Research Data Management through Research Data Repositories in the field of Computer Sciences

Prerna Prashar and Harish Chander

The present study described the research data repositories landscape all over the globe briefly. The main objective of the paper is to describe and analyse the research data repositories indexed in re3data.org in the subject of Computer Science. About 98 repositories were indexed in re3data.org. These repositories were analysed for country, software, language of the repository, content type, database access, data access, data upload, data license, policies and metadata standards followed by these repositories. The result will highlight the position of computer science repositories in the registry and further provide insights to the repository developer for further research in this area.

Introduction

Due to emerging technology like Big data, Predictive analysis, there emerged a data deluge like situation. Enormous amount of data is generated through research activities in research laboratories and Universities globally. This resulted in emergence of new jobs mainly dedicated for data managing, organising and analysis. Academia and libraries are not remained untouched from such developments. The main role of the libraries is the organisation of information, management, access and long term preservation. Research is an intensive part of the research bodies and it is the main responsibility of the research labs and Universities to produce new research. Research thus produced huge volumes of data. Role of libraries come here to manage data of different variety and in different volumes. Such research intensive and data driven environment created new job profiles for librarians like data scientist, data curator, data visualization expert, data manager and data archivist with further enhancement in their skills (Uzwyshyn, 2016). Many Research Organisations in International or National sphere which funded the research of scholars urged them to share their research results in public domain. The Open Access (OA) movement supported by three B's namely Budapest, Bethesda Statement of open Access and Berlin declaration also advocated for the free access of information. The formation of Directory of Open Access Journals, Directory of Open Repositories and Directory of Open Access of Books are the outcomes of the open access movement (Suber, 2012). The main objective of the Open access movement and open data initiatives was to rise above the legal, technical and organizational hurdles in accessing the data (Kim, 2018). This movement further triggered the formation of institutional repositories where researchers deposited their research content for wider visibility and citations. Further, the concept of Research Data Repositories (RDR) emerged. RDR are the repositories which host the datasets of the research outcomes to the wider audience. Research Data Repositories as defined by Uzwyshyn (2016) "Online research data repositories are large database infrastructures set up to manage, share, access, and archive researchers'datsets. Repositories may be specialized and relegated to aggregating disciplinary data or more general, collecting over larger knowledge areas, such as the sciences or social sciences. Online repositories may also aggregate experts' data globally or locally, collecting a

university or consortium of universities researcher's data for mutual benefit". The datasets in these repositories helped the researcher in their research and avoid duplication of research.

2. Literature Review

The research papers of Arora and Chakravarty (2021) and Gohain (2021) dwelled upon the mapping of the entire Registry of Research data repository. Their analysis represented that country wise contribution, subject, content, metadata standards, repository language, repository software, repository type, quality management, persistent identifier and license type. Bhardwaj (2019) and TR et.al (2018) described the landscape of Indian research data repositories indexed in re3data.org. Both of the paper highlighted the status of Indian repositories used content analysis method. Pal and Singh (2019) advocated for the formation of Indian Research Data Repository and urged the higher education bodies of India to frame guidelines for the data management of research data. TA (2018) described the importance of research data repositories for libraries and also explained the librarian's view about the development of data repositories. Austin et.al (2015) surveyed the 32 online data platforms to know their hardware and software requirements. Maximum of the data repositories platform were multidisciplinary in nature. 69% of the repositories were used metadata schema or locally created metadata for description of datasets. Kim (2018) discussed about the functional requirements for data repositories Author used Data Curation Profiles available from Purdue University, USA for this study. 13 categories and 75 functional requirements were emerged after the analysis. The 13 categories were includes metadata, identifiers, authentication and authorization, data access, policy support, publication, submission ingestion management, data organization, location, integration/networking, preservation and sustainability, user interface, data and product quality. This categorization will help the future data repository developer in designing the data repository. Cho (2019) described the 152 Asian data repositories indexed in re3data.org on different parameters. Author arranged the countries in four groups using cluster analysis based upon their operation. Pathfinder Network (PFNET) was performed on extraction of keywords. Similarly **Misgar, Bhat and Wani (2020)** analysed the research data repositories of BRICS countries. Result of the study inferred that among the BRICS nations India has the highest and South Africa has least number of data repositories.

3. Data Repositories Directories

Data repository directories represent the compilation of different repositories of different subject areas in one platform. These repositories then further categorised into disciplinary, institutional or other categories as per the data requirements. The most widely accepted and used data directories indexing different types of data repositories are discussed in brief in the following section:

3.1 Open Access Directory (OAD)

The Open Access Directory (OAD) is hosted by the School of Library and Information Science at Simmons University**. "**It is a compendium of simple factual lists about open access (OA) to science and scholarship,

maintained by the OA community at large" as cited in the OAD web page. The OAD is a wiki thus depend upon the users for addition of information, curation about data repositories and keeping it update. OAD platform provides a complete list of repositories available in open domain thus ensures easy discovery and allow users to use them for their specific research purpose. Users can select repositories of their choice from the Table of contents or otherwise type their query in the search box for easy retrieval.

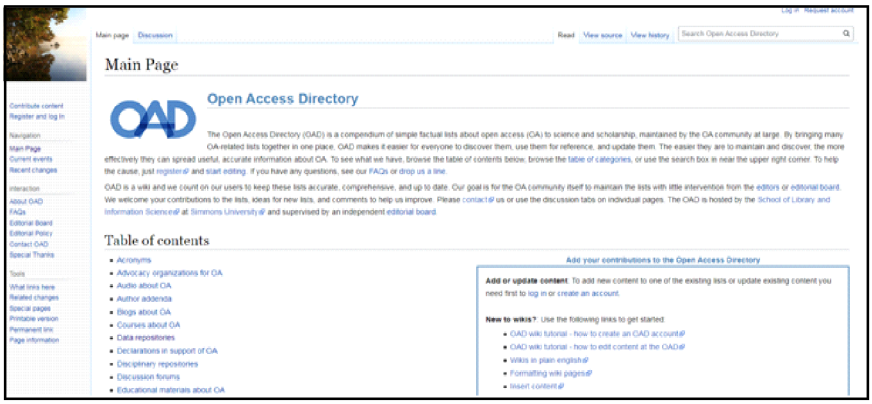

Figure 1: Screenshot of the Open Access Directory (OAD)

3.2 Re3data.Org

Re3data, a global repository came into being in 2012 and wholly funded by German Research Foundation (DFG). The Re3data compiled all the data repositories in one platform and provides access and permanent storage of the content. The main objective of the repository is to provide better visibility, access and sharing to research data. European Commission's "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020" emphasized upon the use of re3data. This project is the collaboration of Berlin School of Library and Information Science, Research Centre for Geosciences, the KIT Library at the Karlsruhe Institute of Technology (KIT) and the libraries of the Purdue University (https://re3data.org/).

Figure 2: Screenshot of re3data.org

3.3 Repository Finder

Repository Finder is a platform which helps in searching appropriate repository to users for deposit of their data. This is the pilot project of "Enabling FAIR Data Project" led by the American Geophysical Union (AGU) in partnership with DataCite and the Earth, space and environment sciences community. Laura and John Arnold Foundation supported this project. The main aim of this project is to implement FAIR principles in the European research data providers and repositories. The platform is hosted on DataCite and searches the queries related re3datarepsoitories. Repository Finder does not have its own data and mainly thrives on the data of re3data.

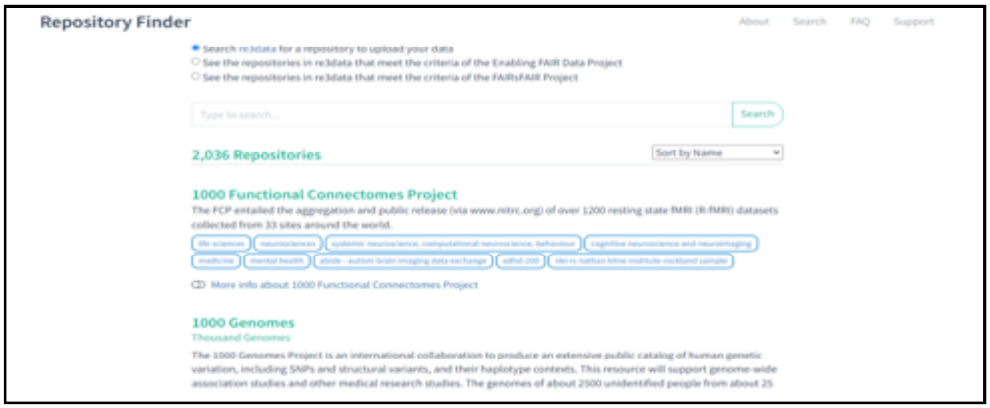

Figure 3: Screenshot of the Repository Finder

4. Methodology

The present study provides the information about the data repositories indexed in re3data.org in the field of Computer Sciences. A Content Analysis technique was used to find out the information about these repositories. A link was provided to each repositories in re3data.org. Link provided the all information pertaining to the website under different sub headings. Ninty Eight repositories in the field of Computer Sciences were indexed in the registry of research data repositories. The computer Science field included the repositories devoted to the following areas like:

- Theoretical Computer Science
- Software Technology
- Operating, Communication and Information Systems
- Artificial Intelligence, Image and Language Processing
- Computer Architecture and Embedded Systems

The main objective of the study is:

- √ To find out the country wise positioning of theses repositories;
- √ To find out the repository type of these repositories;
- √ To find out the repository size of theses repositories;
- √ To find out the software used by these repositories in hosting the content;
- √ To find out the provision of these repositories as data provider and/or service provider;
- √ To find out the language wise distribution of theses repositories;
- √ To find out the metadata Standards and database access types followed in these data repositories;

5. Analysis

5.1 Country Wise Positioning of the data repositories

Figure 4 highlights the country wise positioning of computer Science data repositories. About 28 Data repositories in the field of computer science represented by United States of America and thus USA stood at position one in this ranking. UK is at second position with 9 repositories followed by Germany with 8. South Africa, Singapore, Korea, Republic of, Italy, European Union, France, and Australia represented 2 repositories each. Countries which represented only one repository clubbed into the others category. 41 repositories fall in this category. India represented only one data repository namely GTS AI Data Collection.

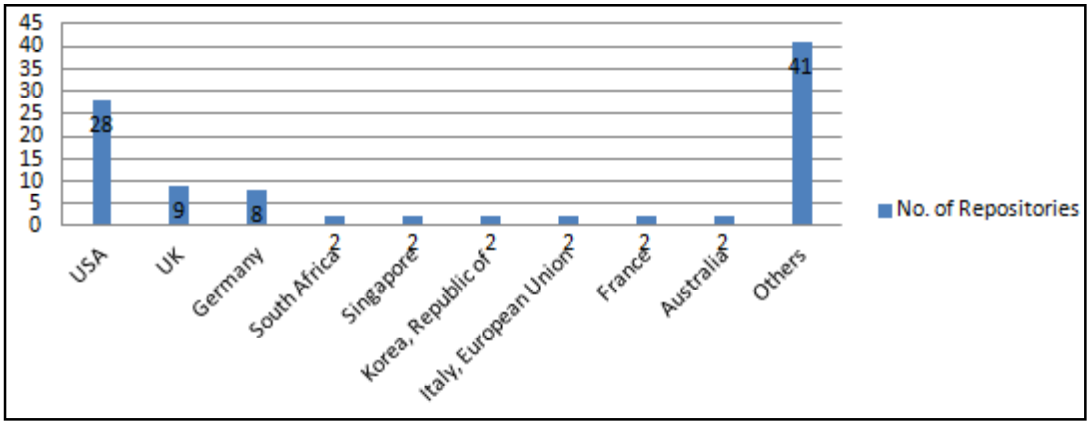

Figure 4: Country Wise Positioning of RDR of Computer Science

5.2 Repository Type

Disciplinary repositories represent the content of any particular discipline**.** It is clear from the analysis in Figure 5 that disciplinary repository was utmost used by the repository manger for organising and managing the content discipline-wise. 46 Disciplinary repository in computer science were indexed in the registry

followed by 34 Institutional . Some of the repositories served dual purpose as institutional cum disciplinary and disciplinary cum institutional. Institutional cum disciplinary repositories (6) and disciplinary cum Institutional (5). The repositories which didn't defined their category considered as other. Their number is four and they are namely; BitBucket, SourceForge, Github and Launchpad. In the category, Institutional, Other; Institutional, Disciplinary, Other; and Disciplinary, Other in each category there is one repository representation.

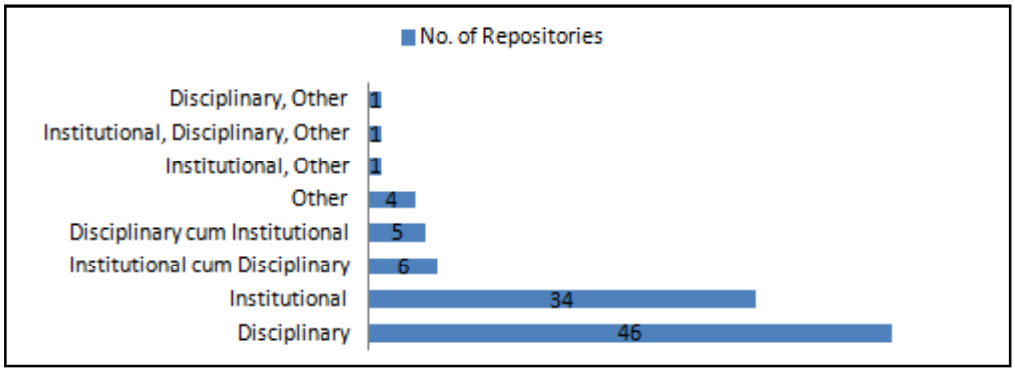

Figure 5: Repository Type of Computer Science Repositories

5.3 Repository Size

From figure 6 it is apparent that 58% of the repositories defined the repository size while 42% of the repository didn't define their repository size. Repository size represents the quantity of the content available in the repository. The analysis showed that repositories had diversified data content and thus size of the repositories was so large in nature. FLOSSmole repository had the largest of the content size i.e, several terabytes followed by Academic Torrents had 83TB of research data, AIDA Data Hub 4.32 TB of the data still growing, ARM Data Center 3 petabytes of data, UK Government Web Archive 3 billion URLs, SourceForge 2.1 million developers and 502000 projects, MindBigData 1867623 brain signals, 491836511 data points and Tierstimmenarchiv-Museum fur Naturkunde Berlin had 12000 recordings of animal voices.

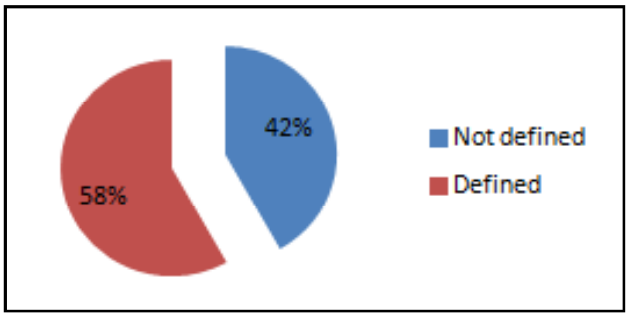

Figure 6: Repository Size of Computer Science Repositories

5.4 Repository Software

75% of the repositories didn't define the software on which they have hosted their repository content as visualized from figure 7. (8%) of the repositories were using dataverse software for hosting their content followed by DSpace (7%), Fedora (4%) and CKAN (3%). of the repositories were using Fedora Software. Defra Data services, Mysql, Eprints were used by only 1% of the repositories each. Dataverse software is used by the repositories namely IIT dataverse, Yale-NUS dataverse, TRR170-DB, Propylaeum@heiDATA, Stockholm University Library Dataverse, Dalhousie University Dataverse@Borealis, Wilfrid Laurier University Dataverse, University of Guelph Research Data Repository Dataverse. GTS AI Data Collection repository hosted by India also didn't define the repository type.

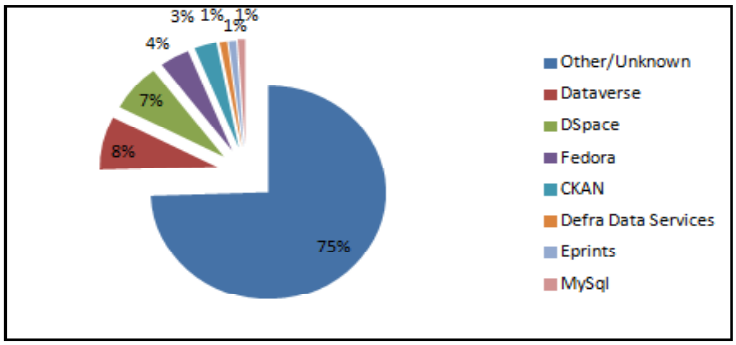

Figure 7: Repository Software used by Computer Science Repositories

5.5 Data and /or Service Provider

This section explained the status of the repository as data provider or service provider. From the figure 8, it is inferred that 70 repositories were mainly data provider in nature and only 9 repositories served as service provider. One repository has undefined this parameter. 18 repositories served dual purpose as data provider and service provider.

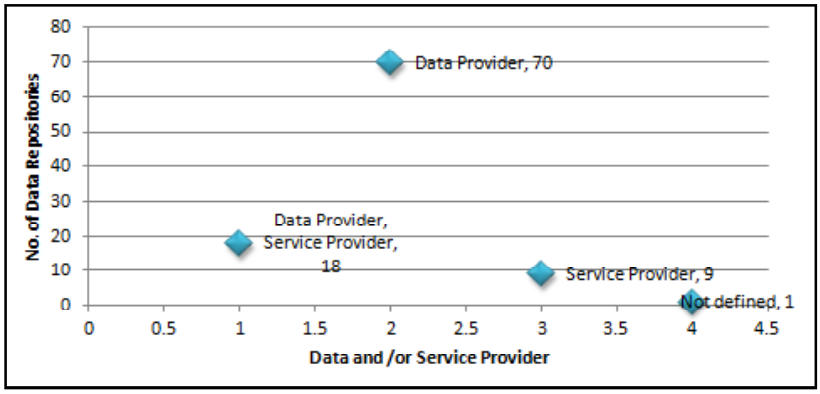

Figure 8: Repository Software used by Computer Science Repositories

5.6 Language of the Data Repository

Majority of the repositories had their content in English Language i.e., (97%). Only 3% of the repositories have content in other languages as mentioned in Figure 9. These repositories were SLAPIS (French), HilData (German) and RepositorioInstitucional UCASAL (Spanish, Castilian).

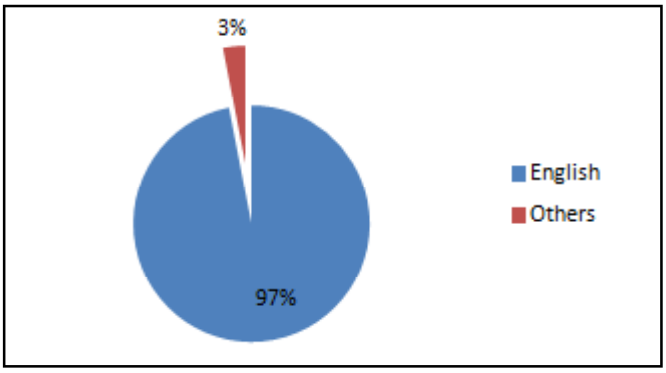

Figure 9: Repository Software used by Computer Science Repositories

This parameter was further analysed on lingual basis and interpretation developed that majority of the repositories (77%) had content in one language i.e., in English, German and French. Bilingual repositories had (20%) share followed by (3%) repositories had multilingual nature as per Figure 10.

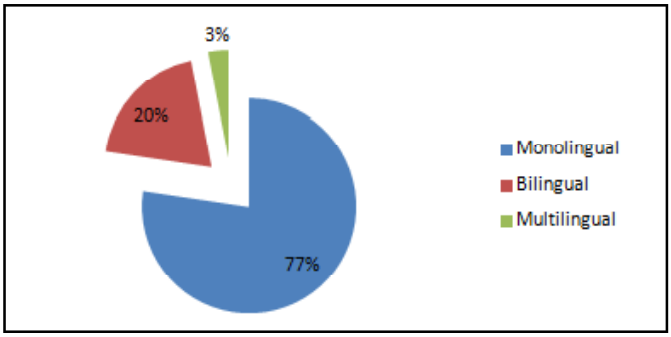

Figure 10: Repository Software used by Computer Science Repositories

5.7 Policy and Database Access Type

Figure 11, indicated that 88 repositories had defined their policy related to data management while 10 repositories didn't define their policy with respect to data management. Further, the analysis revealed that repository database provided access to its content in three ways, open, closed and restricted. Figure 12 indicates that 95 repositories came under the open category which means anyone can access its content without any hassle. The access to 2 repositories was restricted. These were

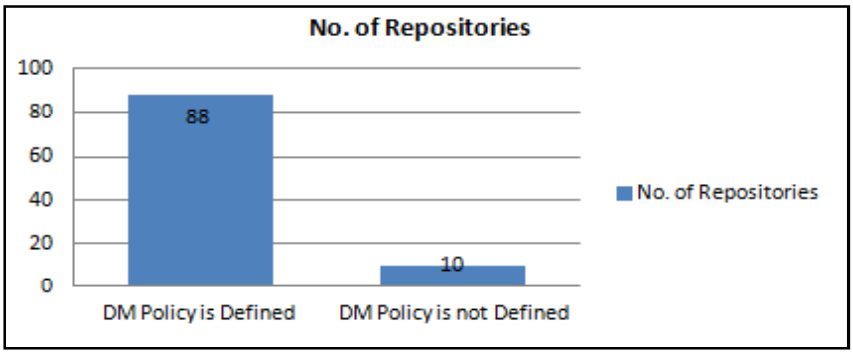

Figure 11: Policy status of Computer Science Repositories

HilData and ScienceData. The status of one repository was closed namely GTS AI Data Collection. Further the analysis result revealed that (87%) of the repositories had restriction in data upload for users. Registration (33) and institutional membership (33) as major restriction type for data uploading in repositories. 13 repositories didn't specify their restriction for data upload as indicated in figure 13. In one repository namely Informatics Research Data Repository fee is required for data upload.

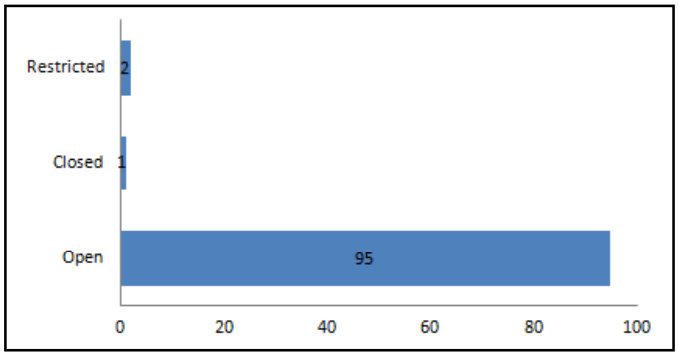

Figure 12: Database Access Type of Computer Science Repositories

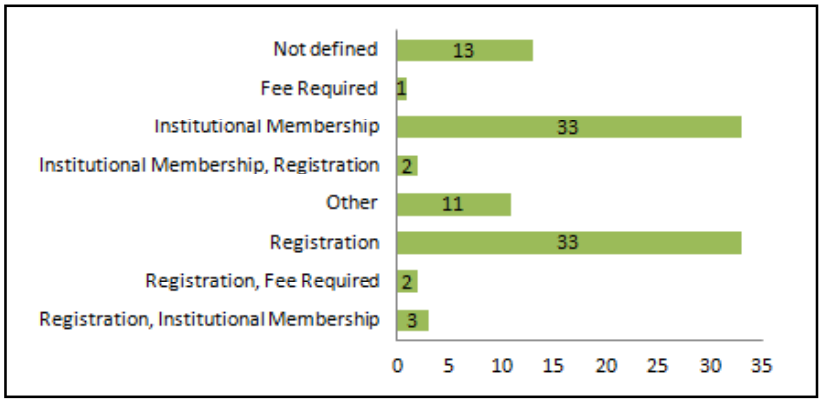

Figure 13: Database Access Type of Computer Science Repositories

5.8 Persistent Identifier System

Persistent identifier helps in easy location and access of datasets if assigned. Figure 14 revealed that DOI was the maximum used persistent identifier by the data repositories followed by hdl (13%), DOI, hdl (4%) and other (3%). 36% of the repositories didn't specify the persistent identifier they used. One repository was used DOI, PURL namely CLARIN-LV Repository. Persistent Identifier URN was used by FLUIR Data. URN, DOI was used by BABS repository.

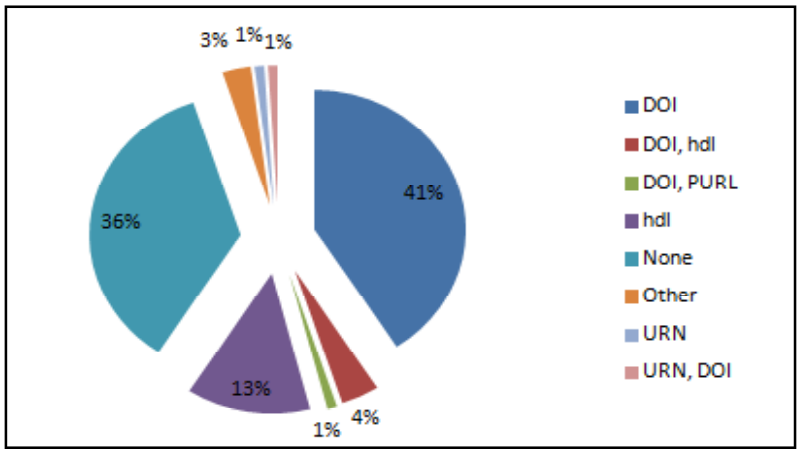

Figure 14: Persistent Identifier of Computer Science Repositories

5.9 Metadata Standards

It was found out that 52% of the repositories had defined the metadata standard for defining their metadata content and the remaining (48%) didn't defined the metadata standard. Metadata Standard analysis revealed that majority of the repositories which defined their metadata content followed these metadata standards namely Dublin Core, DataCite Metadata Schema, Data Documentation Initiative, OAI-ORE (Open Archives Initiative-Object Reuse Exchange), RDF Data Cube Vocabulary, Repository Developed Metadata Schemas, ISA-TAB, FGDC/CSDGM Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata, DCAT- Data Catalog Vocabulary, CSMD-CCLRC Core Scientific Metadata Model, Climate and Forest Metadata Conventions (CF), AVM-Astronomy Visualisation Metadata, Darwin Core and EML-Ecological Metadata Language.

6. Conclusion and Summary

The present study focused on to know about the research data repositories in the field of computer science indexed in re3data. The result inferred that USA stood first in having maximum number of data repositories in computer science. Only 48% of the repositories defined their repository size. Maximum of the computer science repositories are disciplinary in nature, thus presenting the datasets devoted to one particular discipline. Dataverse software was used by repositories for hosting their content. This analysis thus confirmed

that such kind of studies give insights to repository developer to minimise the shortcomings and develop the better repository. Data repository helped the researcher in reducing duplication of research, use of authentic dataset and validation of research data. The development of such repositories help the research community worldwide and further the data citation implication and trend will increase.

References

1. Bhardwaj, Raj Kumar (2019). A Content Analysis of Indian Research Data Repositories Prospects and Possibilities. DESIDOC Journal of Library and Information Technology. 39 (6):280-289. https://doi:10.14429/djlit.39.6.15137

2. Gohain, Rashmi Rekha (2021). Status of global research data repository: an exploratory study. Library Philosophy and Practice (e-journal) 5193. https://digitalcommons/unl/edu/libphilprac/5193

3. Pal, Birender and Singh, Sanjay Kumar (2019). Indian Academic Research data repository (IARDR) with INFLIBNET: a futuristic plan [Paper Presentation]. 12th International CALIBER-2019, KIIT, Bhubaneswar, Odisha, 28-30 November, 2019.

4. Arora, Surbhi and Chakravarty, Rupak (2021). Preserving Global research data: Role and Status of Re3data in RDM. Library Philosophy and Practice (e-journal) 5550. https://digitalcommons/unl/edu/libphilprac/5550

5. TA, Abdul Azeez (2018). Open Research Data Repositories: A Librarian's Perspective. Kelpro Silver Jubilee Souvenir. 47-54.

6. Uzwyshyn, Ray (2016). Research Data Repositories: The What, When, Why and How. Retrieved from: https://infotoday.com

7. Austin et.al (2015). Research Data Repositories: Review of Current features, Gap Analysis, and Recommendations for minimum requirements. IASSIST Quarterly 39(4):24-38.

8. Kim, Suntae (2018). Functional Requirements for Research Data Repositories. International Journal of Knowledge Content Development and Technology 8(1): 25-36.

9. TR, Manu et.al (2018). Analysis of Research Data Repositories in India. Presented in Knowledge Organisation in Academic Libraries (I-KOAL-2018):312-322.

10. Suber, Peter (2012). Open access. Cambridge: MIT Press.

11. Cho, Jane (2019). Study of Asian RDR based on re3data. The Electronic Library 37(2):302-313. https://doi.org/10.1108/EL-01-2019-0016.

12. Misgar, S.M., Bhat, Ajra and Wani, Z. A. (2020). A study of Open Access Research Data Repositories developed by BRICS Countries. Digital Library Perspectives 38(1): 45-54. https://doi.org/10.1108/DLP-02-2020-0012.

Keywords: Data Repositories; Persistent Identifier; Open Access; Dataverse; Research Data Management

About Authors

Mrs. Prerna Prashar
Assistant Librarian
Central University of Jammu and Research Scholar, Guru Nanak Dev University, Amritsar
Email: prerna@cujammu.ac.in

Mr. Harish Chander
Assistant Professor
Guru Nanak Dev University, Amritsar
Email: harish.libsc@gndu.ac.in