Data Science 2 (2019) 177–179
DOI 10.3233/DS-190023
IOS Press

Editorial: Special Issue on Scholarly Data Analysis (Semantics, Analytics, Visualisation)

Alejandra Gonzalez-Beltran [a], Francesco Osborne [b], Silvio Peroni [c] and Sahar Vahdati [d]
[a] *Scientific Computing Department, Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Campus, Didcot, OX11 0QX, United Kingdom*
E-mail: alejandra.gonzalez-beltran@stfc.ac.uk; ORCID: https://orcid.org/0000-0003-3499-8262
[b] *Knowledge Media Institute, The Open University, Walton Hall, MK7 6AA, United Kingdom*
E-mail: francesco.osborne@open.ac.uk; ORCID: https://orcid.org/0000-0001-6557-3131
[c] *Digital Humanities Advanced Research Centre (DHARC), Department of Classical Philology and Italian Studies, University of Bologna, Via Zamboni 32, 40126, Italy*
E-mail: silvio.peroni@unibo.it; ORCID: https://orcid.org/0000-0003-0530-4305
[d] *Department of Computer Science, University of Oxford, Parks Rd, OX1 3QD, United Kingdom*
E-mail: sahar.vahdati@cs.ox.ac.uk; ORCID: https://orcid.org/0000-0002-7171-169X

Abstract. The increasing interest in analysing, describing, and improving the research process requires the development of new forms of scholarly data publication and analysis that integrates lessons and approaches from the field of Semantic Technologies, Science of Science, Digital Libraries, and Artificial Intelligence. This editorial summarises the content of the Special Issue on Scholarly Data Analysis (Semantics, Analytics, Visualisation), which aims to showcase some of the most interesting research efforts in the field. This issue includes an extended version of the best papers of the last two editions of the "Semantics, Analytics, Visualisation: Enhancing Scholarly Dissemination" (SAVE-SD 2017 and 2018) workshop at The Web Conference.

Keywords: Semantic publishing, visual analytics, ontologies, data mining, scholarly data

In recent years, we have seen an increasing interest in analysing, describing, and improving the research process with the aim of tackling the reproducibility crisis [6], informing research policies [2], enriching representation of research papers semantically [7], identifying promising research directions [9], monitoring the evolution of research topics [4], and, ultimately, accelerating the scientific progress [3]. These efforts require the development and publishing of new forms of scholarly artifacts and their analysis which integrates lessons and approaches from the field of semantic technologies, science of science, digital libraries, and artificial intelligence.

While nowadays mostly set on the digital world, the document-based nature of current scholarly communication comes from the long-running practice followed by scholars of exchanging knowledge in form of paper-based articles. Therefore, most of the information is still stored in textual form, or unstructured artifacts, such as PDF files. Describing research knowledge and outputs in a more machine-actionable way can revolutionise how we perform and communicate research, but it is still an open challenge. Indeed, while several stakeholders (e.g., publishers, libraries, project managers, researchers) are moving towards providing services in modern technologies such as scholarly knowledge graphs, in practice, we are still far from the knowledge-based communication vision stated by Vannevar Bush in

his influential 1945 essay "As We May Think" [1]. Primary meta-research tasks with complex combination of metadata such as tracing citations back or forward, retrieving publications of a particular topic with a focus on a particular component e.g., methodology or evaluation, exploration of relevant research groups, grants, and other research-related artifacts are still time consuming with no totally satisfactory results.

The latest web technologies – the Semantic Web, Knowledge Graphs, Question Answering, Management Systems and Recommendation-based services – combined with Artificial Intelligence are expected to provide the required pillars for addressing the aforementioned issues. We are in a stage where scholarly data are being increasingly released as Linked Open Data (e.g., ScholarlyData [5] and OpenCitations [8]) by using Semantic Web formats that facilitate the integration and connection between different repositories. In particular, seminal works suggest to push the on-going digital publishing revolution towards its natural next step, i.e., Semantic Publishing [10,11]: the semantic enhancement of scholarly data, their publication on the Web, and the creation of novel and smart services built on top of them. In order to push the current state of research on scholarly data to the next level, we need to broaden the range of technologies to be explored, e.g.: applying Web data mining techniques to support better use of scholarly data; applying more thorough and large scale analytics to understand the pattern of data usage; and, finally, exploring personalisation and analysing the user's behaviour in interacting with these data to provide the best user experience. The realisation of such vision requires community involvement in facilitating the creation and curation of scholarly metadata and of scholarly knowledge bases.

These topics have attracted high priority attention from international funding councils and community forums, such as FORCE11 and the Research Data Alliance (RDA). Supporting open and accessible data is one of the major themes in the EU Horizon2020 programme and the US National Institutes of Health Big Data to Knowledge (BD2K) initiative. It also attracted interest from industrial companies. We already see the provision of scholarly data as a service by area leaders like Thomson Reuters, Google, Microsoft and Elsevier, support for data publication and preservation (e.g., Figshare, Dryad, Zenodo), offering of alternative metrics for assessing academic performance (e.g., Altmetrics), and many other innovative services based on research data (e.g., Index Labs, Linknovate). This interest from the business world presents an unprecedented opportunity for rapidly transforming academic knowledge into practice and achieving the data-driven science and innovation that is promised by the Big Data era.

This special issue aims to showcase some of the most recent and interesting research efforts in the field and start a conversation about the next steps that we will have to take as scientific community in the next few years. This special issue contains the extended version of the best selected papers of the last two editions of the "Semantics, Analytics, Visualisation: Enhancing Scholarly Dissemination" (SAVE-SD) workshop at the Web Conference. The aim of SAVE-SD was to bring together publishers, technology companies and researchers from different fields (including Document and Knowledge Engineering, Semantic Web, Natural Language Processing, Scholarly Communication, Bibliometrics, Human-Computer Interaction, Information Visualisation, Bioinformatics, and Life Sciences) in order to bridge the gap between the theoretical/academic and practical/industrial aspects in regards to scholarly data.

The special issue consists in four contributions.

In *"Geographical trends in academic conferences: an analysis on authors' affiliations"*, the authors present a study on conference proceedings that illustrates insightful geographical trends and highlights the unbalanced growth of competitive research institutions worldwide in the 1996–2016 period. The analysis shows that the annual and overall turnover rate of country rankings is extremely low and steadily declines over time, suggesting an alarmingly static landscape in which new entries struggle to emerge.

In *"Enabling text search on SPARQL-endpoints through OSCAR"*, the authors introduce the latest version of OSCAR (Version 2.0), the OpenCitations RDF Search Application, which has several improved features and extends the query workflow comparing with the previous version. OSCAR is a user-friendly search platform that can be used to search any RDF triplestore providing a SPARQL endpoint, while hiding the complexities of SPARQL, making the searching operations operable by those who are not experts in Semantic Web technologies with required backgrounds.

In *"Identifying PIDs playing FAIR"*, the author presents a view on the findability and interoperability of PIDs and their compliance with the FAIR data principles. The paper discusses the failure to use PIDs for citation and suggest to add context and meaning to PIDs, making them easier to identify through namespace prefixes and object types.

Finally, in *"Towards a scientific data framework to support scientific model development"*, the authors present an innovative framework to automatically take advantage of large amounts of scientific data extracted from the literature for supporting research and specifically scientific model development. The paper discusses a set of use cases in the field of combustion kinetics and a prototype of the framework that can be generalised to other domains.

We would like to thank all the people that contributed to this special issue. We thank the Editors-in-Chief, Michel Dumontier and Tobias Kuhn, which gave us the opportunity to run this special issue and supported us during the whole process. We thank all the authors for their insightful contributions that represent an excellent example of the efforts that are taking place in this field. Finally, we would like to thank all the reviewers for their suggestions and discussions that contributed greatly to improve the quality of this issue: Riccardo Albertoni, Phil Archer, Aliaksandr Birukou, Simon Cox, Hanna Ćwiek-Kupczyńska, Scott Edmunds, Patricia Feeney, Jeff Grethe, Laurel L. Haak, John Kunze, Cameron Neylon, Eric Prud'hommeaux, Stian Soiland-Reyes, Sarala Wimalaratne and anonymous reviewers.

References

[1] V. Bush, As we may think, *The Atlantic Monthly* **176**(1) (1945), 101–108. https://www.ias.ac.in/article/fulltext/reso/005/11/0094-0103.

[2] B. Daniel, Big data and analytics in higher education: Opportunities and challenges, *British journal of educational technology* **46**(5) (2015), 904–920. doi:10.1111/bjet.12230.

[3] S. Fortunato, C.T. Bergstrom, K. Börner, J.A. Evans, D. Helbing, S. Milojević, A.M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi et al., Science of science, *Science* **359**(6379) (2018), eaao0185. doi:10.1126/science.aao0185.

[4] A. Mannocci, F. Osborne and E. Motta, The evolution of IJHCS and CHI: A quantitative analysis, *International Journal of Human–Computer Studies* **131** (2019), 23–40. doi:10.1016/j.ijhcs.2019.05.009.

[5] A.G. Nuzzolese, A.L. Gentile, V. Presutti and A. Gangemi, Conference linked data: The ScholarlyData project, in: *International Semantic Web Conference*, Springer, 2016, pp. 150–158. doi:10.1007/978-3-319-46547-0_16.

[6] R. Peng, The reproducibility crisis in science: A statistical counterattack, *Significance* **12**(3) (2015), 30–32. doi:10.1111/j.1740-9713.2015.00827.x.

[7] S. Peroni, F. Osborne, A.D. Iorio, A.G. Nuzzolese, F. Poggi, F. Vitali and E. Motta, Research articles in simplified HTML: A web-first format for HTML-based scholarly articles, *PeerJ Computer Science* **3** (2017), e132. doi:10.7717/peerj-cs.132.

[8] S. Peroni, D. Shotton and F. Vitali, One year of the OpenCitations Corpus, in: *International Semantic Web Conference*, Springer, 2017, pp. 184–192. doi:10.1007/978-3-319-68204-4_19.

[9] A.A. Salatino, F. Osborne and E. Motta, AUGUR: Forecasting the emergence of new research topics, in: *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, ACM, 2018, pp. 303–312. doi:10.1145/3197026.3197052.

[10] D. Shotton, Semantic publishing: The coming revolution in scientific journal publishing, *Learned Publishing* **22**(2) (2009), 85–94. doi:10.1087/2009202.

[11] S. Vahdati, A. Dimou, C. Lange and A. Di Iorio, Semantic publishing challenge: Bootstrapping a value chain for scientific data, in: *International Workshop on Semantic, Analytics, Visualization*, Springer, 2016, pp. 73–89. doi:10.1007/978-3-319-53637-8_9.