# Exploring Metadata Quality for Scientific Data in Indian Research Data Repositories: A Survey

**Sanghamitra Dalbehera**

Librarian, Siskha 'O' Anusandhan University, Bhubaneswar, Odisha

**Abstract**

*The study aims to trace the development of Indian Research Data repositories (RDRs) indexed in Registry of Research Data Repositories(re3data.org) with their types, subject coverage, software tools, standards and specification used for implementation. The study strives to achieve the following objectives such as: to analyze different aspects of Indian Research Data Repositories(RDRs), to identify data licenses, data upload and access restriction policies and to ascertain quality of scientific metadata being used in RDRs in India. The result of the survey is presented by examining the collected data from the libraries of Research Data Repositories. The result of the study will help to find the effective and qualitative research data in various discipline in Indian subcontinent.*

**Keywords:** India, Metadata, Quality Assessment, Re3data.org, Research Data Repositories, Scientific data

## 1. Introduction

Due to rapid advances in digital technology and emergence of scholarly publishing, many changes have been seen in managing scientific data in data repositories across the world. Scientific data repositories are emphasizing on a unique platform to serve to deposit, share and access research data for the scholarly communities in India. Being heterogeneous in nature across disciplines scientific data needs a workflow to be implemented in a FAIR modelling of research process in order to be accessible among the scientific team. Under the FAIR Data project re3data.org is an eminent platform that provides information about research data repositories around the world for the researchers, publishers, libraries and funding organizations. Therefore, it is necessary to create Research Data Repositories (RDRs) which collect, upload and retrieve research data in India. Today, Indian research data repositories are the primary source to store scholarly resources in a variety of digital media and depending on the parent institution type, discipline, specialization and access policies. It is necessary to integrate scientific datasets into repository collections to provide access to faculty members, research scholars and students. In this context, metadata schema provides the metadata properties which helps in recommending a standard for describing the Research Data Repositories by providing the basis for interoperability between re3data.org and RDRs, which intends to move towards shared standards and practices. Metadata sustain core functionalities of research data repositories as they use different metadata elements to describe data. Therefore, high metadata quality is needed to perform the core bibliographic functions of discovery, use, provenance, currency, authentication and administration.

Corresponding Author: Dr. Sanghamitra Dalbehera, Email:sang2016nayak@gmail.com

For assessing metadata quality a list of core criteria are used either to individual metadata elements or to entire metadata collections. In this paper, a case study was conducted on Research Data Repositories indexed in re3data.org registry developed and maintained by Indian research and academic institution.

## 2. Review of Literature

Witt(2012) explored a number of academic and research libraries which are taking a more active role in data management to find data and integrate into their learning, teaching and research by adapting library practices to describe research datasets, to develop data collections, data literacy and data repositories through assisting researchers funder-required data plans .According to Karcher. Kirilova and Weber(2017),"Research data repositories are the main infrastructure for depositing, sharing and reusing of research data". Antonio et.al.(2020) analyzed that qualitative research are conducted and shared among multi-institutional and geographically separated researchers through secure data management policy of research data repositories. There is a significant differences related to data management practices, attitudes and interest in support services among the faculty members in different research domain; (Akers & Doty, 2013). A study conducted by Kim and Yoon(2017) reported that data reuse is influenced by availability of data repositories at the disciplinary level. A similar study by Faniel and Yakel(2017) found that data processing, metadata availability, trust in repositories and data selection play an important role in reusing data. But the possibilities of reusing data in new contexts might loose important information about the data when it is moved from one context to another (Borgman,2015; Leonelli,2015; Loukissas,2019). In the context of scholarly and scientific research, sharing of data is an essential component which improves the result of research data analysis and generate new ideas (Parr and Cummings 2005). Research data should be organized and stored in a structured way so that developers, policy makers and users can access metadata automatically and seamlessly and take correct decisions in the data repository lifecycle(Grunzke, R. et.al 2019). A case study was conducted on re3data.org research data repositories by Pampel et.al.(2013).He identified the differences between the four repository types e.g. institutional, disciplinary, multidisciplinary and project-specific and the features of re3data.org project which provides research data to the appropriate user in need. According to Mayernik(2015), "Research data repositories are important actors in metadata management which are taking an active role in research data management". The study revealed that analysis of metadata elements related to data sharing, format, availability and coverage in 5 repositories stated heterogeneity in the number of supported metadata elements, the obligation levels and also in the use of controlled vocabularies.

A study of metadata of different research data repositories by Kindling et.al.(2017) reported that the nature of repositories are heterogeneous depending on the parent institution type, discipline, specialization, access policies ,data sharing and metadata quality assessment. The metadata provides the standard properties about interoperability between research data repositories and re3data.org which helps data repositories move towards shared standards and practices(Rucknagel et.al.2015).Metadata quality is defined as "functional requirements" related to the purpose of bibliographic control in facilitating discovery, identification, selection and use of research data(Guy, Powell and Day 2004). According to Wieczoreketal et.al(2012).” The DataCite metadata Schema which is generic is used for retrieval and citation for a large set of heterogeneous datasets, whereas a discipline specific metadata schema such as DarwinCore is used in more detailed descriptions of datasets”. While evaluating metadata quality by using core dimensions such as completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness and accessibility, many metrics are used for the conformity to a set of requirements (Park and Lu 2008). He analyzed that completeness, accuracy and consistency are the most commonly used dimensions for measuring metadata quality which reflects the functional property of metadata.

## 3. Conceptual Framework of Metadata Quality

From the literature study, it reveals that the metadata practices in scientific repositories are heterogeneous as there are significant differences in metadata requirements such as the use of controlled vocabularies, number of schema used and the obligation levels of metadata elements within and across disciplines. This variety leads to difficulty in consistent use of metadata and entry of metadata values which causes data integration between several institutions. For example ,there are different software used in Indian RDRs. some repositories use Dublin core element to describe the dimension of an item while other repositories use MODS (Metadata Object Description Schema) for their descriptive metadata. For this reason, the functional perspective of metadata quality is lost due to different metadata schema and different conformance standards used in RDRs. So, it is necessary to assess and evaluate metadata quality by using some core criteria either to the individual element or to entire metadata collections. In this study. a conceptual framework is described for metadata quality assessment in research data repositories in India. The framework consists of four sections and the corresponding core criteria /dimensions of each section which are given below:

❖ **General section** which includes the types, formats and granularity of the metadata to assess provision. Under this section four metadata core criteria are available such as completeness, comprehensiveness, appropriateness and accessibility. Completeness consists of the use of individual metadata elements are described completely i.e. the number of metadata elements used in a metadata record in relation to the number of a metadata elements available. The individual metadata elements used here indicates how frequency a metadata element is used in the sample of metadata records. Comprehensiveness of metadata description deals with the use of element description i.e. the number and combined character length of descriptions in a metadata record. Accessibility consists of the metadata used can be easily accessed without any difficulties.

❖ **Tools and technique** which deals with the structure, application of semantic web technologies, indexing and use of terminologies to assess the metadata. In this section the main criteria are accuracy, discoverability, interoperability, extendibility etc. Accuracy refers to the metadata elements used in different scientific databases and resource content should be described accurately.

❖ **Usability** section refers to the presence in repositories, application of semantic mappings, metadata standards and cross-walks provision. It consists of Conformance to expectation i.e. the metadata is described in such a way to meet the expectations of the user. Another criteria is Logical consistency and coherence which means the metadata elements are consistent with standard definitions and description should be coherent across collection.

❖ **Management and Curation** section deals with two parts. First is the creation and version of the metadata being used and second the creation and version information used itself i.e. meta-metadata of the quality assessment itself .The main criteria in this section are timeliness, versionability and meta-metadata.

**Table 1: mentioned the metadata quality criteria used in this study**

| **Metadata quality criteria** | **Description** |
|---|---|
| **Completeness** | Use of individual metadata elements are described completely i.e. the number of metadata elements used in a metadata record in relation to the number of a metadata elements available. The individual metadata elements used here indicates how frequently a metadata element is used in the sample of metadata records |
| **Accessibility** | Extent to which metadata can be easily accessed without any difficulties. |
| **Comprehensiveness** | Use of element description i.e. the number and combined character length of descriptions in a metadata record. |
| **Appropriateness** | Metadata and data documentation to appropriately describe data |
| **Accuracy** | Metadata elements are described correctly. |
| **Discoverability** | How the metadata are easily found. |
| **Conformance to expectation** | Metadata is described in such a way to meet the expectations of the user |
| **Logical consistency and coherence** | Metadata elements are homogeneous and constant. They are consistent with standard definitions and description should be coherent across collection. |
| **Open data licence** | Data are assigned with an open licence |
| **Reuse potential** | The dataset is analyzed by others in future. |
| **Interoperability** | Extent to which metadata can be exchanged and used without any problem |
| **Timeliness** | Metadata is current having temporal information. |
| **Versionability** | Extent to which a new version may be easily created. |
| **Meta-metadata** | Metadata about the metadata. |

In order to describe the core elements of metadata quality framework for scientific research data, this study focuses on the use of a generic metadata schema for describing diverse research data. The DataCite metadata schema is one of the most comprehensive sources for metadata on research data. It allows uniform statements about heterogeneous research data which is designed to recommend a standard for describing RDRs; provide interoperability between RDRs and re3data.org and included a list of metadata properties for consistent identification of data to cite and retrieve purposes. Table-2 gives the DataCite metadata properties Version 4.3.

**Table 2: DataCite metadata elements**

| **Element name** | **Obligation level** | **Mandatory type** |
|---|---|---|
| Identifier | Mandatory | Descriptive |
| Creator | Mandatory | Descriptive |
| Title | Mandatory | Descriptive |
| Publisher | Mandatory | Descriptive |
| Publication year | Mandatory | Descriptive |
| Resource type | Mandatory | Technical |
| Subject | Recommended | Descriptive |
| Contributor/s | Recommended | Descriptive |
| Related identifier | Recommended | Structural |
| Date | Recommended | Descriptive |
| Description | Recommended | Descriptive |
| Geolocation | Recommended | Descriptive |
| Language optional | Optional | Descriptive |
| Alternate identifier | Optional | Structural |
| Size | Optional | Technical |
| Format | Optional | Technical |
| Version | Optional | Structural |
| Rights | Optional | Rights |
| Funding references | Optional | Descriptive |

## 4. Objectives of the Study

The study includes the following objectives:-

1. To explore research data repositories in India indexed in re3data.org.
2. To analyze the yearwise growth of RDRs.
3. To examine various aspects of data access policies, persistent identifiers and application programming interfaces used in Indian RDRs.
4. To know the metadata standards and schemas used in RDRs.
5. To identify different criteria of metadata quality for assessment.

## 5. Methodology

In this paper, a case study was conducted on the research data repositories in India indexed in Registry of Research Data Repositories (re3data.org).The list of RDRs registered in re3data.org was downloaded from the website which is given in Appendix-1. A Questionnaire wis sent to 350 respondents from March to April 2023 to collect data about RDRs such as type, subject coverage, software used, persistent identifiers, API and dimensions of metadata quality assessment. The questionnaire received from 45 nos. of research data repositories all over India was 286 i.e. 81% . The data collected was analyzed and the result of the analysis is as follows.

## 6. Results of Study

The data analysis of the collected data reveals the following results:

### a) Yearwise growth of Research data repositories in India

There are total 51 nos. of research data repositories registered in re3data.org in India. Figure.1 shows the growth of RDRs from 2008 to 2023 which indicates that there is a large increase in the number of RDRs in the year 2015. The growth remains stagnant until 2021.

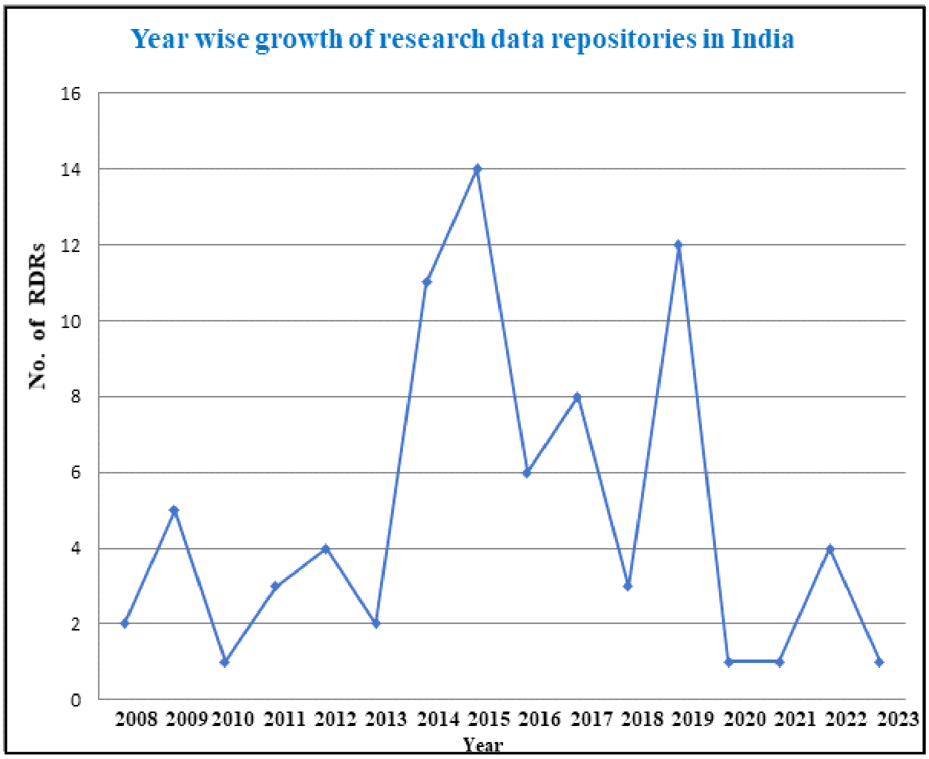
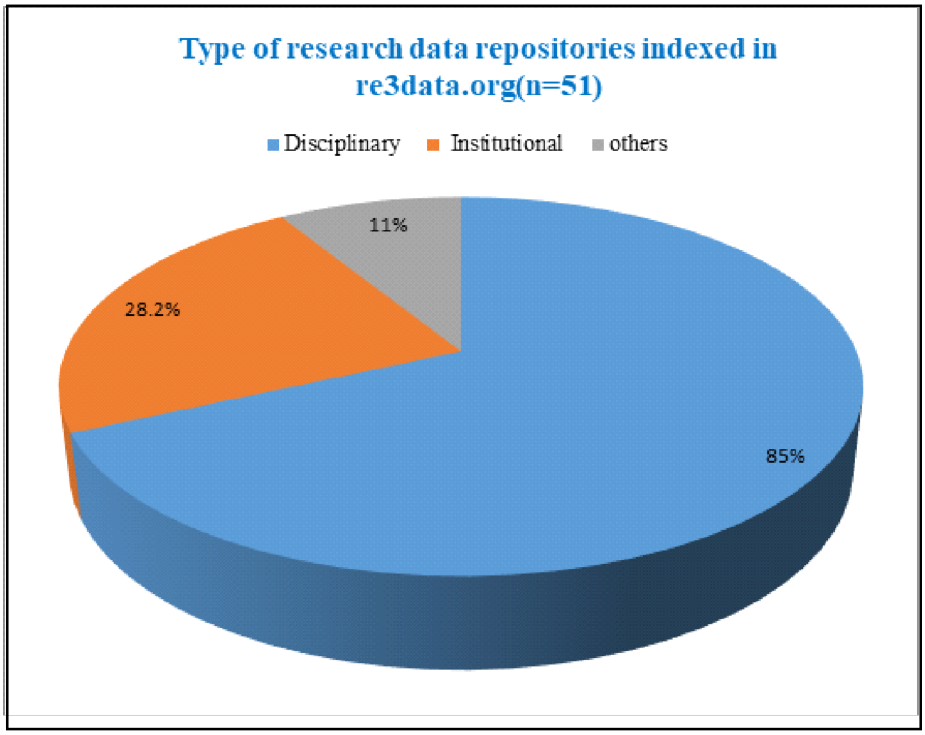


**Figure 1: Year wise growth of research data repositories in India**

### b) Types of Research Data Repositories in India

There are three main types of research data repositories in India which is depicted in figure.2. Majority(85%) type of RDRs are disciplinary, followed by institutional(28.2%) and other(11%) criteria.



.

**Figure 2: Types of Research Data Repositories indexed in re3data.org**

### c) Research Data Repositories according to subjectwise coverage in India

From the analysis of data on the basis of subjects, it was found that ‘Engineering Science’ holds majority percentage i.e. 51% among research data repositories in India. Next highest subject ‘Life Science’ consists of 45% followed by ‘Agriculture, Forestry, Horticulture and Veterinary Sciences’ which hold 41%. The fourth highest(38%) subject coverage comes under ’Health and Medicine’. Figure.3 presents the distribution of RDRs according to subject coverage in India.

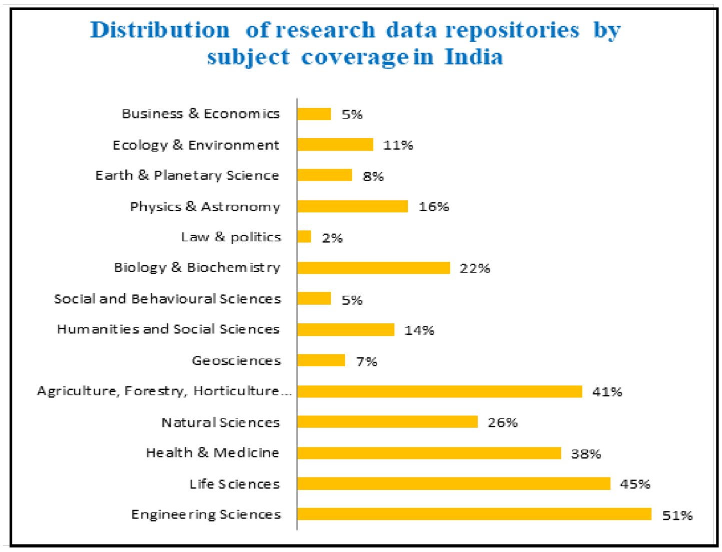


**Figure 3: Distribution of Research Data Repositories by Subject coverage in India**

### d) Research data repositories with Persistent Identifiers

There are different Persistent Identifiers(PID) assigned in RDRs of India to identify, retrieve and access data. Figure.4 demonstrates the use of PIDs which reveals that the most commonly used persistent identifiers are Digital Object Identifiers(DOI) with 21% followed by Handles(11%).The next highest PID is Uniform Resource Name(URN) 2.2% followed by Archival Resource Key(ARK) and Persistent Uniform Resource Locator(PURL) which consists of 1.75% and 1.4% respectively.

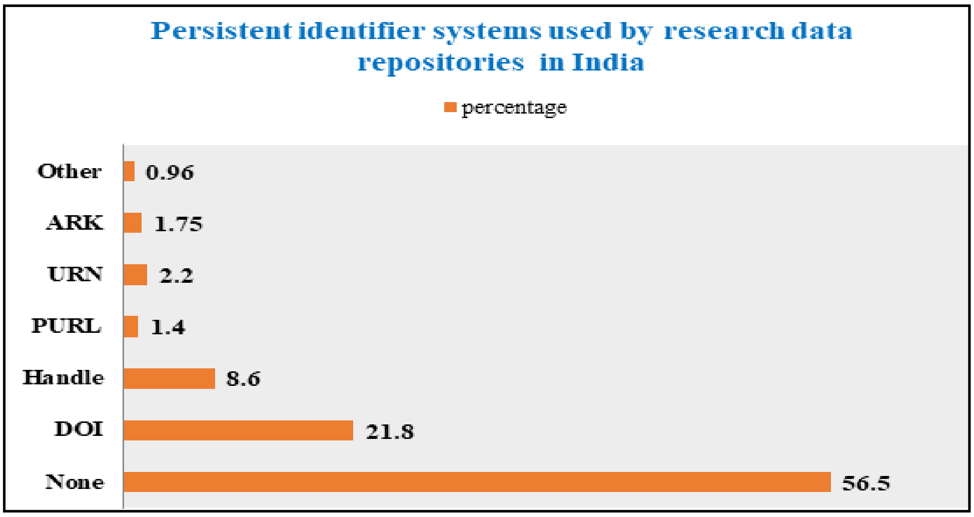


**Figure 4: Research Data repositories with Persistent Identifiers**

### e) Software used in Research Data Repositories

Figure 5 illustrates the repository software used in the RDRs of India. The result shows that DSpace(9.8%) is the most common used software followed by DataVerse(8.54%), EPrints(7.8%), Digital Commons(7.41%) and Fedora(5.23%).

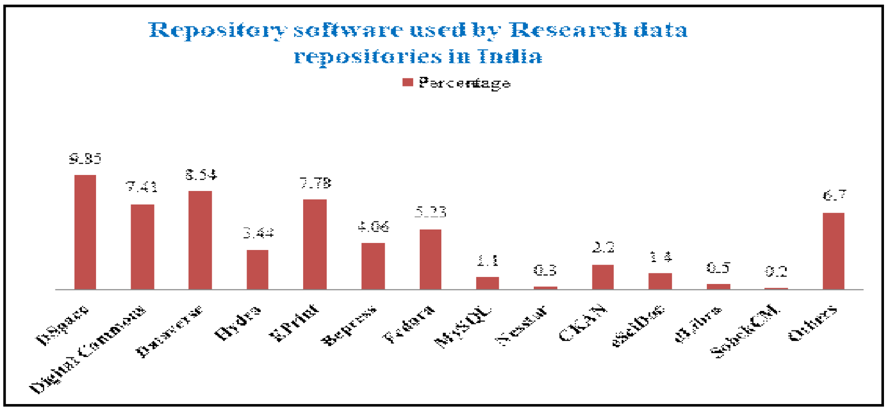


**Figure 5: Repository software used by Research data repositories in India**

### f) Research Data Repository distribution according to API

There are many Application Programming Interface(API) created by data service providers. The analysis of data indicates that REST(Representation State Transfer) is used highest 35(29.3%) among the Research Data Repositories in India followed by OAI-PMH(Open Archives Initiative Protocol for Metadata Harvesting) 23(25.6%) and FTP(File Transfer Protocol)18(21.5%). Other types of API used in RDRDs are NetCDF(Network Common Data Form)5.32%, SWORD(Simple Web-service Offering Repository Deposit)3.87% , SOAP(Simple Object Access Protocol)3.81%, OpenDAP(Open source Project for a Network Data Access Protocol)4.23% and SPARQL(SPARQL Protocol and RDF Query Language)2.65%. Figure.6 illustrates the use of API in research data repositories of India.

.

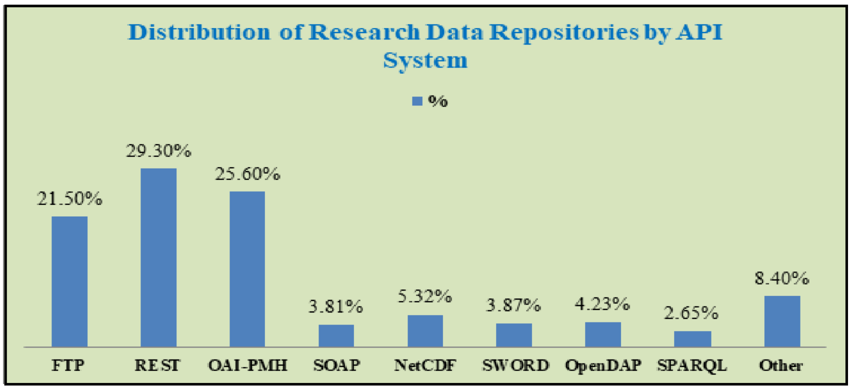


**Figure 6: Distribution of RDRs by API in India**

### g) Metadata Standards used in Research data repositories

In this study, it was found that the most common metadata standard used in Indian RDRs is Dublin Core(256), followed by Data Documentation Initiative(DDI)(181) and DataCite Metadata Schema(104). The other metadata standards e.g. ISO19115(161), Federal Geographic Data Committee Content Standards for Digital Geographic Metadata- FGDC/CSDGM(87), Directory Interchange Format-DIF(41), Climate and Forecast-CF(43), Ecological Metadata Language-EML(35) are used. Figure 7 demonstrate the metadata standards used in RDRs of India.

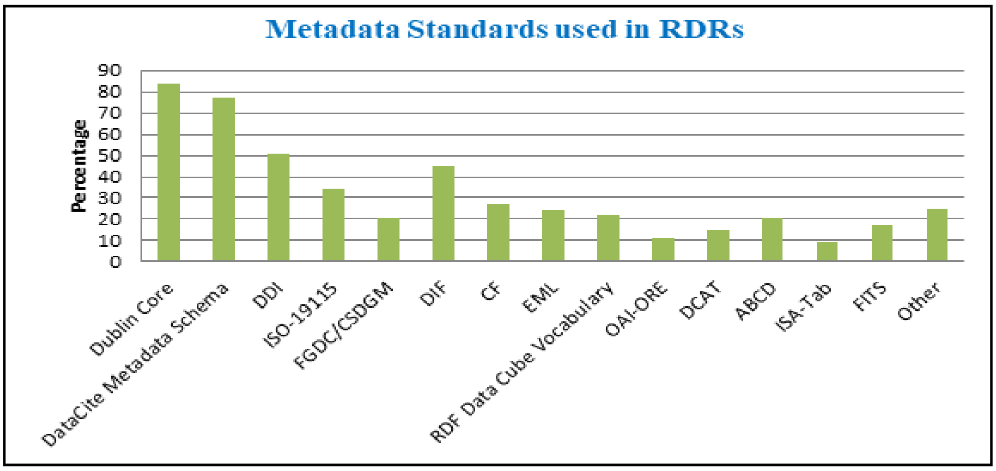


**Figure.7: Metadata Standards used in Research Data Repositories of India**

### h) Metadata Quality Assessment

The respondents were asked to rank the ten aspects of metadata quality criteria. Based on their response, the most important dimension is accuracy(65.8%) followed by accessibility(53.5%), comprehensiveness(44.2%) and discoverability(32.6%). The least important quality dimension is extendibility(11.2%)

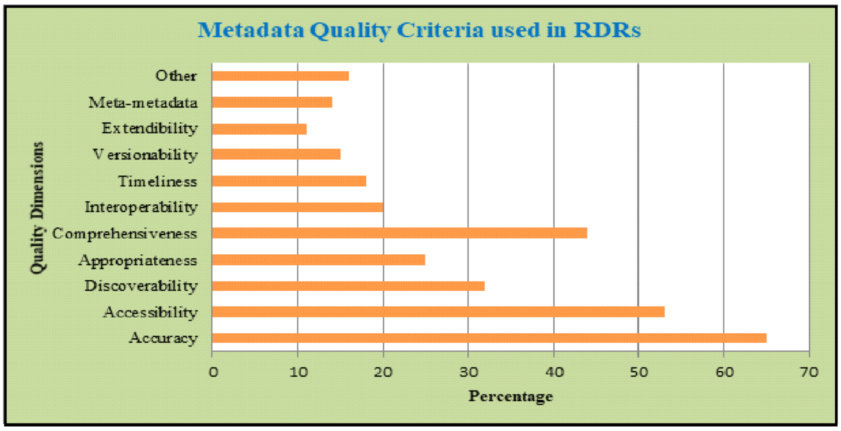


**Figure 8: Dimensions of metadata quality aspects in RDRs of India**

## 7. Conclusion

The re3data.org covers research data repositories from all disciplines. Its goal is to promote access, data sharing and better visibility of scientific research data. The RDRs help research scholars to find scholarly institutions, publishers and funding agencies for their research need, This study analyzed Indian RDRs on the basis of subject, software, persistent identifiers used as well as metadata standards and quality criteria used for assessment. The result of the analysis shows statistically significant differences in the use of metadata elements, the comprehensiveness and completeness of metadata quality across RDRs of different types and certification status. The study discusses the difficulties in using generic metadata schema for describing diverse research data. Some repositories implement successful metadata practices and workflows but some metadata elements remain unused. Further investigation of metadata quality is required to identify factors behind inaccurate, inconsistent and incomplete metadata creation. Moreover, metadata quality evaluation technique should be incorporated as a platform independent method of assessing metadata quality in the Research data repositories.

## References

Assante, M., Candela, L., Castelli, D., & Tani, A. (2016). Are scientific data repositories coping with research data publishing? Data Science Journal, 15(6), 1–24. https://doi.org/10.5334/dsj-2016-006

Antonio, M. G., Schick-Makaroff, K., Doiron, J. M., Sheilds, L., White, L., & Molzahn, A. (2020). Qualitative data management and analysis within a data repository. Western Journal of Nursing Research, 42(8), 640-648. https://journals.sagepub.com/doi/abs/10.1177/0193945919881706

Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. International Journal of Digital Curation, 8(2), 5-26. http://ijdc.net/index.php/ijdc/article/view/8.2.5

Grunzke, R. et.al.(2019).The MASI repository service- Comprehensive, metadata-driven and multi-community research data management. Future Generation Computer System, 94,879-94.

Bruce, T. and Hillmann, D.(2004).The Continuum of Metadata Quality: Defining, Expressing, Exploiting in Metadata in Practice. American Library Association:Chicago.p.238-256.

Park, J. and Caimei Lu.(2008).Metadat Professionals :Roles and Competencies as Reflected in Job Announcement,2003-2006.Cataloging & Classification Quarterly,47(2):145-160.

Guy M., Powell, A and Day, M.(2004).Improving the Quality of Metadata in E-Print Archives,Ariadne,38. http://www.ariadne.ac.uk/issue38/guy/

Kim, Y., & Yoon, A. (2017), Scientists' data reuse behaviors: A multilevel analysis, Journal of the Association for Information Science and Technology, 68(12), pp. 2709-2719.

Faniel, I. M., & Yakel, E. (2017), Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation, Curating research data, volume one: Practical strategies for your digital repository, 1, pp.103-126.

Karcher, S., Kirilova, D. & Weber, N.(2017).Beyond the matrix: Repository services for qualitative data, IFLA Journal, 42(4). DOI:http://doi.org/10.1177/0340035216672870

Kindling, M.; Pampel, H.; Sandt, S. van de, et al. (2017). The landscape of research data repositories in 2015: a re3data analysis. In D-Lib Magazine 23(3). DOI: 10.1045/march2017-kindling.

Parr, C. S., and Cummings, M. (2005). Data sharing in ecology and evolution. Trends in Ecology & Evolution20(7), 362-363. http://www.sciencedirect.com/science/article/pii/S0169534705001308

Pampel, H., Paul, V., Frank, S., Roland, B., Maxi, K., Jens, K., Hans-Jürgen, G., Jens, G., Peter, S. & Uwe, D. (2013). Making research data repositories visible: the re3data.org registry. PloS one, 8(11), e78080. https://doi.org/10.1371/journal.pone.0078080

Mayernik, Matthew S.(2015). Research data and metadata curation as institutional issues. In: Journal of the Association for Information Science and Technology 67.4, pp.973–993. doi:10.1002/asi.23425.

Wieczorek, Johnetal. (2012).Darwin Core : An Evolving Community- Developed Biodiver-sity Data Standard. In: PLOS ONE 7.1, e29715.doi:10.1371/journal.pone.0029715.

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge infrastructures in science: Data, diversity, and digital libraries. International Journal on Digital Libraries, 16(3–4), 207–227.

Loukissas, Y. A. . All data are local: Thinking critically in a data-driven society. The MIT Press, 2019.

Leonelli, S. (2015). What counts as scientific data? A relational framework. Philosophy of Science, 82(5), 810–821.

DataCite Metadata Schema (n.d.). DataCite. https://schema.datacite.org/ visited on October 10, 2022.

Witt, M.(2012).Co-designing, co-developing and co-implementing an institutional data repository service, Journal of Library Administaration,52(2),172-188. http://doi.org/10.1080/01930826.2012.655607

Rucknagel,J, et.al.(2015). Metadata Schema for the description of Research Data Repositories: version 3.0,p.29.DOI:http://doi.org/10.2312/re3.008.

Si, L., Xing, W., Zhuang, X., Hua, X. & Zhou, L.(2015).Investigation and analysis of research data services in university libraries, Electronic Libraray,33(3),417-449. doi:10.1108/EL-07-2013-0130