International Journal of Research in Library Science (IJRLS)

ISSN: 2455-104X

DOI: 10.26761/IJRLS.11.1.2025.1839

Volume 11, Issue 1 (January-March) 2025, Page: 460-471, Paper ID: IJRLS-1839

Research Data Repositories of India Registered in re3data.org: A Bibliometric Analysis (2012–2024)

Manish Sharma[1]; Dr. Praveen Babel[2]

Research Scholar[1]; Faculty & Research Supervisor[2], Department of Library and Information Science, Banasthali Vidyapith, Rajasthan, India[1,2]

lismanishdadhich@gmail.com; praveenbabeludr@gmail.com

ABSTRACT

This paper conducts a bibliometric and infrastructural analysis of the registries of research data repositories (RDRs) situated in India, as cataloged by the Registry of Research Data Repositories (re3data.org) from 2012 through 2024. re3data.org, as the premier global, cross-disciplinary registry of research data repositories, cataloged over 3,000 research data repositories worldwide as of 2023. Utilizing the re3data.org API, the research team extracted, refined and assessed data pertinent to research data repositories of Indian research institutions. The research team obtained a final dataset of 55 research data repositories that contained an Indian institutional or a geographic connection. This study exhibits that by 2012, the total number of RDRs from India increased from a negligible number to 14 RDRs with a declared founding year. The majority of the remaining RDRs (24 out of 55) did not contain a machine-readable field for the founding year, thus, significantly limiting the research. Of the 55 RDRs, 36 were disciplinary, 15 were institutional, and 1 was governmental. Also, 52 of the 55 RDRs were classified as data-provider RDRs, while the remaining 13 were classified as service-provider RDRs. The most represented subjects RDRs were the Humanities and Social Sciences (28 RDRs) and the Life Sciences (26 RDRs); followed by the Natural Sciences (23 RDRs) and the Engineering Sciences (8 RDRs). Discernibly, RDRs have poor adoption of Persistent Identifiers. Of the 55 RDRs surveyed, only 5 RDRs had been assigned a DOI, while 27 RDRs explicitly stated they had no PID.The study identifies structural deficits regarding the completeness of metadata, licensing, and certification (none of the repositories in the sample certify a data repository certification such as CoreTrustSeal) and analyzes the effects on India's open science and the National Data Sharing and Accessibility Policy (NDSAP). The study presents multiple recommendations directed towards repository managers, funding bodies, and data policy bodies geared towards improving discoverability, interoperability, and the FAIR principles in the Indian research data infrastructure.

.

KEYWORDS: re3data.org; research data repositories; India; bibliometric analysis; data repository registry.

1. INTRODUCTION

Research data is increasingly recognized as an output of equal standing to research publications. Addressing open research demands from funding bodies and an increasing number of journals require data citation and data-sharing declarations. This has highlighted the need for a core component of scholarly infrastructure: the ability to identify trustworthy research data repositories. The Registry of Research Data Repositories, re3data.org, began in 2012 as part of a project funded by the German Research Foundation to assist researchers, funding bodies, publishers, and librarians in locating data repositories across all disciplines. Repositories are described in a standard format that allows for a systematic analysis and comparison of repositories across different countries and research disciplines.

In its first years, the re3data.org project described over 2,400 research data repositories, an increase from a survey of 1,381 repositories in 2015. This increase included over 3,000 repositories in 2023. The rapid increase in the number of open research data and the number of research data repositories around the world resulted in the registry's use by researchers, librarians, and research institutions as the primary means of reference. Despite the increased number of repositories, the geographical coverage of the repositories remains uneven, concentrated primarily in North America and Europe, and less in South Asia and India in particular.India presents a rare opportunity for studying the effects of an infrastructural audit. With India being one of the largest producers of research and having an active government policy of open data with the National Data Sharing and Accessibility Policy (NDSAP) and Open Government Data (OGD) Platform India, there is justified reason for India to create FAIR compliant, accessible data repositories. A previous study of the current state of Indian Data Repositories, included in the re3data registry, found that only 41 of the 51 (80%) repositories, listed under the “India” section, were operational, indicating a lack of metadata and maintenance. This research is a continuation of this study and provides an update, in the form of a new extraction of the re3data registry, drawing statistics for 2024, to investigate growth, distribution of subject and type, provider characteristics, the type of content, and persistent identifiers for the complete set of India associated data repositories.

1.1 Objectives

- ✓ Analyze the increase of research data repositories associated with India that are found on re3data.org for the years 2012-2024.
- ✓ Categorize these repositories according to type and provider type, as well as the subject and the type of content.
- ✓ Evaluate the presence of persistent identifiers and certification as indicators of FAIR compliance.
- ✓ Examine the implications on policies and practices of the open science and research data infrastructure of India.

2. LITERATURE REVIEW

Pampel et al. was the first to describe the aims of re3data.org. As search engines help users find data, re3data.org helps users find research data repositories. The research search engines that re3data.org helps users find work by cross-referencing a multitude of disparate research data repositories. Kindling et al. first quantitatively researched re3data.org in 2015. They analyzed the first 1,381 research data repositories that re3data.org listed. They categorized the research data repositories by discipline and by the types of data they contain. Other researchers used this data to

analyze research data repositories by country and field. In 2023, Vierkant et al. researched the past ten years of the activity of re3data.org. They researched that over 3,000 research data repositories had been listed on the site and confirmed that re3data.org was the largest site of its kind. Their research complemented the research done by the authors of Information Discovery and Delivery. The site was a repository of research data repositories from most of the countries in the world. However, the site was most populated by countries that have a large amount of research activity. There has been a rise in country- and region-specific bibliometric studies on re3data.org since its inception, focusing on the repository landscape in particular countries to identify the local strengths and weaknesses in the research data infrastructure. This study is related to a study on the existing state of Indian data repositories listed on re3data. This study found that, of the 51 repositories that were associated with India, 41 repositories (80%) were still functioning, and also found gaps in topic areas, licensing, and the certification of repositories. This is in line with the critique of open-data policies, which argue that while India's research data policies rely on the NDSAP policies, the policies do not take steps to provide support to align the research data with the internationally accepted principles of open data that are the FAIR principles (Findable, Accessible, Interoperable, Reusable).Beyond the data available in the literature on re3data, other related registry services like DataCite have published their own growth studies and stated that, as of 2024, there are over 3,000 data repositories in their registry service and over 500 repositories are actively registering and issuing DOIs. This illustrates the global growth to adopting Persistent Identifiers in research data infrastructure and services and helps to show a baseline to compare the state of India's data infrastructure and services. Considering all of the above, the available literature gives a clear picture of the documented state of research data infrastructure in a cross-national study available on re3data.org, and shows that when country studies are performed, the studies find low levels of certification of repositories, low metadata completeness, and poor alignment with the FAIR principles. This study seeks to determine whether these patterns hold with an updated extraction of records for India from re3data.org conducted in 2024.

3. METHODOLOGY

A quantitative study using a bibliometric registry-mining design is undertaken here. The population of interest is research data repository records in re3data.org that are institutionally or geographically connected to India.

3.1 Data Source and Extraction

Repository metadata were obtained from the re3data.org API, which offers repository records in the re3data.org Schema for Description of Research Data Repositories. An initial keyword-based search of the API for records that mentioned "India" returned 61 candidate repository identifiers. The API's per-repository endpoint was accessed again to obtain the complete metadata of all 61 repositories. XML documents returned by the API were parsed to collect the repository name and web address, start year, type of repository and provider, subject and content type, and the affiliations and locations of the hosting institutions. The parsed XML documents were also used to extract metadata on repository licenses, available APIs, tools used to implement the repository, systems used to manage persistent identifiers, and certification.

3.2 Filtering and Data Cleaning

Out of the 61 candidate repositories that were first listed, 55 were found to have at least one associated repository with metadata that has an ISO country code of IND (India). The other 6 records were removed because they had the keyword "India" in a description of repositories or in content subjects (for instance, repositories that include India-

related datasets that are hosted by US, French, or Kenyan organizations that are not Indian), so there were no Indian institutions in them. The final analytic sample therefore consists of 55 repositories that are linked to India.

3.3 Variables and Analytical Approach

Across five analytical dimensions, descriptive bibliometric techniques were implemented. These dimensions were: (i) annual and cumulative counts of repositories according to their registries' listed start years from 2012 to 2024, representing the temporal growth of the repositories; (ii) classification of repository types according to the re3data scheme, including disciplinary, institutional, governmental, or other repositories; (iii) differentiation between data and service providers; (iv) classification of subject domains according to the four principal DFG subject categories as well as their sub-disciplines; and (v) mapping of the infrastructural and FAIR-related dimensions, such as the types of content that may be expressed, persistent identifiers, and certification. In Python's Pandas, frequency counts and cross-tabulations were used. The repositories that were associated with various types or subject domains were considered once per tag (non-mutually exclusive multiple-response coding). This was done in line with the multi-valued characteristic of the underlying schema fields.

3.4 Limitations of the Method

There are three important methodological limitations that should be brought to attention. To begin with, the re3data keyword search and institutional-country filter, although consistent with similar country-level re3data studies, relies on re3data's self-reported institutional metadata accuracy and completeness. As a result, certain India-based repositories could be missing country metadata and thus be unreported. The second limitation is that the licensing, software, and certification fields returned empty for the entire sample in a machine-readable format. Licensing, software, and certification may not appear in institutional records, or there is a limitation in the way these fields are populated in the schema and exposed via the API. I consider this a significant issue regarding the completeness of the metadata and place a substantial confidence level on this finding and suggest it be examined via a manual review of the registries. The final limitation is that this research is a snapshot of the registries in the year 2024, and does not account for new repositories that have been added, deprecated, or delisted in the time that has elapsed since the snapshot was taken and this reading.

4. RESULTS

4.1 Overview of the India-Linked Repository Population

The completed analysis includes 55 research data repositories associated with at least one Indian institution, based on an initial selection of 61 India-related records in re3data.org. This situates India's repository counts at a small percentage of the registry's global total that surpassed 3,000 research data repositories by 2023indicating the incipient state of formal, registry-verified research data infrastructure in India compared to advanced countries like Germany and the United States, and generally matching the national repository count of other single countries such as Spain (23 repositories as of December 2017).

4.2 Growth Trend, 2012–2024

Among 31 repositories with a documented beginning of registry year, two growth waves were observed. The first cluster, concentrated in 2009–2012, was related to repositories of the genomic, oceanographic, and economic-

statistics data and the following, larger wave, was related to the national implementation of Open Government Data (OGD) portals in the Indian states, and occurred in the years 2014–2017. Focusing on the years 2012–2024, the number of repositories with announced begin-year of the registry increased from 3 in 2012 to 14 in 2019. No other begin-year of the registries was noted in the sample from 2019 to 2024.

Table 1. Repositories with Declared Start Year, by Year (2012–2024)

Year	New Repositories Registered	Cumulative Total
2012	3	3
2013	0	3
2014	2	5
2015	1	6
2016	5	11
2017	2	13
2018	0	13
2019	1	14
2020	0	14
2021	0	14
2022	0	14
2023	0	14
2024	0	14

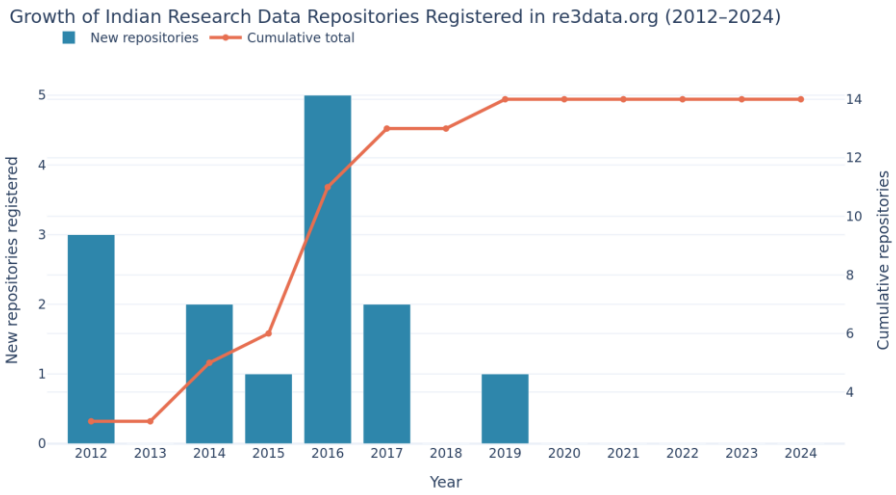

Figure 1. Growth of Indian research data repositories with declared start years registered in re3data.org, 2012–2024.

This trend must be interpreted very cautiously: in the sample, 24 of the 55 repositories (43.6%) do not have an explicit start year in their registry metadata. Hence, the actual growth curve of India-linked repositories is definitely more pronounced than what can be estimated from the declared-year data. The lack of registry-recorded launches after 2019 is more likely due to underdeveloped metadata for the more recently added repositories, rather than representing a wane in the development of the Indian data infrastructure, particularly in light of the new state-level Open Government Data portals and new sector-based repositories (agriculture, judiciary, genomics, etc.), which have been observed and contain blank start date fields.

4.3 Repository Type and Provider Type

Disciplinary repositories, those organized around particular research domains instead of a singular research institution, make up the largest part of the sample with 36 instances; next in line are 20 “other” tagged instances, 15 institutional repositories, and 1 tagged governmental repository. Looking by provider type, out of the 55 surveyed repositories, 52 are classified as data providers (repositories that host and disseminate primary datasets), while 13 are classified as service providers (repositories that provide discovery and/or aggregation and value-added services based on externally hosted datasets). Also, several repositories are tagged with both provider types.

Table 2. Repository Type and Provider Type Distribution

Category	Count of Tagged Repositories
Disciplinary	36
Other	20
Institutional	15
Governmental	1
Data provider	52
Service provider	13

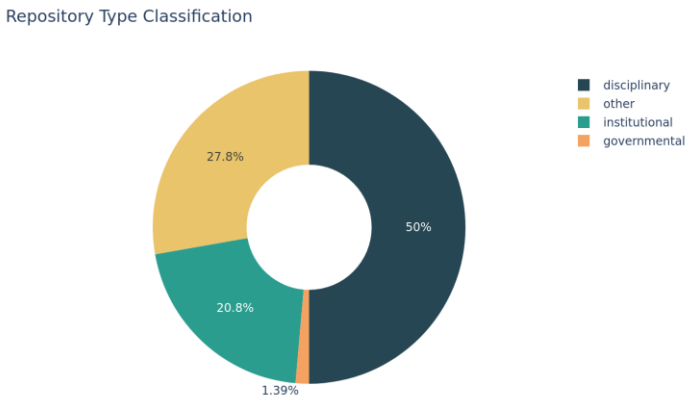

Figure 2. Distribution of repository type classifications among India-linked repositories.

Most disciplinary repositories are examples of research data infrastructure built for specific domains, highly specialized in areas like agriculture, genomics, oceanography, and public health. Many of them are maintained by central state research institutions like the Indian Council of Agricultural Research and state oceanographic research institutes. The very low count of governmental repositories, i.e., only 1 is explicitly indicated, is also notable since a significant number of records in the sample are state-level Open Government Data portals. Most of these are instead categorized under "other" or "institutional" in the registry, indicating a lack of uniformity of schema for government-hosted open data.

4.4 Subject Domain Distribution

Subject classification is based on the DFG (German Research Foundation) subject scheme applied throughout re3data.org. At the highest tier, Humanities and Social Sciences is the most commonly assigned category (28 repository-tag instances), followed by Life Sciences (26), Natural Sciences (23), and Engineering Sciences (8). Many repositories assign tags to numerous top-level categories. This demonstrates the interdisciplinary nature of national data infrastructures relating to statistics, the environment, and genomics.

Table 3. Top-Level Subject Domain Distribution (DFG Scheme)

Top-Level Subject Domain	Number of Repository Tags
1 Humanities and Social Sciences	28
2 Life Sciences	26
3 Natural Sciences	23
4 Engineering Sciences	8

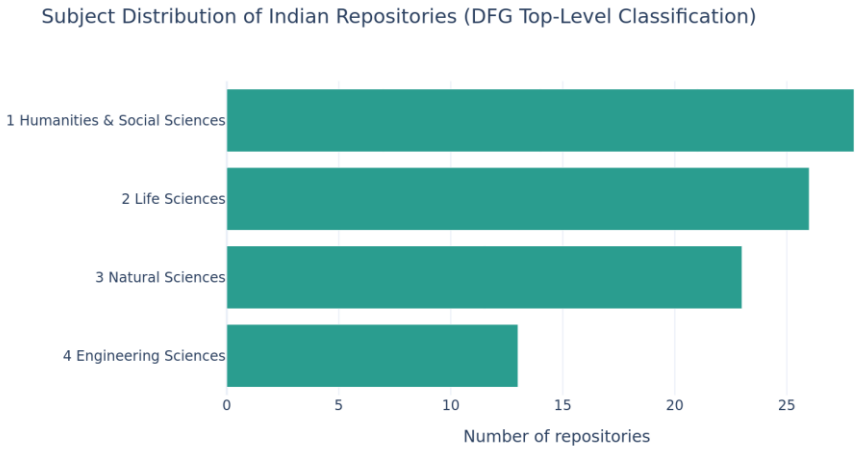

Figure 3. Distribution of India-linked repositories across DFG top-level subject domains.

At the sub-discipline level, the highest tag counts were for Economics (15 tags), Social Sciences (13), Statistics and Econometrics (9), Biology (11), and Medicine (11). Geosciences, including Geography (17 tags), formed the largest sub-discipline as a result of the prominence of remote sensing, satellite, and geospatial data portals such as Bhuvan and Bhoonidhi developed by the Indian Space Research Organisation. This trend is largely in line with the previous results that the Indian repository infrastructure is considerably more focused on government statistical and earth observation data in comparison to biomedical or engineering data.

Table 4. Top 12 Sub-Discipline Subject Tags

Sub-Discipline (DFG Subject Tag)	Number of Repositories
12 Social and Behavioural Sciences	22
34 Geosciences (including Geography)	17
112 Economics	15
111 Social Sciences	13
23 Agriculture, Forestry, Horticulture and Veterinary Medicine	12
21 Biology	11
22 Medicine	11

205 Medicine	9
11205 Statistics and Econometrics	9

4.5 Content Types Supported

Across the sample, the most common content types include standard office documents (38 repositories), scientific and statistical data formats (35), structured graphics (22), raw data (15), and images (15). This suggests that Indian repositories are generally more accommodating of tabular and/or document-style data products, as opposed to scientific and other specialized data formats having computer code and/or data in networked format and having less than three repositories.

Table 5. Content Types Supported by India-Linked Repositories

Content Type	Number of Repositories
Standard office documents	38
Scientific and statistical data formats	35
Structured graphics	22
other	16
Raw data	15
Images	15
Plain text	14
Databases	10
Audiovisual data	7
Archived data	5
Structured text	5
Software applications	2
Source code	2
Network based data	2

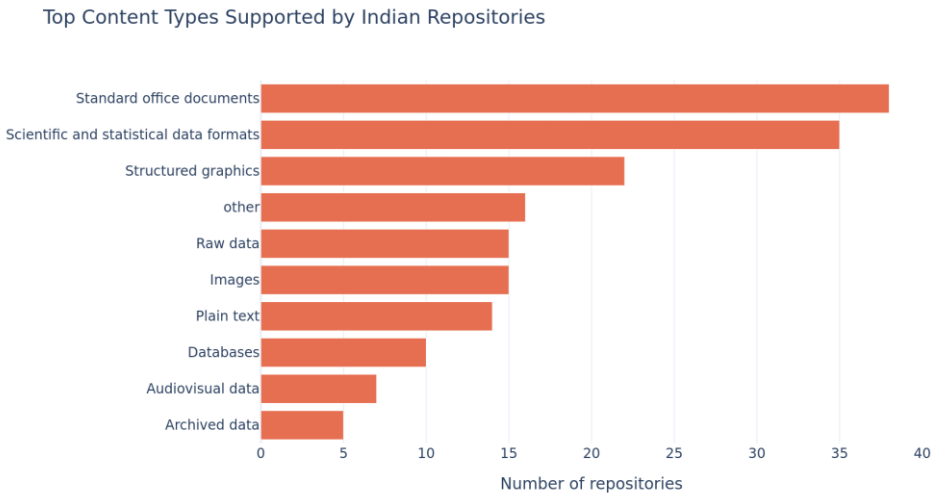

Figure 4. Top content types supported by India-linked repositories in re3data.org.

4.6 Persistent Identifiers, Licensing, and Certification

There is strikingly low adoption of Persistent Identifiers (PIDs). Of the repositories with a PID field, twenty-seven say "none." Five mentions that they assign DOIs. Two mention use of the Handle (HDL) System, while one mentions an unknown PID System. The remaining repositories leave the field blank. In addition, none of the sampled repositories provided a machine-readable database license, a data license, a software license, or a certification value through the API extraction. This may mean that the fields mentioned above are left blank for Indian repository records, or that repositories leave the fields blank and provide licensing and certification in free-text description fields instead of the dedicated fields in the schema designed for automated system harvesting.

Table 6. Persistent Identifier (PID) System Usage

PID System	Number of Repositories
none	27
DOI	5
hdl	2
other	1

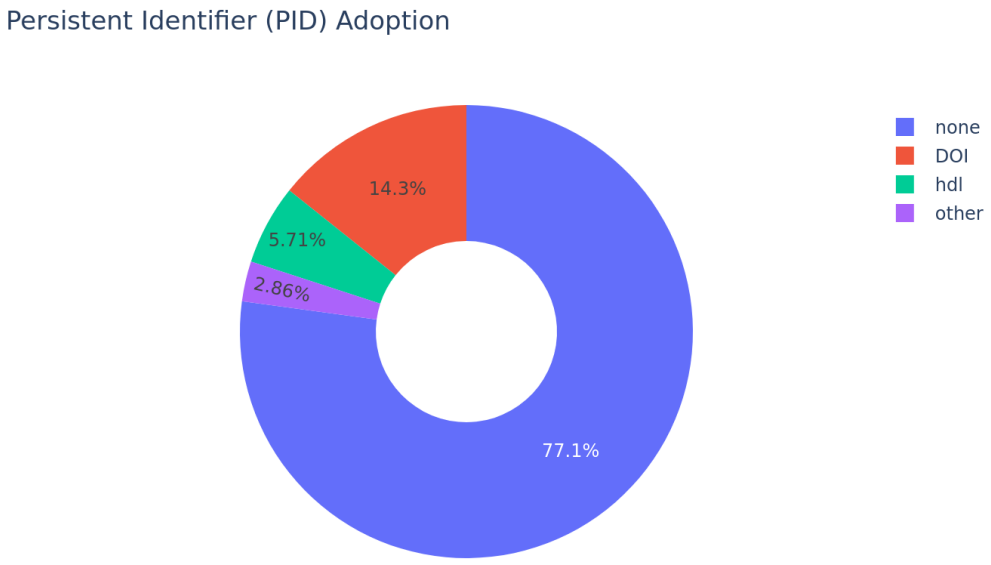

Figure 5. Persistent identifier system adoption among India-linked repositories.

This observation reaffirms that repositories in India on re3data have low structured metadata completeness and licensing transparency. In contrast to the situation presented by DataCite, whose active global repository network has exceeded 3,000 DOI-registered repositories, and with increasing new repositories adopting DOI assignment as a normal practice at registration, Indian repositories on re3data have low structured completeness in metadata and low licensing transparency.

4.7 Institutional Contributors

Entities of the Government of India and central ministries are the dominant institutional contributors of India-related repositories. Among the most frequently tagged responsible institutions are the Government of India (6 direct tags), Ministry of Electronics and Information Technology (4), and the Open Government Data Unit of the National Informatics Centre (4). The Indian Space Research Organisation along with its National Remote Sensing Centre account for 4 tagged cases. This reflects the significance of Earth observation and the infrastructure of the geospatial data in Section 4.4. A few records show international collaboration with the Bill & Melinda Gates Foundation as a jointly responsible institution for 2 repositories in health and agriculture.

Table 7. Leading Institutional Contributors (Top 10)

Institution	Number of Repository Tags
Government of India	6
Ministry of Electronics and Information Technology	4
National Informatics Centre, Open Government Data Division	4
Indian Space Research Organisation	2
National Remote Sensing Centre	2
Government of India, Ministry of Science & Technology, Department of Biotechnology	2
National Institute of Oceanography, Bioinformatics Centre	2
Government of India, Ministry of Science and Technology, Department of Biotechnology	2
Government of India, Ministry of Statistics and Programme Implementation	2
Government of India, National Knowledge Commission	2

5. DISCUSSION

The findings validate and build upon previous observations of India's presence on re3data.org. India's presence is non-trivial but remains small in comparison to the re3data.org global scale of 30,000 global data repositories. Most repositories have a considerable amount of the underlying metadata describing the start dates, licensing, certification, and persistent identifiers that are either absent or default to "none." Three major themes emerge that demand attention.

5.1 A Government-Led, Statistics and Earth Observation Heavy Ecosystem

Examining the prominence of Humanities and Social Sciences and Life Sciences subject tags and the active involvement of central government ministries, the National Informatics Centre’s Open Government Data Division, and the Indian Space Research Organisation as contributors, the public administration, statistics, as well as satellite and geospatial data programs, constitute the primary public data programs shaping India’s registry-indexed research data infrastructure. This is in contrast to repository systems that are documented in the original 2015 registry-wide landscape study, where repository systems tend to be built by universities and laboratories. Thus, the research data

repository systems in India are largely shaped by public administration, statistics, and geospatial data, and are unlike the repository systems in the leading countries of the world.

5.2 Low FAIR-Readiness Signals

The scant availability of organized licensing data, the absence of certified repositories, and the observation that only five out of fifty-five repositories issue DOIs cumulatively imply that most repositories in the registry that are linked to India do not meet the Findable, Accessible, Interoperable, and Reusable (FAIR) criteria. The formal requirements for data-sharing infrastructures recognize the importance of reliable and certified data. In the case of Indian repositories, the lack of persistent identifiers and certified data means there is much room for improvement which can be realized in a short time frame compared to other repository infrastructures.

5.3 Metadata Completeness as a Barrier to Accurate Measurement

Almost half of the sample (24 of 55 repositories) lacked a machine-readable start date. The lack of newly recorded launches post-2019 is more likely a metadata duration artifact than a real slowing of the pace of India's data infrastructure. The policy momentum of the state Open Government Data Portals and the targeted scientific repositories (which are present in the sample) also support this. This adds to the themes in the existing re3data work. The first challenge in using the registry as a source of bibliometric data is the completeness of the self-reported metadata of the repositories. This was mentioned as a challenge in the first landscape assessments of the registry and can still be seen in the India-focused reviews almost a decade later.

LIMITATIONS

Limitations of this study must be identified and considered in relation to the findings. First, the analysis refers to a single snapshot of re3data.org, meaning repositories that were registered/delisted after the snapshot, along with modifications to metadata, are not captured. Also, repositories are likely to be missed, or inaccurately identified, where there are incomplete or missing affiliations to Indian institutions. Additionally, if a repository had a partial, or historically affiliated, Indian institution, and its primary operational base has since relocated, it can be assumed that the repository will most likely be identified. Second, if a repository has no declared start year, then growth-trend analysis shown in Section 4.2, will be limited to only 31 of 55 repositories. Of the 31 repositories, 43.6% no declared start year. For these repositories, the growth-trend analysis of 2020 to 2024, will be especially limited. Lastly, as stated in previous work, no values were found in the licensing, software platform, and certification fields for any of the sampled repositories. Through the API, this is a significant finding. However, it is possible that some of this information is available in free-text descriptions of the repositories, or on the repositories themselves, but remains uncaptured in the structured fields.

CONCLUSION

Between 2012 and 2024, this bibliometric study focuses on 55 India-linked research data repositories in re3data.org. These repositories illustrate strong government and Earth observation interests, a dominance of disciplinary repositories, a focus on Humanities/Social Sciences and Life Sciences, and a lack of persistence in structured metadata for licensing, certification, and persistent identifiers. Post 2012, and notably after the 2014–2017 Open

Government Data launch, India's registry-indexed data infrastructure has expanded significantly. However, the analysis shows major gaps in metadata that, among other things, affect the bibliometric tracking and FAIR-readiness of the repositories. The study significantly updates and supports earlier studies of Indian repositories in re3data. The study aims to highlight realistic and feasible options for the Indian open research data community, especially the improvement of persistent identifiers and metadata, and offers an opportunity for substantial improvement in the integration of India's research data infrastructure in the global open research data system mapped by re3data.org.

REFERENCES

[1] Archana, S. N., & Padmakumar, P. K. (2023). The status quo of Indian data repositories indexed in re3data registry. Digital Library Perspectives, 39(4), 496–515. https://doi.org/10.1108/DLP-02-2023-0017

[2] DataCite. (2024, January 21). DataCite's thriving community: 3000 repositories and counting [Blog post]. https://datacite.org/blog/datacites-thriving-community-3000-repositories-and-counting/

[3] GFZ German Research Centre for Geosciences. (n.d.). re3data – Registry of Research Data Repositories (DFG-funded project description). https://www.gfz.de/en/section/data-and-information-management/projects/re3data-registry-of-research-data-repositories-dfg

[4] Khan, A. M., Loan, F. A., Parray, U. Y., & Rashid, S. (2023). Global overview of research data repositories: An analysis based on re3data registry. Information Discovery and Delivery, 52(1), 53–65. https://doi.org/10.1108/IDD-08-2022-0083

[5] Kindling, M., Pampel, H., van de Sandt, S., Rücknagel, J., Vierkant, P., Kloska, G., Witt, M., Schirmbacher, P., Bertelmann, R., &Scholze, F. (2017). The landscape of research data repositories in 2015: A re3data analysis. D-Lib Magazine, 23(3/4). https://www.dlib.org/dlib/march17/kindling/03kindling.html

[6] Koley, M., Kaur, P. J., & Khan, F. (2022). Open research data for India: Exploring infrastructure and governance [Policy brief]. DST Centre for Policy Research, Indian Institute of Science. https://dstcpriisc.org/iisc/wp-content/uploads/2022/06/Open-Research-Data_Policy-Brief.pdf

[7] Pampel, H., &Weisweiler, N. L. (2023). re3data – Indexing the global research data repository landscape since 2012. Scientific Data, 10, 587. https://doi.org/10.1038/s41597-023-02462-y

[8] Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P., & Dierolf, U. (2013). Making research data repositories visible: The re3data.org registry. PLOS ONE, 8(11), e78080. https://doi.org/10.1371/journal.pone.0078080

[9] re3data.org. (n.d.-a). About [Registry mission and scope]. https://www.re3data.org/about

[10] re3data.org. (n.d.-b). Browse by subject [DFG subject classification scheme]. https://www.re3data.org/browse/by-subject/

[11] University of Edinburgh Digital Research Services. (n.d.). Registry of Research Data Repositories – re3data.org. https://digitalresearchservices.ed.ac.uk/resources/re3data-org

[12] Vierkant, P., Weisweiler, N. L., Pampel, H., et al. (2023). re3data – Indexing the global research data repository landscape since 2012. Scientific Data, 10, 571. https://doi.org/10.1038/s41597-023-02465-9
