MAPPING OF OPEN DATA AND OPEN SCIENCE RESEARCH OUTPUT USING SCOPUS DATABASE

S.B.Patil
M. G. Pathan
Y. G. Jadhav

S.B.Patil
Assistant Professor, Department of Library & Information Science, Shivaji University, Kolhapur-416004, Maharashtra, India, E-Mail: sbp.lib@unishivaji.ac.in
Corresponding Author

M. G. Pathan
Student, Department of Library & Information Science, Shivaji University, Kolhapur-416004, Maharashtra, India, E-Mail: pathanmg1@gmail.com

Y. G. Jadhav
Assistant Professor, Department of Library & Information Science, Shivaji University, Kolhapur-416004, Maharashtra, India, E-Mail: ygj.lib@unishivaji.ac.in

The present study analyzed 3335 global publications on open data and open science published from 1980-2021(42Years). The data from Scopus was exported in BibTeX (.bib) file format. This file was further used as the data source in the biblioshiny app. After data analysis, it is found that the highest contribution in open data and open science research takes the form of conference papers. It is observed that the highest research publications were published during the period 2009-2021. The USA was found to be the most productive country in open science research. Professor M. Janssen of the Delft University of Technology, Netherlands, was the most prolific author. Springer publication has published more articles than any other source in the top ten prolific sources. The study reported open data, linked open data, semantic web, etc. as the significant author keywords in the publications.

Keywords - Open data, Open science, Scopus database, R Statistical Package, Biblioshiny app, Research output, Scientometrics

INTRODUCTION

The meaning itself tells that ‘Open Data’ and ‘Open Science’ is open to all. The use of open data and open access is an integral element of open science. UNESCO in its recommendation on open science (2021) described open science as “ an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community. It comprises all scientific disciplines and aspects of scholarly practices, including basic and applied sciences, natural and social sciences and the humanities, and it builds on the key pillars viz. open scientific knowledge, open science infrastructures, science communication, open engagement of societal actors and open dialogue with other knowledge systems.”

Many studies on open data and open science highlighted its different aspects (Refer to 'Literature Review' section). The present study attempts to analyze the research publications on open data and open science indexed by the Scopus database.

REVIEW OF LITERATURE

Pontika, Knoth, Cancellieri, and Pearce (2015) described the FOSTER (Facilitate Open Science Training for European Research) approach in structuring the Open Science domain for educational purposes. Allen and Mehler (2019) discussed the challenges and benefits of open science. They gave suggestions from the perspective of Early Career Researchers (ECRs) for moving towards open science. The challenges include restrictions on flexibility, the time cost and incentive structure, while the benefits include greater faith in research, new helpful systems and investment in an individual's future. The study pointed out that adopting open practices requires a change in attitude and productivity expectations, and academics and funders at all levels should consider this.

Zarate, Buckle, Mazzanti & Samec (2019) studied open science publications from CONICET Digital, an open access institutional repository of the National Council of Scientific and Technical Research (CONICET). The Linked Data life cycle approach given by Villazon Terrazas et al. is adopted in the present study. The study carried out by Zuiderwijk and Spiers (2019) intends to provide in-depth insight into the complex interaction of factors influencing motivations for sharing and re-using open research data within the discipline of astrophysics.

Raffaghelli and Manca (2019) explored the practices of open data publications and sharing in the field of Educational Technology. Wolfram, Wang, Hembree & Park (2020) identified 617 Open Peer Review (OPR) journals that published at least one article with open identities or open reports as of 2019 and analyzed their wide-ranging implementations to derive emerging OPR practices. The study observed a steady growth in OPR adoption since 2001.

The article by Borgerud and Borglund (2020) investigated how Swedish universities and public authorities work with archiving and implementing open research data and their perceptions on open access. The study found that research data management was not part of an overall record-keeping strategy. Chiware (2020) has presented a literature review on research data management services in African academic and research libraries on the background of the advancing open science and open research data infrastructures. This review depicts that open science in Africa is still in its development stages. Research infrastructures face funding and technical challenges while data management services are in the formative stages.

International Science Council (2020) published a draft paper about Open science in 21st century. It describes the grounds for and the origins of the modern open science movement, its dimensions and its applications. The paper recommended that there should be a persistent discussion between public and private sectors about how the scope of open science principles and priorities can be enlarged and mutually shared. Shmagun, Oppenheim, Shim, Choi, and Kim (2021) identified factors that South Korean and Australian experts view as enablers or barriers to

Open Science (OS) practices in public health emergencies. Pagliaro (2021) investigated the citation patterns of preprints published in three preprint servers, i.e. bioRxiv, ChemRxiv, and Research Square, from 2017 to 2020. The study noted that preprints are frequently cited in peer-reviewed journal articles, books, and conference papers. The results of this investigation further validate the value of open science in relation to citation-based metrics on which the evaluation of scholarship continues to rely.

Open data report (2021) described the findings of a complementary methods approach to examine the practices, motivations, obstacles to data sharing and perceived advantages among researchers across different disciplines worldwide. The study employed a bibliometric analysis, a survey and case studies. The study found that researchers publish less than 15% of their data in data repositories. The study reported that most of the researchers are willing to allow others to access their research data.

OBJECTIVES OF THE STUDY

The main objective of this study is to analyze the open data and open science related publications indexed by Scopus database. In particular, the study aims to find:

- Scientometric Profile of the Publications
- Growth of Publications
- Most Cited Countries
- Highly Productive Countries
- Author Impact
- Most Relevant Sources
- Most Global Cited Documents
- Significant Author Keywords

METHODOLOGY

A scientometric method was used for the present study. All years and all types of publications, are considered for the analysis. The study period is from 1980-2021(42 Years). To retrieve the dataset for conducting the present study following search strategy was used : TITLE ("Open data" OR "Open science") AND (LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "ch") OR LIMIT-TO (DOCTYPE , "re") AND (EXCLUDE (PUBYEAR , 2022)) AND (LIMIT-TO (LANGUAGE , "English") AND (EXCLUDE (LANGUAGE ,"French") OR EXCLUDE (LANGUAGE, "German") OR EXCLUDE (LANGUAGE "Spanish") OR EXCLUDE (LANGUAGE ,"Chinese") OR EXCLUDE (LANGUAGE, "Italian") OR EXCLUDE (LANGUAGE, "Portuguese") . As of the date 6 October 2021, 3335 records in BibTeX (.bib) file format were retrieved from the Scopus database. Then, this file was uploaded into the Biblioshiny interface. The biblioshiny app from R Statistical Package (https://bibliometrix.org/Biblioshiny.html) was used for comprehensive science mapping analysis. Finally, as per the study's objectives, excel files were downloaded and used for further analysis.

Table 1: Main Information about Collection

S.N.	Description	Results
Main information about collection		
1.	Time span	1980:2021
2.	Sources (Journals, Books, etc)	1488
3.	Documents	3335
4.	Citations	32490
5.	Average citations per documents	9.742
6.	References	96461
Document types		
7.	Article	1291
8.	Book chapter	159
9.	Conference paper	1765
10.	Review	120
Document contents		
11.	Keywords plus (id)	11298
12.	Author's keywords (de)	6248
Authors		
13.	Authors	11219
14.	Authors of single-authored documents	472
15.	Authors of multi-authored documents	10747
Authors collaboration		
16.	Single-authored documents	544
17.	Multi-authored documents	2791
18.	Documents per Author	0.30
19.	Authors per Document	3.36
20.	Collaboration index	3.85

DATA ANALYSIS

Scientometric Profile of Open Data & Open Science Publications

Table 1 depicts the scientometric profile of global 'Open Data and Open Science' research. These results are retrieved through the biblioshiny app of bibliometrix R package (Version 2.3.2 released on 23/11/2019). It is interesting to note that the highest contribution in open data and open science research takes the form of conference papers. The conference papers accounted for 52.92 % share among the total output, followed by journal articles (38.71%), book chapters (4.77%) and reviews (3.60%). It is observed that the majority of the publications are multi-authored, which reveal the multi-authorship trend in open data and open science research field Growth of Publication.

Table 2: Annual Production

S.N.	Year	TP	% of TP	Mean TC per Article	Mean TC Per Year	Citable Years
1.	1980	1	0.03	0	0	41
2.	1985	1	0.03	15	0.42	36
3.	1992	1	0.03	3	0.10	29
4.	1995	1	0.03	3	0.12	26
5.	1996	3	0.09	10.33	0.41	25
6.	1997	1	0.03	0	0	24
7.	1998	1	0.03	99	4.30	23
8.	2000	1	0.03	0	0	21
9.	2001	1	0.03	6	0.3	20
10.	2002	1	0.03	3	0.16	19
11.	2004	5	0.15	35	2.06	17
12.	2005	4	0.12	28.75	1.80	16
13.	2006	6	0.18	10.83	0.72	15
14.	2007	13	0.39	205	14.64	14
15.	2008	16	0.48	21.94	1.69	13
16.	2009	20	0.60	12.45	1.04	12
17.	2010	48	1.44	14.73	1.34	11
18.	2011	73	2.19	22.77	2.28	10
19.	2012	132	3.96	20.59	2.29	9
20.	2013	190	5.70	12.99	1.62	8
21.	2014	270	8.10	15.25	2.18	7
22.	2015	299	8.97	10.94	1.82	6
23.	2016	343	10.28	10.81	2.16	5
24.	2017	391	11.72	9.17	2.29	4
25.	2018	402	12.05	9.02	3.01	3
26.	2019	405	12.14	3.87	1.94	2
27.	2020	416	12.47	2.47	2.47	1
28.	2021	290	8.70	0.88	--	0
Total		**3335**	**100%**	--	--	--
TP - Total Publications; TC - Total Citations						

The global output in open data and open science research accumulated 3335 publications during 1980-2021. The highest publications (416) on open data and open science were appeared in the year 2020, while the lowest publications (1) were published in the year 1980, 1985, 1992, 1995, 1997, 1998, 2000, 2001, and 2002.

Most Cited Countries

It is found that the highest 'mean total citations per article' (205) and 'mean total citations per year' (14.64) for the 13 publications on open data and open science appeared in the year 2007. According to mean total citations per article count, the years 1998, 2004, 2005, 2007, 2008, 2011, 2012 and 2014 are found to be the most productive (Table 2).

Table 3: Top Ten Most Cited Countries

Rank	Country	Total Citations	Percentage (%)
1.	USA	4277	24.35
2.	Germany	3272	18.63
3.	Netherlands	2846	16.20
4.	United Kingdom	2468	14.05
5.	Italy	1212	6.90
6.	Canada	1127	6.42
7.	Australia	697	3.97
8.	China	691	3.93
9.	Spain	526	2.99
10.	Ireland	449	2.56
Total		**17565**	**--**
N ≥ 449; N - No. of the Citations			

The top 10 countries accounted for as much as (54.06%) of the global citations output. It is found that the United States of America (USA) with 4277 (24.35%) citations ranked first in the list. The other productive countries were Germany with 3272 (18.63%) citations, followed by the Netherlands with 2846 (16.20%) citations and the United Kingdom with 2468 (14.05%) citations. It is significant to mention that the majority of highly cited documents belong to European countries. China is the only Asian country that is included the list of highly cited countries (Table 3).

Highly Productive Countries

Table 4: Country Scientific Production

Rank	Country	Publications	Percentage (%)
1.	USA	1577	30.70
2.	UK	608	11.84
3.	Germany	571	11.12
4.	Italy	542	10.55
5.	China	389	7.57
6.	Canada	321	6.25
7.	Spain	298	5.80
8.	France	280	5.45
9.	Netherlands	279	5.43
10.	Japan	272	5.29
Total		**5137**	**100%**
Total Publications of Top 10 Countries = 5137 **(**N ≥272 ; N - No. of the Publications)			

The research output in open data and open science is originated from 100 countries during 1980-2021, but its distribution across publishing countries is highly skewed. Among these top 10 productive countries in the domain of open data and open science research, the highest publications, i.e. 1577 (30.70%), are contributed by the United States, followed by UK and Germany with 608 (11.84%) and 571(11.12%) publications respectively. Table 4 demonstrates that the highest publications on open data and open science are published from European countries. The study conducted by Zhang, Hua, and Yuan (2017) supports this observation. Two Asian countries, i.e. Japan and China, are included in this list (Table 4).

Author Impact

Table 5: Prolific Authors

S.N.	Author	NP	PY_start	h_index	g_index	m_index	TC
1.	Janssen, M.	52	2012	21	49	2.1	c
2.	Zuiderwijk, A.	45	2012	20	45	2	2374
3.	Ojo, A.	25	2014	9	15	1.125	230
4.	Hyvonen, E.	20	2013	6	8	0.667	86
5.	Charalabidis, Y.	17	2012	7	17	0.7	1068
Total		**159**	--	--	--	--	**6214**
NP - Number of Publication; TC -Total Citations; PY - Publication Year							

Table 5 shows the performance of the top five productive authors based on their publications and citation metrics. Together, these authors contributed 4.76 % share in the world's cumulative publications output during the reported period. Professor M. Janssen from the Delft University of Technology, Netherlands, contributed the highest publications (52), citations (2456), h-index (21), g-index (49) and m-index (2.1). Among the total global citations output, the top five authors have shared 19.12% of citations (Table 5).

Most Relevant Sources.

Table 6: Top Ten Productive Sources

S.N.	Sources	TP	Percentage (%)
1.	Lecture Notes in Computer Science (Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)	254	28.73
2.	CEUR Workshop Proceedings	221	25
3.	ACM International Conference Proceeding Series	182	20.59
4.	Communications in Computer and Information Science	67	7.58
5.	Advances in Intelligent Systems and Computing	30	3.39
6.	Lecture Notes in Business Information Processing	29	3.28

7.	Journal of Physics: Conference Series	28	3.17
8.	IFIP Advances in Information And Communication Technology	25	2.83
9.	Government Information Quarterly	24	2.71
10.	Proceedings of the Annual HAWAII International Conference on System Sciences	24	2.71
Total		**884**	**100%**

- Total Publications of Top 10 Sources = 884 (26.50%)
- Total output of 1488 Sources = 3335
- N ≥24 ; TP – Total Publications

In all, 3335 records on open data and open science have been published in 1488 sources. Table 6 represents the top 10 sources (N ≥24) preferred by the researchers in the domain of open data and open science. These ten prominent sources account for 26.50 % share among the Most Global Cited Documents.

total publications output. The highest 254 (28.73%) publications are contributed by 'Lecture Notes in Computer Science', followed by 'CEUR Workshop Proceedings' with 221 (25%) papers and 'ACM International Conference Proceeding Series' with 182 (20.59%) papers (Table 6).

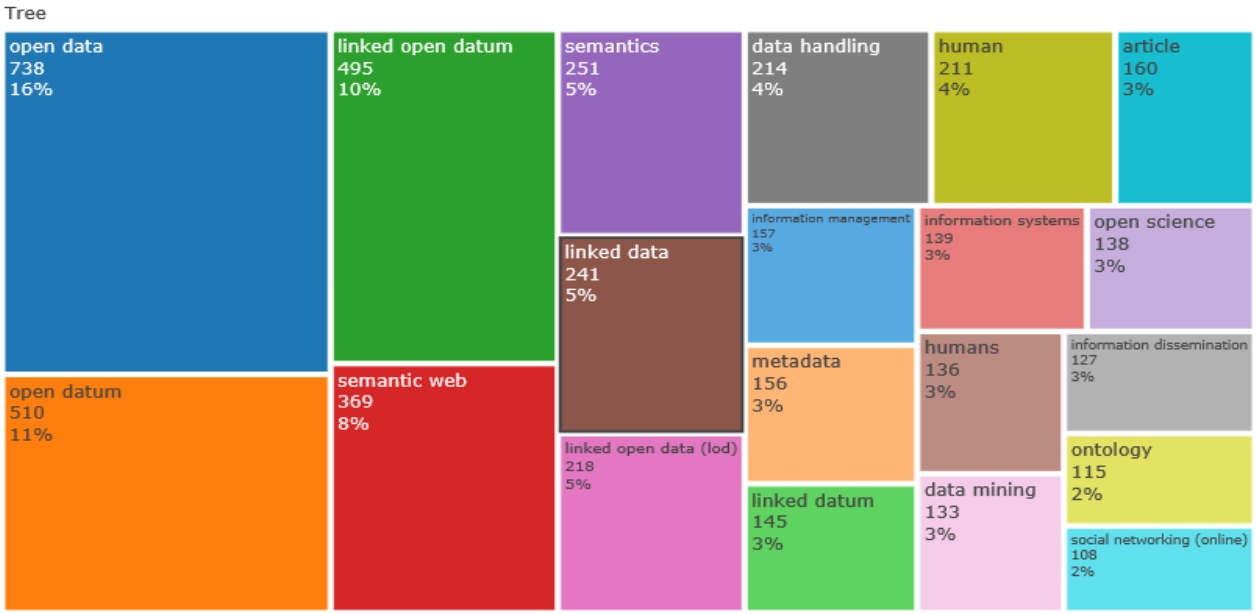
Table 7: Highly Cited Documents

S.N.	Details of Publications	Total Citations
1.	Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z. (2007) DBpedia: A Nucleus for a Web of Open Data. In: Aberer K. et al. (eds) The Semantic Web. ISWC 2007, ASWC 2007. *Lecture Notes in Computer Science*, Vol. 4825. Springer, Berlin, Heidelberg. doi:10.1007/978-3-540-76298-0_52	2100
2.	Price-Whelan, A., Sipőcz, B., Günther, H., Lim, P., Crawford, S., & Conseil, S. et al. (2018). The astropy project: Building an open-science project and status of the v2.0 core package. *The Astronomical Journal*, *156*(3), 123. doi: 10.3847/1538-3881/aabc4f	1242
3.	Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *29*(4), 258-268. doi: 10.1080/10580530.2012.716740	925
4.	Reichman, O., Jones, M., & Schildhauer, M. (2011). Challenges and opportunities of open data in ecology. *Science*, *331*(6018), 703-705. doi: 10.1126/science.1197962	419
5.	Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., & Roy, A. et al. (2007). The open science grid. *Journal of Physics: Conference Series*, *78*, 012057. doi: 10.1088/1742-6596/78/1/012057	315

6.	Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, *31*(1), 17-29. doi: 10.1016/j.giq.2013.04.003	312
7.	Brynn Hibbert, D., & Thordarson, P. (2016). The death of the job plot, transparency, open science and online tools, uncertainty estimation methods and other developments in supramolecular chemistry data analysis. *Chemical Communications*, *52*(87), 12792-12805. doi: 10.1039/c6cc03888c	303
8.	Piwowar, H., & Vision, T. (2013). Data reuse and the open data citation advantage. *Peerj*, *1*, e175. doi: 10.7717/peerj.175	267
9.	McKiernan, E., Bourne, P., Brown, C., Buck, S., Kenall, A., & Lin, J. et al. (2016). How open science helps researchers succeed. *Elife*, *5*. doi: 10.7554/elife.16800	228
10.	Stadler, C., Lehmann, J., Höffner, K., & Auer, S. (2012). LinkedGeoData: A core for a web of spatial open data. *Semantic Web*, *3*(4), 333-354. doi: 10.3233/sw-2011-0052	214
Total		**6325**
Where, N≥214 (N - No. of citations)		

Table 7 displays the ten most frequently cited papers during 1980-2021. The top 10 highly cited global articles (N≥214) lead to 6325 citations. The citation count was obtained from the Scopus database. The research paper based on the Significant Author Keywords. DBpedia project, authored by Auer et al. (2007), received the highest citations (2100). These top ten articles contribute 19.46% share in total citations output of the world publications.

Fig. 1: Tree map of Significant Author Keywords



Top 20 frequently occurred author keywords (N ≥108) with their frequencies, is shown in Figure 1. The main advantage of selecting author keywords is that it provides insight into main topics and research trends in publications. The Sum of the 20 productive keywords is 4791. The study found open data (738), open datum (510), Linked Open datum (495) and semantic web (369) etc. as the significant author keywords in the publications (Figure 1).

DISCUSSION AND CONCLUSION

The results (Table 2) of a study show that, the highest 3279 (98.32%) research publications on open data and open science were published during 2009-2021. Globally, different initiatives (open data portals, repositories, preprint servers, projects, policies etc.) were announced during this period. Data.gov (2009) is the federal government's open data site that aims to make the government more open and accountable. Anyone, especially local, state, and foreign governments, can borrow the code behind Data.gov. As of 15th November 2021, 48 U.S. states, 48 U.S. cities and countries, 53 international countries, and 165 international regions have launched their open data sites using it. It is worth noting that all of the highly cited (Table 3) and productive countries (Table 4) have started their open data portals using the platform of data.gov. Major open data repositories initiated during 2009-2021 include Dryad (2009), Figshare (2012), Re3data (2012), Zenodo (2013), Mendeley Data (2016) etc. Many prominent preprint servers, for example, bioRxiv (2013), OSF Preprints (2016), Research Square (2018), medRxiv (2019) etc., were launched to speed up scholarly publishing. The initiatives, for instance, the PolyMath Project (2009), OpenAIRE (2010), Amsterdam Call for Action on Open Science (2016), FOSTER Plus (2017), Open Research Europe (2021) etc. geared up the open science movement.

It is significant to mention that, in terms of publication and citations, the European countries are leading than their Asian and African counterparts (Table 3 & 4). This emphasizes the need for increasing open data and open science related publications in Asian and African countries. Professor M. Janssen of the Delft University of Technology, Netherlands, is found to be the most prolific author in the domain of open data and open science research (Table 5). He has contributed articles on open government data, open data policies, big data, and socio-technical impediments of open data etc. His two papers titled as 'Benefits, adoption, barriers and myths of open data and open government' and 'Open data policies, their implementation and impact: A framework for comparison' received 925 & 312 citations respectively in the Scopus database (Table 7). In terms of the top ten prolific sources, it is Springer publication, which has published more articles than any other sources. It accounts for 12.14% of total publications output (Table 6). Articles based on four open data projects received the highest citations (Table 7). It includes DBpedia, the Astropy Project, the Open Science Grid (OSG), and the LinkedGeoData project. The hierarchical treemap (Figure 1) indicates the emerging trends and prominent areas related to open data and open science.

Though the present study provides an insight into various scientometric aspects of the global research publications on open data and open science, still it has some limitations. First, the present study is confined to the Scopus database.

It has its selection policy to include and exclude the sources based on the predefined parameters. Second, the data is collected up to 6 October 2021 using the search string mentioned in the methodology section. There are many inconsistencies for search strings in the Scopus database (Islam & Roy, 2021). Therefore, there is a possibility of variations in the results with the different search strings. Even importing and analyzing data from other data sources might give a different picture. Hence, it is suggested to conduct a comparative study with the different citation databases, which will be greatly helpful. With such intensive studies, we will arrive at a concrete understanding of the nature of the research publications in the open data and open science research field. The present study is a step towards such a situation.

REFERENCES

1. Allen, C., & Mehler, D. (2019). Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, *17*(5), e3000246. doi: 10.1371/ journal.pbio. 3000246
2. Aria, M. & Cuccurullo, C. (2017) bibliometrix: An R-tool for comprehensive science mapping analysis, *Journal of Informetrics*, 11(4), 959-975.
3. Borgerud, C., & Borglund, E. (2020). Open research data, an archival challenge? *Archival Science*. doi: 10.1007/s10502-020-09330-3
4. Chiware, E. (2020). Open research data in African academic and research libraries: a literature analysis. *Library Management*, *41*(6/7), 383-399. doi:10.1108/lm-02-2020-0027
5. International Science Council, (2020). Open Science For the 21st Century: A Draft ISC Working Paper. Available at:https://council.science/publications/open science-for-the-21stcentury/
6. Islam, M., & Roy, P. (2021). Bibliometric study of scholarly productivity of library and information science research in Bangladesh from 1971-2020. *DESIDOC Journal of Library & Information Technology*, *41*(03), 213-225. doi: 10.14429/djlit.41.03.16854
7. Open data report (2021). Available at: https://www.elsevier.com/open-science/ research-data/ open-data-report
8. Pagliaro, M. (2021). Did you ask for citations? An insight into preprint citations en route to open science. *Publications*, *9*(3), 26. doi: 10.3390/publications9030026
9. Pontika, N., Knoth, P., Cancellieri, M., & Pearce, S. (2015, October). Fostering open science to research using a taxonomy and an eLearning portal. In *Proceedings of the 15th international conference on knowledge technologies and data-driven business* (pp. 1-8).
10. Raffaghelli, J., & Manca, S. (2019). Is there a social life in open data? The case of open data practices in educational technology research. *Publications*, *7*(1),9. doi:10.3390/ publications7010009
11. Shmagun, H., Oppenheim, C., Shim, J., Choi, K. N., & Kim, J. (2021, January). Open Science at a time of the COVID-19 pandemic: a new opportunity to improve emergency response. In *Proceedings of the 54th Hawaii International Conference on System Sciences* (p. 2275).

12. UNESCO (2021). UNESCO Recommendation on Open Science. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en
13. Wolfram, D., Wang, P., Hembree, A., & Park, H. (2020). Open peer review: promoting transparency in open science.
14. Zarate, M., Buckle, C., Mazzanti, R., & Samec, G. (2019). Improving open science using Linked Open Data: CONICET Digital use case. *Journal of Computer Science and Technology*, *19*(01), e05. doi: 10.24215/16666038.19.e05
15. Zhang, Y., Hua, W., & Yuan, S. (2017). Mapping the scientific research on open data: A bibliometric review. *Learned Publishing*, *31*(2), 95-106. doi:10.1002/leap.1110
16. Zuiderwijk, A., & Spiers, H. (2019). Sharing and re-using open data: A case study of motivations in astrophysics. *International Journal of Information Management*, *49*, 228-241. doi: 10.1016/j.ijinfomgt.2019.05.024