J Data Sci. Info. Citation Studies., 2025; 4(1):96-102.
https://www.jcitation.org

Short Communication

Mining Indian Scholar's Information: Data Overview, Benefits and Challenges

Kiran Sharma, Ziya Uddin*

School of Engineering and Technology, BML Munjal University, Sidhhrawali, Gurugram, Haryana, INDIA.

ABSTRACT

Understanding the growth and impact of science in any nation is crucial and scholarly publishing databases play a significant role in this endeavour. Currently, the most popular scholarly databases include Scopus, Web of Science and Dimensions, which are subscription-based. Additionally, there are some open-access databases available, but their accessibility is limited. India also has its own scholarly information dataset called VIDWAN, managed by INFLIBNET, which encompasses the research profiles of scholars. This database is designed to recommend peers based on expertise, facilitating collaboration within the country. However, its limitation lies in the fact that users can only view the information and cannot access the data directly. To overcome this limitation, the paper scraped data from the VIDWAN portal and organized it into a structured format. The dataset contains information on scholars' profiles such as name, gender, position, affiliation, publication details sourced from Scopus and experiences such as PhD handles and research funding. The paper describes the coverage of data from various aspects, including funding details. Typically, such information is gathered from multiple sources to analyze individual performance. Therefore, this Indian scholar dataset serves as a valuable asset for assessing individual performance. Finally, it highlights the limitations encountered and proposes possible solutions to overcome them.

Keywords: Bibliometrics, Data scrapping, Indian database, Indian Scholars, Scientometrics.

Correspondence:

Dr. Ziya Uddin
School of Engineering andTechnology,
BML Munjal University, Sidhhrawali,
Gurugram, Haryana, INDIA.
Email: ziya.uddin@bmu.edu.in

Received: 06-11-2024;
Revised: 24-01-2025;
Accepted: 12-04-2025.

INTRODUCTION

Research plays a pivotal role in shaping the trajectory of a country's growth. Access to information about on-going research endeavours, prominent scholars in specific fields and funding agencies dedicated to advancing research in particular domains is of paramount importance. Such knowledge significantly influences collaborative research efforts going forward. In the digital age, scholarly database platforms play a crucial role in facilitating access to research outputs, fostering collaboration among scholars and driving the advancement of scientific knowledge across diverse fields. Well-known databases such as Scopus (Burnham, 2006), PubMed (Canese and Weis, 2013), Web of Science (Birkle, *et al.*, 2020), IEEE Xplore, ScienceDirect, JSTOR, Google Scholar (Orduña-Malea, *et al.*, 2019), ProQuest, OpenAlex, CiteScore, Dimension, EBSCOhost, etc. provide access to journals, newspapers, conference proceedings, dissertations and other content (Mikki, 2009; Aksnes and Sivertsen, 2019). However, these databases often lack comprehensive information about authors, including gender, designation, PhD status, research projects, funding, organizational affiliations and other database IDs. This research paper aims to conduct a thorough examination of the Indian scholarly database platform "VIDWAN".

DOI: 10.5530/jcitation.20250140

Publishing Partner : Manuscript Technomedia. [www.mstechnomedia.com]

A Brief on VIDWAN Database

VIDWAN (https://vidwan.inflibnet.ac.in/), an Indian scholars database managed by INFLIBNET, is an author-specific platform that showcases the scientific profiles of scholars, researchers and faculty members affiliated with Indian academic institutions and R&D organizations (Sab, *et al.*, 2019). The database includes details about their current organization, academic position, gender, type of organization, research area, expertise, research funding (if any), professional experience and accomplishments. Each profile may list Scopus ID, Researcher ID, Google Scholar ID, or ORCID ID. Publication and research metrics are automatically fetched from the Scopus database.

The primary goal of the VIDWAN database is to facilitate information sharing among experts in the same field, thus promoting potential collaborations. VIDWAN simplifies the search for professionals by allowing users to sort profiles based on subject categories. Individual profile pages are provided for researchers and faculty members who have registered on the portal and meet a certain score threshold. Those who do not meet the required score are registered as users but do not receive a profile page.

Study Objectives

The profile information provided by VIDWAN is crucial for various purposes, such as analyzing individual performance, assessing an institution's performance, identifying key experts in a given discipline, etc. Generally, gathering scholars' profiles can be challenging, but this interface offers a centralized source of author information. However, a limitation of this interface is that it only allows viewing the data without direct access to download it. Therefore, the main goal of this work is to scrape all web pages, organize the data for practical use and highlight the extent of the data coverage.

METHODOLOGY

Data Scrapping and Filtering

First, we conducted a systematic web scraping analysis of academic profiles hosted on the VIDWAN platform. Our methodology involved utilizing Python libraries, such as requests for HTTP interactions and BeautifulSoup for HTML parsing. The scraping process was organized into a series of functions, with the main function responsible for fetching HTML content from VIDWAN profile URLs. We extracted various data points, including author details, publication statistics and personal information such as gender and department. Error handling mechanisms were implemented to address potential issues during the scraping process, ensuring robust data extraction.

The collected data was organized into a Python list and subsequently converted into a Pandas DataFrame for structured storage. This DataFrame, serving as a tabular representation of the scraped data, included key variables such as author details, citation counts, publication statistics and personal information. This structured dataset facilitates further analysis and visualization.

We adhered to ethical web scraping practices by making requests responsibly and respecting the terms of service of the VIDWAN platform. The focus was strictly on publicly available information and we acknowledged the importance of data privacy and platform guidelines. The structured dataset lays the foundation for meaningful insights into the academic landscape represented on the VIDWAN platform. To manage the scraping process efficiently, we divided it into two fragments based on the type of relationship between the data:

- Atomic Data: Data with a one-to-one relationship.
- Projects and Funding Data: Data with a one-to-many relationship.

Understanding these relationships helps maintain the integrity of the data throughout the scraping process.

This entire process scraped 213,352 author profiles. Among these, 190,442 profiles had no funding information, while 22,910 profiles included funding information. We further filtered the data, removing duplicates and incomplete profiles. This resulted in a total of 192,243 unique author profiles.

What does this database offer?

This section provides a detailed overview of the 192,243 filtered unique author profiles where 56.53% are male and 33.62% are female profiles. As previously discussed, the filtered data includes information on authors' names, affiliations, types of organizations, academic positions, gender, subject expertise, research funding and any Scopus, Researcher, Google Scholar and ORCID IDs.

Institutional Categorization

The database provides the names of affiliated institutions and universities. The type of institution is mapped through an external database and categorized into: colleges, national institutes, private universities, public universities, research centres, research institutes and medical institutes. All institutions of national importance, such as IITs, IIITs, NITs, IISc and IISER, are categorized under "National Institutes." Similarly, all government-funded central and state universities fall under the "Public University" category. All medical institutes, hospitals and clinics are classified as "Medical Institutes." National research centres like ICSSR, CSIR and ICMR are grouped under "Research Centres," while self-funded institutes and universities are categorized as "Private Universities." Finally, all affiliated colleges are listed under the "Colleges" category.

Figure 1 shows the distribution of various types of institutions in the VIDWAN database. The majority of profiles are from college professionals (29.5%), followed by scholars from national institutes (25.7%), private universities (21.2%) and public universities (14.5%). Approximately 3% of the profiles come from research institutes, research centres and medical institutes. This distribution highlights the prominence of general educational institutions like colleges and national institutes, while research-oriented and medical institutions appear to be less active on VIDWAN database.

Table 1 presents the gender distribution across different organizational types. The proportion of female scholars is higher in colleges (38.59%) compared to males (24%), while the proportion of male scholars is higher in national institutes (28.6%) compared to females (20.8%). Similarly, in public universities, the male proportion is higher than the female proportion. Overall, this data highlights that there are more females in private institutions than males. This trend is noteworthy for further study on gender inequality in premier institutions.

Disciplines Categorization

Figure 2 shows the distribution of profiles based on their areas of expertise. The expertise categories include Engineering and Technology, Physical Sciences, Social Sciences, Arts and

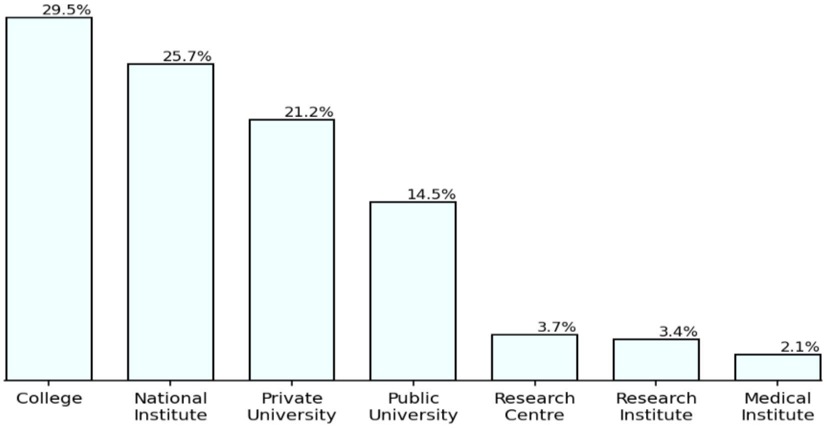

Figure 1: Expertise distribution as per Institutions.

Table 1: Number of female and male profiles as per institution type and discipline type.

Institution Category	Female Profiles		Male Profiles	
	Count	In %	Count	In %
College	27666	38.59	28972	24.03
National Institute	14974	20.89	34491	28.61
Private University	15631	21.8	25118	20.84
Public University	8211	11.45	19633	16.29
Research Centre	1559	2.17	5538	4.59
Research Institute	2161	3.01	4300	3.57
Medical Institute	1491	2.08	2498	2.07
Discipline Category				
Engineering and Technology	25843	34.38	50572	41.95
Physical Sciences	11375	15.13	21244	17.62
Social Sciences	11678	15.54	15352	12.73
Arts and Humanities	11254	14.97	10888	9.03
Health and Medical Sciences	8045	10.7	10649	8.83
Agricultural Sciences	4118	5.48	8345	6.92
Biological Sciences	2856	3.8	3500	2.9

Humanities, Health and Medical Sciences, Agricultural Sciences and Biological Sciences. The majority of author profiles are from the Engineering and Technology discipline (39.1%), followed by Physical Sciences (16.6%), Social Sciences (13.8%), Arts and Humanities (11.3%) and so on. The least represented discipline is Biological Sciences (3.2%). It is important to note that these profiles are dynamic and may increase in the future, depending on how many institutions and individual authors create their profiles in VIDWAN.

Table 1 shows the gender distribution of expertise across various disciplines. In the Engineering and Technology discipline, female profiles are less represented (34.3%) compared to males (41.9%). Females are more represented in Social Sciences, Arts and Humanities and Health and Medical Sciences than males, while males have higher representation in Engineering and Technology, Physical Sciences and Agricultural Sciences.

Research Funding Information

The other part of the work involves examining the funding information provided by scholars in their profiles. This information is useful for understanding the types of funding and funding agencies. While the amounts of the funds are also mentioned, there are many discrepancies, so we filtered out that information. The funding information is useful for analyzing the trends of various agencies and provides insights into significant funding patterns. This understanding helps us comprehend the current research needs in the field. Identifying these trends is essential for pinpointing key research opportunities and streamlining efforts to pursue related research potential. To

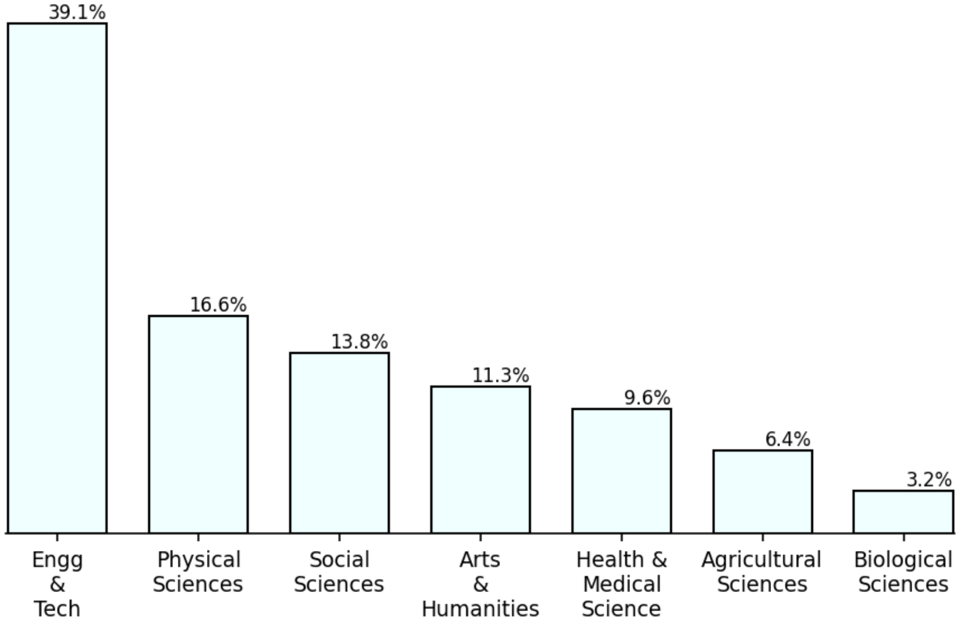

Figure 2: Expertise distribution as per discipline.

associate the funding information with its relevant funding agency, we developed a cleaning algorithm, as shown in Figure 3.

K-Means Clustering is a primary method used for analyzing text patterns within funding agency data. Recognizing its effectiveness, the study explores minor adjustments to the algorithm's parameters, such as incorporating cosine similarity and text vectorization. These tweaks aim to improve the accuracy of the clustering process, resulting in approximately 85% reliability in the obtained results. Outliers are managed by placing them within separate cluster bounds, ensuring a comprehensive understanding of the funding landscape. Table 2 provides an overview of various funding agencies and the number of research projects they support, each with more than 100 funded projects. The table lists both the absolute count and the percentage of total projects associated with each funding agency. The University Grants Commission (UGC) leads with the highest number of projects, totalling 7,089, which represents 19.53% of all projects. The Department of Science and Technology (DST) follows, funding 5,185 projects, accounting for 14.28%. Internal Funding ranks third, supporting 4,536 projects (12.5%), while the Science and Engineering Research Board (SERB) contributes to 3,421 projects (9.42%). The All-India Council for Technical Education (AICTE) funds 2,745 projects, comprising 7.56% of the total. Other notable agencies include the Ministry of Agriculture and Farmers Welfare (2,323 projects, 6.4%) and the Department of Biotechnology (DBT) (2,319 projects, 6.39%). State-funded projects account for 2,122 projects (5.85%). Funding from specialized councils, such as the Indian Council of Social Science Research (ICSSR) (1,182 projects, 3.26%) and the Council of Scientific and Industrial Research (CSIR) (1,099 projects, 3.03%), also contribute significantly. Smaller but still noteworthy contributions come from agencies like the Department of Education (DOEd) with 891 projects (2.45%), the Indian Council of Medical Research (ICMR) (791 projects, 2.18%) and the Defence Research and Development Organization (DRDO) (778 projects, 2.14%). The Department of Space (DOS) funds 503 projects (1.39%) and Private Funding supports 299 projects (0.82%). A few agencies have less than 1% of the total projects, such as the Ministry of Electronics and Information Technology (MeitY) with 138 projects (0.38%), Tata Trusts with 130 projects (0.36%), the Ministry of Environment and Forests with 120 projects (0.33%) and the Department of Atomic Energy (DAE) with 119 projects (0.33%). This distribution demonstrates the strong support provided by government bodies, with UGC and DST contributing the majority of funding.

Table 3 shows the distribution of research projects across various thematic areas, detailing both the number and percentage of projects associated with each theme. The Engineering and Technology theme has the largest share, with 10,669 projects, making up 29.39% of the total. Close behind is Health and Medicine, with 10,577 projects, representing 29.14%. Together, these two themes account for more than half of all projects, indicating a strong focus on technological and medical research. Environment and Sustainable Development ranks third with 6,475 projects, or 17.84%, showing significant interest in ecological and sustainability-related research. Agricultural Research follows with 2,820 projects, making up 7.77% of the total, while Science and Technology has 2,617 projects, representing 7.21%. Social Sciences account for 1,558 projects (4.29%), highlighting research in societal and cultural fields, while Education and

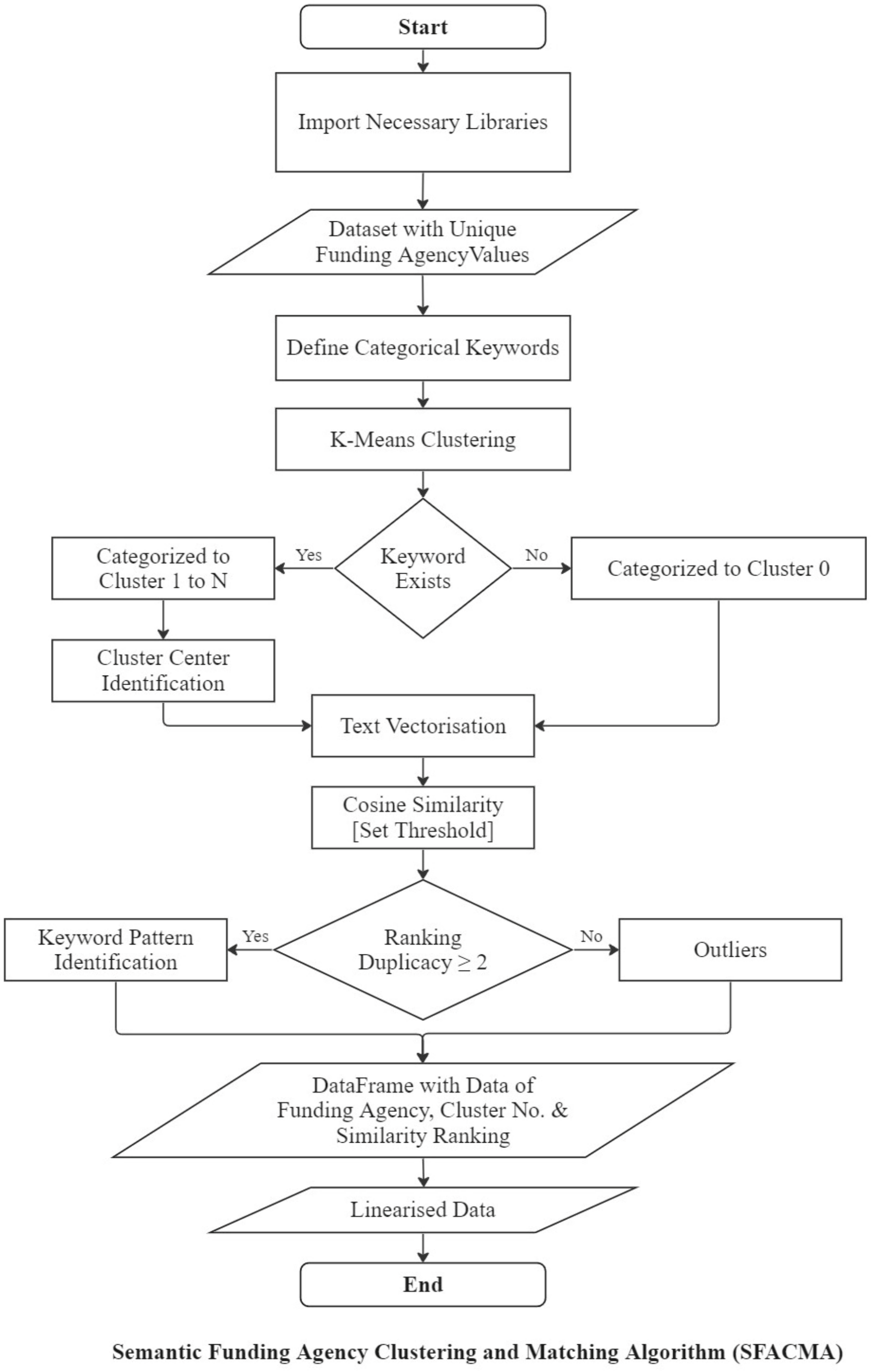
Start
Import Necessary Libraries
Dataset with Unique Funding AgencyValues
Define Categorical Keywords
K-Means Clustering
Keyword Exists
Yes
No
Categorized to Cluster 1 to N
Categorized to Cluster 0
Cluster Center Identification
Text Vectorisation
Cosine Similarity [Set Threshold]
Ranking Duplicacy ≥ 2
Yes
No
Keyword Pattern Identification
Outliers
DataFrame with Data of Funding Agency, Cluster No. & Similarity Ranking
Linearised Data
End
Semantic Funding Agency Clustering and Matching Algorithm (SFACMA)

Figure 3: Algorithm flow to identify the funding agency.

Learning comprises 1,154 projects, making up 3.18%. Smaller categories include Business and Economics with 324 projects (0.89%) and Food Safety and Quality with 78 projects (0.21%). The themes with the fewest projects are Research and Innovation (15 projects, 0.04%) and Arts and Culture (13 projects, 0.04%), suggesting comparatively lower research activity in these areas. Overall, the data indicates a strong emphasis on projects related to engineering, medicine and sustainability, with less focus on business, food safety and cultural studies.

DISCUSSION AND CONCLUSION

VIDWAN is India's first scholarly information database that connects peers based on their expertise from numerous fields. As discussed, this database provides a viewing facility to users

Table 2: List of funding agencies and the number of projects (>100).

Funding Agency (India)	No of Projects	
	Count	In %
University Grants Commission (UGC).	7089	19.53
Department of Science and Technology (DST).	5185	14.28
Internal Funding	4536	12.5
Science and Engineering Research Board (SERB).	3421	9.42
All India Council for Technical Education (AICTE).	2745	7.56
Ministry of Agriculture and Farmers Welfare.	2323	6.4
Department of Biotechnology (DBT).	2319	6.39
State Funded	2122	5.85
Indian Council of Social Science Research (ICCSR).	1182	3.26
Council of Scientific and Industrial Research (CSIR).	1099	3.03
Department of Education (DOEd).	891	2.45
Indian Council of Medical Research (ICMR).	791	2.18
Defence Research and Development Organization (DRDO).	778	2.14
Department of Space (DOS).	503	1.39
Private Funding	299	0.82
Ministry of Electronics and Information Technology (MeitY).	138	0.38
Tata Trusts	130	0.36
Ministry of Environment and forests.	120	0.33
Department of Atomic Energy (DAE).	119	0.33

Table 3: Number of projects listed under various themes.

Funding Themes	Count	In %
Engineering and Technology.	10669	29.39
Health and Medicine.	10577	29.14
Environment and Sustainable Development.	6475	17.84
Agricultural Research.	2820	7.77
Science and Technology.	2617	7.21
Social Sciences	1558	4.29
Education and Learning.	1154	3.18
Business and Economics.	324	0.89
Food Safety and Quality.	78	0.21
Research and Innovation.	15	0.04
Arts and Culture.	13	0.04

but does not allow data downloads. To analyze research trends and individual performance in science, there is a pressing need for a comprehensive scholarly database. Typically, scholars' publication information is sourced from Scopus, Web of Science, Google Scholar, or other databases. However, these sources do not provide profile information of scholars along with their expertise and gender. Researchers often need to map multiple databases to extract relevant information.

India has, for the first time, offered comprehensive information about scholars, encompassing personal details, research particulars and professional experience, all centralized on a single platform. This development presents fresh opportunities for researchers. Nonetheless, the extraction of information from this portal remains a challenge. Hence, we conducted web scraping to compile all pertinent details about Indian scholars in a structured manner.

However, this dataset presents several limitations. Firstly, the data provided by VIDWAN is self-declared, increasing the likelihood of errors as there is no external validation during data entry. Secondly, the entered data lacks consistency and missing information diminishes its reliability and usability. Thirdly, since the data is self-reported, individuals may not regularly update their information, resulting in potential outdatedness.

To resolve these issues, the following steps are recommended:

- Each individual should be responsible for entering accurate details.

- Each institute should ensure the reliability and completeness of their faculty profiles. They could include this as part of their faculty appraisal process.

- A standardized data format should be established.

In conclusion, VIDWAN is a valuable source of information on Indian scholars. However, due to its limited viewing properties and incomplete user profiles, it lacks popularity among users. Despite this, it is a powerful source of information, providing Indian scholar researcher IDs, Scopus IDs, Google Scholar IDs and ORCID IDs, which can be used to extract data from other databases like Scopus, Web of Science, etc.

DATA AVAILABILITY STATEMENT

The data and code is accessible from authors upon request by users.

ACKNOWLEDGEMENT

The authors would like to acknowledge the efforts of Ajeet Kumar Verma, Mahi Singhal, Monika Raghav, Sai Varshith Gali and K. Sri Sai Sri Vamsi Krishna.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest.

REFERENCES

Aksnes, D. W., & Sivertsen, G. (2019). A criteria-based assessment of the coverage of Scopus and Web of Science. Journal of Data and Information Science, 4(1), 1–21. htt ps://doi.org/10.2478/jdis-2019-0001

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. Quantitative Science Studies, 1(1), 363–376. https://doi.org/10.1162/qss_a_00018

Burnham, J. F. (2006). Scopus database: A review. Biomedical Digital Libraries, 3, 1. ht tps://doi.org/10.1186/1742-5581-3-1

Canese, K., & Weis, S. (2013). The NCBI handbook, 2(1). PubMed: PubMed: the bibliographic database.

Delgado López-Cózar, E., Orduña-Malea, E., & Martín-Martín, A. (2019). Springer handbook of science and technology indicators (pp. 95–127). Google Scholar as a data source for research assessment.

Mikki, S. (2009). A literature review. Nordic Journal of Information Literacy in Higher Education, 1(1), Google Scholar compared to web of science.

Sab, C. M., Kumar, P. D., & Biradar, B. S. (2019). Network System (IRINS): An Overview. Research Information. Academia.edu

Cite this article: Sharma K, Uddin Z. Mining Indian Scholar's Information: Data Overview, Benefits and Challenges. Journal of Data Science, Informetrics, and Citation Studies. 2025;4(1):96-102.