Available online at www.eccomasproceedia.org
Eccomas Proceedia COMPDYN (2023) 1266-1275

ECCOMAS
Proceedia

COMPDYN 2023
9th ECCOMAS Thematic Conference on
Computational Methods in Structural Dynamics and Earthquake Engineering
M. Papadrakakis, M. Fragiadakis (eds.)
Athens, Greece, 12-14 June 2023

A BIBLIOMETRIC OVERVIEW OF OPEN SCIENCE RESEARCH

Nikolaos P. Bakas[1], Andreas Athenodorou[2], Nana Anastasopoulou[1], Katerina Kyprianou[1], George Katsikatsos[1], George Markou[3]

[1] National Infrastructures for Research and Technology – GRNET
7 Kifisias Avenue, 11523, Athens, Greece
e-mail: {nibas,nana,kkyprianou,katsikatsos}@admin.grnet.gr

[2]Computation-based Science and Technology Research Center, The Cyprus Institute
20 Konstantinou Kavafi Street, 2121, Aglantzia Nicosia, Cyprus.
e-mail: a.athenodorou@cyi.ac.cy

[3]Civil Engineering Department, University of Pretoria
Private Bag x 20, Hatfield, 0028
e-mail: george.markou@up.ac.za

Abstract.

Scientific research exhibits a vast increase in terms of published works, as well as produced data, computer codes, and scientific methods. Accordingly, openness is a modern demand in science, and many efforts occur in this direction. The reproducibility of results, data sharing, transparency, and replication are emerging challenges in research, as by following "open science" principles, the research community, as well as society, obtains beneficial outcomes in the best possible way. In this work, we computationally analyse a large volume of papers related to open science. Particularly, we use the Scopus database and search for papers with the term "open science" in the title, abstract or keywords, within the past ten years. In total 4878 papers were identified, in the following categories: Article (N=2770), Conference paper (N=724), Review (N=646), Editorial (N=188), Note (N=176), Book chapter (N=119), Erratum (N=112), Conference review (N=54), Data paper (N=30), Letter (N=29), Book (N=16), and Short survey (N=14). We start with descriptive statistics of the keywords appearing in the papers, and continue with inter-items' associations, by computing the contingency matrix. Accordingly, we calculate the objects' potions on the bibliometric map, by using a novel multidimensional scaling algorithm. In order to verify the associations, we permute the contingency matrix to minimise its bandwidth and plot the resulting clusters of keywords. Finally, in order to evaluate the evolution of the terms in recent years, we compute the timeseries of the occurrence of keywords and select the ones exhibiting higher trends, based on linear as well as nonlinear model fit. We also compute the normalised timeseries, in order to subtract the overall increase in scientific output. We discuss significant conclusions on open science research, based on a rigorous computational procedure.

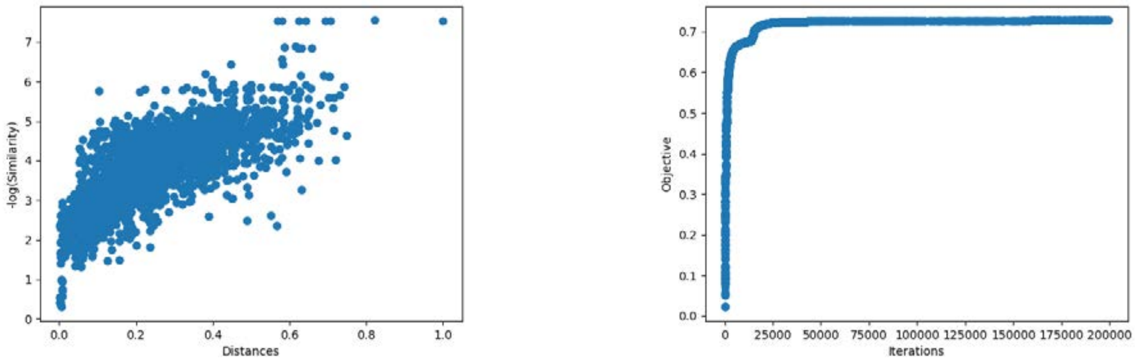
Keywords: Open science, Open access, Reproducibility, Open data, Data sharing, Transparency, Replication

ISSN:2623-3347 © 2023 The Authors. Published by Eccomas Proceedia.
Peer-review under responsibility of the organizing committee of COMPDYN 2023.
doi: 10.7712/120123.10475.21297

1 Introduction

Open Science (OS) is a modern demand in scientific publishing [1, 2], aiming to make publications and data open to the broad scientific community [3, 4]. Open science is a major policy goal for the European Commission [5], and initiatives like Plan S [6] that aim to avoid journals' paywalls. Policymakers can also benefit from open science practices [7] by using open data, methods and tools [8] for evidence-based decision-making. OS facilitates the dissemination of information among researchers [9, 10], and scholarly communication [11, 12], while social media can also be considered as a communication channel [13]. Teaching how to publish should also comprise open science concepts [14] in order to increase the open scientific output. Reproducibility of scientific outcomes is necessary to avoid "re-inventing the wheel" practices and is considered as a quality criterion for research output [15, 16, 17]. Furthermore, in medicine, replicability can be improved by OS practices, like preregistration, data sharing and preprints [18].

The purpose of the work is to analyse a vast amount of papers through bibliometric techniques [19, 20], present the landscape of publications regarding OS and identify future trends. For the needs of this research work, a computationally rigorous procedure is used to analyse the papers' keywords and their time series trends and review the most related papers with the top-frequent keywords.



(a) Optimal distances on bibliometric map d_{ij}, versus dissimilarities $ds_{ij} = -log(s_{ij})$. With the proposed methodology, a clear assessment of the representation of dis-similarities with distances is obtained.

(b) Optimisation history of objective function $cor(d_{ij}, ds_{ij})$. The algorithm converges after a few iterations while the computing time derived from the use of a standard personal computer is in the order of seconds.

Figure 1: Optimisation Algorithm Results

2 Methodology

The method presented in [21, 22] is extended and utilized to analyse the obtained database of papers. To define the contingency matrix **c**, we denote the elements of the matrix as c_{ij}, where $i, j \in [N] = \{1, 2, \dots, N\}$, $[N]$ the iterator for the number of bibliometric objects with order N. The values in the contingency matrix c_{ij} correspond to the co-occurrence counts of the objects.

In this work, for the similarity matrix s, instead of using $s_{ij} := \frac{c_{ij}}{\max \mathbf{c}}$, we utilise the following expression:

$$s_{ij} \leftarrow \frac{c_{ij}}{c_{ii} + c_{jj} - c_{ij}} \forall i, j \in [N]. \tag{1}$$

Furthermore, instead of minimising the differences among the distances and dissimilarities as proposed in [21], the Pearson correlation is maximised among the distances and dissimilarities by defining dissimilarities as:

$$ds_{ij} := -log(s_{ij}). \tag{2}$$

It was found that this approach provided a measurable metric of the success of the algorithm, as well as high values of the objective function $cor(d_{ij}, ds_{ij})$, with Pearson values $R > 0.7$ (Figure 1a).

Ultimately, a k-means clustering [23] is applied on the map to obtain a graphical representation of the clusters of the most frequent bibliometric objects. A computer code which was written in Python [24] is implemented for the retrieval of the papers from *.bib files and optimising the positions on the map and scikit-learn [25, 26] for clustering.

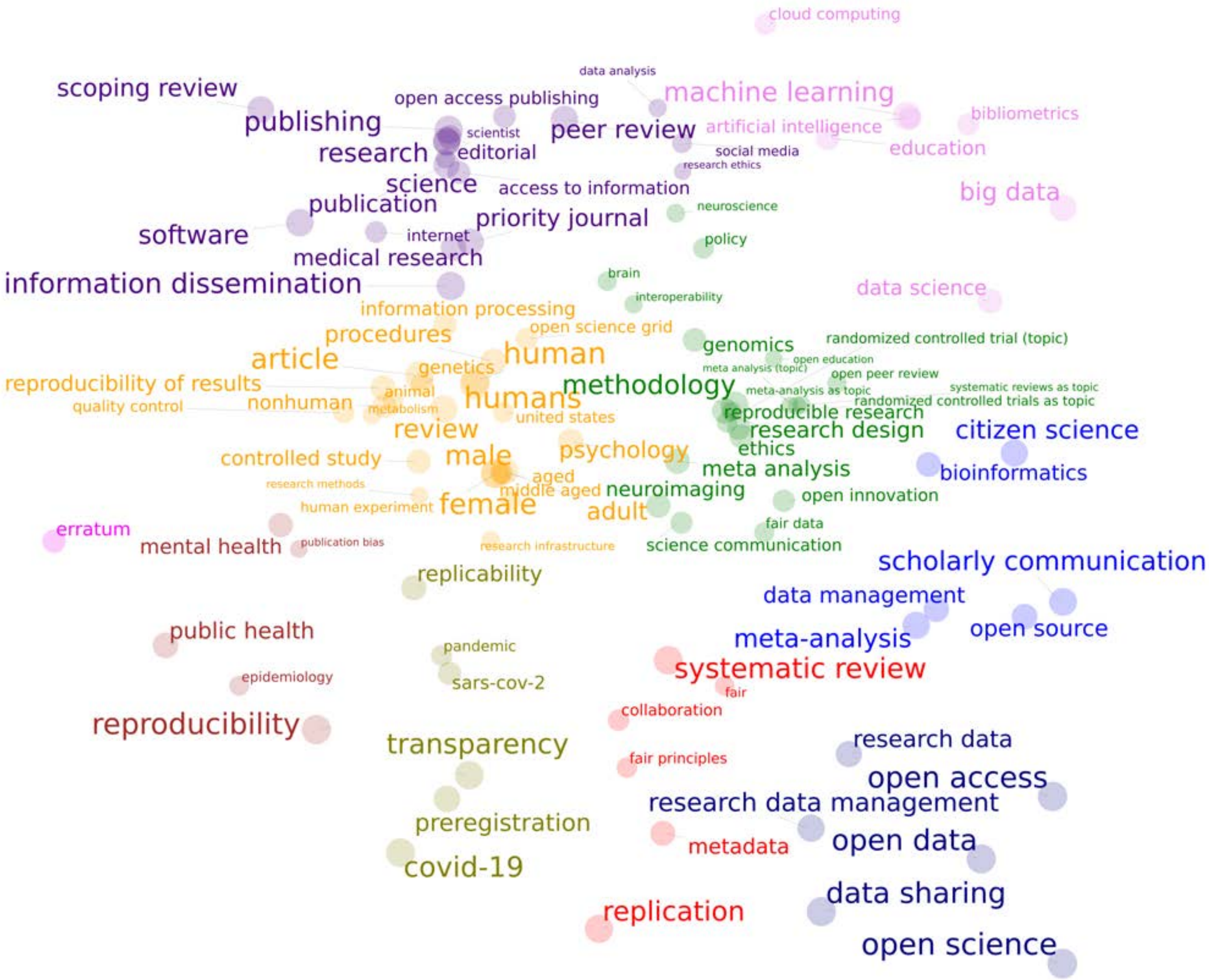


Figure 2: Map of the 100 most frequent keywords.

3 Top Keywords

In Figure 2, the map of the 100 most frequent keywords is presented. The cluster in navy colour represents the most frequent keywords of OS, like open access, open data, data sharing, research data management, and research data. The cluster in orange color, refers to health sciences, with keywords like: humans, female, male, review, adult, controlled study, genetics, aged, open science grid, quality control, animal, human experiment, and metabolism. The cluster in indigo color, is correlated to information dissemination, with keywords like: publishing, software, research, peer review, science, scoping review, priority journal, publication, editorial, access to information, open access publishing, internet, social media, scientist, data analysis, and research ethics. Furthermore, the cluster in violet color regards data science, where the following keywords can be seen: machine learning, big data, artificial intelligence, bibliometrics, cloud computing. Finally, in the cluster in olive colour, keywords related to the pandemic are shown, such as covid-19, and sars-cov-2.

According to the findings from Figure 2 and the numerical investigation performed for the needs of this research work, the following conclusions resulted in relation to the different types of keywords, with references to the corresponding 4878 publications that were investigated herein.

3.1 Open Access Publishing

Open Access (OA) publishing has been adopted by many universities worldwide, however, the definitions of each OA publication type (e.g. green OA) should be further improved [3]. Although academics value the merit of OA publishing, it has not yet been widely adopted [27]. Various research areas have adopted OA practices, with mathematics, natural and social sciences being pioneers, while health sciences have also increased the ratio of OA publishing [28]. Preregistration, the process of announcing the aims of a research study in a public space prior to conducting research, was found to promote transparency and decrease false positives [29].

3.2 Open Peer Review

The process of making reviewers' reports and identities open, named Open Peer Review (OPR), is another trend in OS [30]. However, some argue that disclosing reviewers' identities might lead to less strict reviews. Specific challenges have to be addressed, such as the papers' databases do not always index OPR journals [31], and OPR reviewing model is not yet adopted broadly [32].

3.3 Open Data

Making data open is a vastly significant process [33, 34] in a variety of scientific disciplines, such as Education [35], Ecology [36], and Technology [37]. Open data and information sharing facilitate transparency [38, 39], is an essential principle in any scientific endeavour.

3.4 FAIR Principles

The FAIR (Findability, Accessibility, Interoperability, Reproducibility) [40, 41] principles appear in the foundational concepts of OS. Reproducibility is important for statistical analyses [42] and indicated research quality [16]. Interoperability regards the integration of two or more datasets into a whole [43] promoting re-usage [44]. Making data FAIR involves metadata definition and management [45], which is necessary for multidisciplinary research works [46]. FAIR principles are considered in data infrastructures [47], enhancing data quality [48] and teaching FAIR principles has been materialised by integrating FAIR principles in curricula [49].

3.5 Data Science

OS is highly associated with evolving data science [50] and machine learning [51] fields, as the sharing of codes and data accelerates research output. Replication is also significant for machine learning research [52], being a major challenge [53, 54]. Replication usually refers to fields involving data and analytics [55], however, it is also important for social sciences [15, 56, 57].

3.6 Health Sciences

Open data and publications in health sciences are imperative for systematic reviews [58] and meta-analyses [59], and also considered an integral part of information science for health [60, 61]. Similarly, research infrastructures and data [62, 63] support bioinformatics research. During the COVID-19 pandemic, it accelerated the productivity of scientific results [64, 65] by promoting data collection [66] from various sources and proposing actions to avoid future pandemics [67].

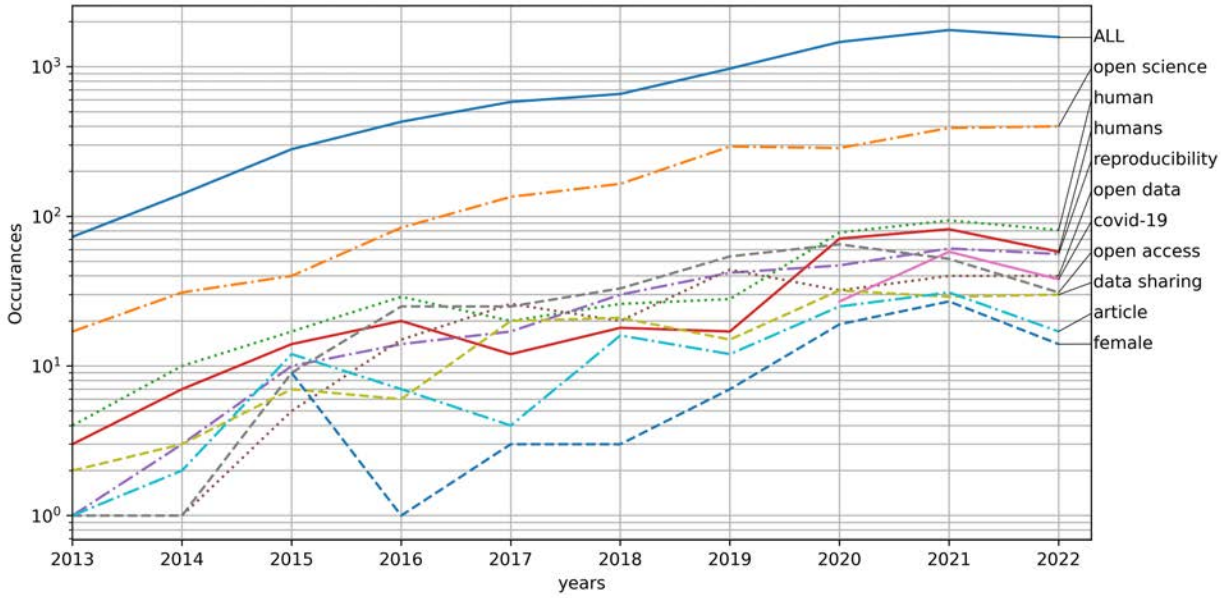


Figure 3: Timeseries of the top 10 most frequent keywords.

3.7 Trends in OS

Figure 3 demonstrates the evolution of the keywords during the last 10 years (2013-2022), as well as the total number of keywords' occurrences per year (shown in the graph as "ALL"). In Figure 3, the y-axis is in logarithmic scale; hence we see an almost exponential increase in terms of scientific output. It is also easy to depict a clear increase in **reproducibility**, highlighting the need for replication of research output. Furthermore, the keyword **human** also increases, indicating the importance of OS in health sciences.

4 Conclusions

Through computational analysis, a total of 4878 research works were analyzed on OS that were obtained from the Scopus database. The articles were from multiple research disciplines, as openness is a modern demand in all scientific fields. Clusters of the most frequent keywords were identified, where their evolution over time was demonstrated. According to the findings of this research work, an exponential increase in the aggregated keywords, with a particular inflation of reproducibility was observed. It was also found that OS is vastly significant in health sciences, as well as data sciences, machine learning and artificial intelligence. Furthermore, making data FAIR is important for identifying and replicating research results. Open research infrastructures are essential in various fields, such as biological sciences and high-performance computing, accelerating the production of scientific output in engineering, physics and artificial intelligence. The communication of scientific output is also highlighted as a significant part of science, which can be strengthened by OS mechanisms.

REFERENCES

[1] A. Athenodorou, E. Bennett, J. Lenz, and E. Papadopoullou, “Open science in lattice gauge theory community,” *arXiv preprint arXiv:2212.04853*, 2022.

[2] A. Grand, “Open science,” *Journal of Science Communication*, vol. 14, no. 4, 2015.

[3] N. Robinson-Garcia, R. Costas, and T. van Leeuwen, “Open access uptake by universities worldwide,” *PeerJ*, vol. 2020, no. 7, 2020.

[4] S. Akterian, “Towards open access scientific publishing,” *Biomedical Reviews*, vol. 28, pp. 125–133, 2017.

[5] C. Ramjoué, “Towards open science: The vision of the european commission,” *Information Services & Use*, vol. 35, no. 3, pp. 167–170, 2015.

[6] H. Else, “A guide to plan s: the open-access initiative shaking up science publishing.” *Nature*, 2021.

[7] M. Cruz, N. Dintzner, A. Dunning, A. Kuil, E. Plomp, M. Teperek, Y. Velden, and A. Versteeg, “Policy needs to go hand in hand with practice: The learning and listening approach to data management,” *Data Science Journal*, vol. 18, no. 1, 2019.

[8] K. Anderson, J. Harris, L. Ng, P. Prins, S. Memar, B. Ljungquist, D. Fürth, R. Williams, G. Ascoli, and D. Dumitriu, “Highlights from the era of open source web-based tools,” *Journal of Neuroscience*, vol. 41, no. 5, pp. 927–936, 2021.

[9] C. Haeussler, L. Jiang, J. Thursby, and M. Thursby, “Specific and general information sharing among competing academic researchers,” *Research Policy*, vol. 43, no. 3, pp. 465–475, 2014.

[10] Y. Matsubara, “Information and communications technology in disaster mitigation technology,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E99A, no. 8, pp. 1504–1509, 2016.

[11] E. Kulczycki, "Rethinking open science: The role of communication," *Analele Universitatii din Craiova, Seria Filozofie*, vol. 37, no. 1, pp. 81–97, 2016.

[12] J. Duart and S. Mengual-Andrés, "Impact of the knowledge society in the university and in scientific communication," *RELIEVE - Revista Electronica de Investigacion y Evaluacion Educativa*, vol. 20, no. 2, pp. 1–11, 2014.

[13] E. Kulczycki, "Transformation of science communication in the age of social media," *Teorie Vedy/ Theory of Science*, vol. 35, no. 1, pp. 3–28, 2013.

[14] M. Englesbe and J. Markovac, *Teaching Publishing in Medical Education-An Overview*. Elsevier Inc., 2018.

[15] J. Moody, L. Keister, and M. Ramos, "Reproducibility in the social sciences," *Annual Review of Sociology*, vol. 48, 2022.

[16] S. Leonelli, "Rethinking reproducibility as a criterion for research quality," *Research in the History of Economic Thought and Methodology*, vol. 36B, pp. 129–146, 2018.

[17] S. Feger and P. Woźniak, "Reproducibility: A researcher-centered definition," *Multimodal Technologies and Interaction*, vol. 6, no. 2, 2022.

[18] M. Munafo, "Open science and research reproducibility," *ecancermedicalscience*, vol. 10, 2016.

[19] M. Ochsner, *Relationship between peer review and bibliometrics*. De Gruyter, 2020.

[20] L. Bracco, "Promoting open science through bibliometrics: A practical guide to building an open access monitor," *LIBER Quarterly*, vol. 32, no. 1, pp. 1–18, 2022.

[21] D. Koutsantonis, K. Koutsantonis, N. P. Bakas, V. Plevris, A. Langousis, and S. A. Chatzichristofis, "Bibliometric literature review of adaptive learning systems," *Sustainability*, vol. 14, no. 19, p. 12684, 2022. [Online]. Available: https://www.mdpi.com/2071-1050/14/19/12684/htm

[22] N. Bakas, D. Koutsantonis, V. Plevris, A. Langousis, and S. Chatzichristofis, "Inverse transform sampling for bibliometric literature analysis," in *The Thirteenth International Conference on Information, Intelligence, Systems and Applications. Ionian University, Corfu, Greece, 18-20 July 2022*. IISA 2022, 2022. [Online]. Available: http://easyconferences.eu/iisa2022/

[23] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[24] python, "Python programming language." [Online]. Available: https://www.python.org/

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[27] Y. Zhu, "Who support open access publishing? gender, discipline, seniority and other factors associated with academics' oa practice," *Scientometrics*, vol. 111, pp. 557–579, 2017.

[28] A. Severin, M. Egger, M. P. Eve, and D. Hürlimann, "Discipline-specific open access publishing practices and barriers to change: an evidence-based review," *F1000Research*, vol. 7, 2018.

[29] J. P. Simmons, L. D. Nelson, and U. Simonsohn, "Pre-registration: Why and how," *Journal of Consumer Psychology*, vol. 31, no. 1, pp. 151–162, 2021.

[30] EDITORIAL, "Nature will publish peer review reports as a trial," 2020. [Online]. Available: https://doi.org/10.1038/d41586-020-00309-9

[31] D. Wolfram, P. Wang, A. Hembree, and H. Park, "Open peer review: promoting transparency in open science," *Scientometrics*, vol. 125, no. 2, pp. 1033–1051, 2020.

[32] J. Teixeira da Silva, "Challenges to open peer review," *Online Information Review*, vol. 43, no. 2, pp. 197–200, 2019.

[33] R. Inkpen, R. Gauci, and A. Gibson, "The values of open data," *Area*, vol. 53, no. 2, pp. 240–246, 2021.

[34] W. Thompson, J. Wright, and P. Bissett, "Open exploration," *eLife*, vol. 9, 2020.

[35] J. Logan, S. Hart, and C. Schatschneider, "Data sharing in education science," *AERA Open*, vol. 7, 2021.

[36] W. Michener, "Ecological data sharing," *Ecological Informatics*, vol. 29, no. P1, pp. 33–44, 2015.

[37] J. Maienschein, J. Parker, M. Laubichler, and E. Hackett, "Data management and data sharing in science and technology studies," *Science Technology and Human Values*, vol. 44, no. 1, pp. 143–160, 2019.

[38] L. Lyon, "Transparency: The emerging third dimension of open science and open data," *LIBER Quarterly*, vol. 25, no. 4, pp. 153–171, 2016.

[39] K. Elliott, "A taxonomy of transparency in science," *Canadian Journal of Philosophy*, vol. 52, no. 3, pp. 342–355, 2022.

[40] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[41] M. Barker, N. Chue Hong, D. Katz, A.-L. Lamprecht, C. Martinez-Ortiz, F. Psomopoulos, J. Harrow, L. Castro, M. Gruenpeter, P. Martinez, and T. Honeyman, "Introducing the fair principles for research software," *Scientific Data*, vol. 9, no. 1, 2022.

[42] V. Stodden, "Reproducing statistical results," *Annual Review of Statistics and Its Application*, vol. 2, pp. 1–19, 2015.

[43] M. Campos, L. Campos, and N. Barbosa, "The challenges of semantic interoperability in the era of escience on the web," *Knowledge Organization*, vol. 47, no. 8, pp. 680–695, 2020.

[44] Y. Sakai, Y. Miyata, K. Yokoi, Y. Wang, and K. Kurata, "Data integration as the major mode of data reuse," *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, 2020.

[45] C. Lortie, C. Vargas Poulsen, J. Brun, and L. Kui, "Tabular strategies for metadata in ecology, evolution, and the environmental sciences," *Ecology and Evolution*, vol. 12, no. 8, 2022.

[46] A. Child, J. Hinds, L. Sheneman, and S. Buerki, "Centralized project-specific metadata platforms: toolkit provides new perspectives on open data management within multi-institution and multidisciplinary research projects," *BMC Research Notes*, vol. 15, no. 1, 2022.

[47] J. Neumann, "Fair data infrastructure," *Advances in Biochemical Engineering/Biotechnology*, vol. 182, pp. 195–207, 2022.

[48] Z. Triki and R. Bshary, "A proposal to enhance data quality and fairness," *Ethology*, vol. 128, no. 9, pp. 647–651, 2022.

[49] H. Shanahan, N. Hoebelheinrich, and A. Whyte, "Progress toward a comprehensive teaching approach to the fair data principles," *Patterns*, vol. 2, no. 10, 2021.

[50] P. Wittenburg, "Open science and data science," *Data Intelligence*, vol. 3, no. 1, pp. 95–105, 2021.

[51] M. Braun and C. Ong, *Open science in machine learning*. CRC Press, 2014.

[52] C. Emmery, A. Kadar, T. Wiltshire, and A. Hendrickson, "Towards replication in computational cognitive modeling: a machine learning perspective," *Computational Brain and Behavior*, vol. 2, no. 3-4, pp. 242–246, 2019.

[53] D. Willer and P. Emanuelson, "Theory and the replication problem," *Sociological Methodology*, vol. 51, no. 1, pp. 146–165, 2021.

[54] M. da Luz Antunes, T. Sanches, C. Lopes, and J. Alonso-Arévalo, *Publishing within open science challenges*. Nova Science Publishers, Inc., 2019.

[55] H. Fraser, A. Barnett, T. Parker, and F. Fidler, "The role of replication studies in ecology," *Ecology and Evolution*, vol. 10, no. 12, pp. 5197–5207, 2020.

[56] J. Chin and K. Zeiler, “Replicability in empirical legal research,” *Annual Review of Law and Social Science*, vol. 17, pp. 239–260, 2021.

[57] W. Pridemore, M. Makel, and J. Plucker, “Replication in criminology and the social sciences,” *Annual Review of Criminology*, vol. 1, pp. 19–38, 2018.

[58] K. Gunnell, V. Belcourt, J. Tomasone, and L. Weeks, “Systematic review methods,” *International Review of Sport and Exercise Psychology*, vol. 15, no. 1, pp. 5–29, 2022.

[59] M. Hagger, “Meta-analysis,” *International Review of Sport and Exercise Psychology*, vol. 15, no. 1, pp. 120–151, 2022.

[60] S. Hunt and C. Bakker, “A qualitative analysis of the information science needs of public health researchers in an academic setting,” *Journal of the Medical Library Association*, vol. 106, no. 2, pp. 184–197, 2018.

[61] A. Flahault, A. Geissbuhler, I. Guessous, P. Guérin, I. Bolon, M. Salathé, and G. Escher, “Precision global health in the digital age,” *Swiss Medical Weekly*, vol. 147, 2017.

[62] R. LeDuc, M. Vaughn, J. Fonner, M. Sullivan, J. Williams, P. Blood, J. Taylor, and W. Barnett, “Leveraging the national cyberinfrastructure for biomedical research,” *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 195–199, 2014.

[63] L. Candela, G. Coro, L. Lelii, G. Panichi, and P. Pagano, “Data processing and analytics for data-centric sciences,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12003, pp. 176–191, 2020.

[64] M. Cardenas-Gonzalez and E. Alvarez-Buylla, “The covid-19 pandemic and paradigm change in global scientific research,” *MEDICC Review*, vol. 22, no. 2, pp. 14–18, 2020.

[65] Z. Yang, M. Wang, Z. Zhu, and Y. Liu, “Coronavirus disease 2019 (covid-19) and pregnancy: a systematic review,” *Journal of Maternal-Fetal and Neonatal Medicine*, vol. 35, no. 8, pp. 1619–1622, 2022.

[66] E. Martínez Beltrán, M. Quiles Pérez, J. Pastor-Galindo, P. Nespoli, F. García Clemente, and F. Gómez Mármol, “Convida: Covid-19 multidisciplinary data collection and dashboard,” *Journal of Biomedical Informatics*, vol. 117, 2021.

[67] M. Hoque, G. Faisal, F. Chowdhur, A. Haque, and T. Islam, “The urgency of wider adoption of one health approach for the prevention of a future pandemic,” *International Journal of One Health*, vol. 8, no. 1, pp. 20–33, 2022.