Status of Indian Research Data Repositories: A Study based on Research Data Registry

Vinayak Wadhwa and Manoj Kumar Joshi

A research data repository is a location to preserve and store research data virtually. Maintaining research data in repositories promotes research collaboration opportunities and increases data transparency. The primary objective of this study is to examine the different research data repositories in India listed in the Research Data Repositories Registry with their subject coverage, access policies, data formats, standards & specifications used for implementation. This global registry presents data in typological classifications like institutional, disciplinary, and multidisciplinary project-specific repositories.

Introduction

During the research, the researchers generate lots of data. Once the study is over, the entire dataset is either discarded or stored in personal or institutional hard drives. Sharing data supports open scientific investigation, encourages diversity of studies and opinions, encourages new scopes of work, and helps explore topics not covered by the initial investigators. Besides, unrestricted access to research data increases the returns from public investment in research and development.

Defining research data is challenging because data is heterogeneous. OECD (2007) described "research data as factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings." with an increasing number of digital research data, to make research more reliable, the research data need to be conserved. The most suitable way to store and share research data is with a data repository. A repository is an online database that authorises research data to be conserved and helps others to find and archive the data (RLG-OCLC Report 2002). Apart from archiving research data, a repository will assign a unique identifier for the citation of each uploaded data.

Research data management supports an open-access ecosystem to share research data through open-access data repositories, especially research funded by public funding agencies. Recently government research funding agencies like the Department of Science and Technology, Indian Council of Agricultural Research (ICAR), Indian Council of Social Science Research (ICSSR), Department of Biotechnology, and International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), have formulated data management and sharing policies for research data generated from government-funded research to mandate proper data management throughout the research lifecycle stating guidelines for sharing of research data through research data repositories to make this data publicly accessible.

2. National Data Policy of India

The first initiative for open governance and data sharing was the tenth Five Year Plan (2002-2007) by the Government of India when it announced to become an 'S.M.A.R.T' (Simple, Moral, Accountable, Responsible and Transparent) state. In 2005, the Right to Information (RTI) Act presented and became a milestone in transparent and citizen-driven governance.

There are many challenges while the opening of government data like many of government departments was unclear in what formats most government data is stored. The government of India framed the National Data Sharing and Accessibility Policy in 2012. The policy aims to share government authority-owned data through an efficient network with an accessible format for machines and men (NDSAP, 2012). However, there continues to be a state of urgency to observe research data preservation value for the growth of India's social, economic, and scientific development.

3. Registry of Research Data Repositories(re3data.org)

Re3data (www.re3data.org) was launched in December 2012 and funded by the German Research Foundation (DFG). Re3data.org is the most comprehensive registry to search and identify data repositories. It enables a practice of sharing, accessing, and better visibility of research data. It allows researchers to find different data repositories that can help researchers to access various types of data. Educational institutions and publishers also can recommend researchers for data archiving purposes. This Registry of Research Data Repositories (re3data) indexes 2938 research data repositories worldwide, as on dated 31/08/2022. Of 2938, India has 50 research data repositories in the registry and presents data in typological classifications in institutional, disciplinary, and multidisciplinary project-specific repositories.

4. Review of Literature

In India, research data management is an emerging area that has drawn the attention of researchers in higher education institutions. The literature review is organised into two parts, the theoretical framework and research studies.

The selected global literature review was undertaken on RDM in academic libraries to understand the policies, conceptual framework, and the research data life cycle of RDM for higher educational institutes and the Indian repositories for research data.

Several authors have described a conceptual framework for managing research data in higher educational institutions. Tripathi&Pandy(2018) explained the framework of the workflow of the data life-cycle in its different stages from creation, storage, organisation, sharing, and usage. Singh, Monu, & Dhingra (2018) discussed some critical issues associated with research data management and the roles and responsibilities of the institutions and presented a policy framework for managing research data at the institutional level. Patel(2016) proposed a conceptual framework for organising research data at an institutional level and proposed a model for a National Repository of Open Research Data (NRORD) and the workflow of the functioning of NRORD. Meena & Balasubramania (2020) analysed the difficulties faced in research information management and explored librarians' awareness of RDM in university libraries of Tamil Nadu. Tripathi, Awasthi& Payal(2019) explore the researcher's need, importance, and behaviour in different domains regarding data sharing and the role of libraries in developing data management services. Birender & Sanjay (2019) investigate the present status of research data management policy in India. It proposed research data management policies for research institutions in India to build the anticipated "Indian Academic Research Data Repository."

Effective management of research data in universities and research institutions is being emphasized in literature worldwide. The selected literature review mainly focuses on research data management in research institutions covering aspects such as research data organization, data curation, and preserving raw data for the long term in research data repositories.

Gowen and Meier (2020) analyse the shifts in research data management services and staffing in Association of American Universities (AAU) libraries. The study aimed to identify the development libraries' objectives for research data management. Hamad, Al-Fahed& Al-Soub (2019) uncovered the responsibilities of academic libraries in Jordan and suggested developing RDM policies in collaboration with researchers. Shelly & Jackson (2018) analysed 13 Australian universities' requirements to manage research data, and the study found that libraries significantly provide guidelines, web pages, and hyperlinks to RDM resources. Tripathi, Shukla & Sonker (2017) analysed 47 Central Universities of India as per the list of the University Grants Commission and the best 20 universities as per Times Higher Education 2016-17 ranking to analyse steps initiated to deliver services to their researchers by the libraries and to identify the main components of RDM services. Manu et al. (2018) evaluated research repositories in India with their types, indexed content types, software tools, standards, and specifications used for implementation. They found that Indian research data repositories include only basic research data types. Borkakoti & Singh (2021) studied North East Indian institutes to find library professionals' perceptions about research data management. The study found that most professionals intend to facilitate RDM. Chawinga & Zinn (2020) explored the role of libraries in research data infrastructure at the public libraries in Malawi. The librarians were involved in basic RDM activities. Piracha & Ameen (2018) studied the research data management practices of the university faculty of the "University of Punjab (PU)" and found essential factors, including RDM curation practices, and readiness to share data.

5. Objectives of the study

The objectives of the study are:

i. To identify research data repositories in India,

ii. To identify access policies of Indian data repositories,

iii. To examine data formats available in Indian data repositories, and

iv. To identify how many data sets are available in the data repositories.

6. Data Collection

The registry's data is available by subject, content, and country type. Re3data.org helps various search options. The keyword used in the search is the default option. Other search options, like subject, content, countries, etc., are required to search and analyse all achievable information regarding a repository. After searching, 'country' revealed that there are 50 data repositories in India.

7. Limitations of the study

The analysis is based only on Indian research data repositories indexed in re3data.org. If other Indian research data repositories are available but not listed on re3data, the same are not considered in this study.

8. Results and Discussion

8.1 Subject Coverage

The primary subjects covered in the Indian research data repositories at re3data.org repository are shown below in Figure1. The Life Sciences has the highest coverage in 28 repositories, followed by Medicine -27 and Natural Sciences - 20. Geosciences; Humanities and Social Sciences; Biology; Social and Behavioral Sciences; Economics; Agriculture, Forestry, Horticulture, and Veterinary Medicine; Social Sciences; and Basic Biological and Medical Research are covered in more than ten repositories.

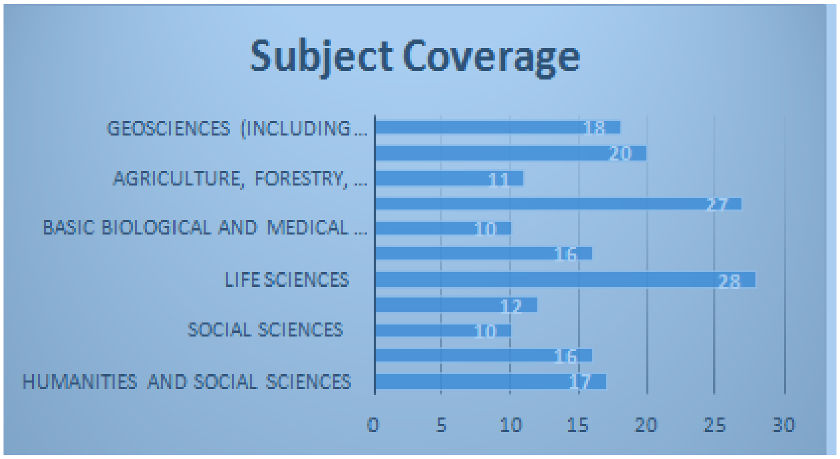

Figure 1: Subject Coverage

8.2 Document Type

Most repositories are of scientific disciplines, thirty-three repositories contain scientific and statistical data, and twenty-six consist of standards office documents. Text, graphics, images and audiovisual data are also stored in these repositories. Figure 2 shows the content type of data maintained by Indian data repositories.

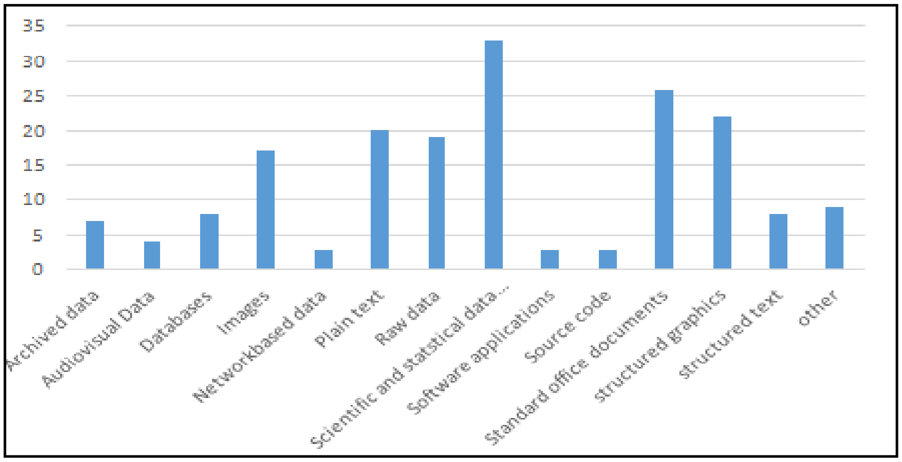

Figure 2: Content type in data repositories.

8.3 Content Access

Out of fifty repositories, thirty-eight provide data by open access. Twenty-one repositories are restricted, three are closed access, and embargoed types are four repositories. Some repositories use all types of access policies depending upon the nature of the data.

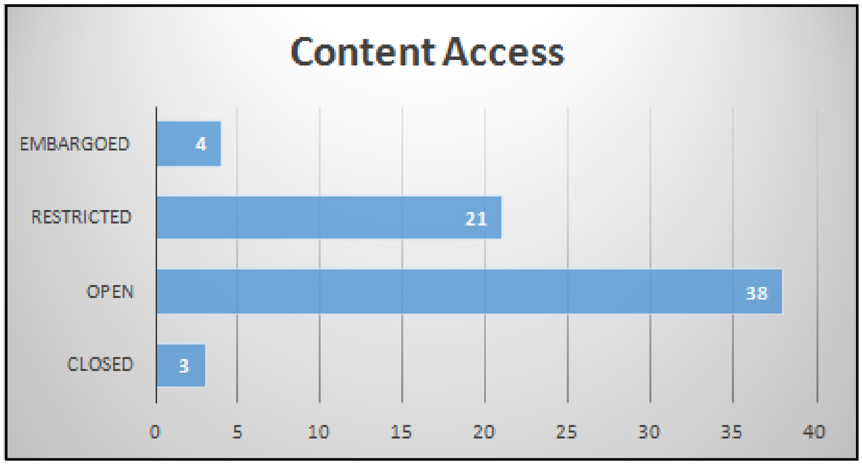

Figure 3: Content Access Type in data repositories.

8.4 Repository Language

All repositories are in English language. Four repositories use the Hindi language for data services besides English. One is in the French language also.

8.5 Document Submission Method

Data managers compile researchers' data through various mediums like email and direct upload using user login. 50% of the data uploading methods are restricted. Out of 50 repositories, 50% (25) use restricted data uploading. Twenty-two repositories use the closed and two repositories use the open data submission method.

8.6 Data Sets Available

After analysing 50 Indian data repositories websites, 22 data repositories public their information about the total number of data sets available in the data repositories. Thus, 22 data repositories with repository sizes are shown below in Table1.

Table 1: Data Repositories with repository size

Sr no.	Name of the Repository	Repository Size
1	MolTable	44 datasets
2	Open Government Data Platform India	577420 resources, 12842 catalogues, 161304 api
3	Open Government Data Portal of Tamil Nadu	36624 resources,4080 catalogues,542 api
4	India Water Portal	335 records
5	Histome-The Histone Infobase	50 histone proteins and ~150 histone modifying enzymes
6	Oral Cancer Gene Database	374 genes
7	"CSISA Data Repository-Cereal Systems Initiative for South Asia (CSISA) Research Data"	24 datasets; 276 files
8	Human Proteinpedia	"2.710 experiments; 15.231 protein entries; 1.960.352 peptide identifications; 4.855.122 MS/MS spectra; 150.368 protein expression, 17.410 post-translational modifications; 34.624 protein-protein ineractions; 2.906 subcellular localization"
9	ACEpepDB: Peptide Database	"135 food sources; 865 peptides; 11 purification methods; 17 assay method references"
10	Biosearch-Marine Biodiversity Database of India	19.760 organism records
11	Open Government Data Portal of Surat City	119 resources
12	Open Government Data Portal of Odisha	17 resources; 10 Catalogues
13	Human Protein Reference Database	"30.047 protein entries; 41.327 protein-protein interactions; 93.710 PTM's; 112.158 protein expression; 22.490 subcellular localization; 470 domains; 453.521 PubMed Links"

14	TBNet India-A National Portal for Tuberculosis InitiativeProtocols Reference Database	8.178 genes
15	Open Government Data Portal of Sikkim	102 datasets
16	India Biodiversity Portal	“58.211 species; 1.479.356 observations; 206 maps; 2596 documents”
17	Indian National Centre for Ocean Information Services	about 1,5 TB
18	Pune Datastore	349 Datasets; 76 Suggested Datasets
19	NetSlim	10 immune signaling and 10 cancer signaling pathways
20	ICSSR Data Service: Social Science Data Repository	138 Datasets; 140 studies
21	“KRISHI - Knowledge based Resources Information Systems Hub for Innovations in Agriculture”	1011 datasets
22	IMEx-The International Molecular Exchange Consortium	“Data:22954,Interactors : 118,924,Interactions : 750,480,Binary Interactions : 1,194,594”

8.7 Data Licenses

The largest number of 42% of data repositories, use self-copyright license data policies. Besides, Open General License (OGL) and Creative Commons (CC) are used as it is comfortable for most repositories to implement. Following are the available data licenses:

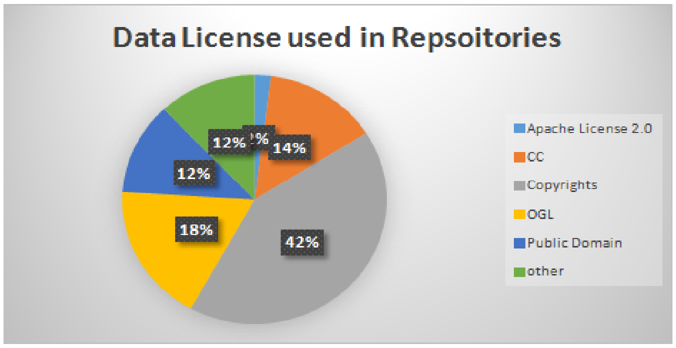

Figure 4: Data License used in data repositories

9. ICSSR Data Service: Social Science Data Repository

ICSSR Data Service is a national data repository launched formally on 20th June 2016, which aims to provide access to a wide range of datasets generated by the MoSPI, New Delhi and social science institutions under the direct purview of ICSSR and other government organisations.

Currently, 140 data sets are provided by ICSSR in the repository with metadata elements. These data sets are categorised into eight significant subjects: "Debt & Investment, Domestic Tourism, Education, Enterprise Survey, Employment & Unemployment, Housing Conditions, House Hold consumers Expenditure and Health Care".

Table 2: Usage of the social science data repository year-wise.

Year	Types of Datasets	No of downloads	No. of users
2018-19	140	576748	1128
2019-20	140	1043425	1221
2020-21	140	545419	923

*Table 2 Data provided by NASSDOC

As per Table 2, the Total number of downloads of data sets from 2018-19 to 2020-21 is more than 2 Million. The usage of the ICSSR social science data repository is quite large by social science researchers for their research.

10. Conclusion

The present study analyses the Indian research data repositories indexed in the most comprehensive research data registry (re3data.org). The registry indexes 2938 research data repositories from around the world; India has 50 research data repositories in the registry and presents data in typological categories like institutional, disciplinary and multidisciplinary project-specific repositories. India has most of the disciplinary repositories. The repositories mainly cover subjects such as Life Sciences, Medicine and Natural Sciences. Most repositories consist of scientific and statistical data. Thirty-eight repositories provide open access to data. The largest number of 42% data repositories, use self-copyright license data policies, and 22 data repositories' websites public their information about the total number of data sets available in the data repositories. The study also finds that ICSSR-Data Service is a national social science data repository to make all social science statistical datasets generated by government and non-government initiatives, public in open access to the entire social science research society. The social science community's usage of this social science repository is quite significant.

Acknowledgement

The Authors gratefully acknowledge the information provided by Mr. Ramesh Yernagula, Director, NASSDOC, about ICSSR Data Service.

References

1. Borkakoti, R., & Singh, S. K. (2021). Research Data Management in Central Universities and Institutes of National Importance: a perspective from North East India. Library Philosophy and Practice.

2. Bunkar, A. R., & Bhatt, D. D. (2020). Perception of Researchers & Academicians of Parul University towards Research Data Management System & Role of Library: A Study. DESIDOC Journal of Library & Information Technology, 40(3), 139-146.

3. Chawinga, W. D., & Zinn, S. (2020). Research Data Management in Universities: A Comparative Study from the Perspectives of Librarians and Management. International Information & Library Review.

4. Cox, A., & Verbaan, E. (2018). Exploring Research Data Management. Facet Publishing.

5. Gowen, E., & Meier, J. J. (2020). Research Data Management Services and Stratergic Planning in Libraries Today:A Longitudnal Study. Journal of Librarianship and Scholarly Communication, 8(General Issue), 1-19.

6. Hamad, F., Al-Fadel, M., & Al-Soub, A. (2019). Awareness of Research Data Management Services at Academic Libraries in Jordan: Roles, Responsibilities and Challenges. New Review of Academic Librarianship.

7. Manu, T., Asjola, V., Gowda, M., AA, S., Chaudhary, P., & Muduli, P. K. (2018). Analysis of research data repositories in India. Knowledge Organisation in Academic Libraries (I-KOAL-2018), (pp. 312-322).

8. Meena, S., & Balasubramanian, P. (2020). A Study on Research Data Management in University Libraries: A Modern Day Scenario. Library Progress (International), 40, No-2(July-December 2020), 328-335.

9. Organization of Economic Co-operation and Development (OECD) (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding, available at: https://www.oecd.org/sti/inno/38500813.pdf (accessed 26 August 2022).

10. Pal, B., & Singh, S. K. (2019). Indian Academic Research Data Repository (IARDR) With INFLIBNET: A Futuristic Plan. 12th International CALIBER-2019 (pp. 36-44). KIIT, Bhubaneswar, Odisha: INFLIBNET Centre, Gandhinagar, Gujarat.

11. Patel, D. (2016). Research data management:a conceptual framework. Library Review, 65(4/6), 226-241.

12. Piracha, H. A., & Ameen, K. (2018). Research Data Management Practices of Faculty Members. Pakistan Journal of Information Management & Libraries, 20, 60-75.

13. Saeed, S., & Naushad Ali, P. (2019). Research Data Management and Data Sharing among Research Scholars of Life Sciences and Social Sciences. DESIDOC Journal of Library & Information Technology, 39(6), 290-299.

14. Shelly, M., & Jackson, M. (2018). Research data management compliance: is there a biggerrole for university libraries. Journal of The Australian Library and Information Association, 67(4), 394-410.

15. Shrivastava, P., & Gupta, D. K. (2018). Research Data Preservation in India: An Analysis based on Research Data Registry. World Digital Libraries, 11(2), 107-121.

16. Singh, N. K., Monu, H., & Dhingra, N. (2018). Research Data Management Policy and Institutional Framework. 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services.

17. Manu,T. (2018). Researchers' Perceptions on Research Data Management: A Survey., International Conference on Exploring the Horizons of Library and Information Science: From Libraries to Knowledge Hubs.Bangalore: Document Research and Training Center(DRTC) , 115-128.

18. Tripathi, D. P., & Pandy, S. R. (2018). Developing a Conceptual Framework of Research Data Management for Higher Educational Institutions. 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services.

19. Tripathi, M., Awasthi, S., & Payal. (2019). A Selective Review of Literature on Research Data Management in Academic Libraries. DESIDOC Journal of Library & Information Technology, 39(6), 338-345.

20. Tripathi, M., Shukla, A., & Sonker, S. K. (2017). Research Data Management Practices in University Libraries: A Study. DESIDOC Journal of Library & Information Technology, 37(6), 417-424.

21. RLG-OCLC (2002). Trusted Digital Repositories: Attributes and Responsibilities An RLG-OCLC Report RLG Mountain View, CA. available at: https://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf.

22. Whyte, A., & Tedds, J. (2011). Making the Case for Research Data Management. DCC.

Keywords: Indian Research Data Repositories; Research Data Registry

About Authors

Mr. Vinayak Wadhwa
Research Scholar
Department of Library & Information Science, Kurukshetra University, Kurukshetra
Email: vinayakexams@kuk.ac.in

Prof. Manoj Kumar Joshi
Professor and Chairman
Department of Library & Information Science, Kurukshetra University, Kurukshetra
Email: manojkj01@yahoo.com