A Bibliometric Review of Publications on Research Data Management

*Preeti   **** Amit Kumar#**

* Research Scholar, Department of Library and Information Science, Maharshi Dayanand University, Rohtak, Haryana - 124001, INDIA; Email: preetidlis97@gmail.com
** Research Scholar, Department of Library and Information Science, Maharshi Dayanand University, Rohtak, Haryana - 124001, INDIA; Email: amit.tkrn@gmail.com
# Corresponding author.

_____

Abstract

Research Data Management (RDM) is the process, policies, and tools used to keep track of research data throughout the research lifecycle. The present study aims to identify the growth of the literature, the most prolific authors, institutes and countries and their collaboration patterns in RDM research. To accomplish the objectives, a bibliometric analysis approach was adopted, and data was collected from the Scopus database. The extracted data was analysed using MS Excel, R-package bibliometrix, and VOSviewer. A total of 1088 documents have been published on this topic in 465 sources, receiving 8088 citations. "Article" was the most prominent document type. The vast majority of the documents were published under the Computer Sciences subject domain. C. Ribeiro from the Universidade do Porto was the most prolific author in this field, while the United States was the most prolific country. The study also found that only 15% of the documents received nearly 75% of the total citations. The results of this study give important information about the growth, authorship, and citation trends of RDM literature, which can help guide future research in this field.

_____

1. Introduction

Academic libraries have been important places for learning and research. They give students and scholars access to valuable resources (Gumpenberger, Wieland, & Gorraiz, 2012). Traditionally, library's research assistance was focused on resource discovery, collection development, and information management (Auckland, 2012). The emergence of technology and economy have changed the way academic libraries serve their users (Ball & Tunger, 2006). In recent times, research data has become increasingly valuable, particularly for businesses and universities. Researchers may use data long after a project's funding has ended, with subsequent studies analyzing and contributing to the data, making it a vital part of the research lifecycle (Stobierski, 2021). Therefore, research data management (RDM) has gained significant interest, and academic libraries are striving to become integrated into the RDM ecosystem (Ashiq, 2020). This marks a "revolutionary" shift for academic libraries (Hswe and Holt, 2011).

Cox et al. (2019) say that RDM includes things like data literacy, repository management, metadata tagging, collection management, and data retrieval. By using RDM, academic

libraries can get into the "black box" of research and provide more than just academic materials and can help in research (Koltay, 2016; Cox and Tam, 2018).

Even though there has been a lot of research on RDM, not much is known about how the current literature is put together and how it is growing. So, the goal of this paper is to use bibliometric methods to look into the "pattern" of scholarly writing about RDM. The main goal of this research is to find out more about the most important publications, most prolific authors, and most important ideas in the field of RDM. By analysing the scholarly literature on RDM, the authors aim to assist researchers, librarians, and other stakeholders in identifying research gaps and developing strategies to improve RDM practises and services. Ultimately, our goal is to promote the effective utilisation of research data in various fields.

2. Review of Literature

The effective management of research data is becoming increasingly important in scholarly communication and research practice. Bibliometric analysis is an important tool to examine the trends and patterns of research in this field. This literature review aims to explore the bibliometric studies on research data management (RDM) and identify key findings and themes.

Several bibliometric studies have been conducted on RDM, including those by Ardito et al. (2019), Gupta et al. (2021), Zhang and Eichmann-Kalwara (2019), and Pradhan and Zala (2021). These studies highlight the growth of research output on RDM, the most cited publications, the most active authors and institutions, and the key research topics and themes. They also reveal the interdisciplinary nature of RDM research, with contributions from various fields such as library and information science, computer science, and business and management.

The studies also identify some of the challenges and gaps in RDM research, including the need for more studies on the use of RDM practices in different disciplines, the development of standardized data sharing policies and infrastructure, and the importance of addressing ethical and legal issues related to data management. Overall, this literature review highlights the importance of bibliometric analysis in understanding the trends and patterns of RDM research, and the need for further research to address the gaps and challenges in this field.

3. Objectives

The primary objective of this study is to examine the growth pattern of RDM literature. The specific objectives of the study are as follows:
    a. To find the type of documents under which the RDM literature published;
    b. To check the subject area under which these papers appeared;
    c. To know the most preferred sources for publication;
    d. To identify the most prolific authors, institutes, countries;
    e. To explore the collaboration pattern at international level;
    f. To examine the citation pattern;
    g. To explore the most frequently used keywords.

4. Research Methodology

The present study employed a research design known as bibliometric analysis, which is a recognized method for evaluating and analysing scientific literature. Bibliometric involves

using quantitative analysis of bibliographic data to uncover patterns, trends, and relationships within a specific research field.

The data for this study was obtained from Scopus, a widely respected indexing and abstracting database that encompasses a wide range of scholarly literature across different disciplines. Scopus was chosen as the data source due to its comprehensive coverage and suitability for bibliometric analysis. To collect the relevant data, a search string was developed: "(TITLE-ABS-KEY ("Research Data Management") OR TITLE-ABS-KEY ("Research Data Service*") OR TITLE-ABS-KEY ("Research Data Admin*") OR TITLE-ABS-KEY ("Data Management Plan")) AND (LIMIT-TO (LANGUAGE , "English"))". This search string ensured that the retrieved articles were focused on research data management and were written in English. The data collection was collected on December 30, 2022. MS Excel facilitated data cleaning, organization, and basic statistical analysis. For advanced bibliometric analyses, the R-package bibliometrix was used to calculate publication counts, perform citation analysis, analyze co-authorship networks, and identify keyword co-occurrence. VOSviewer was employed to generate visual representations.

5. Data Analysis

Table-1 provides a brief summary of the literature on research data management. The first RDM publication dates back to 1974. Since then, 1088 publications have been published in 465 sources. These papers were written with the aid of 30965 references and received an average of 7.43 citations per paper, with a total of 8088 citations. A total of 3374 authors contributed to these papers with an average of 3.89 authors per paper. There were only 213 single-authored documents.

Table 1: Publication Summary

Description	Results
MAIN INFORMATION ABOUT DATA	
Publication years	1974-2022
Sources (Journals, Books, etc.)	465
Documents	1088
Annual Growth Rate %	1.42
Average citations per document	7.434
References	30965
Author's Keywords (DE)	2190
Authors	3374
Single-authored documents	213
Co-Authors per Documents	3.89
International co-authorships (%)	14.06

5.1 Literature Growth

Figure-1 illustrates the growth of the RDM literature. 1974 was the year when the first publication on RDM appeared. Prior to 2010, the number of publications each year was low. In 2010, only two papers were published in this area of research. After that the number of publications in a year has increased. In 2022, the highest number of publications (143) were published.
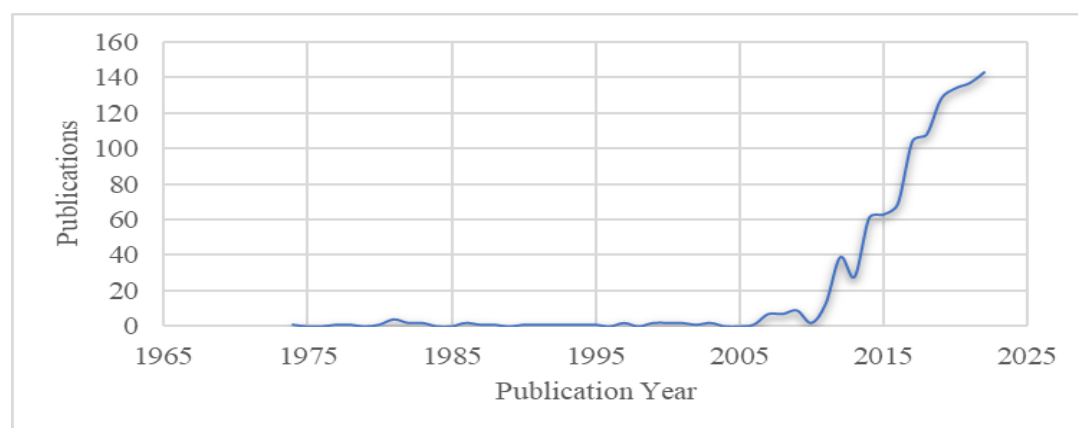
Figure 1: Literature Growth

5.2 Types of Document

The document types of the publications under the study are shown in Table-2. Out of 1088 publications, 615 (56.53%) were Articles, 318 (29.23%) were Conference Paper, and 55 (5.06%) were Book Chapters. Nearly ninety percent of all publications were published in these three categories. Articles received the highest number of citations (6505), followed by Conference Paper (956) and Review papers (442). The citations on articles were nearly three-quarters of the total citations. The average number of citations per paper was highest for articles (10.58), followed by books (10.00) and review articles (5.00).

Table 2: Documents Type

Type of Documents	Number of Publications	Share (%)	Total Citations	Citations Per Paper
Article	615	56.53	6505	10.58
Conference Paper	318	29.23	956	3.01
Book Chapter	55	5.06	68	1.24
Review	52	4.78	442	8.50
Conference Review	27	2.48	0	0.00
Note	10	0.92	83	8.30
Editorial	5	0.46	14	2.80
Book	2	0.18	20	10.00
Erratum	2	0.18	0	0.00
Data Paper	1	0.09	0	0.00
Short Survey	1	0.09	0	0.00
Total	**1088**	**100.00**	**8088**	**7.43**

5.3 Subject Category

The research on RDM is not limited to any specific subject. The top 10 subject categories are listed in Table *3*. The highest number of papers (498) were published in the subject category 'Computer Science', which received the highest number of citations (2732), followed by 'Social Sciences' (140) with 961 citations. Documents published in the 'Agricultural and Biological Sciences" category received the highest number of citations per document (33.85), followed by 'Biochemistry, Genetics and Molecular Biology' (27.26).

Table 3: Subject Category

Subject	Total Publications	Total Citations	Citation per Document
Computer Science	498	2732	5.49
Social Sciences	140	961	6.86
Medicine	132	1571	11.90
Engineering	113	363	3.21
Mathematics	113	592	5.24
"Arts and Humanities"	55	170	3.09
"Biochemistry, Genetics and Molecular Biology"	54	1472	27.26
Decision Sciences	49	250	5.10
Environmental Science	41	396	9.66
Agricultural and Biological Sciences	40	1354	33.85
Total	**1235**	**9861**	**7.98**

5.4 Most Preferred Sources

The top ten sources that have published the highest number of articles on RDM are shown in Table *4*. The source with the highest number of articles was "*Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics*" (19), followed by "*Data Science Journal*" (38) and "*Communications In Computer And Information Science*" (32). The articles published in "*The IFLA Journal*" have received the highest number of citations (207) and have the highest APCC (11.5).

Table 4: Most Preferred Sources

Sources	Total Publication	Total Citation	Citation Per Paper	CiteScore (2021)
Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics	49	112	2.29	2.1
Data Science Journal	38	143	3.76	2.5
Communications In Computer and Information Science	32	89	2.78	0.9
Ceur Workshop Proceedings	28	40	1.43	1.1
ACM International Conference Proceeding Series	18	23	1.28	--
IFLA Journal	18	207	11.50	1.4
Proceedings of The Association for Information Science And Technology	15	19	1.27	1.1
Journal of Academic Librarianship	14	152	10.86	2.9
Liber Quarterly	14	77	5.50	2.1
Procedia Computer Science	14	84	6.00	3.6
Total	**240**	**946**	**3.94**	**--**

5.5 Most Prolific Authors

The most prolific authors on research data management are shown in Table *5*. Among these 10 authors, 3 were from Portugal, 2 form Germany and the USA, each. C. Ribeiro has contributed 23 papers, followed by J.A. Castro (16 articles), and J.R. Da Silva (11). J.A. Castro and J.R. Da Silva were from the same institute. T. Koltay received the highest citations (180) in the list with the highest average citation per paper (20.00).

Table 5: Most Prolific Authors

Author's Name	Affiliation	Total Publication	Total Citation	Citation Per Paper	h-index
Ribeiro, C.	Universidade do Porto, Portugal	23	139	6.04	7
Castro, J.A.	Institute for Systems and Computer Engineering, Technology and Science, Portugal	16	124	7.75	6
Da Silva, J.R.	Institute for Systems and Computer Engineering, Technology and Science, Portugal	11	41	3.73	4
Cox, A.M.	The University of Sheffield, Information School, United Kingdom	10	459	45.90	10
Miksa, T.	Technische Universität Wien, Austria	10	31	3.10	3
Heuer, A.	Universität Rostock, Germany	9	21	2.33	3
Koltay, T.	Institute of Learning Technologies, Hungary	9	180	20.00	6
Auge, T.	Universität Rostock, Germany	8	20	2.50	3
Chard, K.	The University of Chicago, United States	8	121	15.13	5
Foster, I.	The University of Chicago, United States	8	121	15.13	5
Total		**112**	**1257**	**11.22**	

5.6 Most Productive Institutes

Table-6 reveals the top 10 most productive institutes working on RDM. These institutes produced 173 papers receiving 1395 citations. Universidade do Porto, Portugal has produced 32 papers followed by "Institute for Systems and Computer Engineering, Technology and Science" (31 papers) which received 133 and 146 citations simultaneously. The University of Sheffield has produced only 13 papers on RDM but these papers received the highest citations per paper in the list.

Table 6: Most Productive Institutes

Institute	Total Publication	Total Citations	Citation Per Paper
Universidade do Porto, Portugal	32	133	4.16
Institute for Systems and Computer Engineering, Technology and Science	31	146	4.71
Karlsruher Institut für Technologie	21	91	4.33
Rheinisch-Westfälische Technische	19	64	3.37

Hochschule Aachen				
The University of Sheffield	13	506	38.92	
Georg-August-Universität Göttingen	13	126	9.69	
The University of Edinburgh	12	151	12.58	
University of Illinois Urbana-Champaign	11	98	8.91	
Purdue University	11	58	5.27	
Universität Rostock	10	22	2.20	
Total	**173**	**1395**	**8.06**	
Note: One paper may appear under two or more institutes.				

5.7 Most Productive Countries and Collaboration

The top 10 most productive countries are shown in Table *7*. It was observed that 108 countries have produced all these 1088 papers. Alone the USA contributed to nearly one-third papers (29.04%) of the total publications, followed by Germany (19.3%), and the UK (10.66%). Only these three countries produced around 60% of the publications. The top ten most productive countries have published 803 (73.81%) papers. Although the number of publications counted in the table was 928 papers, that shows that some papers were produced in collaboration.

Table 7: Most Productive Countries

Rank	Country	Total Publication (N=1088)	Percentage (%)	Citations	Citation Per Paper
1	United States	316	29.04	4206	13.31
2	Germany	210	19.30	1125	5.36
3	United Kingdom	116	10.66	1376	11.86
4	Netherlands	54	4.96	412	7.63
5	Canada	48	4.41	492	10.25
6	Australia	44	4.04	550	12.50
7	Portugal	43	3.95	168	3.91
8	France	37	3.40	241	6.51
9	Italy	31	2.85	318	10.26
10	India	29	2.67	111	3.83
Total number of papers published by top 10 countries		803	73.81	7039	8.77

The collaboration amongst the countries which has more than 10 publications were only selected for this present study. Figure *2* illustrates the collaboration pattern among the different countries.
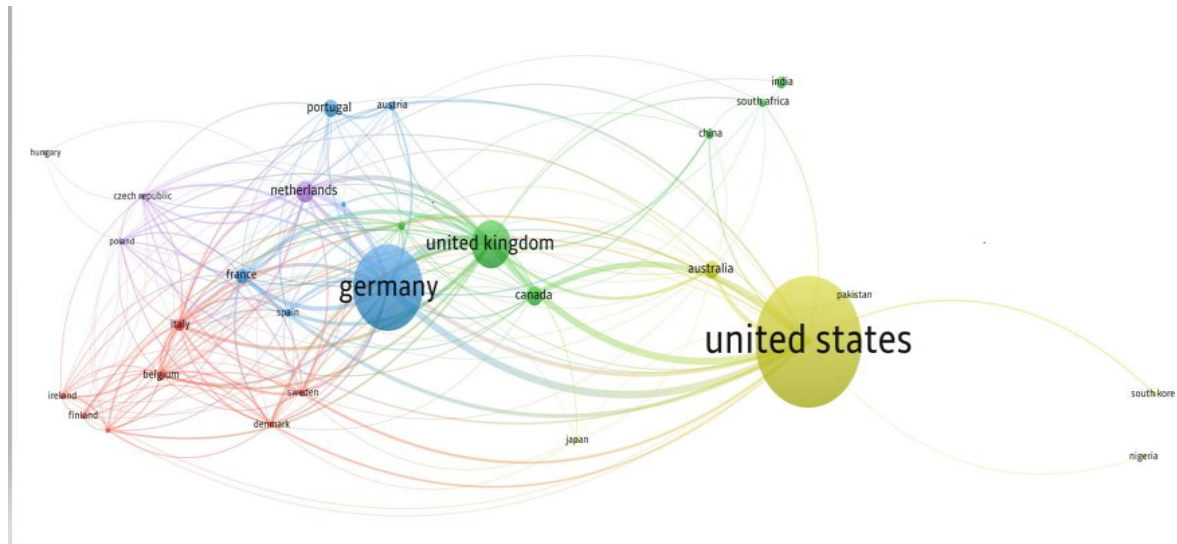
Figure 2: International Collaboration among the Countries

5.8 Citation Pattern

Citation pattern on the publications on RDM is shown in Table *8*. It was observed that only 10 papers (which received more than 100 citations) received 2139 (26.45%) citations. Nearly 15% (which received more than 10 citations) received approximately 75% of the total citations. Around one-third papers received no citations.

Table 8: Citation Profile

No. of Citations	Total Publications	Percentage (%)	Total Citations	Percentage (%)
Zero	337	30.97	0	0.00
1	171	15.72	171	2.11
2	105	9.65	210	2.60
3	69	6.34	207	2.56
4	61	5.61	244	3.02
5	56	5.15	280	3.46
6	39	3.58	234	2.89
7	27	2.48	189	2.34
8	25	2.30	200	2.47
9	15	1.38	135	1.67
10	13	1.19	130	1.61
11-15	56	5.15	721	8.91
16-20	33	3.03	568	7.02
21-30	36	3.31	885	10.94
31-40	16	1.47	578	7.15
41-50	6	0.55	284	3.51
50-100	13	1.19	913	11.29
≥100	10	0.92	2139	26.45
Total	**1088**	**100.00**	**8088**	**100.00**

5.9 Most Frequently Used Keywords

Figure-3 illustrates the most frequently used keywords. The most frequently used keywords were "Research Data Management" besides of that "Data Management", "Open Science", "Data sharing", "Research data" "data curation" "Open data", "fair data", "data repository" etc were also frequently used in the context of RDM.
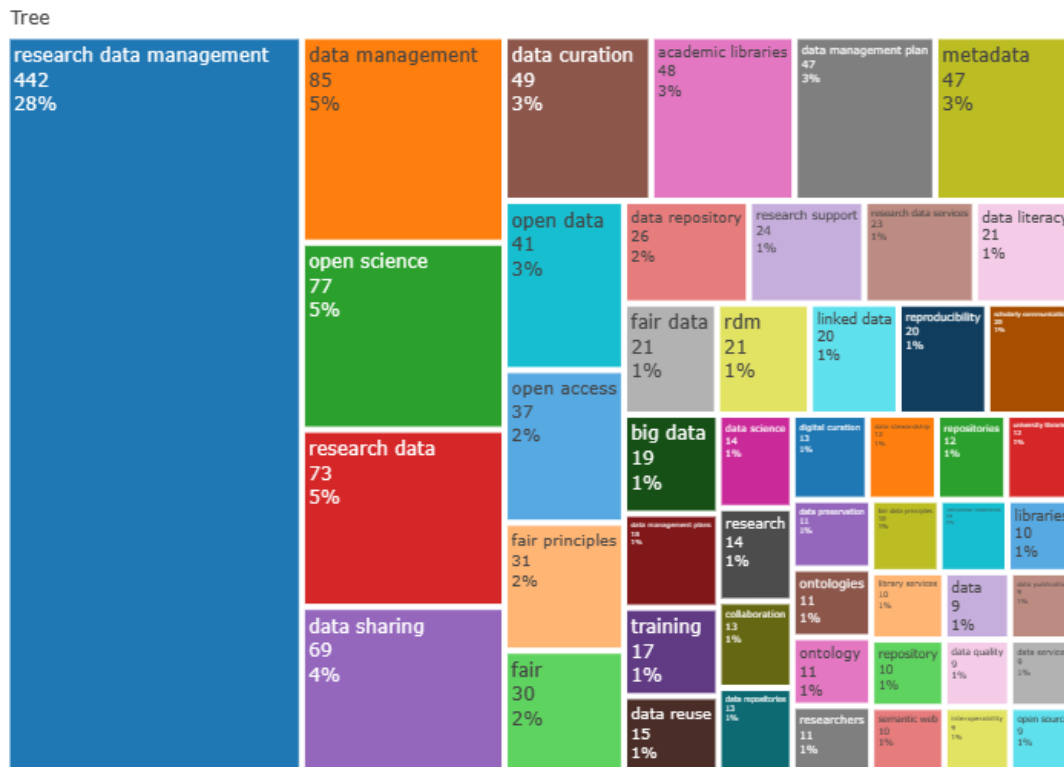

Figure 3: Most Frequently Used Keywords

5.10 Research trend in RDM

Figure-4 shows the most recent trends in research related to RDM. The figure shows that machine learning, AI, fair principles, fair data, open science, data management plan are the current area of research in Research Data Management. Looking towards the future, emerging trends in research related to RDM include data ethics, privacy preservation, blockchain applications, semantic technologies, and federated data sharing for enhanced collaboration and reproducibility.
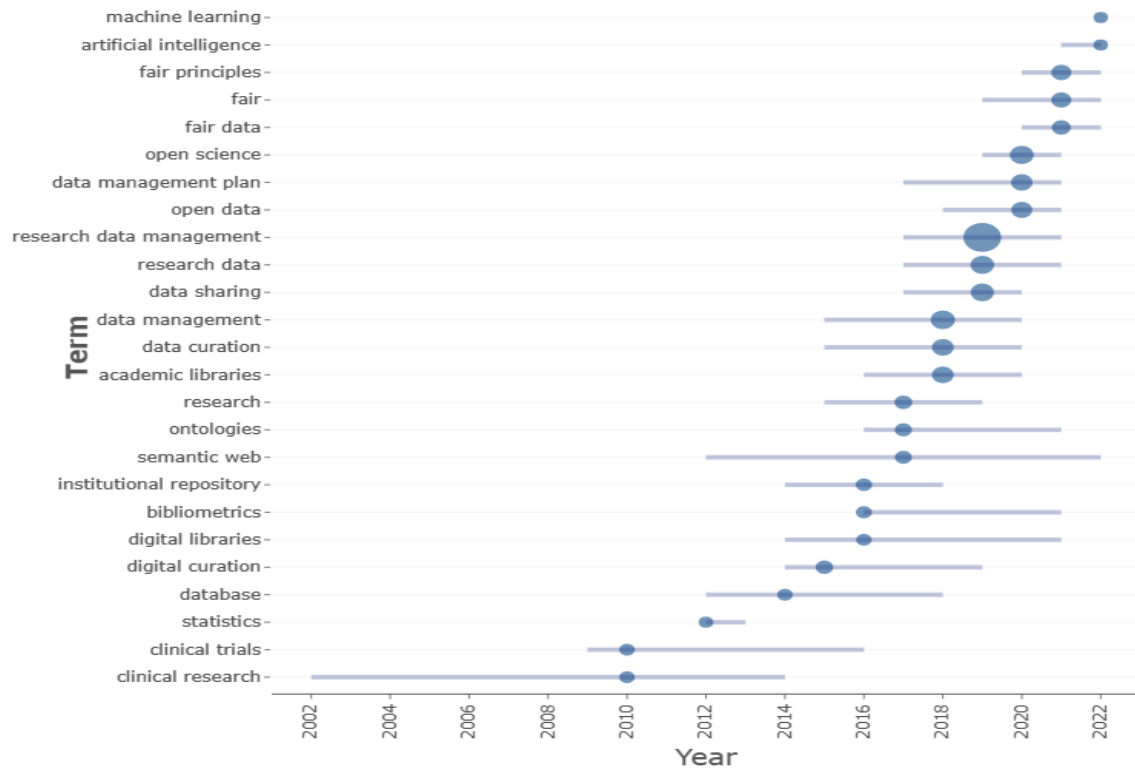
Figure 4: Topic Trend

6. Discussion

The findings of this study provide valuable insights into the research landscape of "Research Data Management" (RDM). Over the span of 48 years until 2022, a total of 1088 papers on RDM were published in Scopus. These papers originated from 465 different sources, indicating a broad interest and involvement across various platforms. These papers collectively received 8088 citations, reflecting the impact and influence of RDM research.

One notable trend observed in the data is the exponential increase in the annual publication rate of RDM papers since 2010. This suggests a growing recognition of the importance of RDM and an increased focus on conducting research in this field. Such a trend may be attributed to factors like advancements in data management practices, evolving research requirements, and the increasing availability of research data.

In terms of document types, the most prominent types were Articles and Conference papers, which is consistent with the scholarly nature of RDM research. Furthermore, it was found that Articles received the highest citations per article, indicating their significance and impact within the RDM community.

The subject category "Computer Science" emerged as the dominant field within which RDM research was published. This finding aligns with the growing reliance on computational methods and technologies for data management. However, it is noteworthy that papers published under the subject category "Agricultural and Biological Sciences" received the highest citations per paper, highlighting the interdisciplinary nature of RDM and its relevance to diverse domains.

In terms of sources, "Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics" and "Data Science Journal" emerged as the most preferred sources, demonstrating their contribution to the RDM literature. These sources have provided platforms for researchers to disseminate their work and contribute to the advancement of RDM knowledge.

Examining author contributions, "Ribeiro, C." from "Universidade do Porto, Portugal" emerged as the most prolific author with 23 publications that garnered 139 citations. This highlights the significant contribution made by this individual in advancing RDM research. Additionally, the University of Porto emerged as the leading institution in terms of the number of papers published, indicating their institutional commitment and expertise in RDM.

In terms of citations received, the University of Sheffield stood out with an average of 38.92 citations per paper, highlighting the impact and recognition of their research output. Moreover, the United States emerged as the leading country in both the number of papers published and the number of citations received, emphasizing its influential role in shaping the field of RDM.

Interestingly, the citation analysis revealed that a small proportion of highly influential papers shaped the field of RDM. Only 15% of publications that received more than 10 citations accounted for 75% of the total citations. This suggests the presence of key influential papers that have significantly impacted the understanding and development of RDM.

7. Conclusion

This study provides valuable insights into the research landscape of Research Data Management (RDM). The analysis of 1088 papers published over 48 years reveals a broad interest in RDM across 465 different sources. The exponential increase in annual publication rates since 2010 suggests growing recognition of RDM's importance. Articles and Conference papers are the most prominent document types, with Articles receiving the highest citations per article. While RDM research is predominantly published in the field of Computer Science, papers in Agricultural and Biological Sciences receive the highest citations per paper, indicating interdisciplinary relevance. "Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics" and "Data Science Journal" are preferred sources. Notably, "Ribeiro, C." from "Universidade do Porto, Portugal" is the most prolific author, and the University of Porto leads in institutional contributions. The University of Sheffield receives the highest number of citations, and the United States leads in both publications and citations. The presence of a small number of highly influential papers shaping the field highlights their significant impact on RDM's understanding and development.

References

1. Ardito, L., Scuotto, V., Del Giudice, M., & Petruzzelli, A.M. (2019). A bibliometric analysis of research on big data analytics for business and management. *Management Decision, 57*(8), 1993-2009. doi:10.1108/MD-07-2018-0754
2. Ashiq, M., Usmani, M.H., & Naeem, M. (2022). A systematic literature review on research data management practices and services. *Global Knowledge, Memory and Communication, 71*(8-9), 649-671. doi:10.1108/GKMC-07-2020-0103
3. Auckland, M. (2012). Re-skilling for research: An investigation into the roles and skills of subject and liaison librarians required to effectively support the evolving information

needs of researchers. London: RLUK Research Libraries UK. Retrieved October 22, 2012, from http:/ /www.rluk.ac.uk/content/re-skilling-research.

4. Ball, R., & Tunger, D. (2006). Bibliometric analysis – a new business area for information professionals in libraries? Support for scientific research by perception and trend analysis. *Scientometrics, 66*, 561–577.

5. Cox, A.M. & Tam, W.W.T. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management, 70*(2), 42-157.

6. Cox, A.M., Kennan, M.A., Lyon, L., Pinfield, S. & Sbaffi, L. (2019). Maturing research data services and the transformation of academic libraries. *Journal of Documentation, 75*(6), 1432-1462.

7. Gumpenberger, C., Wieland, M., & Gorraiz, J. (2012). Bibliometric practices and activities at the University of Vienna. *Library Management, 33*, 174-183.

8. Gupta, N., Arora, S., & Chakravarty, P.R. (2021). Science mapping and visualization of research data management (RDM): Bibliometric and scientometric study. *Library Philosophy and Practice, 2021*, 1-23.

9. Hswe, P. & Holt, A. (2011). Joining in the enterprise of response in the wake of the NSF data management planning requirement. *Research Library Issues, 274*, 11-17.

10. Koltay, T. (2016). Are you ready? Tasks and roles for academic libraries in supporting research 2.0. *New Library World, 117*(1), 94-10.

11. Pradhan, P., & Zala, L.N. (2021). Bibliometrics analysis and comparison of global research literatures on research data management extracted from scopus and web of science during 2000 - 2019. *Library Philosophy and Practice, 2021.*

12. Stobierski, T. (2021). 8 Steps in the data life cycle. Retrieved from https://online.hbs.edu/blog/post/data-life-cycle. On 28th December 2021.

13. Zhang, L., & Eichmann-Kalwara, N. (2019). Mapping the scholarly literature found in scopus on research data management: A bibliometric and data visualization approach. *Journal of Librarianship and Scholarly Communication, 7*(1), eP2226. doi:10.7710/2162-3309.2266.

■■■