September 2021

Science Mapping and Visualization of Research Data Management (RDM): Bibliometric and Scientometric Study

Nidhi Gupta Research Scholar
Panjab University, Chandigarh, India, nidhigupta@pu.ac.in

Surbhi Arora
Panjab University, Chandigarh, India, arora.surbhi03@gmail.com

Prof. Rupak Chakravarty
Panjab University, Chandigarh, India, rupak@pu.ac.in

Science Mapping and Visualization of Research Data Management (RDM): Bibliometric and Scientometric Study

(i) **Nidhi Gupta**, Department of Library & Information Science, Panjab University Chandigarh 160014, https://orcid.org/0000-0003-4244-4470

(ii) **Surbhi Arora**, Department of Library & Information Science, Panjab University,Chandigarh.

(iii) **Prof. Rupak Chakravarty,** Department of Library & Information Science, Panjab University Chandigarh 160014, https://orcid.org/0000-0001-5046-1663

Abstract:

Research Data Management (RDM) is an ever-evolving phenomenon focused to augment and manage the research capital of an organization, especially in context with developing countries. This study aims at to understand and highlight the state of art of RDM literature between 1989 and 2021 using Web of Science Core Collection (WoSCC) database and VOSviewer software with a bibliometric or scientometric approach thereby highlighting the most influential authors, countries, journals, institutions, and to their co-authorship pattern, co- citation pattern, bibliographic coupling pattern, the co-occurrence of keywords pattern in the field of RDM.The data comprises of a total of 797 documents ,further analysed by using VOSviewer software. Visualization analysis reveales that the number of publications related to RDM are increasing year by year, reaching a peak in 2020. The most productive author in this field is Rafael Alexandrine-benavent and USA is the most productive country published on RDM. The Zeitschrift fur Bibliothekswesen und Bibliographie Journal ismost influential journal. The most frequently used keywords are data sharing, research data management, research data, science and metadata. The analysis shows collaboration relation between authors, countries and sources. The visualizations conducted on this topic offer exploratory information on current status and to identify the main trends in RDM research and its future research initiatives.

Keywords: Research data management, Science mapping. Co-citation. Bibliographic coupling. Co-occurrence

Introduction

Data is a treasure trove for organizations, especially Higher Education Institutions (HEIs). Data often have a longer lifespan than the research project that creates them. Researchers may continue to work on data after funding has ceased, follow-up projects may analyse or add to the data, and data may be re-used by other researchers. From its development for a study to its delivery and reuse, the data lifecycle (DLC) represents all stages of data's life. The DLC starts with researchers creating a study concept; once a study concept is created, data for that study is collected. Generation, Collection, Processing, Storage, Management, Analysis, Visualization, and Interpretation are the eight common stages, steps, or phases of this DLC [[12]].

Data is collected, stored, and distributed so that it can be archived and used by other researchers at a later time. When data enters the dissemination stage of the lifecycle, it is saved in a place (i.e., archive, registry) where other researchers can locate and access it. Data repurposing follows data discovery, creating a continuous chain back to the original data processing point, where the repurposed data is archived and dispersed for discovery [[4]]. Data that are well-organised, registered, maintained, and shared are essential for advancing scientific inquiry and expanding opportunities for learning and innovation. Data management is crucial for making data discoverable, available, and understandable, and making them discoverable, accessible, and understandable is an important part of what librarians do [[13]]. Data is the starting point of the research journey as input as-well-as, it is the most cared for outcome of the research process in the form of study findings. Research as a systematic, planned process, where data are being captured, analyzed and produced.

Background:

Research 2.0 is a modern research methodology that needs more attention to activities such as Research Data Management (RDM) [[9]]. The RDM is associated with several significant factors. Such factors include more visibility and citations, opportunities for collaborations, career recognition, enables new research and new insights on the data, increase transparency and trust in science.

Research Data Types (Forms and Formats)

Research data can be categorized for different purposes in different categories. Its forms might be:

- laboratory notebooks, field notebooks, diaries
- documents, spreadsheets
- questionnaires, transcripts, codebooks
- audiotapes, videotapes

- photographs, films
- test responses
- slides, artefacts, specimens, samples
- collections of digital outputs
- database contents (video, audio, text, images)
- models, algorithms, scripts
- contents of an application (input, output, logfiles for analysis software, simulation software, schemas)
- methodologies and workflows
- standard operating procedures and protocols
- Scientific and statistical data formats
- Standard office documents
- Images
- Raw Data
- Plain Data
- Structured graphics
- Structured text
- Archived data
- Text
- Numeric
- Multimedia
- Development code
- Syntax

In addition to research results, researchers often must think about how to handle the corresponding research documents during and after the life of a project-that is, administrative materials and supporting documentation-during a research project. Some examples of these research records are given in the list below:

- Correspondence (electronic mail and paper-based correspondence)

- Grant applications
- Ethics applications
- Technical reports
- Technical appendices
- Research reports
- Research publications
- Signed consent forms
- Social media communications (blogs, wikis, tweets, etc.)

The RDM is a task of various activities where data are neither static nor isolated. Data management ensures that the narrative of a researcher's data collection process is structured, understandable, and transparent. RDM is the organisation, management, publication and preservation of the products of research. Research products result from our scientific and scholarly research and may include: field observations, specimens or other materials collected, computer software, simulation models, imagery, and metadata (information about these data). Being crucial for the research footprint and research lifeline of a research-driven academic organisation, it is important to understand the research trends in the field of RDM, which can be well undertaken using the scientometric or bibliometric analysis.

Bibliometrics is a field of study that quantifies bibliographic metrics and offers broad overviews of a collection of documents. Scientometric analysis is used to map the scientific knowledge field critically, wherein the critical examination is used to classify research trends and problems based on the scientometric findings [17]. Scientometrics is a sub-field of bibliometrics. Scientometrics and bibliometrics are methodological approaches in which the scientific literature itself becomes the subject of analysis. In a sense, they could be considered a science of science [14]. These two methods for quantitative studies were used to explore the research field of RDM for assessing performance and science mapping. Performance analysis is used to evaluate the citation impact of the scientific publications of different actors in the field of science. Science mapping, or bibliometric mapping, aims to visually present and display the conceptual, social or intellectual structures of scientific research, and the evolution, development and dynamics thereof. Scientometrics and bibliometrics has made substantial advancements in recent years due to the development of computers, internet connections, databases, algorithms and programs used to perform bibliometric analyses.

Literature review

Palsdottin (2021) study data management literacy activities in the Arts, Technology, and Social Science disciplines. The ultimate aim of the study was to look into the management of research data activities and the sharing of research data habits among academics and doctoral students at the University of Iceland. A Lime Survey was used to gather data for the research, which was based on a quantitative study. The overall response rate was 18%, with graduate students receiving a response rate of 24 percent and academic researchers receiving a response rate of 15%. Most participants were confused about DMPs, servers, metadata requirements, and study data was the most common practice, according to the findings. Data management was also not a common practice in any discipline, according to the report. Respondents were keeping data on their own computers for a long time. The study also found that participants had a poor understanding of research data literacy. The need for training assistance was demonstrated by the respondents. The findings showed that researchers' knowledge and understanding of the value of data management approaches needed to be strengthened immediately.

Lefebvre, Bakhtiari and Spruit (2020) explored challenges when implementing RDM. The study explored the technical and organizational challenges faced during the RDM planning stage, in particular during the grant submission process. The study was focused on primary data, and an explanatory case analysis was conducted in the Netherlands to gather evidence. Data were collected using semi-structured interviews face-to-face or Skype. The interviews were recorded, transcribed and analyzed from NVivo software (version 12.5). Results found in 3 main categories i.e., the grant application process, the quality of data management plans and challenges of Research DMP. Study noticed various grant application systems of various method forms, and researchers were unable to initiate a project until DMP was accepted by the grant application process. Study concluded with a complex task about soundness and effectiveness to generate reusable data. there was need to provide proper guidelines to researchers on FAIR Principles, DMP etc.

Zhang and Eichmann-Kalwara (2019) used the Scopus database to conduct a bibliometric analysis on RDM. The aim of the study was to explore and interpret the intellectual structure and progression of the current RDM literature using visual perception. The analysis used bibliometric techniques and a visualisation method (CiteSpace) to classify and analyse the trends of scholarly publications about RDM, which were focused on research questions. The bulk of the RDM literature was released after 2002, according to the findings. Within this interdisciplinary area, major research clusters include "Science Collaboration,""research support service," and "data literacy," with

"scientific collaboration" being the most involved. In the RDM literature, terms like "digital curation" and "information processing" appeared frequently. Several particular topics, such as "digital library, "big data," and "data sharing," also saw a substantial rise. The study concluded that the increasing trend of RDM literature would aid researchers in better understanding this rapidly changing environment, identifying research patterns and gaps, and improving capabilities across various fields.

Prado (2016) conducted a bibliometric study involving various information fields on multivariate analysis of credit risk and bankruptcy research results (1968-2014). The study's aim was to find and explain how multivariate data analysis methods were used in credit risk and bankruptcy scenarios. Between 1968 and 2014, data was collected in publications indexed in Thomson Reuters' web of science database. The findings confirmed information from the literature and previous bibliometric reviews, as well as highlighting other indicators of the research field's construction and growth. The increased use of neural networks as a study object in publications had made them important since the 1990s. In journals and Web Science databases, this area tended to be multidisciplinary, encompassing fields such as business and economics, organizational analysis, management, mathematics, data processing, engineering, and statistics. Another important result was the rise in the number of publications on this subject following the 2008 financial crisis. According to the results, studies focused on neural networks have dominated since the 1990s. The research came to a conclusion of important findings and patterns.

Onyancha (2016) conducted a bibliometric study on open research data in Sub-Saharan Africa. The study's goal was to look into the current state of research data sharing among researchers in Sub-Saharan Africa (SSA) and around the world. An advanced search technique was used to retrieve data from the data citation index (DCI), which was limited to publication years between 2009 and 2014. The number of data records by country, organization, subject category, year of publication, and document form, as well as the number of citations, were calculated using the data. To assess the relationship between the data records and research papers, a spearman's correlation analysis was used. Just 20 countries in Sub-Saharan Africa (out of 50) created at least one data record in the DCI, with South Africa leading the way with 539 (61.39 %) records, followed by Kenya, Cameroon, and Ghana. In contrast to 1.4 % of the world's research papers, SSA contributes just 0.03 percent of the world's research results. The main contributors of research data were research institutions and universities, which mostly concentrated on Genetics and Heredity (61.3 %), Biochemistry and Molecular Biology (61.3 %), Agriculture (29.2 %), and Forestry (29.2 %) (27.3 %). The study data had earned fewer citations on average than the papers. According to the findings, the study reached on conclusion that

correlational analysis of the data showed a strong association between data publication and research publications.

Corrall, Kennan and Afzal (2013) found emerging trends in library research support through bibliometrics and RDM Services. The aim of the study was to learn about libraries' changing roles in providing research support services, such as bibiometrics, RDM services, and future plans. A survey of 140 libraries in Australia, New Zealand, Ireland, and the United Kingdom was conducted using an online multiple-choice questionnaire to determine existing and expected programmes, target audiences, service constraints, and staff training needs. According to the report, the majority of respondents provided some sort of bibliometrics service. The most famous resources were bibliometrics training or literacy, citation reports, and research impact calculations. RDM services were also offered at a lower level in all countries, but at the highest level in the United Kingdom. Overall, the study found that differences in expertise, skills, and trust were major roadblocks, with nearly universal support for bibliometrics and, in particular, data management, being included in technical education and continuing development programmes. A majority of respondents offered or planned bibliometrics training, citation reports, and impact calculations, according to the study, but there were significant differences between countries. The majority expected to be involved in the future, especially in the areas of technology assistance, data deposit, and policy growth. According to the results, librarians need a multifaceted understanding of the research environment.

It can be observed from the review of literature on RDM, that to date studies on the evolution of RDM related research area from a global point of view on bibliometric or scientometric analysis over time are still insufficient. The present study on bibliometric or scientometric analysis of RDM literature from 1989 to 2021 using mathematical and statistical approaches, can be used to address this knowledge gap.

Purpose of the study:

The purpose of this study is to understand, from a bibliometric or scientometric perspective, the state of art of RDM research field from 1989-2021. Bibliometric analysis allows to evaluate developments in knowledge on a specific subject and assesses the scientific influence of researches and sources. The study attempts to analyze the interactions between scientific publications, research organizations, scientific sources, countries, researchers, keywords or terms. It will help to identify the past and current trends in the core areas of RDM; identify the relationships of keywords, countries, authors, and

sources through co-authorship, co-occurrence, citation, bibliographic coupling, and co-citation analysis reflecting upon the future directions of research on RDM.

Objectives of the Study:

The main objective of the study are given below:

- To analyze the year wise publications and growth trends in RDM.
- To analyze the highly productive authors, their number of publications, citations and h-Index and their co-authorship pattern.
- To analyze the most prolific countries and their Bibliographic coupling patterns.
- To identify the most popular sources in terms of publications and their co-citation pattern.
- To analyze the institutional productivity and their citation pattern.
- To analyze the co-occurrence pattern of all the keywords in RDM.

Scope and Limitations

The Scope of the present study is to evaluate the research outcomes of RDM literature from 1989-2021.

The information is retrieved from the WoSCC database having the most stringent indexing policy thus generating lesser number of records as compared to Scopus or Google Scholar (GS).

Material and Methods:

This study attempts to "analyze and visualize the literature on the field of RDM employing the principles of science mapping. A map can be created based on a network that is already available, but it is also possible to first construct a network. VV can be used to construct networks of scientific publications, scientific journals, researchers, research organizations, countries, keywords, or terms. Items (a node or a vertex) in these networks can be connected by co-authorship, cooccurrence, citation, bibliographic coupling, or co-citation links. To construct a network, bibliographic database files (i.e., Web of Science, Scopus, Dimensions, and PubMed files) and reference manager files (i.e., RIS, EndNote, and RefWorks files) can be provided as input to VV. Alternatively, VV can download data through an API (i.e., Microsoft Academic API, Crossref API, Europe PMC API, and several others).

Input data for this study were collected from the A/I and Citation database Web of Science Core Collection (WoSCC) in .txt file format from 1989 till march 2021. An advanced search was conducted for the retrieval of data i.e., Title Search TI= ((“Research data management”) OR (“Research data services”) OR (“Research Data”)) from the WoSCC database. A total 731 publications/records were retrieved. No limiters were applied in the retrieved results on the basis of type or language of documents. Retrieved data were exported (.txt) into a folder in batches, each having 500 records. the upper-limit of records that can be exported at a time. In order to process the exported results, they were merged into a single file using command line.

Results and Discussions:

Analysis of Publication Distribution

Figure 1. presents the annual publication distribution of research literature productivity on RDM. As seen in the figure, there is a growth in the number of documents published on RDM research area after 1997, peaking at 1999 with about 21 documents. Post this, a decline can be noted for several years (2000–2012), followed by surge which increased number of publications. In 2018, 72 documents were published related to RDM, becoming the highest year of productivity in the field of RDM equaling 2020 with same number of research. While in 2019 only 61 papers were authored and published.

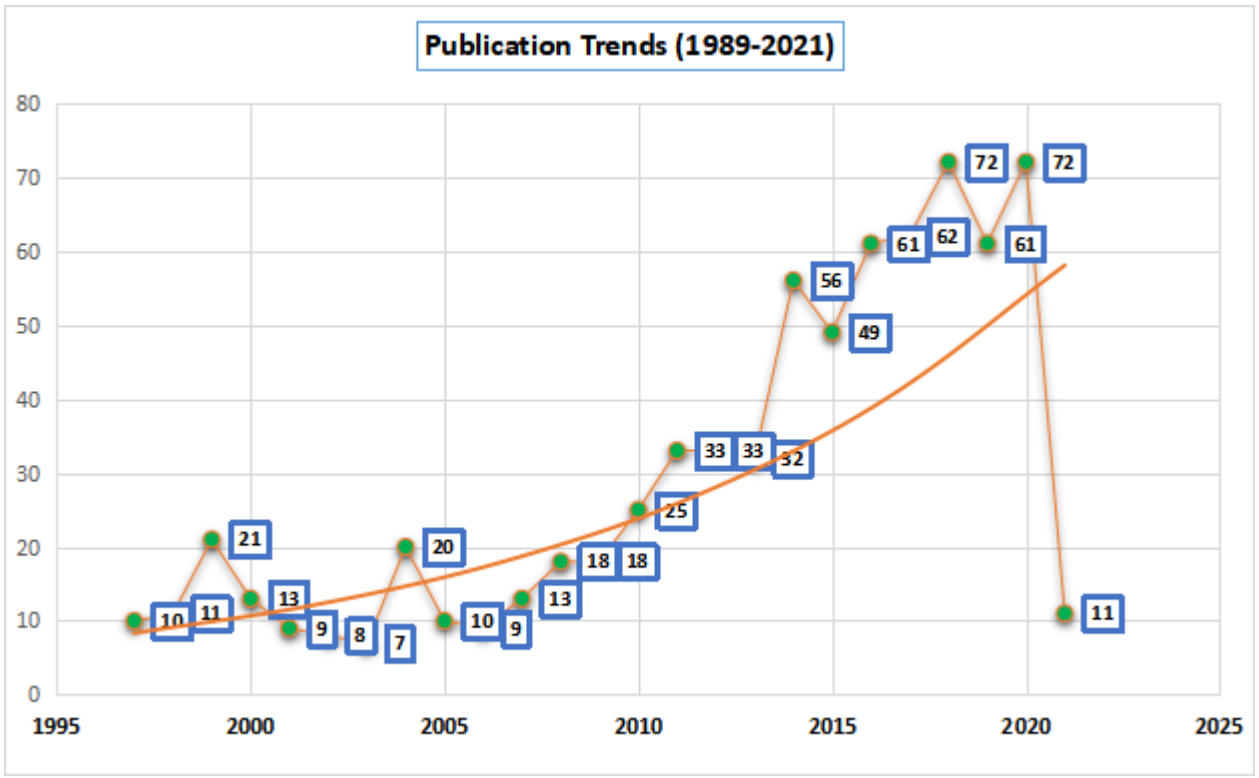

Fig 1. Distribution of annual publications on RDM

Most productive authors and their co-authorship analysis

A total of 2325 authors have contributed in 797 publications. The authors are being ranked on the basis of the total publications, the total citations and h-index. Table.1 shows top 10 authors in the field of RDM from 1989 to march 2021. The author with the highest number of publications is Rafael Alexandrine-benavent (n=8, TC = 39 and h-index = 4). Susan bull and Andrew M. Cox are in the 2nd and 3rd positions (n = 6). Stephen pinfield spot the 5th rank in author-productivity with 6 publications but is the highly influential with TC=161 and h-index of 26.83. Andrew M.Cox (TC=145) and Michael Parker (TC=124) (Ranked 2nd and 4th respectively as most productive authors according to citation). In the h-index, there are total three authors with value 6. First one's are Andrew M. Cox, Michael Parke and Stephen pinfield followed by Susan Bull with h-index of 5. The core authors primarily are from the England, Spain and USA suggesting that the dominant authors are from prolific countries and organizations.

Table.1 The top 10 authors

SN	Author	Country	Number of Publications (n)	Total Citations (TC)	Average citation per pub. (CPP)	Average Pub.Year (APY)	h-index	Total link strength (TLS)
1	Rafael Alexandrine-benavent,	Spain	8	39	4.875	2015.88	4	51
2	Susan Bull	England	6	124	20.6	2015.50	5	119
3	Andrew M. Cox.	England	6	145	24.16	2016.83	6	53
4	Michael Parker	England	6	124	20.66	2015.50	6	119
5	Stephen	England	6	161	26.83	2015.83	6	50

	Pinfield							
6	Fernanda Peset	Spain	5	18	3.6	2014.80	3	20
7	David E. Cohen	USA	4	17	4.25	2016	3	0
8	Felix Engel	Germany	4	5	1.25	2017.75	2	4
9	Rajiv Jalan,	Spain	4	17	4.25	2016	3	0
10	Vicki Marsh	Keney	4	72	18	2015.75	3	54

Figure 2. shows the co-authorship pattern amongst the authors in RDM is examined with VV. The circle size shows the volume of publications of an author. The curved connections represent cooperation between authors. The colour scales indicate the collaboration clusters. The minimum number of documents of an author has been established on one publication. Of the 2325 authors, 2325 meet the thresholds. For each of the 2325 authors, the total strength of the co-authorship links with other authors has been calculated. Some of them are not connected to the other ones, so they are not shown on the graphic. After excluding unconnected authors, the remaining 32 authors are included for the visualization map. From the visualization map, there are four clusters identified. Cluster 1. Red Cluster 2. Green, Cluster 3. Blue, and Cluster 4. Yellow. All the clusters are connected with each other. Cluster yellow has more links with the cluster green and blue as compare to Red which means that the yellow, blue and green clusters have more closeness or connected with each other. As seen in the fig 3b. it is observed from the map that cluster yellow includes most productive or influential authors that shows Susan bull and Michael parker who are collaborated more closely. The largest nodes are of Susan bull and Michael parker with 119 TLS, each and Vicki Marsh with 54 TLS.

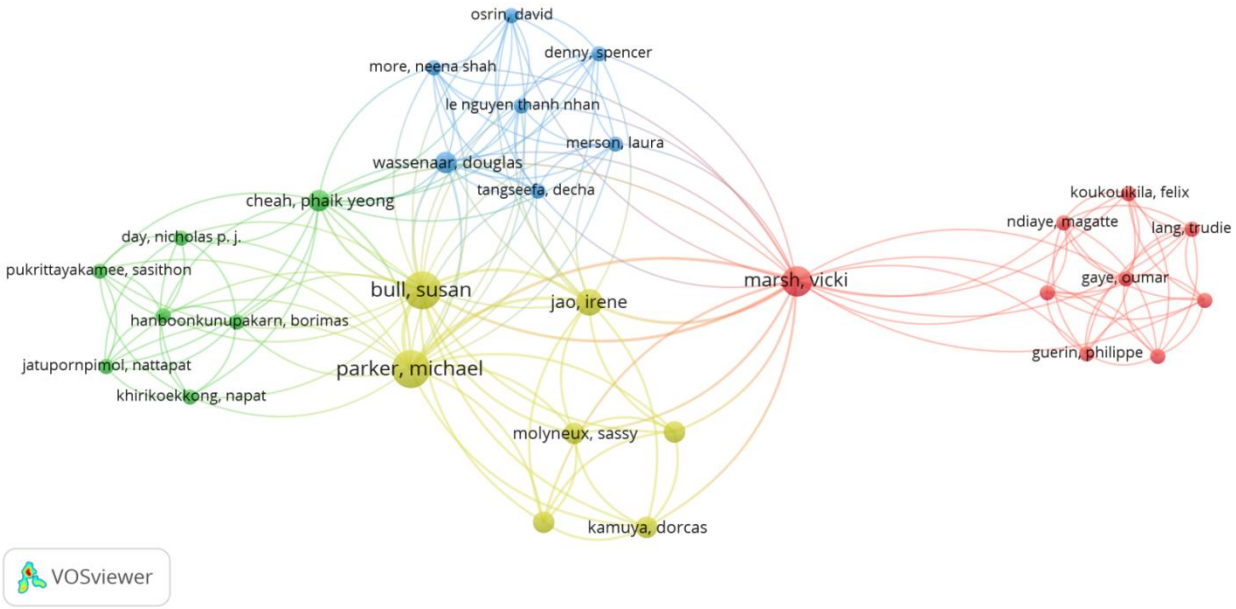

Fig 2. The co-authorship network of Authors by VV (network visualization)

Leading Countries and their bibliographic coupling analysis

A total of 69 countries have participated in the 797 publication on RDM literature. The countries are being ranked on the basis of the total number of publications produced by the country. Table 2. shows the top 10 countries in the field of RDM from 1989 to march 2021. The highest number of documents are produced by USA (n=290) ranked 1st in the most productive countries. England (n=107) ranked 2nd followed by Germany (n= 88), Australia (n=40) and Netherlands (n= 34) are the most productive countries.

Table 2. The top 10 leading countries

SN	Country	Number of Publications (n)	Total link strength (TLS)	Total Citations (TC)	Avg. citation (AC)	Avg. Pub Year (APY)
1	USA	290	277	4258	14.68	2011.71
2	England	107	231	1727	16.14	2014.09
3	Germany	88	53	2461	27.96	2015.65

4	Australia	40	83	633	15.82	2015.40
5	Netherlands	34	41	525	15.44	2015.18
6	Canada	31	51	468	15.09	2013.62
7	Spain	24	45	218	9.08	201.78
8	France	21	23	111	5.2	2014.62
9	Switzerland	18	43	82	4.55	2015.61
10	Italy	13	15	119	9.15	2016.25

Figure 3. shows the bibliographic coupling network of countries. The bibliographic coupling pattern of countries is examined with VV. The circle size shows the volume of countries with publication. The curved connections represent cooperation between countries. The color scales show the collaboration clusters. A country's minimum number of publications was 5. Of the 69 countries, 25 met the threshold. For all of the countries, the number of publications, the number of citations, and total link strength were calculated. The countries with the greatest total link strengths were selected. USA is on number one place (n=290; TC=4258; TLS=277). For the other countries, the first numbers stand for the number of publications, the second one is the number of citations, and the third ones are the total link strengths. The other countries are; England (n=107; TC=1727; TLS=231), Germany (n=88; TC=2461; TLS=53), Australia (n=40; TC=633; TLS=83), Netherlands (n=34; TC=525; TLS=41), Canada (n=31; TC=468; TLS=51), Spain (n=24; TC=218; TLS=45), France (n=21; TC=111; TLS=23), Switzerland (n=18; TC=82; TLS=43), and Italy (n=13; TC=119; TLS=15).

Figure 3. shows four different color of clusters that are more frequently linked with each other. The Cluster.1(Red) composed of Austria, Brazil, Germany, Italy, Japan, Netherlands, People of China, Poland South Korea, and Spain. The Cluster.2 (Green) composed of England, India, Kenya, Russia, South Africa, and Sweden. The Cluster.3 (Blue) composed of Belgium, France, Greece and Switzerland. The Cluster.4 (Yellow) composed of Australia, Canada, Scotland and USA. The countries in the same clusters cite each other more frequently.

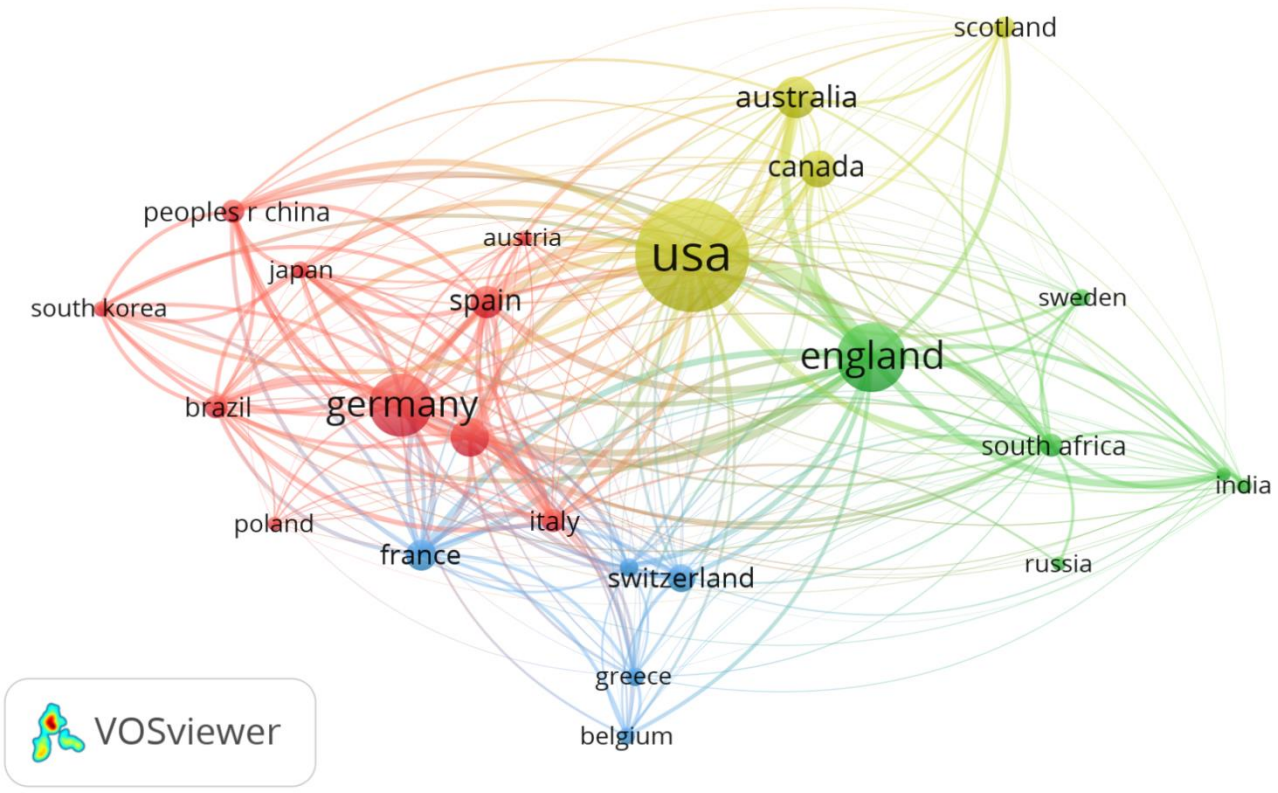

Fig.3 Bibliographic coupling network of leading countries by VV (network Visualization)

Leading Sources and their co-citation analysis

A total of 479 Sources have published 731 publications on RDM literature. The Sources are being ranked on the basis of the total number of publications. Table 3 shows the top 10 Sources in the research area of RDM from 1989 to march 2021. The highest number of documents are published in Zeitschrift fur Bibliothekswesen und Bibliographie Journal (n=21) ranked 1st as the most productive journal published literature on RDM. PLOS One (n=20) ranked 2nd followed by Journal of the American Medical Informatics Association (n=18), Lancet (n=11) and British Medical Journal (n= 8) publications.

Table 3 Top 10 most productive Sources

SN	Sources	Country	Number of Publications (n)	Total Citations	TLS Total link	Avg. citation	SJR-19	H-index	Impact factor-20

				(TC)	streng th (TLS)				
1	Zeitschrift fur Bibliothekswesen und Bibliographie	Germany	21	12	3	0.57	0.11	5	5.87
2	PLoS One	USA	20	1011	34	50.55	1.02	300	2.740
3	Journal of the American Medical Informatics Association	US	18	346	4	19.22	1083	142	4.112
4	Lancet	UK	11	505	21	45.9	14.55	747	60.392
5	British medical journal	UK	8	45	3	5.62	2.05	412	30.223
6	Journal of academic librarianship	UK	8	40	5	5	1.17	55	1.235
7	Program-Electronic library and information systems (Data technologies	UK	8	74	12	9.25	0.38	28	0.556

	and applications)								
8	Journal of Empirical research on human research ethics	USA	7	129	16	18.42	0.62	26	1.253
9	Scientometrics	Netherlands	7	95	10	13.57	1.21	106	2,8670
10	BM-british medical journal	UK	6	10	1	1.6	2.05	412	30.223

Figure 4. shows the co-citation network of cited-sources. The co citation pattern of cited sources is examined with VV. The circle size shows the volume of cited-sources with publication. The curved connections represent cooperation between circles (sources). The color scales show the citation clusters. The minimum number of citations of a source was 5. Of the 8119 sources 526 met the threshold. PLOS One is the most co-cited source (TC= 334; TLS=9707) followed by Nature TC= 147T; TLS=3437) followed by Science (TC=136; TLS=3794), Journal of the American Medical Informatics Association (TC=114; TLS=1683), New England Journal of Medicine (NEJM) (TC=114, TLS=2514). The sources in the same cluster cite each other more frequently. Cluster red shows top cited sources which are also citing each other more frequently.

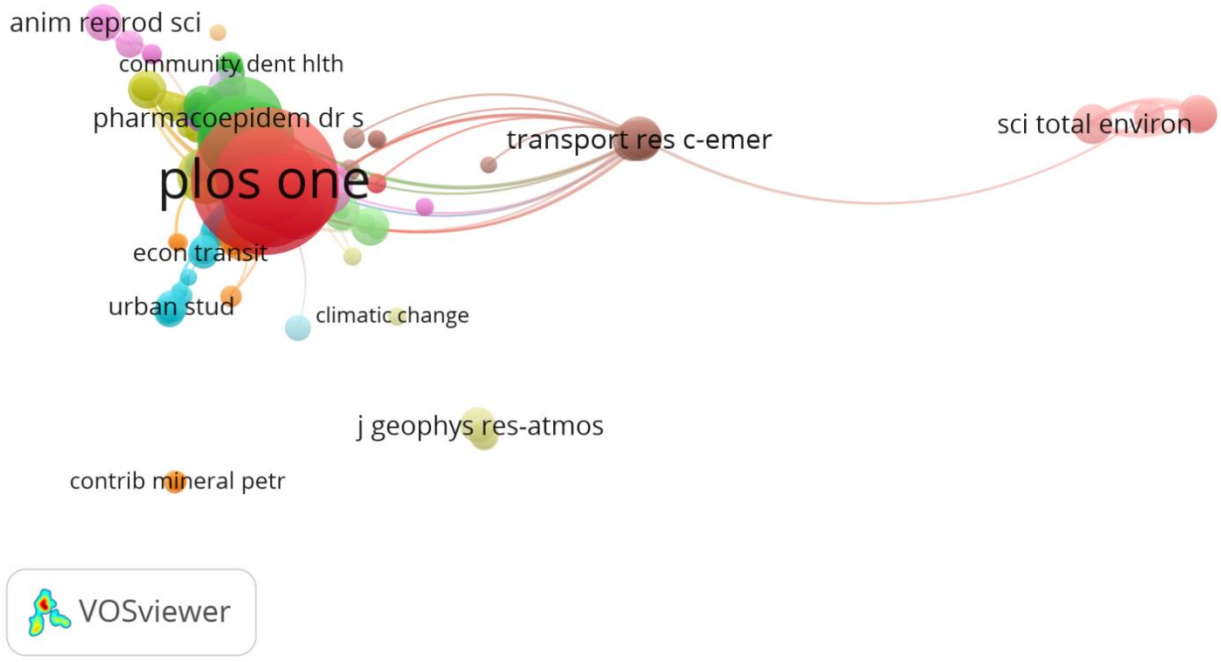

Fig.4 Co-citation analysis of cited- sources by VV (network visualization)

Organizations and their citation pattern

A total of 954 organizations have participated in the 797 publication on RDM literature. The organizations are being ranked on the basis of the total number of publications. Table 4 shows the top ten organizations in the field of RDM from 1989 to march 2021. The highest number of documents are produced by The University of Sheffield (n=15) ranked 1st as the most productive organization. University Pittsburgh (n= 14) ranked 2nd followed by University of Oxford (n=13), UCL (n=11) and Duke University (n=10).

Table 4. The top 10 Organizations

SN	Organization	Country	Documents	TLS	Citations	Avg. Citations	Avg.Pub Year
1	University of Sheffield	UK	15	33	291	19.4	2016.13
2	University	USA	14	52	625	44.64	2011.64

	Pittsburgh						
3	University of oxford	UK	13	37	457	35.15	2016.38
4	UCL	UK	11	16	136	12.36	2010.36
5	Duke university	USA	10	6	215	21.5	2012.80
6	Stanford univ	USA	9	13	230	25.5	2014.33
7	UPV University Politecn Valencia	Spain	9	32	80	8.8	2014.44
8	University Valencia	Spain	9	35	41	4.5	2015.56
9	University Washington	USA	9	4	241	26.7	2006.89
10	Columbia University	USA	8	7	159	19.8	2014.38

Figure 5. shows the citation network of organizations. The citation pattern of organizations is examined with VV. The circle size shows the volume of organizations with publication. The curved connections represent cooperation between organizations. The color scales show the citation clusters. The minimum number of citations of an organization was 5. Of the 944 organizations 33 met the threshold. The University of Pittsburgh (TC= 625; TLS=52) is the most cited organization followed by University of Oxford (TC=457; TLS= 37), University of Sheffield (TC=291; TLS=33) University of Washington (TC=241; TLS=4).

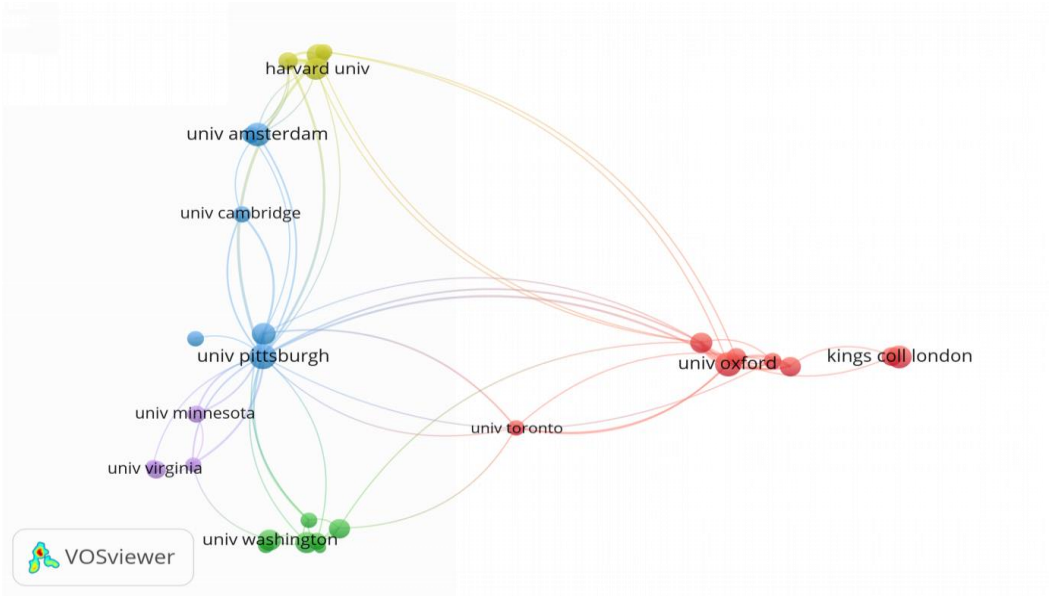

Fig 5. Citation network of organization by VV (network visualization)

Mapping of All keywords Occurrences:

A total of 1906 keywords were retrieved from 791 documents, minimum number of occurrences of keyword is 5 out of which, 64 met the threshold. The network visualization map shows the co-occurrence relations of all the keywords (**Figure 6**). The size of the circle indicates the occurrence of keywords. The curved connections represent cooperation between circles (keywords). The color scales show the collaboration clusters. The high frequency keywords are concentrated in 6 clusters.

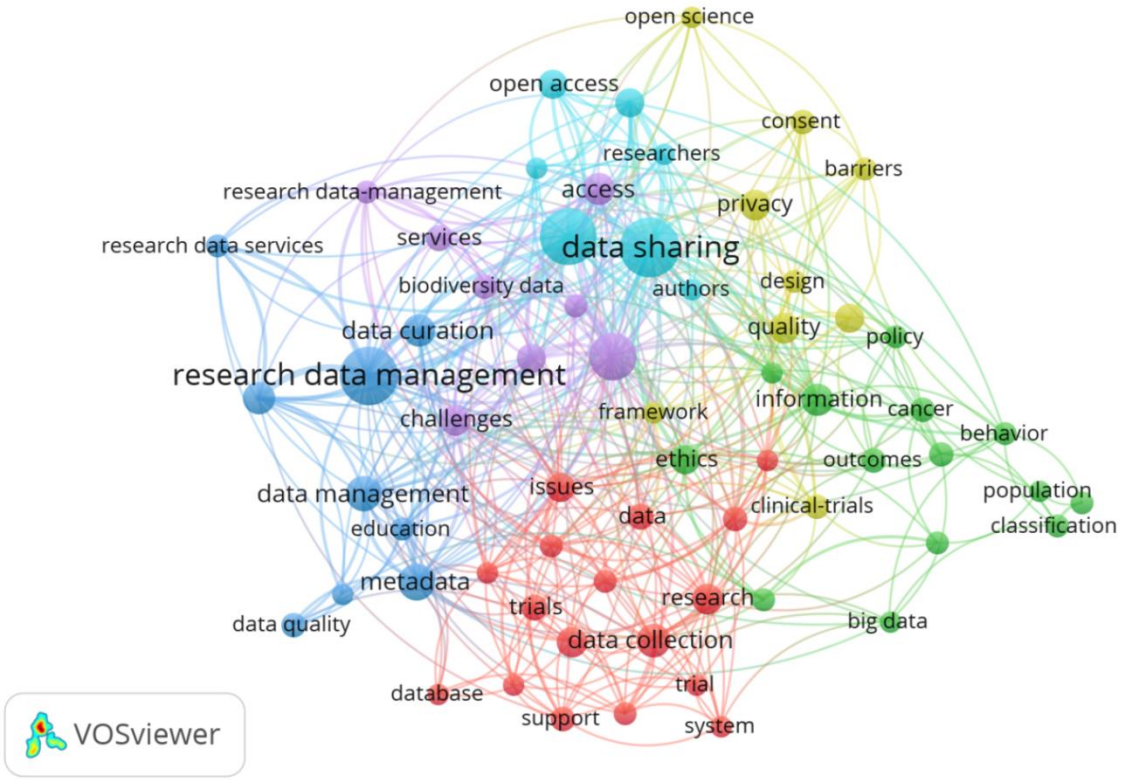

Fig.6 Co-occurrence analysis of all the keywords by VV (network Visualization)

Cluster 1. (Red) cluster represents the Data Management Plan (DMP) DMP i.e., the initial stage of RDM during research. It is the activity of planning regarding how data will be managed by responsible personnel. The activity includes agreement, schedule, preservation, ethics etc. The other most frequent keyword in this cluster is care, data, database, experiences, health, internet, issues, network, research, research ethics, research methods, social media, support, system, trial and trials.

Cluster 2. (Green) cluster represent the Data Management (DM) approach that shows the behavior of DM activity. DM is an activity of different kind of stages of research data lifecycle, that starts form the DMP followed by data collection, preservation, modification and sharing. The most frequently used keyword in the cluster is information followed by behavior, big data, cancer, classification, ethics, impact, metanalysis, outcomes, performance, policy, population, prevalence, and risk.

Cluster 3. **(Blue) shows** Data Literacy and Lifecycle approach. It is the ability to understand, synthesized and doable activity to manage research data during different stages of research data lifecycle. The most used keyword in the cluster is research data management followed by other keywords like academic libraries, data curation, data management, data quality, education, metadata, research data services and training.

Cluster 4. (Yellow) represents the privacy and quality of data. It is a critical stage of RDM to collect qualitative data with a high level of authenticity and to protect the legal privacy of data if it appears to be harmful. Whereas most frequently used keyword in the cluster is privacy and quality followed by other keywords such as: barriers, clinical-trails, consent, design, framework, informatics and open science.

Cluster 5. (Purple) represents the data science and literacy. It is scientific field that ensures and derived knowledge from both structured and unstructured data. The most frequently used keyword in the cluster is science followed by followed by other keywords such as: access, biodiversity data, challenges, Sources, knowledge, research data management, science and services.

Cluster 6. **(Turquoise)** represents the about the data preservation options and sharing. It is the act of exchanging data between owners and others under the terms of a legal agreement that guarantees privacy and authenticity. The most occurrence keyword in the cluster is data sharing. The other keywords included are authors, institutional repositories, open access, open data, research data, researchers.

The most frequently occurred keywords each with frequency, average publication year and total link strength are Data sharing (38;2016.92;91), Research data management (36;2017.08;60), Research data (35;2015.74), Science (25;2016.16;62) and Metadata (15;2017.53;31). The average published year of these keywords, including data management (14), data collection (13), access (12) and data curation (12) and academic libraries (11) is between 2014-2018. However, in the recent years, an increasing

number of researchers have given attention to research data management, social media and open data, training and social media topics indicating that these topics will be the focus of future research.

Conclusion

In this study, we have presented the visualization and science mapping analysis of RDM literature. The data were collected from the WOS database and VV software was used to analyze the data. The study includes the science mapping analysis of the publication trends, most productive authors, institutions, journals, countries. This study analyzes and visualize the co-authorship network analysis of authors, bibliographic coupling analysis of countries, co-citation analysis of sources, citation analysis of organizations and the Keyword analysis of all the keywords, network visualization or maps. The result indicates an increase in publication trends, especially from 2018. In terms of most productive author Rafael Alexandrine-benavent is the most prolific author and USA is the most productive country published on RDM. The most influential journal is the Zeitschrift fur Bibliothekswesen und Bibliographie .The University of Sheffield is the leading institution in the research area of RDM. The co-authorship analysis of authors shows that Susan bull and Michael parker who are collaborated more closely. The bibliographic coupling analysis of countries shows USA has more connection with Australia, Canada and England. PLOS One is the most co-cited source followed by Nature and Science. Keyword analysis indicated that keywords that are more frequently co-occurred in the area of RDM are data sharing, research data management, research data, science and metadata.

From the above results, we have obtained useful information on the literature of RDM through science mapping analysis. The study takes a broad look of RDM as a whole, addressing all of the topics that make up this research issue. RDM is a field that experienced significant growth after 1997, peaking in 1999. It has uses in a variety of fields, including medical sciences, banking, and defence etc. RDM makes research outputs more visible, supports fresh ideas, discoveries, speedier collaborations, and maintains the value of data, therefore the finding could be useful for researchers.

References

1. Baier-Fuentes H, Cascón-Katchadourian J, Sánchez ÁM, Herrera-Viedma E, Merigó J. A Bibliometric Overview of the International Journal of Interactive Multimedia and Artificial Intelligence. International Journal of Interactive Multimedia and Artificial Intelligence. 2018;5(3):9.

2. Cabrera LC, Talamini E, Dewes H. What about scientific collaboration in agriculture? A bibliometric study of publications about wheat and potato (1996-2016). Pesquisa Brasileira em Ciência da Informação e Biblioteconomia. 2018 Jun 28;13(1).

3. Corrall S, Kennan MA, Afzal W. Bibliometrics and Research Data Management Services: Emerging Trends in Library Support for Research. Library Trends. 2013;61(3):636–74.

4. Data Lifecycle | NNLM [Internet]. nnlm.gov. 2015 [cited 2021 Apr 20]. Available from: https://nnlm.gov/data/thesaurus/data-lifecycle#:~:text=The%20data%20lifecycle%20represents%20all

5. Elango B. A Bibliometric Analysis of Franchising Research (1988–2017). The Journal of Entrepreneurship. 2019 Jul 24;28(2):223–49.

6. Gajdács M. A bibliometric analysis of Acta Pharmaceutica Hungarica (1965-2018). Acta Pharmaceutica Hungarica. 2019 Apr 9;89(1):23–9.

7. Gil-Doménech D, Berbegal-Mirabent J, Merigó JM. STEM Education: A Bibliometric Overview. Modelling and Simulation in Management Sciences. 2019 Mar 19;193–205.

8. Heersmink R, van den Hoven J, van Eck NJ, van den Berg J. Bibliometric mapping of computer and information ethics. Ethics and Information Technology. 2011 Apr 19;13(3):241–9.

9. Koltay T. Data governance, data literacy and the management of data quality. IFLA Journal. 2016 Nov 30;42(4):303–12.

10. Onyancha OB. Open Research Data in Sub-Saharan Africa: A Bibliometric Study Using the Data Citation Index. Publishing Research Quarterly. 2016 May 31;32(3):227–46.

11. do Prado JW, de Castro Alcântara V, de Melo Carvalho F, Vieira KC, Machado LKC, Tonelli DF. Multivariate analysis of credit risk and bankruptcy research data: a bibliometric study involving different knowledge fields (1968–2014). Scientometrics. 2016 Jan 11;106(3):1007–29.

12. Stobierski T. 8 Steps in the Data Life Cycle | HBS Online [Internet]. Business Insights - Blog. 2021 [cited 2021 Apr 20]. Available from: https://online.hbs.edu/blog/post/data-life-cycle

13. Surkis A, Read K. Research data management. Journal of the Medical Library Association : JMLA [Internet]. 2015 Jul 1;103(3):154–6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4511058/

14. What are Scientometrics and Bibliometrics? [Internet]. Provalis Research. [cited 2021 Apr 20]. Available from: https://provalisresearch.com/solutions-2/applications/scientometrics-bibliometrics-software/

15. University of Leeds Library. Research data management explained [Internet]. Leeds.ac.uk. 2020. Availablefromhttps://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained

16. Zhang L, Eichmann-Kalwara N. Mapping the Scholarly Literature Found in Scopus on “Research Data Management”: A Bibliometric and Data Visualization Approach. Journal of Librarianship and Scholarly Communication. 2019 Apr 25;7(1).

17. Zhong B, Wu H, Li H, Sepasgozar S, Luo H, He L. A scientometric analysis and critical review of construction related ontology research. Automation in Construction [Internet]. 2019 May [cited 2019 Dec 12];101:17–31. Available from https://www.sciencedirect.com/science/article/pii/S0926580518305648