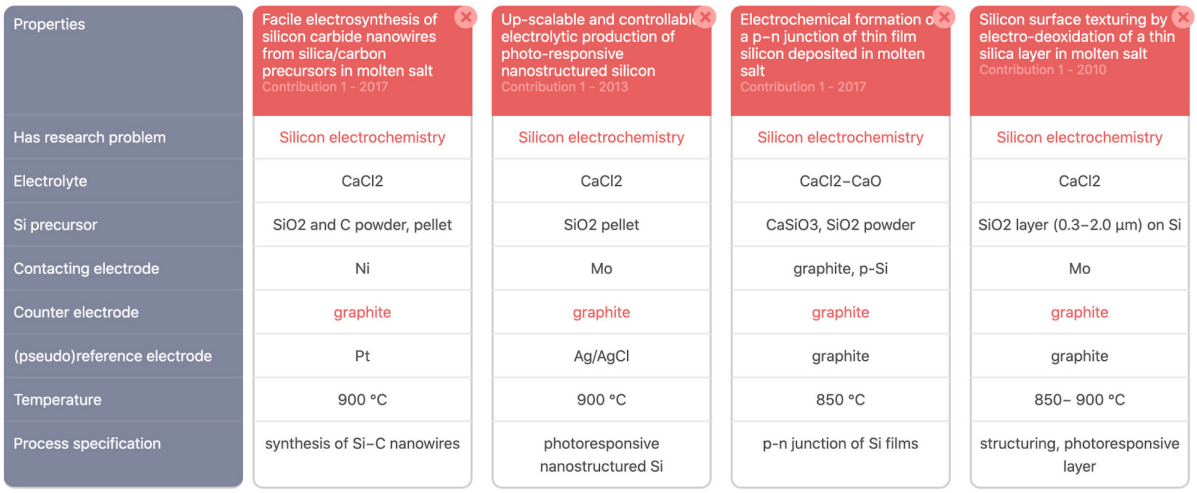


Figure 9: Visualization of the R0 values and the 95% confidence interval over time as data analysis on the ORKG knowledge graph

3.3 Materials Science

Finally, we present an example from electrochemistry, materials and engineering sciences. Silicon is an important element of modern technology. It is widely used in the production of metal alloys, optical fibers, solar elements, advanced ceramics, batteries, microchips and numerous other beneficial applications. For solar and electronic devices, there is a demand for solar-grade silicon (SoG-Si) with a purity of 99.9999 percent or electronic silicon with an even higher purity. Silicon electrochemistry in molten salts has recently attracted much attention due to its potential to produce Solar Grade Silicon with a negligible carbon footprint. The comparison shown in Figure 10 provides a comprehensive overview of several parameters such as silicon dioxide precursor, electrolyte, contact electrode or temperature of the experimental conditions of silicon electrochemical reduction in molten electrolytes. In this way, researchers can easily analyze relevant parameters used in the process specifications for the generation of silicon surface structures.



Properties	Facile electrosynthesis of silicon carbide nanowires from silica/carbon precursors in molten salt Contribution 1 - 2017	Up-scalable and controllable electrolytic production of photo-responsive nanostructured silicon Contribution 1 - 2013	Electrochemical formation of a p–n junction of thin film silicon deposited in molten salt Contribution 1 - 2017	Silicon surface texturing by electro-deoxidation of a thin silica layer in molten salt Contribution 1 - 2010
Has research problem	Silicon electrochemistry	Silicon electrochemistry	Silicon electrochemistry	Silicon electrochemistry
Electrolyte	CaCl2	CaCl2	CaCl2–CaO	CaCl2
Si precursor	SiO2 and C powder, pellet	SiO2 pellet	CaSiO3, SiO2 powder	SiO2 layer (0.3–2.0 µm) on Si
Contacting electrode	Ni	Mo	graphite, p-Si	Mo
Counter electrode	graphite	graphite	graphite	graphite
(pseudo)reference electrode	Pt	Ag/AgCl	graphite	graphite
Temperature	900 °C	900 °C	850 °C	850– 900 °C
Process specification	synthesis of Si–C nanowires	photoresponsive nanostructured Si	p-n junction of Si films	structuring, photoresponsive layer

Figure 10: ORKG comparison for work on silicon electrochemical reduction in molten electrolytes in materials science

4 Integration of library communities and services

Libraries can play a central role in the further development and curation of scholarly knowledge graphs. Not only do scientists need to be supported by AI and text mining procedures, but the information science core competencies in knowledge organization and formal indexing are also indispensable for the success of scholarly knowledge graphs. The Open Research Knowledge Graph is therefore intended to become a comprehensive and broad knowledge infrastructure for the collaboration of specialists, information scientists, librarians and users. In order to involve as many stakeholders and multipliers as possible, we are currently working on various strategies:

- ORKG is a radically open research infrastructure according to the principles of open science, open data and open source: all software, information and data are available under open licenses. The ORKG provides a comprehensive programming interface (API), which can easily be used to implement specific applications based on the ORKG infrastructure.
- In subject-specific observatories, specialist scientists organize research work on relevant research questions in a clearly defined field together with expert speakers. We want to make their commitment visible and recognized by prominently featuring the persons and organizations involved in the respective entries in the ORKG.
- We plan to publish comprehensive and well-documented comparisons of the state of research on a specific research question as independent publications in an Open Access journal. Already now, such comparisons can be provided with a DOI in the ORKG and can be published citable.
- (Semi-)automatic procedures are based on good authoritative reference data, this part of the intellectual development remains the task of library departments. Various ontologies and standards data can then help to structure knowledge in graphs.

We would be very pleased if we could win further collaborators in the library community who, together with specialist scientists, maintain observatories, link existing infrastructures with the ORKG or implement new subject-specific applications (e.g. elements for specialist information services) based on the ORKG infrastructure.

5 Related Scholarly Knowledge Graph Projects

In addition to the ORKG, there are a number of other scholarly knowledge graph projects, such as the Microsoft Academic Graph (Sinha et al., 2015), the Springer Nature SciGraph[1], Papers with Code[2] and ResearchGraph (Aryani 2014). However, these initiatives primarily focus on integrating and organizing bibliographic metadata. Another related initiative targeting the deep analysis of publications is SemanticScholar (Ammar 2018). Semantic Scholar functions mainly like a search engine but focuses primarily on highly automatized artificial intelligence approaches to augment the key-word search and thus lack highly structured and semantically rich descriptions of the research contributions.

Properties	Papers with code Contribution 1	Data integration and disintegration: Managing Springer Nature SciGraph with SHACL and OWL Contribution 1 - 2017	Research Linking Initiative: Toward Interoperability of Research Data Contribution 1 - 2014	An Overview of Microsoft Academic Service (MAS) and Applications Contribution 1 - 2015
Api support	✓	REST API	*Empty*	REST API
Data	*Empty*	1.5 billion triples	*Empty*	8 billion triples
Graph	Papers with code	Springer Nature SciGraph	Research graph	Microsoft academic graph
Has research problem	Scholarly communication	Scholarly communication	Linked research data Scholarly communication	Scholarly communication
Has url	https://paperswithcode.com/	https://www.springernature.com/gp/researchers/scigraph	https://researchgraph.org/	https://academic.microsoft.com/home
Research infrastructure	*Empty*	*Empty*	Crossref DataCite ORCID Research Data Australia	*Empty*
Supports rdf	✗	✓	✓	*Empty*
Metadata	Partially	*Empty*	✓	✓
Semantic representation	✗	*Empty*	✗	✓
Supports research data	✓	✓	✓	✓
Journals	*Empty*	5000	*Empty*	48965

Figure 8: Comparison of different scholarly knowledge graphs in the Open Research Knowledge Graph[3]

6 Conclusion

We need to go beyond organizing scholarly communication just on the surface by managing and integrating bibliographic metadata. Going further with knowledge graphs, scientific findings can be represented semantically in a human and machine readable way according to the FAIR principles. Concepts that are currently still deeply hidden in unstructured publications are given an unique identification with persistent identifiers and can be

[1] https://www.springernature.com/gp/researchers/scigraph
[2] https://paperswithcode.com
[3] https://www.orkg.org/orkg/comparison/R50014

semantically linked to other relevant concepts or artifacts. The semantically structured ORKG content can be easily accessed in many ways via the web, API, data dump or query interfaces and the open licenses.

Despite the potential and benefits of leveraging knowledge graphs for scholarly communication, we are still at the very beginning of development and there are still many open questions to be answered: How can we increasingly involve specialist scientists in the curation process? Do the semantic curation techniques scale for broad subject areas and can a continuous evolution of semantic representation be achieved? How can incentive systems for scientists be adapted or transformed in a meaningful way for knowledge graph contributions? Which tasks are performed by libraries?

Acknowledgements

Parts of the work described in this article have been co-funded by the European Research Council for the project ScienceGRAPH (GA: 819536). We would also like to thank Irina Sens, Oliver Koepler and Philipp Dillschneider, who have contributed to this work.

References

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V.A., Kinney, R.M., Kohlmeier, S., Lo, K., Murray, T.C., Ooi, H., Peters, M.E., Power, J.L., Skjonsberg, S., Wang, L.L., Wilhelm, C., Yuan, Z., Zuylen, M.V., & Etzioni, O. (2018). Construction of the Literature Graph in Semantic Scholar. NAACL-HLT.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 243-246. DOI=http://dx.doi.org/10.1145/2740908.2742839

Aryani, Amir (2014): Research Linking Initiative: Toward Interoperability of Research Data. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.1170011.v1

Brack, A., D'Souza, J., Hoppe, A., Auer, S., Ewerth, R. (2020): Domain-Independent Extraction of Scientific Concepts from Research Articles. European Conference on Information Retrieval ECIR (1) 2020: 251-266

Corrêa, A. S., & Zander, P. O. (2017). Unleashing tabular content to open data: A survey on pdf table extraction methods and tools. In Proceedings of the 18th Annual International Conference on Digital Government Research (pp. 54-63).

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). Educational research: An introduction. Longman Publishing

Lopez, P. (2009). GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In International conference on theory and practice of digital libraries (pp. 473-474). Springer, Berlin, Heidelberg.

Baker, M. (2016): 1,500 scientists lift the lid on reproducibility, Nature.

Oelen, A., Jaradeh, M. Y., Stocker, M., & Auer, S. (2020a). Generate FAIR Literature Surveys with Scholarly Knowledge Graphs. JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, 97–106. https://doi.org/10.1145/3383583.3398520

Oelen, A., Stocker, M., & Auer, S. (2020b). Creating a Scholarly Knowledge Graph from Survey Article Tables. ICADL '20: The 22nd International Conference on Asia-Pacific Digital Libraries (in press)

NSF NCSES (2018). Science and Engineering Publication Output Trends, NCSES InfoBrief, National Science Foundation, 2018. https://www.nsf.gov/statistics/2018/nsf18300/nsf18300.pdf

Jaradeh, M. Y., Stocker, M., Auer, S. (2020): Question Answering on Scholarly Knowledge Graphs. 24th International Conference on Theory and Practice of Digital Libraries (TPDL 2020): 19-32


© TIB/C. Behrens

Prof. Dr. Sören Auer
TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften
Direktor der TIB
Welfengarten 1 B
D-30167 Hannover
auer@tib.eu