an open access journal

Citation:

DOI:
Supporting Information:
Received:
Accepted:
Published:

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:

Handling Editor:

The MIT Press

EDITORIAL

New Trends in Scientific Knowledge Graphs and Research Impact Assessment

Paolo Manghi[1,5] and Andrea Mannocci[1] and Francesco Osborne[2] and Dimitris Sacharidis[3] and 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: Scholarly Data, Knowledge Graphs, Research Impact, Bibliographic Data, Research Dynamics

1 INTRODUCTION

In the last decades, we 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: i) those related to the organisation 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 ii) 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, organising and representing scholarly knowledge through Scientific Knowledge Graphs (SKG). They are large networks describing the actors (e.g., authors, organisations), 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, analysing, and making sense of research dynamics. Some examples include Microsoft Academic Graph (MAG) (Sinha et al., 2015), AMiner (Tang et al., 2008), Open Academic Graph (Sinha et al., 2015; Tang et al., 2008), ScholarlyData.org (Nuzzolese, Gentile, Presutti, & Gangemi, 2016), Semantic Scholar (Ammar et al., 2018), PID Graph (Fenner & Aryani, 2019), Open Research Knowledge Graph (Jaradeh et al., 2019), OpenCitations (Peroni, Shotton, & Vitali, 2017), and the OpenAIRE research graph (Manghi et al., 2019). Despite their popularity, the field of SKGs has a lot of open challenges, such as: i) the design of ontologies able to conceptualise scholarly knowledge, model its representation, and enable its exchange across different SKGs, ii) the extraction of entities and concepts, integration of information from heterogeneous sources, identification of duplicates, finding connections between entities, identifying conceptual inconsistencies, and iii) the development of services that exploit knowledge as provided by one or more SKGs to discover, monitor, measure, and consume research outcomes (Aryani, Fenner, Manghi, Mannocci, & Stocker, 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, Sacharidis, Dalamagas, & Vassiliou, 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, FAIR-ness, readability). Nowadays, due to the growing popularity of Open Science initiatives, a large amount 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 late 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, Kanellos, Dalamagas, & Tryfonopoulos, 2020).

The proposal of 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 the 25th of August 2020. SKG 2020 offered a forum to discuss about the themes surrounding the first set of challenges namely i) methods for extracting entities and relationships from research publications, ii) data models for the description of scholarly data, iii) methods for the exploration, retrieval and visualisation of scientific knowledge graphs and iv) applications for making sense of scholarly data. On the other hand, AIMinScience 2020 focused on the second set of challenges which include i) scientometrics and bibliometrics, ii) applications utilising scientific impact and merit to provide useful services to the research community and the industry, iii) data mining and machine learning approaches to facilitate research assessment, and iv) insightful visualisation techniques that utilise or facilitate research assessment.

Given that the themes of both workshops are interlinked, because scientific knowledge graph can indeed support the 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 catalysed 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 13th of 2021, co-organised 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 ten contributions, equally balanced between advances on Scientific Knowledge Graphs and Research Impact Assessment. The papers in the first category introduced several innovative knowledge graphs that enrich classic metadata about articles, patents, and software with further information for exploring more efficiently these documents, identifying insights,

[1] Sci-K: https://sci-k.github.io/2021/
[2] The Web Conference 2021: https://www2021.thewebconf.org/

and creating more comprehensive analyses of research trends. The articles on impact assessment proposed new approaches for key challenges in this field such as modelling the evolution of credit over time, citing datasets, analysing research trends on social networks, and predicting citation-based popularity. The contributions addressed a variety of scientific domains including computer science, phenomenon-oriented studies, opioids, and COVID-19. In the follow, we briefly summarise each contribution.

Menin 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 "COVID-19 Open Research Dataset" (Lu Wang et al., 2020), then they enriched it using entity linking and argument mining, and finally provided an interface, the "Linked Data Visualizer" (LDViz), which assists the querying and visual exploration of the referred dataset.

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

Angioni, Salatino, Osborne, Recupero, and Motta (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 the Computer Science Ontology (CSO) (Salatino, Thanapalasingam, Mannocci, Osborne, & Motta, 2018). Currently, AIDA describes 21M publications and 8M patents according to the research topics drawn from CSO. In addition, 5.1M publications and 5.6M patents are also characterised according to the type of the author's affiliations (e.g., academia, industry) and 66 industrial sectors (e.g., automotive, financial, energy, electronics).

Buneman, Dosso, Lissandrini, and Silvello (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 (pre-print and peer-reviewed). Usually, authors prefer that the citations all versions receive are merged. The second challenge is the ability to cite datasets as a whole (single object) and also its 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, Chatzopoulos, Karidi, and Dalamagas (2021) introduce BIP4COVID19, an open dataset 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 analyse and measure the impact of phenomenon-oriented research fields. Specifically, they analysed the field of migration research, which focuses on conceptualising, 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, Potnis, and Tahamtan (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 on Twitter users tend to use more generic terms compared to the ones used within publications.

Ghosal, Tiwary, Patton, and Stahl (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, Kanellos, Dalamagas, and Tryfonopoulos (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.

ACKNOWLEDGEMENTS

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 the Quantitative Science Studies Journal 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., Downey, D., ... Etzioni, O. (2018). Construction of the literature graph in semantic scholar.

Angioni, S., Salatino, A. 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), YY–ZZ. doi: 10.1162/qss_a_XXXXX

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 International Publishing.

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

Buneman, P., Dosso, D., Lissandrini, M., & Silvello, G. (2021). Data citation and the citation graph. *Quantitative Science Studies*, *2*(4), YY–ZZ. doi: 10.1162/qss_a_XXXXX

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).

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), YY–ZZ. doi: 10.1162/qss_a_XXXXX

Fenner, M., & Aryani, A. (2019). Introducing the pid graph. doi: 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), YY–ZZ. doi: 10.1162/qss_a_XXXXX

Galligan, F., & Dyas-Correia, S. (2013). Altmetrics: Rethinking the way we measure. *Serials review*, *39*(1), 56–61.

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

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), YY–ZZ. doi: 10.1162/qss_a_XXXXX

Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., Kismihók, G., … Auer, S. (2019). Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th international conference on knowledge capture* (p. 243–246). New York, NY, USA: Association for Computing Machinery. doi: 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*.

Kelley, A., & Garijo, D. (2021). A framework for creating knowledge graphs of scientific software metadata. *Quantitative Science Studies*, *2*(4), YY–ZZ. doi: 10.1162/qss_a_XXXXX

Lu Wang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., … 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., La Bruzzo, S., … Summan, F. (2019). *Openaire research graph dump*. Zenodo. doi: 10.5281/zenodo.3516918

Menin, A., Michel, F., Gandon, F., Gazzotti, R., Cabrio, E., Corby, O., … 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), YY–ZZ. doi: 10.1162/qss_a_XXXXX

Nuzzolese, A. G., Gentile, A. L., Presutti, V., & Gangemi, A. (2016). Conference linked data: The scholarlydata project. In P. Groth et al. (Eds.), *The semantic web – iswc 2016* (pp. 150–158). Cham: Springer International Publishing.

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 International Publishing.

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), YY–ZZ. doi: 10.1162/qss_a_XXXXX

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 Int. Publishing.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., & Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web* (p. 243–246). New York, NY, USA: Association for Computing Machinery. doi: 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* (p. 990–998). New York, NY, USA: Association for Computing Machinery. doi: 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), YY–ZZ. doi: 10.1162/qss_a_XXXXX