EasyChair Preprint

№ 10863

On the Form of Research Publications for Use in Scientific Knowledge Graphs

Leon Martin, Robin Jegan and Andreas Henrich

August 7, 2024

On the Form of Research Publications for Use in Scientific Knowledge Graphs

Leon Martin
leon.martin@uni-bamberg.de
University of Bamberg
Bamberg, Bavaria, Germany

Robin Jegan
robin.jegan@uni-bamberg.de
University of Bamberg
Bamberg, Bavaria, Germany

Andreas Henrich
andreas.henrich@uni-bamberg.de
University of Bamberg
Bamberg, Bavaria, Germany

ABSTRACT

Current research proposes scientific knowledge graphs to support various research activities by acquiring and integrating scientific information including research publications. The accompanying envisaged shift towards knowledge graph based research motivates rethinking the form of research publications since the traditional form of research publications, i.e., self-contained documents, may leave opportunities unused. This paper investigates different publication forms that are used in scientific knowledge graphs, identifies their flaws from the authors', providers', and readers' perspectives subsequently, and finally outlines a first set of requirements that a publication form tailored for use in scientific knowledge graphs should fulfill.

KEYWORDS

scientific knowledge graphs, research publications, publication forms, requirements

1 INTRODUCTION

Motivated by the growing number of scientific communities and research publications [2], the interest in Scientific Knowledge Graphs (SciKGs) [11], also known as scholarly knowledge graphs [13, 14] or research knowledge graphs [8], i.e., knowledge graphs that acquire and integrate scientific information in a knowledge base [1, 4], is on the rise. In this regard, the TIB Leibniz Information Centre for Science and Technology and the L3S Research Centre in Hannover are important contributors. One of their recent papers [2] presents a thorough requirements analysis for their Open Research Knowledge Graph (ORKG) [8][1], an already operative scientific knowledge graph, that is intended to facilitate typical research activities like finding related work, assessing relevance, and reproducing results among others.

Nevertheless, for many decades, research publications have come in the form of self-contained documents, so-called papers. The shift away from document-centric research towards the envisaged knowledge graph based research, however, also includes reconsidering the form of research publications as traditional papers may not take full advantage of the new paradigm's opportunities. Therefore, this paper first investigates current publication forms that are used in SciKGs (section 2), identifies their flaws (section 3), and finally outlines requirements for a publication form tailored for use in SciKGs (section 4), before drawing a conclusion in section 5.

[1] https://www.orkg.org (accessed 19/10/2021)

2 CONTEXT & RELATED WORK

Knowledge graphs leverage the Resource Description Framework (RDF) to represent information as triples, each comprising a subject (an entity), a predicate (a property), and an object (an entity or a literal) [3]. A set of RDF triples span a graph, i.e., the RDF or knowledge graph. Ontologies, based on which knowledge graphs are constructed, formally define what entities mean in a given domain, what features they possess, and thereby how entities, properties—both identified via Internationalised Resource Identifiers (IRIs), a generalization of Uniform Resource Identifiers (URIs) [3]— and literals can be arranged in RDF triples [7, 17–25].

In the context of SciKGs, one key challenge is the integration of research publications, which carry a significant part of the available scientific knowledge. From the perspective of SciKG providers, the goal here is to achieve a high coverage of research publications and gather a large user base, whereas authors want to increase the visibility of their publications with little additional effort[2]. Since ontologies provide the formal base of SciKGs as well, they determine how research publications are integrated and thereby what publication forms are supported by the knowledge graph. At the same time, the standard publication form determines the ontology since providers of SciKGs want to maximize the number of potential contributors. As a result, ontologies of SciKGs and publication forms must evolve together due to their mutual influence.

Depending on the intended use cases, SciKGs' ontologies can be designed to focus on the representation of *contextual* information for describing research publications or even to allow the representation of their *contentual* information (cf. Figure 1), e.g., the publications' contributions[3]. For this purpose, the usage of knowledge graph cells [15] is an option. Representing contentual information, however, imposes various challenges like the expression of opinion forming [1] that have to be addressed in the future.

There are different forms of research publications that are already used or lend themselves to be used in SciKGs. The obvious option is to retain the traditional self-contained documents, herein called *document-based publications*. This approach provides the benefit that authors can prepare their publications in the way they are used to. Currently, the ORKG supports this publication form, i.e., document-based publications can be added via Digital Object Identifiers (DOIs) or manually to the knowledge graph. Following the addition, contributors can establish links to other entities in the knowledge graph and add the contributions provided in their publication as new entities based on templates.

[2] As prominently stated on the ORKG's homepage[1] and in [14], aspects like compliance to the FAIR data principles are also important for the success of a SciKG.

[3] cf. https://www.orkg.org/orkg/paper/R134713 (accessed 19/10/2021): A paper featuring multiple contributions that are explicitly represented in the ORKG.

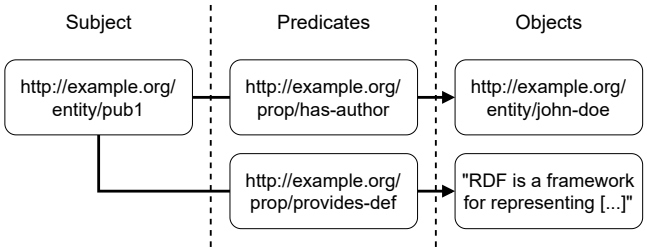

Figure 1: A simple exemplary knowledge graph consisting of two RDF triples. The upper triple provides contextual information, the lower triple contentual information of the publication *pub1*. All non-literal triple members are identified using IRIs.

For indexing and categorization purposes, many research organizations and publishers including the Association for Computing Machinery (ACM), Springer, and Elsevier recommend the usage of keywords in papers. However, these keywords can often be set arbitrarily by the authors, thus impeding their utility for the knowledge graph integration process due to the missing IRIs. To tackle this, more advanced means of classifying research publications are applied. For instance, the ACM promotes the usage of their ACM Computing Classification System (CCS), a poly-hierarchical ontology for classifying research publications in the computing domain, whose current version launched in 2012 [6]. One downside of the CCS is its reliance on proprietary identifiers instead of IRIs which would facilitate connecting to other RDF resources.

To simplify the integration process by closing the gap between document-based publications and knowledge graphs, one option is to leverage OpenIE [5] systems and knowledge graph construction [12] to transform text into knowledge graphs. For the scientific domain, [11] proposes *SciIE*, a unified framework for constructing SciKGs based on scientific literature. Herein, knowledge graphs that result from transforming document-based publications will be referred to as *RDF-transformed publications*. The problem is that knowledge graph construction relies on several to date error-prone Natural Language Processing (NLP) techniques like entity recognition, relation extraction, and coreference resolution, thus yielding sub-optimal results (cf. [11]). Approaches that employ neural networks are able to beat previous approaches (cf. [9]), but the results still remain below the quality required for SciKGs. However, the current interest in NLP and the further investigation of machine learning techniques in this field will eventually result in improved knowledge graph construction approaches, as well. Note that we assume in this paper that knowledge graph construction techniques are able to produce high quality RDF-transformed publications for a fair comparison.

In contrast to traditional document-based publications, another publication form has emerged in recent years, namely *nanopublications* [10]. This publication form consists of one "[...] atomic snippet of a formal statement [...]" [10] accompanied by the origin of this information, also mentioned as provenance, and metadata. The information is formatted as linked data, more precisely as RDF graphs, and thus far mostly used in the Life Science domain. Whereas traditional self-contained documents usually provide accompanying information to the subject such as sections on introduction, related work, and future work among others, nanopublications do not include such contextual information but use links instead.

3 PROBLEM IDENTIFICATION

To conceptualize the requirements of a future publication form tailored for use in SciKGs, it is necessary to first investigate the flaws of the previously described publication forms in this context. For this purpose, three perspectives are considered: The authors' perspective, which represents the group producing research publications. The providers' perspective, which represents the group maintaining the SciKG. The readers' perspective, which represents the group exploring the SciKG.

Authors' perspective. It is common that authors produce publications that overlap to a certain degree. For instance, a researcher in the field of linked data will produce multiple publications that rely on RDF as a concept, or a formal definition of knowledge graphs. For readability, the necessary concepts and definitions must be introduced to the readers in each publication. In the case of document-based publications, this results in passages—typically found in sections called *Related Works* or *Foundations*—that are effectively redundant across multiple papers even though they are often rephrased to avoid plagiarism. As a result, authors are forced to spend valuable time, effort, and space to produce passages that do not provide any contribution[4]. RDF-transformed publications do not mitigate this problem, as they rely on document-based publications prepared in the same manner.

Although nanopublications do not contain redundant information by definition, they require a different amount of time and effort from the author, since their formal nature necessitates the study of the guidelines[5]. The formal structure of nanopublications is a restricting factor in another way, meaning the ontology that is used as its basis. Authors without experience in nanopublications or who study in domains without any entities already present in a system based thereon, would need to construct the definitions, concepts, and other data relevant to their publication, thus creating an ontology on their own, before being able to publish a nanopublication.

Providers' perspective. The problems that arise from a providers' perspective are related to the integration of research publications into the knowledge graph. The integration of new publications requires the identification and addition of entities and relations that are not yet represented in the graph, e.g., contributions the publication provides, as well as the recognition and linkage of entities and relations that are already represented in the graph, e.g., an author who is already part of the graph. Document-based publications do not contain useful information to decide whether a mentioned entity or relation is present in the knowledge graph or not. Of course, many publications today include ORCID identifiers to disambiguate authors or DOI information to identify referenced documents, but they are just an intermediate step to obtain the actual IRIs of entities, which are required for the integration. As already discussed in Section 2, keywords and similar systems like

[4]Note that adjustments made to said concepts and definitions, or contextualizing information provided by the authors are not meant here as they represent actual contributions.

[5]http://nanopub.org/guidelines/working_draft (accessed 19/10/2021)

the CCS do not suffice for properly and conveniently categorizing document-based publications.

In contrast to document-based publications, RDF-transformed publications provide the benefit that mentioned entities are assigned their correct IRIs as long as the entity linking within the knowledge graph construction process succeeds. If necessary, new entities can be created and added to the SciKG, too. Relations are extracted as well such that RDF-transformed publications include them, thus further simplifying the integration process.

Similar to RDF-transformed publications, using nanopublications can improve the integration of new research publications into an existing knowledge graph. However, the infrastructure necessary for the formal setup of nanopublications would have to be offered, maintained and developed by the providers. Furthermore, the effort in establishing such an infrastructure including definitions and other information relating to a domain without prior usage of nanopublications would be substantial, especially when considering the training of authors that should use the system. Advantages, on the other hand, would be significant, enabling filtering according to certain subjects, authors or dates.

Readers' perspective. When readers view a research publication that interests them, they may want to investigate other research publications that are related to the topic. This includes publications that are referenced by the publication at hand as well as publications that reference the publication at hand. Document-based research publications only mention referenced publications, thus only supporting a backwards search. In contrast, a forwards search, i.e., the investigation of referencing publications, is only possible using external tools like Semantic Scholar[6].

RDF-transformed publications allow readers to both view the original document-based publication and the generated knowledge graph, given that a suitable knowledge graph visualization is available. The former provides the familiarity and readability of the traditional publication form, while the latter could be leveraged to implement features like a forwards search. However, readers have the additional effort of viewing two representations of the same publications for different use cases instead of one coherent representation, which is not optimal from a usability point of view.

To the best of our knowledge, a system displaying nanopublications with a mature interface does not exist yet, thus decreasing the usefulness for potential readers, since they would have to use scripts or APIs to query publications. Furthermore, citation figures are not available when compared to the other publication forms[7], which enables a quick estimation of how influential or popular a given paper is, thus presenting a simple filter for the reader.

4 REQUIREMENTS

As shown in Section 3, each publication form considered here has advantages as well as disadvantages regarding their use in SciKGs. Based on our findings, we propose a first set of requirements for a future publication form tailored for use in this context. As explained in Section 2, note that the ontology of the SciKG has to be designed in a way compatible to the future publication form.

[6]https://www.semanticscholar.org (accessed 19/10/2021)

[7]The need for new scholarly communication incentive measures that arises from the shift towards knowledge graph based research is also noted in [1].

Table 1 lists the requirements set; the description is given below. As one can see, all three perspectives are covered by some requirements with respect to the flaws we identified from each perspective. That being said, it is difficult to really provide a clear-cut mapping from the requirements to the affected groups since many requirements somewhat influence multiple perspectives as well as each other. Hence, we assigned each requirement to the groups affected the most.

Table 1: A first set of requirements for a future publication form tailored for use in scientific knowledge graphs. The three columns, (A)uthors', (P)roviders', (R)eaders', indicate the affected perspective(s).

#	Requirement	A	P	R
1	Preparation of main contributions in natural language	×		
2	Import of knowledge	×		
3	Markup of knowledge	×	×	
4	Provision of tooling for obtaining IRIs	×	×	
5	Enriched representation of publications at view time			×

Requirement 1 prescribes that authors shall be able to prepare their main contributions in natural language as they are used to in order to minimize training effort. The idea of nanopublications is interesting, but for a general domain SciKG the limited flexibility caused by the strict ontology does not suffice for properly expressing contributions in all scientific fields. However, for domain specific SciKGs with a narrow scope, e.g., a knowledge graph with focus on empirical studies in the Life Science domain, nanopublications may be a viable option.

Requirement 2 addresses the problem of redundancy across publications. To avoid redundant passages, authors shall be able to import, i.e., link, knowledge that is already present in the knowledge graph within their publications. For this purpose, they shall be able to specify the knowledge they rely on using IRIs where appropriate. For example, if authors rely on a certain definition for RDF that is already present in the SciKG, they shall be able to specify the definition's IRI in their publication. In combination with Requirement 5, this eliminates redundant passages across publications. Of course, changes made to, for example, imported definitions or contextualizing information still have to be explicitly provided by the authors.

Due to Requirement 1, the main part of the future publication form will be written in natural language. As a consequence, it is necessary to either mark up original entities and their relations manually or to leverage knowledge graph construction when an automated integration process is the goal. That being said, the best option may be to use both approaches. To this end, Requirement 3 states that knowledge graph construction techniques shall be applied first to generate suggestions for mentioned entities and their relations. Subsequently, authors shall review and adapt the suggestions if necessary. To mark up RDF elements that are already present in the SciKG, an RDFa[8]-like approach could be used.

[8]https://www.w3.org/TR/rdfa-primer (accessed 19/10/2021)

Despite the suggestions provided using knowledge graph construction, reviewing the suggestions and marking up their publications demands additional effort from the authors. Hence, tooling that facilitates activities like searching for IRIs of mentioned entities shall be provided, which represents Requirement 4.

From compliance with Requirement 2, issues regarding the readability arise since IRIs replace actual text written in natural language. To tackle this, we can make use of the capabilities SciKGs provide: At view time, readers shall be provided a single document-based representation of the publication that is enriched using information currently available in the knowledge graph, constituting Requirement 5. The generated version shall comprise the original text by the authors as well as the natural language fragments that are generated by resolving the links to imported knowledge. To enable backwards and forwards search, clickable IRIs to both referenced publications and referencing publications shall be provided. Depending on the information that is represented in the respective SciKG, other information could be included as well, which is worth exploring.

In summary, the goal is to retain the familiarity of document-based publications while exploiting the opportunities that arise from the use of SciKGs with small additional effort. Requirements 2 and 5 represent a positive aspect in this regard, as they allow saving resources that would otherwise be spent on producing redundant passages. In return, the markup process implies additional effort, thus representing an important aspect for future work.

5 CONCLUSION

This paper provided an overview of publication forms used in SciKGs and subsequently described flaws that arise from their use in this context. Then, we outlined a first set of requirements that a publication form tailored for use in SciKGs should fulfill. The next steps are to refine the requirement set by further investigating the use cases and to design a suitable publication form. Regarding the latter, we are investigating custom commands for LaTeX documents for importing knowledge via IRIs and a preprocessing step for their resolution. In a sense, this approach is similar to the famous Project Xanadu[9] which proposes a transclusion mechanism to include parts of documents in other documents.

REFERENCES

[1] Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria-Esther Vidal. 2018. Towards a Knowledge Graph for Science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, WIMS 2018, Novi Sad, Serbia, June 25-27, 2018*. ACM, 1:1–1:6. https://doi.org/10.1145/3227609.3227689

[2] Arthur Brack, Anett Hoppe, Markus Stocker, Sören Auer, and Ralph Ewerth. 2020. Requirements Analysis for an Open Research Knowledge Graph. In *Digital Libraries for Open Knowledge - 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25-27, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12246)*. Springer, 3–18. https://doi.org/10.1007/978-3-030-54956-5_1

[3] Richard Cyganiak, David Hyland-Wood, and Markus Lanthaler. 2014. RDF 1.1 Concepts and Abstract Syntax. (01 2014). Retrieved July 28, 2021 from https://www.w3.org/TR/rdf11-concepts

[4] Lisa Ehrlinger and Wolfram Wöß. 2016. Towards a Definition of Knowledge Graphs. In *Joint Proceedings of the Posters and Demos Track of the 12th International Conference on Semantic Systems - SEMANTiCS2016 and the 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS'16) co-located with the 12th International Conference on Semantic Systems (SEMANTiCS 2016), Leipzig, Germany, September 12-15, 2016 (CEUR Workshop Proceedings, Vol. 1695)*. CEUR-WS.org. http://ceur-ws.org/Vol-1695/paper4.pdf

[5] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Commun. ACM* 51, 12 (2008), 68–74. https://doi.org/10.1145/1409360.1409378

[6] Association for Computing Machinery (ACM). 2021. *Computing Classification System*. Association for Computing Machinery (ACM). Retrieved July 28, 2021 from https://dl.acm.org/ccs

[7] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutiérrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, Roberto Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan F. Sequeda, Steffen Staab, and Antoine Zimmermann. 2020. Knowledge Graphs. *CoRR* (2020). https://arxiv.org/abs/2003.02320

[8] Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open Research Knowledge Graph: Next Generation Infrastructure for Semantic Scholarly Knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*. ACM, 243–246. https://doi.org/10.1145/3360901.3364435

[9] Tianwen Jiang, Tong Zhao, Bing Qin, Ting Liu, Nitesh V. Chawla, and Meng Jiang. 2019. The Role of "Condition": A Novel Scientific Knowledge Graph Representation and Construction Model. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*. ACM, 1634–1642. https://doi.org/10.1145/3292500.3330942

[10] Tobias Kuhn, Albert Meroño-Peñuela, Alexander Malic, Jorrit H. Poelen, Allen H. Hurlbert, Emilio Centeno, Laura I. Furlong, Núria Queralt-Rosinach, Christine Chichester, Juan M. Banda, Egon L. Willighagen, Friederike Ehrhart, Chris T. A. Evelo, Tareq B. Malas, and Michel Dumontier. 2018. Nanopublications: A Growing Resource of Provenance-Centric Scientific Linked Data. *CoRR* (2018). http://arxiv.org/abs/1809.06532

[11] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. Association for Computational Linguistics, 3219–3232. https://doi.org/10.18653/v1/d18-1360

[12] José-Lázaro Martínez-Rodríguez, Ivan López-Arévalo, and Ana B. Ríos-Alvarado. 2018. OpenIE-based approach for Knowledge Graph construction from text. *Expert Syst. Appl.* 113 (2018), 339–355. https://doi.org/10.1016/j.eswa.2018.07.017

[13] Allard Oelen, Mohamad Yaser Jaradeh, Kheir Eddine Farfar, Markus Stocker, and Sören Auer. 2019. Comparing Research Contributions in a Scholarly Knowledge Graph. In *Proceedings of the Third International Workshop on Capturing Scientific Knowledge co-located with the 10th International Conference on Knowledge Capture (K-CAP 2019), Marina del Rey, California , November 19th, 2019 (CEUR Workshop Proceedings, Vol. 2526)*. CEUR-WS.org, 21–26.

[14] Allard Oelen, Mohamad Yaser Jaradeh, Markus Stocker, and Sören Auer. 2020. Generate FAIR Literature Surveys with Scholarly Knowledge Graphs. In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*. ACM, 97–106. https://doi.org/10.1145/3383583.3398520

[15] Lars Vogt, Jennifer D'Souza, Markus Stocker, and Sören Auer. 2020. Toward Representing Research Contributions in Scholarly Knowledge Graphs Using Knowledge Graph Cells. In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*. ACM, 107–116. https://doi.org/10.1145/3383583.3398530

[9] https://xanadu.com (accessed 19/10/2021)