an open access journal

Citation: Manghi, P., Mannocci, A., Osborne, F., Sacharidis, D., Salatino, A., & Vergoulis, T. (2021). New trends in scientific knowledge graphs and research impact assessment. *Quantitative Science Studies*, *2*(4), 1296–1300. https://doi.org/10.1162/qss_e_00160

DOI:
https://doi.org/10.1162/qss_e_00160

Corresponding Author:
Angelo Salatino
angelo.salatino@open.ac.uk

The MIT Press

EDITORIAL

New trends in scientific knowledge graphs and research impact assessment

Paolo Manghi[1,5], Andrea Mannocci[1], Francesco Osborne[2], Dimitris Sacharidis[3], Angelo Salatino[2], and Thanasis Vergoulis[4]

[1]CNR-ISTI – National Research Council, Institute of Information Science and Technologies “Alessandro Faedo”, Pisa, Italy
[2]Knowledge Media Institute, The Open University, Milton Keynes, United Kingdom
[3]Université Libre de Bruxelles, Belgium
[4]“Athena” RC, Greece
[5]OpenAIRE AMKE, Greece

Keywords: bibliographic data, knowledge graphs, research dynamics, research impact, scholarly data

1. INTRODUCTION

In recent decades, we have experienced a continuously increasing publication rate of scientific articles and related research objects (e.g., data sets, software packages). As this trend keeps growing, practitioners in the field of scholarly knowledge are confronted with several challenges. In this special issue, we focus on two major categories of such challenges: (a) those related to the organization of scholarly data to achieve a flexible, context-sensitive, fine-grained, and machine-actionable representation of scholarly knowledge that at the same time is structured, interlinked, and semantically rich, and (b) those related to the design of novel, reliable, and comprehensive metrics to assess scientific impact.

To address the challenges of the first category, new technical infrastructures are becoming increasingly popular, organizing and representing scholarly knowledge through scientific knowledge graphs (SKG). These are large networks describing the actors (e.g., authors, organizations), the documents (e.g., publications, patents), and other research outputs (e.g., research data, software) and knowledge (e.g., research topics, concepts, tasks, technologies) in this space as well as their reciprocal relationships. These resources provide substantial benefits to researchers, companies, and policymakers by powering several data-driven services for navigating, analyzing, and making sense of research dynamics. Some examples include Microsoft Academic Graph (MAG) (Sinha, Shen et al., 2015), AMiner (Tang, Zhang et al., 2008), Open Academic Graph (Sinha et al., 2015; Tang et al., 2008), ScholarlyData.org (Nuzzolese, Gentile et al., 2016), Semantic Scholar (Ammar, Groeneveld et al., 2018), PID Graph (Fenner & Aryani, 2019), Open Research Knowledge Graph (Jaradeh, Oelen et al., 2019), OpenCitations (Peroni, Shotton, & Vitali, 2017), and the OpenAIRE research graph (Manghi, Atzori et al., 2019). Despite their popularity, the field of SKGs has a lot of open challenges, such as the design of ontologies able to conceptualize scholarly knowledge, model its representation, and enable its exchange across different SKGs; the extraction of entities and concepts, integration of information from heterogeneous sources, identification of duplicates, finding connections between entities, and identifying conceptual inconsistencies; and the development of services that exploit knowledge as provided by one or more SKGs to discover, monitor, measure, and consume research outcomes (Aryani, Fenner et al., 2020; Auer, 2018).

With regard to the second category, we seek effective and precise research assessment. In this context, there is a need for reliable and comprehensive metrics and indicators of the impact and merit of publications, data sets, research institutions, individual researchers, and other relevant entities. Research impact refers to the attention a research work receives inside its respective and related disciplines (Kanellos, Vergoulis et al., 2019), the social/mass media (Galligan & Dyas-Correia, 2013), and so on. A research work's merit, on the other hand, is relevant to its quality aspects (e.g., its novelty, reproducibility, compliance with the Findable, Accessible, Interoperable, Reusable [FAIR] initiative for promoting data discovery and reuse, and readability). Nowadays, due to the growing popularity of Open Science initiatives, a large number of useful science-related data sets have been made openly available, paving the way for the synthesis of more sophisticated research impact and merit indicators (and, consequently, more precise research assessment). For instance, in recent years, due to the systematic effort of various developing teams, a variety of large SKGs has been made available, providing a very rich and relatively clean source of information about academics, their publications, and relevant metadata that can be used for the development of effective research assessment approaches (Chatzopoulos, Vergoulis et al., 2020).

The proposal for this special issue originated from the collaboration of two workshops, the *Scientific Knowledge Graphs Workshop* (*SKG 2020*), and the *Workshop on Assessing Impact and Merit in Science* (*AIMinScience 2020*), held (virtually) in conjunction with the 2020 edition of the *International Conference on Theory and Practice of Digital Libraries* (*TPDL*) on August 25, 2020. SKG 2020 offered a forum to discuss about the themes surrounding the first set of challenges, namely methods for extracting entities and relationships from research publications; data models for the description of scholarly data; methods for the exploration, retrieval, and visualization of scientific knowledge graphs, and applications for making sense of scholarly data. On the other hand, AIMinScience 2020 focused on the second set of challenges, which include scientometrics and bibliometrics; applications utilizing scientific impact and merit to provide useful services to the research community and industry; data mining and machine learning approaches to facilitate research assessment; and insightful visualization techniques that utilize or facilitate research assessment.

Given that the themes of both workshops are interlinked, because SKGs can indeed support research impact assessment, it was a joint decision to edit this special issue on *Scientific Knowledge Graphs and Research Impact Assessment*, with the aim of providing all practitioners interested in the scholarly knowledge with the current advances of these particular aspects. In addition, this collaboration catalyzed the creation of the *International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment*[1] (Sci-K), a new joint event that replaced SKG and AIMinScience, focusing on a wider subject and audience. Sci-K aims to explore innovative solutions and ideas for the generation of approaches, data models, and infrastructures (e.g., knowledge graphs), for supporting, directing, monitoring, and assessing scientific knowledge. Its first edition, Sci-K 2021, was held on April 13, 2021, co-organized with The Web Conference 2021[2]. It was a successful event with 11 presented papers and two keynote talks from Prof. Ludo Waltman and Prof. Staša Milojević.

2. CONTRIBUTIONS TO THE SPECIAL ISSUE

This special issue includes 10 contributions, equally balanced between advances on SKGs and research impact assessment. The papers in the first category introduce several innovative knowledge graphs that enrich classic metadata about articles, patents, and software with further information for exploring these documents more efficiently, identifying insights, and creating more comprehensive analyses of research trends. The articles on impact assessment

[1] Sci-K: https://sci-k.github.io/2021/

[2] The Web Conference 2021: https://www2021.thewebconf.org/

propose new approaches for key challenges in this field, such as modeling the evolution of credit over time, citing data sets, analyzing research trends on social networks, and predicting citation-based popularity. The contributions address a variety of scientific domains, including computer science, phenomenon-oriented studies, opioids, and COVID-19. In the follow, we briefly summarize each contribution.

Menin, Michel et al. (2021) introduce Covid-on-the-Web, a tool that assists users in accessing, querying, and sense making of COVID-19-related literature. In this effort, the authors first built a knowledge graph from the "COVID-19 Open Research Dataset" (Lu Wang, Lo et al., 2020), and then enriched it using entity linking and argument mining, finally providing an interface, the "Linked Data Visualizer" (LDViz), which assists the querying and visual exploration of the referred data set.

Färber and Lamprecht (2021) introduce the Data Set Knowledge Graph (DSKG), describing the metadata of data sets for all scientific disciplines. In this knowledge graph, data sets are connected to the relevant articles, modeled in Microsoft Academic Graph (Sinha et al., 2015). DSKG is then further enriched with ORCID IDs and Wikidata.

Angioni, Salatino et al. (2021) introduce the Academia/Industry DynAmics (AIDA) Knowledge Graph, which is generated with an automatic pipeline integrating data from MAG, Dimensions, English DBpedia, GRID, and Computer Science Ontology (CSO) (Salatino, Thanapalasingam et al., 2018). Currently, AIDA describes 21 million publications and 8 million patents according to the research topics drawn from CSO. In addition, 5.1 million publications and 5.6 million patents are also characterized according to the type of the authors, affiliations (e.g., academia, industry) and 66 industrial sectors (e.g., automotive, financial, energy, electronics).

Buneman, Dosso et al. (2021) focus on two key challenges regarding citation graphs. The first is that citation graphs do not appropriately model the evolution of credit over time—for instance, when credit is assigned to the different versions of the same scientific work (preprint and peer reviewed). Usually, authors prefer that the citations of all versions receive are merged. The second challenge is the ability to cite data sets as a whole (single object) and also their constituents. To tackle these challenges, the authors suggest an extension of the current citation graph model, based on citable units and reference subsumption, which will improve the current practices for bibliometric computations.

Kelley and Garijo (2021) present SOftware Metadata Extraction Framework (SOMEF), an approach to automatically extract scientific software metadata from its documentation, and specifically from the readme file. Next, they propose a methodology for structuring the extracted metadata within a knowledge graph of scientific software. Finally, they also provide a tool for browsing and comparing the contents of the generated knowledge graph.

On the other hand, with regard to research impact assessment, Vergoulis, Kanellos et al. (2021) introduce BIP4COVID19, an open data set that offers a variety of impact measures for coronavirus-related scientific articles. These measures can be exploited for the creation or extension of added-value services aiming to facilitate the exploration of the respective literature. In the same context, as a use case, they also provide a publicly accessible keyword-based search interface for COVID-19-related articles, which leverages BIP4COVID19 data to rank search results according to the calculated impact indicators.

Rothenberger, Pasta, and Mayerhoffer (2021) present an approach to analyze and measure the impact of phenomenon-oriented research fields. Specifically, they analyzed the field of migration research, which focuses on conceptualizing, capturing, and documenting an observed phenomenon (i.e., migration). In this analysis, to measure impact within such fields, the authors set up a framework to acknowledge scientific merit using a novel sophisticated citation factor.

Haunschild, Bornmann et al. (2021) investigate which topics in opioid scholarly publications have received public attention on Twitter. The authors generate topic networks (i.e., networks of co-occurring author keywords), from both the tweets and from the publications that are tweeted by the accounts. The results showed that Twitter users tend to use more generic terms compared to those used within publications.

Ghosal, Tiwary et al. (2021) proposed an automatic method that identifies significant citations, and then developed an approach to trace the lineage of given research via transitively identifying the significant citations to a given article. This approach can improve the retrievability of relevant literature, as well as finding the true influence of a given work in the scientific community beyond citation counts.

Finally, Chatzopoulos, Vergoulis et al. (2021) focus on the problem of estimating the expected citation-based popularity (or short-term impact) of papers. State-of-the-art methods for this problem attempt to leverage the current citation data of each paper. However, these methods are prone to inaccuracies for recently published papers, which have a limited citation history. In this context, the authors introduce ArtSim+, an approach that can be applied on top of any popularity estimation method to improve its accuracy, providing more accurate estimations for the most recently published papers by considering the popularity of similar and older ones.

ACKNOWLEDGMENTS

Many people supported us in producing this special issue and we would like to take this opportunity to thank them. First of all, we would like to thank all the reviewers for their outstanding reviews (listed in alphabetical order): Mehwish Alam, Alberto Baccini, Alessia Bardi, Nikos Bikakis, Arthur Brack, Davide Buscaldi, Rodrigo Costas, Marilena Daquino, Danilo Dessì, Angelo Di Iorio, Michael Färber, Patricia Feeney, Marcia Ferreira, Vladimir A. Fomichov, Yannis Foufoulas, Daniel Garijo, Tirthankar Ghosal, Drahomira Herrmannova, Ilias Kanellos, Dimitrios Katsaros, Barbara S. Lancho-Barrantes, Shubhanshu Mishra, Christos Papatheodorou, Andres Pavas, Giovanni Pizzi, David Pride, Paraskevi Raftopoulou, Diego Reforgiato, Daniele Rotolo, Jodi Schneider, Anna Severin, Dimitrios Skoutas, Markus Stocker, Vincent Antonio Traag, Christos Tryfonopoulos, Nikolay Yakovets. We also would like to thank Professor Ludo Waltman for welcoming this special issue to *Quantitative Science Studies* and for his excellent guidance throughout the whole process. Last but not least, we are grateful to the Technische Informationsbibliothek (TIB)—Leibniz Information Centre for Science and Technology for partially covering the APCs of the papers published in this special issue.

REFERENCES

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., … Etzioni, O. (2018). Construction of the literature graph in semantic scholar. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 3 (pp. 84–91). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-3011

Angioni, S., Salatino, A., Osborne, F., Recupero, D. R., & Motta, E. (2021). AIDA: A knowledge graph about research dynamics in academia and industry. *Quantitative Science Studies*, *2*(4), 1356–1398. https://doi.org/10.1162/qss_a_00162

Aryani, A., Fenner, M., Manghi, P., Mannocci, A., & Stocker, M. (2020). Open science graphs must interoperate! In L. Bellatreche et al. (Eds.), *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium* (pp. 195–206). Cham: Springer. https://doi.org/10.1007/978-3-030-55814-7_16

Auer, S. (2018). Towards an open research knowledge graph. *Zenodo*. https://doi.org/10.5281/zenodo.1157185

Buneman, P., Dosso, D., Lissandrini, M., & Silvello, G. (2021). Data citation and the citation graph. *Quantitative Science Studies*, *2*(4), 1399–1422. https://doi.org/10.1162/qss_a_00166

Chatzopoulos, S., Vergoulis, T., Kanellos, I., Dalamagas, T., & Tryfonopoulos, C. (2020). ArtSim: Improved estimation of current impact for recent articles. In *ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium* (pp. 323–334). Cham: Springer. https://doi.org/10.1007/978-3-030-55814-7_27

Chatzopoulos, S., Vergoulis, T., Kanellos, I., Dalamagas, T., & Tryfonopoulos, C. (2021). Further improvements on estimating the popularity of recently published papers. *Quantitative Science Studies*, *2*(4), 1529–1550. https://doi.org/10.1162/qss_a_00165

Fenner, M., & Aryani, A. (2019). Introducing the PID graph. https://doi.org/10.5438/JWVF-8A66

Färber, M., & Lamprecht, D. (2021). The data set knowledge graph: Creating a linked open data source for data sets. *Quantitative Science Studies*, *2*(4), 1324–1355. https://doi.org/10.1162/qss_a_00161

Galligan, F., & Dyas-Correia, S. (2013). Altmetrics: Rethinking the way we measure. *Serials Review*, *39*(1), 56–61. https://doi.org/10.1080/00987913.2013.10765486

Ghosal, T., Tiwary, P., Patton, R., & Stahl, C. (2021). Towards establishing a research lineage via identification of significant citations. *Quantitative Science Studies*, *2*(4), 1511–1528. https://doi.org/10.1162/qss_a_00170

Haunschild, R., Bornmann, L., Potnis, D., & Tahamtan, I. (2021). Investigating the dissemination of scientific information on Twitter: A study of topic networks in opioid publications. *Quantitative Science Studies*, *2*(4), 1486–1510. https://doi.org/10.1162/qss_a_00168

Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., ... Auer, S. (2019). Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture* (pp. 243–246). New York: Association for Computing Machinery. https://doi.org/10.1145/3360901.3364435

Kanellos, I., Vergoulis, T., Sacharidis, D., Dalamagas, T., & Vassiliou, Y. (2019). Impact-based ranking of scientific publications: A survey and experimental evaluation. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1567–1584. https://doi.org/10.1109/TKDE.2019.2941206

Kelley, A., & Garijo, D. (2021). A framework for creating knowledge graphs of scientific software metadata. *Quantitative Science Studies*, *2*(4), 1423–1446. https://doi.org/10.1162/qss_a_00167

Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., ... Kohlmeier, S. (2020). CORD-19: The Covid-19 Open Research Dataset. *ArXiv*, arXiv:2004.10706v2. Retrieved from https://pubmed.ncbi.nlm.nih.gov/32510522

Manghi, P., Atzori, C., Bardi, A., Schirrwagen, J., Dimitropoulos, H., ... Summan, F. (2019). OpenAIRE Research Graph Dump. *Zenodo*. https://doi.org/10.5281/zenodo.3516918

Menin, A., Michel, F., Gandon, F., Gazzotti, R., Cabrio, E., ... Winckler, M. (2021). Covid-on-the-Web: Exploring the COVID-19 scientific literature through visualization of linked data from entity and argument mining. *Quantitative Science Studies*, *2*(4), 1301–1323. https://doi.org/10.1162/qss_a_00164

Nuzzolese, A. G., Gentile, A. L., Presutti, V., & Gangemi, A. (2016). Conference linked data: The Scholarly Data project. In P. Groth et al. (Eds.), *The semantic web – ISWC 2016* (pp. 150–158). Cham: Springer. https://doi.org/10.1007/978-3-319-46547-0_16

Peroni, S., Shotton, D., & Vitali, F. (2017). One year of the OpenCitations Corpus. In C. d'Amato et al. (Eds.), *The semantic web – ISWC 2017* (pp. 184–192). Cham: Springer. https://doi.org/10.1007/978-3-319-68204-4_19

Rothenberger, L., Pasta, M. Q., & Mayerhoffer, D. (2021). Mapping and impact assessment of phenomenon-oriented research fields: The example of migration research. *Quantitative Science Studies*, *2*(4), 1466–1485. https://doi.org/10.1162/qss_a_00163

Salatino, A. A., Thanapalasingam, T., Mannocci, A., Osborne, F., & Motta, E. (2018). The computer science ontology: A large-scale taxonomy of research areas. In D. Vrandečić et al. (Eds.), *The semantic web – ISWC 2018* (pp. 187–205). Cham: Springer. https://doi.org/10.1007/978-3-030-00668-6_12

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., ... Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 243–246). New York: Association for Computing Machinery. https://doi.org/10.1145/2740908.2742839

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 990–998). New York: Association for Computing Machinery. https://doi.org/10.1145/1401890.1402008

Vergoulis, T., Kanellos, I., Chatzopoulos, S., Karidi, D. P., & Dalamagas, T. (2021). BIP4COVID19: Releasing impact measures for articles relevant to COVID-19. *Quantitative Science Studies*, *2*(4), 1447–1465. https://doi.org/10.1162/qss_a_00169