Journal of Scientometric Research
https://www.jscires.org

Exploration of Data Literacy Research using a Network of Cluster Mapping Approach

Naseema Sheriff, R Sevukan*

Department of Library and Information Science, Pondicherry University, Kalapet, Puducherry, INDIA.

ABSTRACT

Data literacy is essential for academics, researchers, and emerging data management professionals. The quality of scientific productivity is determined by the underlying knowledge of data collection or generation during the research process, which impacts data literacy. This research aims to investigate the current state of data literacy by examining bibliographic data with no constraints on themes or time periods. The generation of digital data has increased, and storage capacity and accessing devices have improved in the upgraded version, but where is the proper education for dealing with data? The research team must focus on data literacy to address all of these challenges. This paper will provide an overview and groundwork for data literacy based on previously published literature. This study used scholarly literature from Elsevier's Scopus database to conduct a knowledge network and mapping analysis. The global contributors and significant countries are mapped with institutions and authors. Primary areas are identified using a keyword co-occurrence network. Influential research papers and journals were identified using Document Co-Citation Analysis (DCA) and Journal Co-Citation Analysis (JCA). This paper highlights the insightful features of data literacy by conducting scientometric analysis with CiteSpace and VOSviewer. It reveals that researchers lack basic data management skills and end with complicated and ambiguous research findings. Researchers must emphasize the importance and fundamentals of data literacy skills and some metrics that may be used to assess the acquired skills.

Keywords: Research Data Management, Research Data, Data Quality, Data Visualization, Data Science, Artificial Intelligence.

Correspondence:
R Sevukan
Department of Library and Information Science, Pondicherry University, Kalapet-605014, Puducherry, INDIA.
Email id: naseemaphd3@gmail.com
ORC ID: 0000-0003-1952-5934

Received: 23-05-2022;
Revised: 20-06-2022;
Accepted: 14-07-2022.

INTRODUCTION

Data literacy is becoming more popular and essential in all forms of the education system; every researcher walks through their life about data literacy. Understanding, creating, analyzing, and communicating data as the information is referred to as data literacy.[1] Data literacy encompasses the capacity to acquire, handle, analyze, use, and critically comprehend data.[2,3] Data literacy is essential for 21st-century learners to succeed in a data-rich society.[4] and we are in a data-driven era.[5] The research method is data-driven, and the diversity of data has significantly increased. Data is everywhere, and the amount of data is rising all the time.[6] Data has become intriguing, influential, and vital in research across many disciplines, including the arts, humanities, social sciences, and sciences, whether expressed statistically or qualitatively. It's a rapidly evolving area that covers various topics such as mathematics, statistics, big data, and machine learning. It's not just true only in computer science but also in astronomy, labor market analysis, marketing, medicine, physics, and a wide range of other fields.[7] The role of educators and researchers in data literacy is crucial. In a nutshell, they should have the expertise and literacy to handle data efficiently and scientifically. In everyday life, data literacy makes a difference.[8] Data literacy is becoming more common in formal, informal, and non-formal education systems.[9]

Library and information science is a profession that is undergoing considerable technological advancements. Librarians should teach young researchers data literacy to collect, evaluate, analyze, and visualize data in a more meaningful and scholarly manner. The term "data visualization" implies a set of data analyses. Data visualization is one area where scholars might benefit from refining and empowering themselves scientifically. Data literacy is the most essential contributing and influencing component in research, and it is neither a science nor a mathematical ability. It is a life skill that everyone can learn, and it is a must-have skill in research.

Data literacy is a buzzword in research, although the idea and its significance are rarely articulated. Researchers with a strong

DOI: 10.5530/jscires.12.1.002

Publishing Partner : EManuscript Tech. [www.emanuscript.in]

understanding of data literacy can better grasp the tools and technologies used in methodologies. It impacts all stages of the research process. Many qualitative and quantitative data-related issues confront research scientists. Unfortunately, some researchers lack the confidence and understanding required to properly connect with and comprehend study results. Many researchers cannot critically analyze the different aspects of data quality, such as provenance and how it is structured, presented, and used to make scholarly inferences, despite data being a significant component of research. As a result, data literacy has grown in importance as a needed research skill today and in the future.[10] Individuals who lack data literacy are exposed to social, personal, physical, and financial risks and limited in their ability to be influential citizens in an increasingly data-driven society.[11] The amount of data rises exponentially and becomes the currency of power. Data literacy skills are becoming increasingly important in today's society.[12]

LITERATURE REVIEW

The techniques of bibliometrics, information visualization, and text mining found that the traditional literature review approach could not provide sufficient understanding.[13] Data literacy is a term that encompasses several different research techniques.[14] Data literacy is necessary to assess various data sources in today's digital life.[15] Data literacy is an essential skill in the age of big data.[16] Making data-driven decisions requires data literacy.[17] As a result, data literacy is thought to play a critical role in many views of the world, society, and the future.[18] Data citizenship necessitates a high level of data literacy.[19] Literacy is defined as interpersonal or intrapersonal meaning-making and transcultural.[20] Everyone has their own unique set of literacy skills to improve;[21] it impacts student achievement.[16] For the data revolution, a UN study advocates global data literacy.[18]

Data literacy is essential for anyone wanting to become a data-literate scientific professional, especially researchers and data managers.[22] Data literacy is becoming an increasingly crucial part of "doing economics."[23] In recent years, the focus of data-driven planning has shifted from a single data source (standardized test data) and a single outcome measure (academic achievement) to using a variety of data sources (e.g., pupil voice data, classroom observations, and parent survey data) and a variety of outcome variables (e.g., student learning, achievements, and wellbeing).[24] Data literacy is crucial for teachers and leaders as educational institutions try to understand individual learners better and improve learning outcomes.[25] An increasing amount of evidence supports data literacy's significance in enhancing teacher quality and student success.[26] Using new instruments of educational decision-making, including academicians, practitioners, and policymakers, requires new levels of data literacy.[27] Since data literacy is becoming essential for designers, contemporary design education is attempting to include it.[28]

Data literacy is widespread in various disciplines, including open data efforts, statistics, computer society, coding initiatives, etc.[29] In today's rapidly digitizing world, data and its uses are becoming increasingly pervasive, and students from all disciplines are being pressured to learn new abilities.[30] Schools typically teach data handling skills using limited, personally gathered data sets obtained through scientific study, resulting in a gap between what is taught and what will be required as big data analytics becomes more prominent.[31] As policy and professional development are focal points, data literacy, teacher evaluation literacy, and data-driven/informed decision-making have developed.[32] Educators struggle with data utilization.[33] An academic concern of the twenty-first century is today's society's constantly evolving data literacy landscape. Introducing data literacy knowledge and skills to teachers can assist them in making the essential improvements in student learning and achieving success in their chosen profession.[34] Data literacy abilities are crucial for personnel in all firm divisions, not only data scientists and analysts but also specialist business intelligence jobs. To support data-driven decisions, organizations aim to be data-centric, collecting and preserving data on their operations, customers, competitors, and the market. A collection of data literacy skills that librarians may help integrate into business school curricula, as they have in other professions.[35]

Because of the above, it is inferred that statistical literacy, assessment literacy, educational understanding, and data-driven decision-making are all examples of data literacy. Although data literacy research is minimal, it does provide insight into recommended methods for professional educators prepared to operate in a data-driven workplace.[25] As librarians and others focus on statistical education and data visualization training, data literacy needs to gain more attention in the scientific literature.[36] This study aims to fill the research gap to determine the global contributions in the field of data literacy. In scientometrics, co-citation studies are one of the most commonly used approaches by the scientific community, particularly in quantitative research, to visualize the structure and themes of a whole research domain.

RESEARCH QUESTIONS

This study analyses and describes the current scenario of scholarly communication in data literacy by examining bibliographic records. The following are the research questions:

Which countries do have substantial data literacy research contributions and involvement?

What are the primary areas of data literacy that have been explored and reported in the literature?

Which papers have received much attention or been pivotal in data literacy?

What are the most popular periodicals that publish articles on data literacy?

Who are the key contributors or emerging authors in the field of data literacy?

MATERIALS AND METHODS

Scopus, owned by Elsevier, is an abstract and citation database of peer-reviewed journals, conference proceedings, and books and was utilized to gather data for this study. Scopus was chosen because of its interdisciplinary nature, which is essential because we're looking at data literacy in various sectors. To collect publications on data literacy in this investigation, we used the phrase "data literacy." The search string used was "TITLE-ABS-KEY("data literacy" OR "data Literac*") AND (EXCLUDE (DOCTYPE, "no") OR EXCLUDE (DOCTYPE," ed") OR EXCLUDE (DOCTYPE, "Undefined")) AND (LIMIT-TO (LANGUAGE, "English"))" excluding editorial and indefinite articles and limit to English language only and retrieved results with 516 records on 16th October 2021 for 65 years spanning between 1956 and 2021. CiteSpace and Bibliometric R packages were used to analyze the bibliographic records of data literacy extracted from the Scopus database.

RESULTS

Why Citespace for Data Literacy?

CiteSpace's effectiveness is in identifying the major research areas in a given domain of knowledge, how the micro-level or subgroups are interconnected under a broad field of expertise, and in determining the strength of the interconnected environments to comprehend the insightful similarity between them. This research is being carried out to represent data literacy visually. CiteSpace computational intelligence constructs unique clusters based on the input dataset. A range of structural and temporal metrics are used to evaluate data literacy, such as co-citation networks analysis, silhouette, betweenness centrality, and modularity are structural metrics. In contrast, citation burstiness and betweenness centrality are temporal and hybrid metrics.[37] However, only the larger clusters are highlighted and identified in the graphics.

This tool's inbuilt features to examine the inferences as nodes are References, Cited Author, Cited Journal, Author, Institutions, Country, Term, Keyword, Source, Category, Article, Grant, and Claim. The size of the node reflects its history. All of the clusters are connected by nodes. CiteSpace uses purple trimming to highlight nodes with strong betweenness centrality. Betweenness centrality scores are a valuable statistic for discovering related clusters. Betweenness centrality values in CiteSpace are normalized to the unit interval [0, 1]; pathways between different thematic clusters are identified using this node.[38,39] A rush of citations indicates an active research landmark, and a citation burst is a name for detecting a burst event that can persist for years or just a year. The burst paper has piqued the scientific community's interest. If a cluster contains a vast number of nodes with high citation bursts, the cluster as a whole suggests an active research area or a recent phenomenon. CiteSpace uses Kleinberg's algorithm to detect bursts.[40,41] Kleinberg's burst detection algorithm detects periods when a target event occurs unusually frequently or "bursty."[42] In CiteSpace, a composite metric called sigma is used to evaluate the combined strength of a node's structural and temporal attributes, such as betweenness centrality and citation burst.[37,43] The highest values of this measure tend to be found in Nobel Prize and other award-winning studies.[37] This research seeks to generate a Data Literacy knowledge network and analyze the present scenario, intellectual basis, hotspots, and development patterns. This study provides an in-depth understanding of data literacy research and lays the groundwork for future studies.

Analysis of Contribution by Country

Figure 1 depicts the international competitiveness of countries in the field of data literacy. It assists in determining the correlations between research institutions and universities in these countries, and the country has made significant contributions to the study area. The result inferred modularity Q value is 0.7172, and the weighted mean silhouette score is 0.9122 for the study period of 1956 to 2021 (time slice 2). The circle in Figure 1 depicts the aggregate contributions of various countries, with the pivot point signifying those whose contributions are higher in this research field and are displayed in the larger font. The result shows that the top scientific contributors are the United States, followed by the United Kingdom, Australia, Germany, Belgium, China, Canada, and the Netherlands. Figure 2 illustrates the three-fold plots of the country's contribution and mapping, the names of the institutions, and the significant contributors identified.

Table 1 demonstrates the citation frequency and cited half-life of data literacy publications by country. It was observed that the United States ranked first in obtaining 188 citations, five times

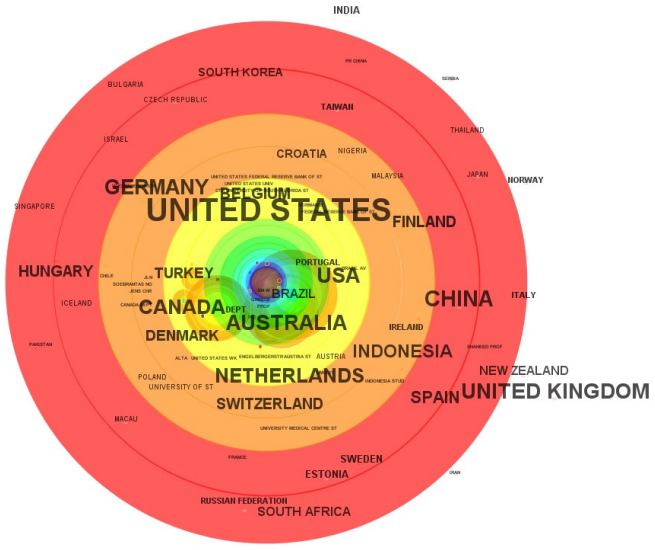

Figure 1: Global Contribution to Data Literacy Literature.

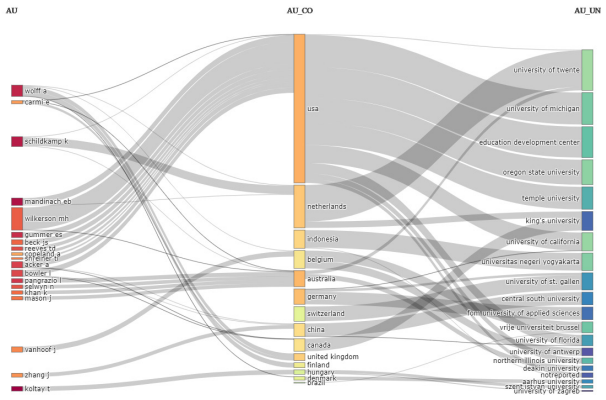

Figure 2: Country mapping with Authors and Institutions.

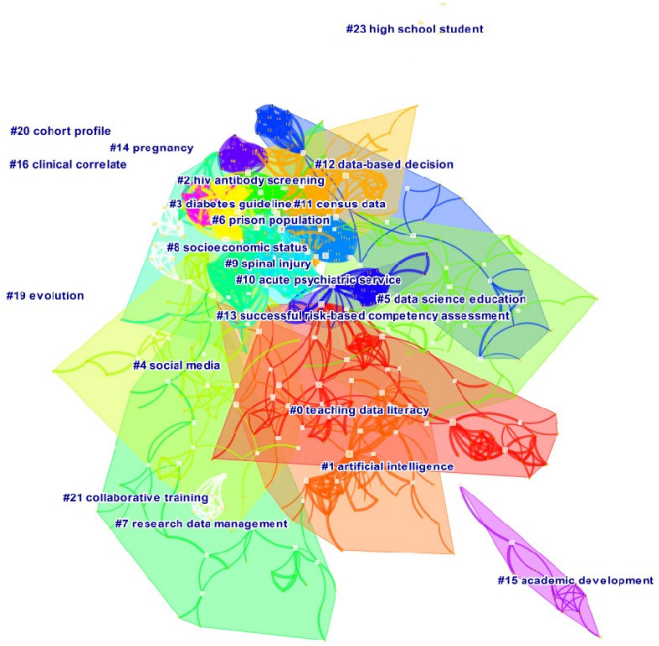

Figure 3: Data Literacy and its Co-Occurrence Clusters.

Table 1: Citation Frequency and Half-life of Data Literacy Literature.

Freq	Country	Half-Life	Cluster-ID
188	United States	32.5	1
38	China	8.5	52
33	Australia	23.5	0
31	United Kingdom	7.5	58
26	Germany	26.5	3
25	Netherlands	11.5	5
24	Canada	12.5	2
18	Indonesia	5.5	6
13	Spain	6.5	10
12	Belgium	8.5	4
11	Hungary	0.5	20
10	Switzerland	5.5	0
10	Finland	16.5	36
8	Denmark	0.5	2
8	Turkey	15.5	2
7	South Africa	-0.5	23
6	Brazil	16.5	0
6	New Zealand	2.5	33
5	South Korea	6.5	7
5	Croatia	12.5	9
4	India	4.5	22

or more than the other countries. The other ranked countries include China (38 Citations), Australia (33 Citations), the United Kingdom (31 Citations), and so on. As far as the half-life of data literacy publications is concerned, again USA tops with the half-life period of 32.5 years, witnessing an enduring relevance of publications contributed by scientists from the USA to data literacy literature. Although Germany was at the fifth rank in terms of citations, the half-life period of 26.5 years sounds better than the other countries such as China, Australia, and United Kingdom. The other notable countries with considerable half-life periods are Australia (23.5 years), Finland and Brazil (16.5 years each), Turkey (15.5 years), and Canada and Croatia (12.5 years each).

Regarding the generation of clusters based on the keyword frequencies, three countries, namely Australia, Switzerland, and Brazil, were in cluster #0, which is the best compared to the other countries. The other notable countries with significant clusters include the United States (Cluster #1), Canada, Denmark, and Turkey (Cluster #2), Germany (Cluster #3), Belgium (cluster #4), and the Netherlands (Cluster #5) and so on. The lower the cluster number, the more significant the density of keyword clusters.

Keyword Co-Occurrence Analysis

Keyword co-occurrence is useful for identifying new research frontiers and hotspots. The cluster established with the node "Keyword" can provide insight into the intellectual key research topics. The interconnections between the clusters are reflected in the modularity Q. Levels of modularity Q of 0.4 to 0.8 are generally considered acceptable, and a mean silhouette value near 1 suggests that references within a cluster have extremely consistent or similar content. The cluster created by the node "Keyword" can provide information about intellectually important study topics. The modularity Q reflects the links between the clusters. A Modularity Q score of 0.4 to 0.8 is considered acceptable, and a mean silhouette value close to 1 indicates that references within a cluster have extremely consistent or similar content. By specifying the following parameters in citespace, Figure 3 shows the significant cluster associated with data literacy. It generated 30 clusters with 614 nodes and 3350 links for the research period of 1956 to 2021 (time slice 1), with a Modularity *Q* value of 0.6819 and a Weighted Mean Silhouette Score of 0.8993. Pathfinder and pruning of the merged network were used to reduce the unweighted links. In total, thirty clusters were identified using 571 records in the dataset, including the eight largest clusters represented in Table 2. Figure 3 exhibits significance clusters, which are labelled as follows: Cluster #0, which was predominantly labelled as "teaching data literacy", Cluster #1 labelled as "artificial intelligence", Cluster #2 labelled as "HIV antibody screening", Cluster #3 labeled as "diabetes guideline", cluster #4 labeled as "social media", Cluster #5 labelled as "data science education", cluster #6 as "prison population", Cluster #7 as "research data management", Cluster #8 as "socioeconomic status". The Locally Linear Regression (LLR) algorithm was used to detect the cluster labels.

For the eight largest clusters, Table 2 shows the cluster size, silhouette score, Label (LLR), label (TFIDF), label (MI), and mean year. The largest cluster (#0) is the most populated, with 57 members and a silhouette value of 0.892. It is classified as "teaching data literacy" by both LLR and TFIDF, and as "text mining analysis" (0.9); "defining data literacy communities" (0.9); "analysis" (0.9); "collaborative training" (0.9); and "research data management survey" by MI (0.9). The second-largest cluster (#1) is the most populous, with 57 members and a silhouette value of 0.933.LLR categorizes it as artificial intelligence, TFIDF categorizes it as data literacy, and MI categorizes it as analysis (4.23), collaborative training (4.23), research data management survey (4.23), text mining analysis (4.23); data library service (4.23). The third-largest cluster (#2) is detected, with 51 members and a silhouette value of 0.831. The fourth-largest cluster (#3) is occupied, with 44 members and a silhouette value of 0.846. The fifth-largest cluster (#4) has 39 members and a silhouette value of 0.922. The sixth-largest cluster (#5) has 38 members and a silhouette value of 0.853. The seventh-largest cluster (#6) has 34 members and a silhouette value of 0.837. The eighth-largest cluster (#7) was identified with 34 members and a silhouette value of 0.946.

Mapping of Document Co-Citation Analysis (DCA)

The mapping in Figure 4 created 338 distinct nodes and 2759 links from 1956 to 2021 (a time slice of 5 years). Document co-citation analysis is primarily based on identifying pairs of frequently cited articles.[44] DCA studies a network of co-cited references.[45-48] Within the cluster network, the relationship between each node and each link has a major impact, emphasizing the coordination of meaningful related clusters. Clusters were given numbers and sorted according to their size, with cluster #0 being the largest. The cluster's size showed how influential the publications were. The greater the cluster means more citations it has. The purple tree rings highlight the burstiness of the author's publishing. DCA revealed 48 clusters in this analysis, with modularity $Q = 0.8986$, Weighted Mean Silhouette S = 0.9448, and Harmonic Mean (Q, S) = 0.9212. The silhouette value of the cluster determines the quality of a clustering configuration. It has a range of-1 to 1 as its value.

The highest value represents the best possible result. To achieve a sound interpretation is recommended to balance the modularity and silhouette scores in CiteSpace simultaneously.[37] Figure 4 depicts the most prominent document citation analysis clusters. The network is organized into 11 co-citation clusters based on their size. The index terms from their citers are used to label these clusters on the left. In terms of node type (references) and the number of member references, Data Literacy (Cluster #0) is the largest cluster. The second-largest cluster is the Data Librarian Role

Table 2: Data Literacy and its Co-Occurrence Clusters.

Size	Silhouette score	Label (llr)	Label (tfidf)	Label (mi)	Mean (cite year)	Cluster-id
57	0.892	Teaching data literacy (145.46, 1.0E-4)	Teaching data Literacy	Text mining analysis (0.9); defining data literacy communities (0.9); analysis (0.9); collaborative training (0.9); research data management survey (0.9)	2013	0
57	0.933	Artificial intelligence (102.68, 1.0E-4)	Data literacy	Analysis (4.23); collaborative training (4.23); research data management survey (4.23); text mining analysis (4.23); data library service (4.23)	2015	1
51	0.831	Hiv antibody screening (70.68, 1.0E-4)	Nation-wide cross-sectional household survey	Sustainable development goal (0.09); characterizing data ecosystem (0.09); open mapping data (0.09); analysis (0.09); collaborative training (0.09)	2005	2
44	0.846	Diabetes guideline (44.47, 1.0E-4)	Diabetes guideline	Data literacy (0.03); venous insufficiency (0.01); western turkey (0.01); veyt-i study (0.01); analysis (0.01)	2004	3
39	0.922	Social media (180.38, 1.0E-4)	Artificial intelligence	Technological sovereignty (0.44); data care (0.44); analysis (0.44); collaborative training (0.44); research data management survey (0.44)	2015	4
38	0.853	Data science education (104.22, 1.0E-4)	Learning analytics	Human trust factor (0.35); analysis (0.35); collaborative training (0.35); research data management survey (0.35); text mining analysis (0.35)	2018	5
34	0.837	Prison population (19.62, 1.0E-4)	Hospital quality	Analysis (0.08); collaborative training (0.08); research data management survey (0.08); text mining analysis (0.08); data library service (0.08)	1996	6
34	0.946	Research data management (264.26, 1.0E-4)	Data literacy	Research data management survey (0.31); analysis (0.31); collaborative training (0.31); text mining analysis (0.31); data library service (0.31)	2018	7

(Cluster #1), the third-largest cluster is Data Quality (Cluster #2), the fourth-largest cluster is Curriculum Design (Cluster #3), and so on. #5 Personal data management, #6 Standardized educational achievement data, #7 Critical race theory perspective, #8 teacher inquiry, #9 School performance feedback use, #10 Information literacy, and # 11 Data Visualization are the significant areas of research and publications in data literacy from 1956 to 2021.

Table 3 represents which source of references had the most citation bursts and the time periods during which the bursts occurred. The burst strength of 4.58 (Carlson J) is the top-ranked item by bursts. The second is the Mandinach EB, which has 3.88 bursts. The third is (Giarlo M), who has 3.76 bursts. (Schneider R) is the fourth, with bursts of 3.59. (Koltay T) is the fifth, with bursts of 3.50. (Calzada Prado J) is the sixth, with bursts of 3.48. (Shulman LS) is the seventh, with bursts of 3.12. (Mooney H) is the eighth, with bursts of 2.98. (Federer L) is the 9th, with bursts of 2.92. (Wolff A) is the tenth, with bursts of 2.92 and so on. (Carlson J) is the top-ranked item by citation frequency counts, with a total of 26 citations. (Mandinach EB) has a citation count of 19, and others are highlighted by citation frequency—the centrality to finding various thematic clusters. Stated sigma values indicate the novelty of the publications in data literacy.[49]

Mapping of Journal Co-Citations Analysis (JCA)

A grouping of Journal Co-Citation Analysis (JCA) associated with data literacy is formed when the node type in CiteSpace is set to "Cited Journal." Figure 5 shows the leading publications that contributed the most to this study based on the dataset used. CiteSpaces' computational intelligence built 29 clusters, 419 nodes, and 1069 linkages from 1956 to 2021 (time slice = 5), with a modularity *Q* value of 0.8679 and a weighted mean silhouette score of 0.9398, and a pathfinder was chosen for the pruning to minimize the unweighted links. This cluster network depicts the most influential journals and the clusters they belong to grasp the in-depth concepts and specificities of data literacy.

Based on co-citations, the network is divided into ten clusters. These clusters are named using the index phrases from their citers. Four of the ten clusters are the largest. Table 4 represents Data visualization (Cluster #0) has 53 references with a mean silhouette value of 0.911, standardized educational achievement data (Cluster #1) has 42 references with a mean silhouette value of 0.887, and future practice (Cluster #2) has 41 references with a mean silhouette value of 0.956, and research libraries (Cluster #3) has 39 references with a mean silhouette value of 0.948. The silhouette score shows cluster homogeneity, and the cluster labels in Figure 5 are determined using the Locally Linear Regression (LLR) technique.

Top Cited Journals Based on Citation Bursts

Table 5 displays the top-ranked cited journal publications based on the strength of the burstness to determine the interference pattern of journals. The top three most robust burst journals are "Determining Data Information Literacy Needs" (4.55), "Journal of Librarianship and Scholarly Communication" (4.51), and "Information literacy competency standards for higher education" (4.03). The two oldest bursts, which began in 1991 and finished in 2010, are "BMI" (2.6) and "Arch Intern Med" (2.58). The most recent busted journals, with publication dates spanning from 2016 to 2021, are "International Journal of Science Education" (3.01), "Beyond Data Literacy" (2.99), "Data-Informed Learning" (2.67), and "What Does It Mean for Teachers to be Data Literate" (2.66).

Performance Analysis of Authors

The node type "Author" was chosen for the investigation to find author productivity, which results in the performance of the authors' data literacy-related works. From 1956 to 2021 (time slice = 1), it calculated 215 clusters using 467 nodes and 600 links, with a Modularity Q score of 0.9731 and a weighted Mean Silhouette Score of 0.7516. Figure 6 illustrates the network of top contributors in the field of data literacy to highlight the leading authors.

Table 6 shows the performance of the top authors based on the strength of the data literacy burst. T Koltay (2015), who appeared in Information Literacy (Cluster #91) with a burstness strength of 6.31, L Bowler (2017), and A Acker (2017), who appeared in Contexts Concept (Cluster #38) with a burstness strength of 2.27. In-School Performance Feedback Use (Cluster #12), J Vanhoof (2011) with 2.25 burstiness, M Valcke (2011), G Verhaeghe (2011), Petegm Van (2011) with 1.69 burst strength appeared. EB Mandinach (2013) with 2.04 burst and ES Gummer (2013) with 1.96 bursts appeared in Data Literacy (Cluster #14). The most prominent cluster labels are detected as "Information Literacy," "Contexts Concept," "School Performance Feedback Use," "Data Literacy," "Developing Personal Data Tactics," and "Human Design."

DISCUSSION

This study has been done to uncover the overall structure by providing an inside look at data literacy. It presents academics with new perspectives on potential collaborators, partnering nations, hotspots, and potential research areas of data literacy research through citation and scientometric analysis. Data literacy needs to gain increasing attention in scholarly literature, primarily as librarians and others focus on statistical education and data visualization training.[36] The Country-wise performance revealed that the United States, the Netherlands, Indonesia, Belgium, Australia, Germany, Switzerland, China, Canada, the United Kingdom, Hungary, Denmark, and Brazil produced the most data literacy publications than any other country study. These countries had a major advantage and significant influence in the field of data literacy, as determined by the number of publications

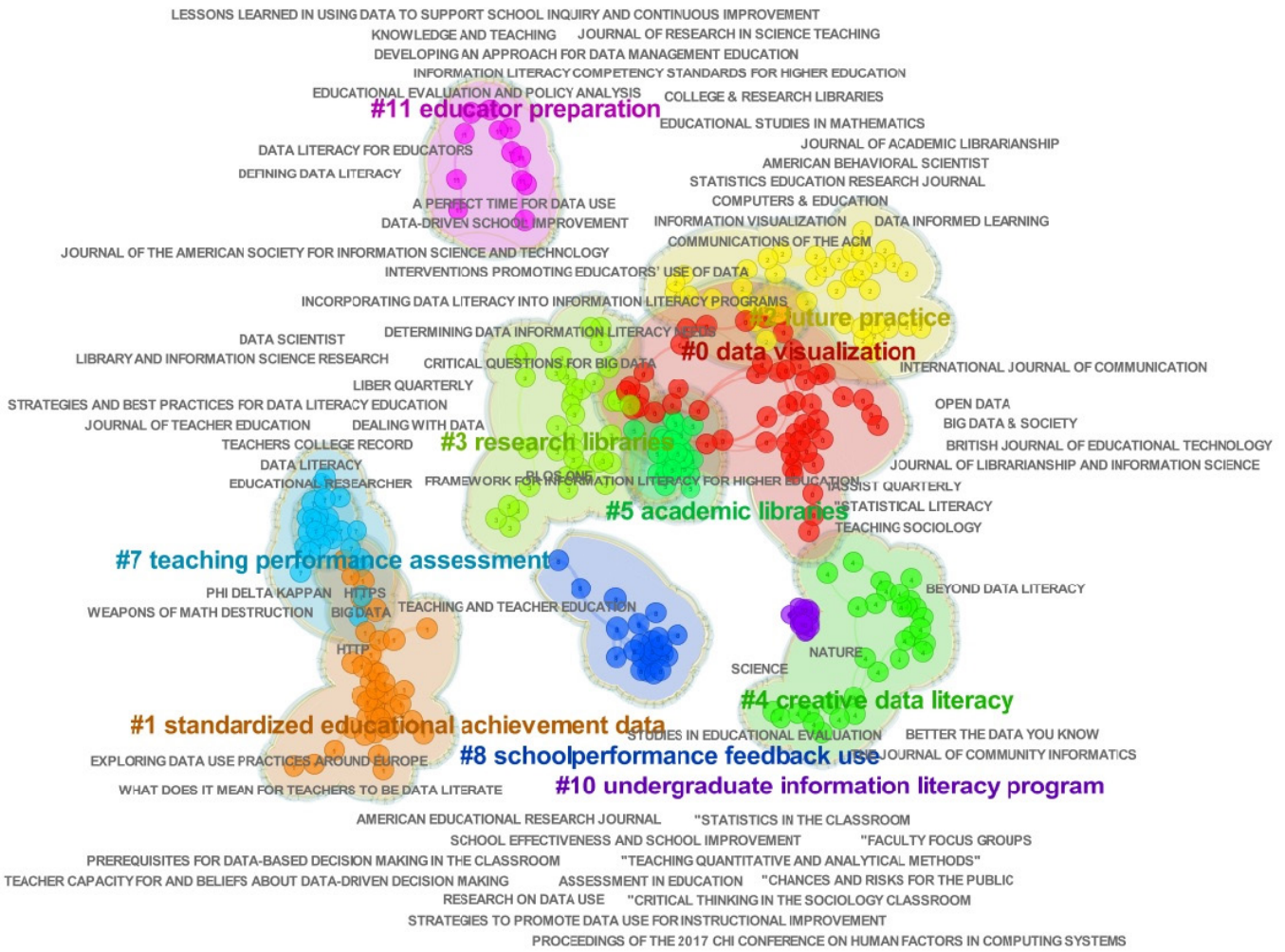

Figure 4: Mapping of Document Co-Citation Analysis (DCA).

and half-life that specifies their contributions to the literature. It facilitates the identification of major universities and research institutions from leading countries, such as the University of Twente, University of Michigan, Oregon State University, Temple University, King's University, University of California, and Northern Illinois University, whose focus and contribution is more on data literacy research, as well as leading contributors from those countries, are Wilkerson MH, Mandinach EB, Schildkamp K, Wolff A, Acker A, and Bowler. This enables the identification of potential data literacy innovators in terms of country, author, and institution.

Finding main keywords from appropriate literature assists researchers in advancing their work and planning for future initiatives. In this case, clustering the keywords by core theme is essential for future work. According to the findings of this study, data literacy, student, big data, human, education, article, female, visualization, adult, and data science are determined based on citation counts. According to burstiness, research data management, student, article, adult, information management, education, human, female, controlled study, and visualization are all predicted. According to centrality, the following terms are identified: adult, article, female, aged, demography, controlled study, human, diabetes mellitus, cross-sectional study, and male. According to sigma, education, adult, data literacy, big data, article, aged, teaching, female, Australia, and diabetes mellitus are observed. Citation counts, burstiness, centrality, and sigma are impactful factors for keywords when using CiteSpace. When using CiteSpace, the most vital metrics to determine the keywords are citation counts, burstiness, centrality, and sigma. Citation counts are bibliometric indicators; Burstness is sudden focus or jumps in the frequency of terms in the literature to find active terms. Centrality is an intermediate node (keyword) that helps connect with various clusters. It indicates the importance of the terms. Sigma score is a combination of Betweenness centrality and citation burst. This all helps to identify the fast-growing terms and topics. The leading clusters are determined and labeled by combining all the term frequencies.

The finding states teaching data literacy, artificial intelligence, HIV antibody screening, diabetes guideline, social media, data science education, prison population, research data management, socioeconomic status, spinal injury, acute psychiatric service, census data, data-based decision, successful risk-based

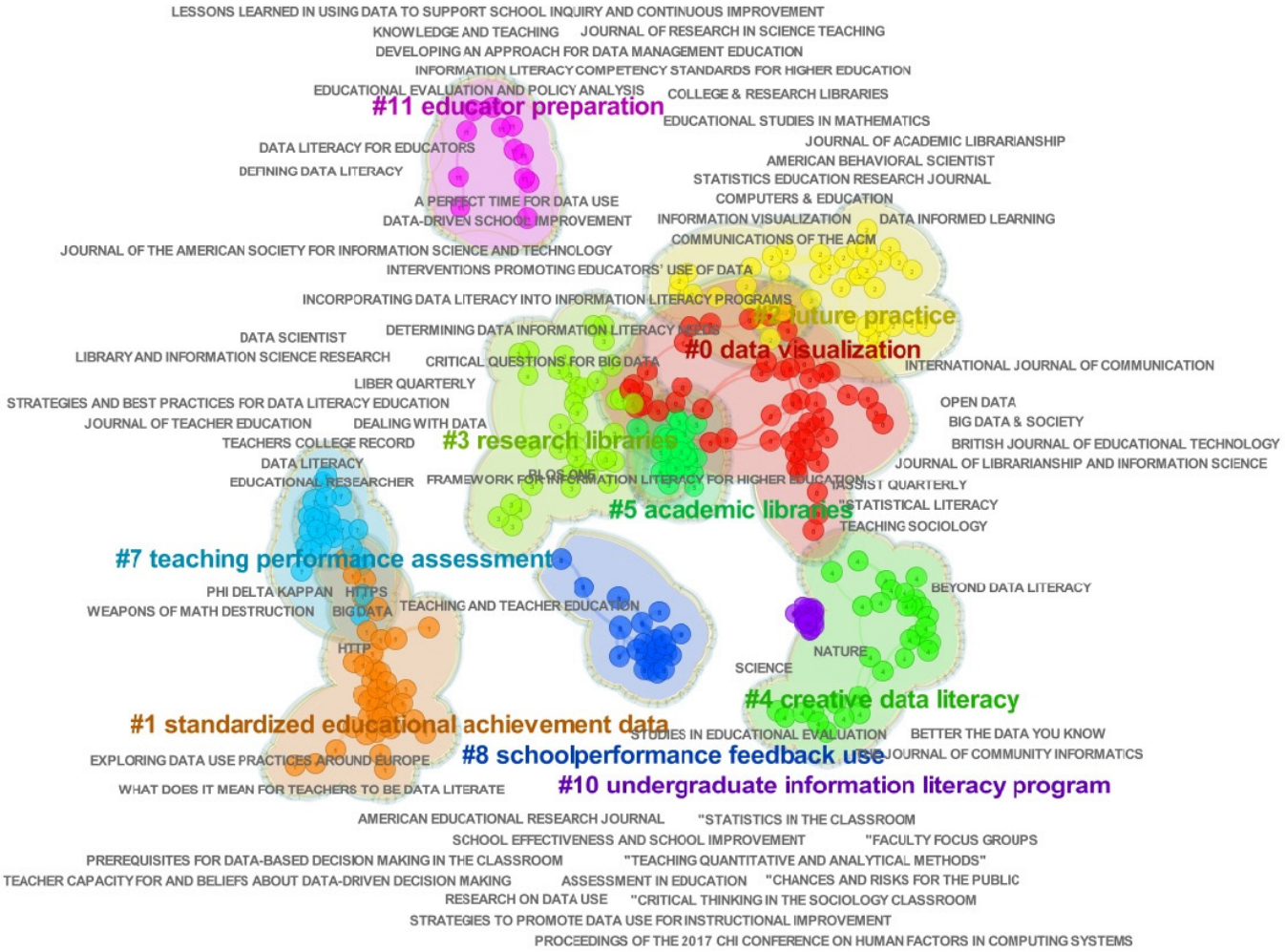

Figure 5: Mapping of Journal Co-Citations Analysis (JCA).

competency assessment, pregnancy, and academic development. As a result, the exciting transdisciplinary subjects engaged in data literacy research have been identified. A network of various clusters exhibits the intellectual structure of a knowledge domain. [37] Therefore, cluster analysis (i.e., determining subject categories) is one method for visualizing knowledge domains. CiteSpace is the most frequently used tool in their field of origin, library, and information.[50]

This study employed Journal Citation Analysis (JCA) to investigate associated scientific journals in the domain of data literacy, implying the prominence and effect of publications published in these core journals. Journal of Librarianship and Scholarly Communication, Library and Information Science Research, Journal of the American Society for Information Science and Technology, International Journal of Science Education, Framework for Information Literacy for Higher Education, Journal of Academic Librarianship, and Information Visualization were featured based on the burstness strength of the citation. Teaching and Teacher Education, LIBRI, ISPRS International Journal of Geo-Information, Journal of Documentation, Big Data and Society, New Media and Society, Cognition and Instruction, Internet and Higher Education, IFLA Journal, Statistical Education Research Journal, Educational Research, Journal of Business Ethics, Qualitative Inquiry, Library Hi Tech among the journals identified across various subjects based are the number of total citations.

The evolution and research dynamics of scientific studies may be inferred by analyzing prolific authors in data literacy. The researchers' scholarly patterns can be determined based on the number of publications and citation frequencies. T Koltay, L Bowler, A Acker, J Vanhoof, Eb Mandinach, Es Gummer, M Valcke, G Verhaeghe, Petegem Van, and L Pangrazio are the most prolific authors with the most robust bursts, according to the dataset. These authors are most active in scientific publications. Mandinach Eb, Gummer Es, Schildkamp K, Pangrazio L, Selwyn N, Koltay T, Vanhoof J, Acker A, and Bowler L are the prolific authors with total citations. These authors published in the study field on a consistent schedule.

Furthermore, if a cluster has many nodes with high citation bursts, the cluster represents an active study area or a new trend. It will display those authors who have rapidly raised their number

Table 3: Mapping of Document Co-Citation Analysis (DCA).

Citation frequency	Burstness strength	Centrality	Sigma	Author	Year	Source	Cluster no
26	4.58	0.01	1.03	Carlson J	2011	Determining data information literacy needs	2
19	3.88	0.01	1.03	Mandinach EB	2016	What does it mean for teachers to be data literate	0
9	3.76	0.07	1.27	Giarlo M	2013	Academic libraries as quality hubs @ Journal of Librarianship and Scholarly Communication	2
7	3.59	0	1	Schneider R	2013	Research data literacy @ ECIL 2013 CCIS	2
25	3.5	0.09	1.36	Koltay T	2015	Data literacy	1
21	3.48	0.11	1.42	Calzada Prado J	2013	Incorporating data literacy into information literacy programs	2
8	3.12	0.12	1.44	Shulman LS	1986	Those who understand	0
7	2.98	0.13	1.44	Mooney H	2012	The anatomy of a data citation	2
7	2.92	0.19	1.67	Federer L	2013	The librarian as research informationist	10
7	2.92	0	1	Wolff A	2016	Creating an understanding of data literacy for a data-driven society @ The Journal of Community Informatics	1
5	2.77	0.08	1.23	Jacobs J	2009	Data literacy	8
13	2.72	0.2	1.65	Mandinach EB	2012	A perfect time for data use	0
5	2.69	0	1	Huffman D	2003	Collaborative inquiry to make data-based decisions in schools @ Teaching and Teacher Education	9
12	2.67	0.04	1.1	Ridsdale C	2015	Strategies and Best Practices for Data Literacy Education	1
4	2.55	0.07	1.19	Cochran-Smith M	2009	Inquiry as stance	8
6	2.5	0.01	1.02	Bawden D	2001	Information and digital literacies	2
6	2.5	0.12	1.32	Buckland M	2011	Data management as bibliography @ Bulletin of the American Society for Information Science and Technology	5
6	2.5	0	1	Bawden D	2009	The dark side of information	2
6	2.3	0	1	Carlson J	2015	Planting the seeds for data literacy	10
7	2.2	0	1	Si L	2013	The cultivation of scientific data specialists	1
4	2.15	0	1	Webber S	2000	Conceptions of information literacy	9
12	2.13	0.07	1.14	Kerr KA	2006	Strategies to promote data use for instructional improvement	9

Table 4: Encapsulated Largest Four Clusters.

Size	Cluster-id	Silhouette score	Cluster label (Llr)	Mean (Citee year)
53	0	0.911	data visualization	2017
42	1	0.887	standardized educational achievement data	2017
41	2	0.956	future practice	2019
39	3	0.948	research libraries	2015

Table 5: Top Cited Journals based on Citation Bursts.

Cited Journals	Strength	Begin	End
Determining Data Information Literacy Needs	4.55	2011	2020
Journal of Librarianship and Scholarly Communication	4.51	2011	2020
Information Literacy Competency Standards for Higher Education	4.03	2011	2020
Dealing with Data	3.5	2011	2020
Intersections of Scholarly Communication and Information Literacy: Creating Strategic Collaborations for a Changing Academic Environment	3.36	2011	2020
Ecil 2013 Ccis	3.32	2016	2020
A Perfect Time for Data Use	3.17	2011	2020
Library and Information Science Research	3.16	2016	2020
Journal of the American Society for Information Science and Technology	3.08	2011	2020
The Cultivation of Scientific Data Specialists	3.05	2016	2020
International Journal of Science Education	3.01	2016	2021
Beyond Data Literacy	2.99	2016	2021
Framework for Information Literacy for Higher Education	2.97	2016	2020
Journal of Academic Librarianship	2.88	2011	2020
Lessons Learned in Using Data to Support School Inquiry and Continuous Improvement	2.86	2011	2015
Conceptions of Information Literacy	2.86	2011	2015
Strategies to Promote Data Use for Instructional Improvement	2.83	2011	2020
Information Visualization	2.69	2016	2020
Jama	2.68	2001	2010
Data Informed Learning	2.67	2016	2021
What Does it Mean for Teachers to be Data Literate	2.66	2016	2021
Bmj	2.6	1991	2010
Arch Intern Med	2.58	1991	2010
Inquiry as Stance	2.57	2011	2015
Evaluation	2.57	2011	2015

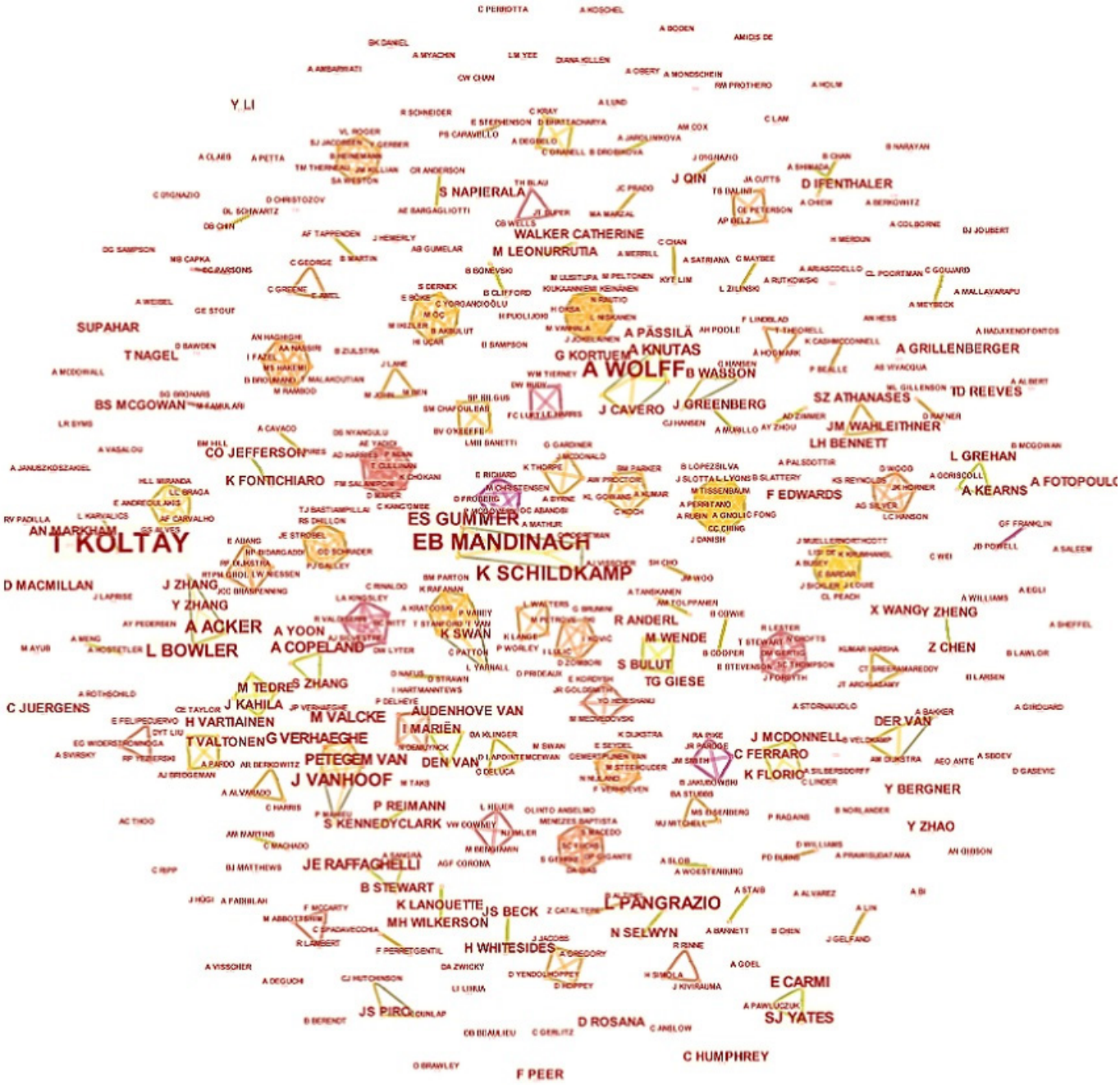

Figure 6: Network of Authors Performance.

of publications for the node sort of good author. The outcomes of this study demonstrate that author clusters are not as well networked as other clusters, suggesting that authors do not devote as much attention to data literacy as other clusters, even though it is an essential topic for research.

CONCLUSION

Based on each node type, the feature considered for analysis is different within the given dataset, so the cluster generated inside each node and its label detection is different and meaningful according to the selection of algorithms like Term Frequency by Inversed Document Frequency (TFIDF), Log-Likelihood Ratio (LLR), and Mutual Information (MI). But this overall discussion is all about hot topics in the main field of study. Increasing data literacy necessitates a multidisciplinary approach across multiple fronts. The clusters represented the essential concepts in the knowledge domain of data literacy based on each finding. It emphasizes specialization in the fields of education, medicine, and technology. Data is the soul substance of the research process; Data is the property and footprint of the technology domain; data is generated in every way in medicine, but in education, imparting knowledge and understanding of data literacy is essential in all facets of society. Data, literacy, and education are all multifaceted and ambiguous concepts.[51]

Table 6: Network of Authors Performance.

Publication frequency	Burstness strength	Author	Year	Clustered	Cluster name Label (llr)
12	6.31	T Koltay	2015	91	Information Literacy
4	2.27	L Bowler	2017	38	Contexts Concept
4	2.27	A Acker	2017	38	Contexts Concept
4	2.25	J Vanhoof	2011	12	School Performance Feedback Use
7	2.04	EB Mandinach	2013	14	Data Literacy
4	1.96	ES Gummer	2013	14	Data Literacy
3	1.69	M Valcke	2011	12	School Performance Feedback Use
3	1.69	G Verhaeghe	2011	12	School Performance Feedback Use
3	1.69	Petegem Van	2011	12	School Performance Feedback Use
4	1.42	L Pangrazio	2018	93	Developing Personal Data Tactics
7	1.33	A Wolff	2015	21	Human Design
3	1.28	A Knutas	2019	21	Human Design

In an increasingly data-driven world, data literacy is becoming more widely recognized as a general talent that all individuals should have. Data literacy is a crucial and essential skill for any researcher. As a result, there's a lot of concern about how it may be taught in schools.[52] It is beneficial to educate stakeholders by conducting workshops, training, seminars, and conferences, as well as publishing a manual for handling data in research by university departments, trying to organize mini-courses, and including a data literacy syllabus for credit scores and quizzes related to this topic by institutions or governing bodies during research. Before educating young researchers, all domain experts should be trained in data literacy. Librarians and information professionals are ideal candidates to impart data literacy to people from many disciplines because they work with people from all walks of life. To give credit to institutions and nations, properly imparting data literacy knowledge to young researchers has a positive impact on quality publications and patent development and can reduce the retractions of publications. Even though we handle big data and analytics and have excelled with Artificial Intelligence, we still lack the fundamentals of data literacy. Providing special courses to young researchers to strengthen their data literacy skills would be beneficial. Finding new strategies will pave the way to teaching the stakeholders about data literacy, which may enable them to become data literate. Such new methods will benefit the researchers, scientists, faculties, and others in data literacy.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Cottone AM, Yoon SA, Coulter B, Shim J, Carman S. Building system capacity with a modeling-based inquiry program for elementary students: A case study. Systems. 2021;9(1):1-21. https://doi.org/10.3390/systems9010009
2. Hachmeister N, Weiß K, Theiß J, Decker R. Balancing plurality and educational essence: Higher education between data-competent professionals and data self-empowered citizens. Data. 2021;6(2):1-15. https://doi.org/10.3390/data6020010
3. Nguyen D. Mediatisation and datafication in the global COVID-19 pandemic: On the urgency of data literacy. Media International Australia. 2021;178(1):210-4. https://doi.org/10.1177/1329878X20947563
4. Rowe S, Riggio M, De Amicis R, Rowe SR. Teacher perceptions of training and pedagogical value of cross-reality and sensor data from smart buildings. Education Sciences. 2020;10(9):1-18. https://doi.org/10.3390/educsci10090234
5. Mariani P, Zenga M. Data Science and Social Research II: Methods, Technologies and Applications. 2021.
6. Gehrke M, Kistler T, Lübke K, Markgraf N, Krol B, Sauer S. Statistics education from a data-centric perspective. Teaching Statistics. 2021;43(S1):S201–15. https://doi.org/10.1111/test.12264
7. Khodabakhsh A, Ari I. Data Science: From Research to Application. 2020;45.
8. Smalheiser NR. Data literacy: How to make your experiments robust and reproducible. Data Literacy: How to Make Your Experiments Robust and Reproducible. 2017. https://doi.org/10.1016/C2016-0-01275-5
9. Rawat KS, Sood SK. Knowledge mapping of computer applications in education using CiteSpace. Computer Applications in Engineering Education. 2021;29(5):1324-39. https://doi.org/10.1002/cae.22388
10. Morrow J. Be Data Literate. KoganPage. Kogan Page Books. 2021.
11. Carmi E, Yates SJ, Lockley E, Pawluczuk A. Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. Internet Policy Review. 2020;9(2):1-22. https://doi.org/10.14763/2020.2.1481
12. Usova T, Laws R. Teaching a one-credit course on data literacy and data visualization. Journal of Information Literacy. 2021;15(1):84-95. https://doi.org/10.11645/15.1.2840
13. Rodrigues SP, Van Eck NJ, Waltman L, Jansen FW. Mapping patient safety: A large-scale literature review using bibliometric visualization techniques. BMJ Open. [cited 2022]. 2014;4(3):e004468. Available from: https://bmjopen.bmj.com/content/4/3/e004468
14. Sander, I. What is critical big data literacy, and how can it be implemented? Internet Policy Review. 2020;9(2):1-22. https://doi.org/10.14763/2020.2.1479
15. Juergens C. Digital data literacy in an economic world: Geospatial data literacy aspects. ISPRS International Journal of Geo-Information. 2020;9(6):373. https://doi.org/10.3390/ijgi9060373
16. Yang N, Li T. How stakeholders' data literacy contributes to student success in higher education: A goal-oriented analysis. International Journal of Educational Technology in Higher Education. 2020;17(1):1-8. https://doi.org/10.1186/s41239-020-00220-3
17. Kippers WB, Poortman CL, Schildkamp K, Visscher AJ. Data literacy: What do educators learn and struggle with during a data use intervention? Studies in Educational Evaluation. 2018;56:21-31. https://doi.org/10.1016/J.STUEDUC.2017.11.001
18. Gray J, Gerlitz C, Bounegru L. Data infrastructure literacy. Big Data and Society. 2018;5(2):1-13. https://doi.org/10.1177/2053951718786316
19. Robertson J, Tisdall EK. The importance of consulting children and young people about data literacy. Journal of Media Literacy Education. 2020;12(3):58-74. https://doi.org/10.23860/JMLE-2020-12-3-6
20. Huang Q, Chen L. Literacy unbound: Multiliterate, multilingual, multimodal. International Journal of Bilingual Education and Bilingualism. 2020. https://doi.org/10.1080/13670050.2020.1835811

21. Sligo F. Literacy and orality at work. Literacy and Orality at Work. 2021. https://doi.org/10.3726/b17476
22. Koltay T. Data literacy for researchers and data librarians. Journal of Librarianship and Information Science. 2017;49(1):3-14. https://doi.org/10.1177/0961000615616450
23. Batt S, Grealis T, Harmon O, Tomolonis P. Learning Tableau: A data visualization tool. Journal of Economic Education. 2020;51(3-4):317-28. https://doi.org/10.1080/00220485.2020.1804503
24. Mandinach EB, Schildkamp K. The complexity of data-based decision making: An introduction to the special issue. Studies in Educational Evaluation. 2021;69:100906. (2020) 2020-2022. https://doi.org/10.1016/j.stueduc.2020.100906
25. Henderson J, Corry M. Data literacy training and use for educational professionals. 2021;14(2):232-44. https://doi.org/10.1108/JRIT-11-2019-0074
26. Riddle DR, Beck JS, Morgan JJ, Brown N, Whitesides H. Making a case for case-based teaching in data literacy. Kappa Delta Pi Record. 2017;53(3):131-3. https://doi.org/10.1080/00228958.2017.1334479
27. Ifenthaler D, Gibson D, Prasse D, Shimada A, Yamada M. Putting learning back into learning analytics: Actions for policy makers, researchers, and practitioners. Educational Technology Research and Development. 2021;69(4):2131-50. https://doi.org/10.1007/s11423-020-09909-8
28. Yi Min Lim D, Yap CEL, Lee JJ. Datastorming: Crafting data into design materials for design students creative data literacy. ACM International Conference Proceeding Series. 2021. https://doi.org/10.1145/3450741.3465246
29. van Audenhove L, Van den Broeck W, Mariën I. Data literacy and education: Introduction and the challenges for our field. Journal of Media Literacy Education. 2020;12(3):1-5. https://doi.org/10.23860/JMLE-2020-12-3-1
30. Lasser J, Manik D, Silbersdorff A, Säfken B, Kneib T. Introductory data science across disciplines, using Python, case studies, and industry consulting projects. Teaching Statistics. 2021;43(S1):S190-200. https://doi.org/10.1111/test.12243
31. Wolff A, Kortuem G, Cavero J. Urban data games: Creating smart citizens for smart cities. Proceedings-IEEE 15th International Conference on Advanced Learning Technologies: Advanced Technologies for Supporting Open Access to Formal and Informal Learning, ICALT 2015. 2015;164-5. https://doi.org/10.1109/ICALT.2015.44
32. Cowie B, Edwards F, Trask S. Explicating the Value of Standardized Educational Achievement Data and a Protocol for Collaborative Analysis of This Data. 2021;6:1-14. https://doi.org/10.3389/feduc.2021.619319
33. Mandinach EB, Schildkamp K. Misconceptions about data-based decision making in education: An exploration of the literature. Studies in Educational Evaluation. 2021;69:100842. (2019) https://doi.org/10.1016/j.stueduc.2020.100842
34. Piro JS, Dunlap K, Shutt T. A collaborative Data Chat: Teaching summative assessment data use in pre-service teacher education. Cogent Education. 2014;1(1):968409. https://doi.org/10.1080/2331186X.2014.968409
35. Pothier WG, Condon PB. Towards data literacy competencies: Business students, workforce needs, and the role of the librarian. Journal of Business and Finance Librarianship. 2020;25(3-4):123-46. https://doi.org/10.1080/08963568.2019.1680189
36. 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;7(1):0-19. https://doi.org/10.7710/2162-3309.2266
37. Chen C, Ibekwe-SanJuan F, Hou J. The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. Journal of the American Society for Information Science and Technology. 2010;61(7):1386-409. https://doi.org/10.1002/asi.21309
38. Brandes U. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology. 2001;25(2):163-77. https://doi.org/10.1080/0022250X.2001.9990249
39. Chen C. The centrality of pivotal points in the evolution of scientific networks. International Conference on Intelligent User Interfaces, Proceedings IUI. 2005;98-105. https://doi.org/10.1145/1040830.1040859
40. Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology. 2006;57(3):359-77. https://doi.org/10.1002/asi.20317
41. Mane KK, Börner K. Mapping topics and topic bursts in PNAS. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(S1):5287-90. https://doi.org/10.1073/pnas.0307626100
42. Kleinberg J. Bursty and hierarchical structure in streams. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002;91-101. https://doi.org/10.1145/775060.775061
43. Aryadoust V, Tan HAH, Ng LY. A scientometric review of rasch measurement: The rise and progress of a specialty. Frontiers in Psychology. 2019;10. https://doi.org/10.3389/fpsyg.2019.02197
44. Garfield E. From Bibliographic Coupling to Co-Citation Analysis via Algorithmic Historio-Bibliography. 2001. Retrieved from http://www.cis.drexel.edu/news/page5.html
45. Small H. Paradigms, citations, and maps of science: A personal history. Journal of the American Society for Information Science and Technology. 2003;54(5):394-9. https://doi.org/10.1002/asi.10225
46. Small H. Co-citation context analysis and the structure of paradigms. Journal of Documentation. 1980. MCB UP Ltd. https://doi.org/10.1108/eb026695
47. Small HG. Cited Documents as Concept Symbols. Social Studies of Science. 1978;8(3):327-40. https://doi.org/10.1177/030631277800800305
48. Small H. Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science. 1973;24(4):265-9. https://doi.org/10.1002/asi.4630240406
49. Ronald R, Leo ERG. Becoming Metric-Wise A Bibliometric Guide for Researchers. 2018. https://doi.org/9780081024744
50. Pan X, Yan E, Cui M, Hua W. Examining the usage, citation, and diffusion patterns of bibliometric mapping software: A comparative study of three tools. Journal of Informetrics. 2018;12(2):481-93. https://doi.org/10.1016/j.joi.2018.03.005
51. Mertala P. Data (il)literacy education as a hidden curriculum of the datafication of education. Journal of Media Literacy Education. 2020;12(3):30-42. https://doi.org/10.23860/JMLE-2020-12-3-4
52. Wolff A, Wermelinger M, Petre M. Exploring design principles for data literacy activities to support children's inquiries from complex data. International Journal of Human Computer Studies. 2019;129:41-54. https://doi.org/10.1016/J.IJHCS.2019.03.006

Cite this article: Naseema S, Sevukan R. Exploration of Data Literacy Research using a Network of Cluster Mapping Approach. J Scientometric Res. 10.5530/jscires.12.1.002.