DESIDOC Journal of Library & Information Technology, Vol. 43, No. 1, January 2023, pp. 39-44, DOI : 10.14429/djlit.43.1.18624

Research Data Curation in Academic Institutions: Challenges & Expectations

Manish Kumar Singh[#*] and Gireesh Kumar T.K.[$]

[#]*Information Scientist, Central Library, Banaras Hindu University, Varanasi- 221 001, Uttar Pradesh, India*
[$]*Assistant Professor, Department of Library and Information Science, Banaras Hindu University, Varanasi- 221 001, Uttar Pradesh, India*
[*]*E-mail: manish.cl@bhu.ac.in*

ABSTRACT

In the activity of Research Data Curation, the unpublished datasets generated during research are curated for possible reuse in future research by any researcher. Use of curated data, in several cases, may become helpful in avoiding repetition of efforts involved in generation of datasets. A large number of academic institutions in India are actively involved in research in various knowledge areas. Apart from the doctoral and post-doctoral research in academic institutions, academicians are involved in research projects sponsored by public or private bodies; thereby generating sizeable primary and secondary unpublished datasets worthy of curation in a data repository for possible reuse in some other research. For several reasons, curation efforts for research datasets in academic institutions in the country are negligible when compared to such efforts in research institutions. The present work makes an attempt to identify the cause of negligible research data curation efforts in academic institutions of India by uncovering the associated challenges and discusses the expectations from Research Data Repositories of academic institutions.

Keywords: Research data curation; Research data management; Data repository

1. INTRODUCTION

The activity of Research Data Curation refers to a group of activities involving acquiring, reformatting, providing metadata, maintenance and delivery of datasets to the researchers. Research workers produce, store, and analyse their research datain much larger volumes than the text, which is the basis of the research reports. The intermediate results during analytical processing can also be a useful dataset. The researchers collect precious parcels of datasets of various sizes, after investing huge amount of time, effort and money. These datasets cannot be considered as disposable after publication of findings. They can be utilised by the same or other researchers for another analytical study. Therefore, for possible reuse in research, the data that our researchers create should be stored in digital form, discoverable and re-used or repurposed, with due citation to the creator.

In an academic institution, the research datasets may be sourced from several sources. Most of the higher education institutions run PhD programmes in various disciplines. The research scholars toil hard to collect and compile their primary or secondary datasets. They need to transfer the copyright of the thesis produced to the respective universities at the time of thesis submission. If the institution has a policy, or may create one, to collect the datasets compiled for provenance of the research output, the datasets thus collected can be curated by the institution library, preferably in an online open data repository, for future reuse without any copyright issue. Research projects are another common feature of higher education institution, where the regular teaching staffs are involved in research work partially funded by the same institution or an external agency. Generally, the external funding agency claims copyright of only the research output of the completed project. The datasets generated in the process of research may be curated by the institutional library in a data repository. Frequently, the analytical outputs of research work are based on some datasets. Such datasets can also be used by the same or some other researcher for performing a different experiment or analysis for obtaining an interesting result. The copyright concerns are same as above. Various external agencies, both government and non-government, may be identified to be in possession of datasets of interest to the researchers of the institution. These can be obtained by satisfying the terms of license under which these datasets are available for use. Closed access datasets possessed by other research institutions may also be obtained under a collaborative research arrangement. It can be observed from the directories of the data repositories, e.g. re3data.org, that none of the academic institutions in India is having its research data repositories, whilst several research institutions have. The preliminary information in the present study also revealed that there is negligible data curation of research data in academic institutions of the country.

In the subsequent sections, relevant literatures are reviewed for knowing the background of developments

Received : 31 October 2022, Revised : 08 January 2023
Accepted : 17 January 2023, Online published : 31 March 2023

in the field of research data curation. The methods adopted for present work are presented in the section of investigation. The output of investigation is discussed later to identify the challenges that has caused negligible research data curation work in academic institutions, before concluding the present work by highlighting the expectations of researchers from academic institutions.

2. BACKGROUND

The data curation research has become popular in LIS literature from the beginning of this century. Importance of data has gained along with refinement and growth in volume of empirical/experimental research. In the beginning, most of the authors focused on describing the new field of research and focused primarily on the conceptual elements, ideas and varied opinions.

The Atkins Report[1] highlighted the importance of trusted and enduring organisations to assume the stewardship for scientific data and stated that "Stewardship includes ongoing creation and improvement of the metadata by people cross-trained in scientific domains and knowledge management". This report emphasised that most of the curation tasks involved can be automated by developing "middleware, standard or interoperable formats, and related data storage strategies". It concludes "each discipline is likely best suited to creating and managing such repositories and tools". It noted that "interoperability with other disciplines is essential". The essence of the report is that by employing technological means much of the problem could be resolved.

Several government agencies, including DST (Govt. of India)[2] and NSF(USA)[3], have brought out reports for data sharing, reuse and accessibility. US National Science Board[4] published a report on long-lived digital data collections in 2005, where it highlighted the challenges of digital preservation and made several recommendations for finding ways to discern collected datasets that may retain their value for the long-term and for devising sustenance strategies. This report emphasised the importance of data management plan in proposals keeping in view the long life of datasets.

Anna Gold[5] wrote about the growing importance of research data in 2007 by stating that data is the currency of science, even if publications are still the currency of tenure. Author emphasised that research data curation is essential for scientific productivity, collaboration and discovery and researcher should be able to exchange, communicate, mine, reuse and review the data associated with research. Brase[6], *et al.*proposed the need of persistent internet identifier for data objects, as well, along with the textual literature and discussed the possibilities of assigning DOI to the data objects. The authors also discussed the visibility of the data objects and suggested that in order to increase the visibility of datasets and their access; the datasets handle should be integrated into the electronic texts, which are most commonly cited.

Some authors including Lewis[7] and Heller[8], *et al.* have expressed their opinion that research libraries are better suited to control direct access to the research data in raw form. Averments are that research libraries traditionally have controlled access to published documentation that contains the research data and extending this service to research data would be a natural extension. Heller[8], *et al.* advocated that large datasets should be managed together collectively in an integrated manner for providing their access, as they require various functions including identification, retrieval, sharing, and recycling, description, organisation, and consistent control. They also require treatment according to harmonised rules, formats, and protocols. These functions are already performed at research libraries for documents. Weber[9], *et al.*, in their paper, highlighted special characteristics of research data which are often complex datasets. They carry different kinds of information, and that they are dependent upon particular domain, context and provenance. The authors emphasised that the functions involved in datasets preservation and maintenance, involving storing and organising, requires scientific knowledge of each domain and advanced technological knowledge of required infrastructure. This is essential for their proper preservation and to facilitate other researchers enable them to effectively query the desired information.

Borgman[10] has delved into the rationale behind sharing of data and identified and examined four rationales. Borgman gave examples from the sciences, social sciences, and humanities to demonstrate each. The four given rationale are: (i) reproducing of the results for verification of research, (ii) the research works that are publicly funded their results should be available to the public, (iii) enabling other researchers to examine and ask new questions from extant data, and (iv) that the state of research and innovation will be advanced through sharing. These rationales are identified and differentiated by Borgman[10] on the basis of varied arguments for sharing, the intended beneficiaries, and various stakeholders' motivations and incentives. This work also enumerated several disincentives to sharing research data. The identified disincentives included (i) lack of reward or credit for sharing, (ii) reusability of datasets require proper documentation of datasets which is highly time consuming job, (iii) misuse and misinterpretation of data is a possibility that may give bad credit, (iv) possible violation of intellectual property, and (v) restrictions over free distribution of data on human subjects and endangered species. Borgman[10] further identified that lack of demonstrated demand for research data apart from subjects of genomics, climate science, astronomy, social science surveys, and a few other subjects, is one of the most significant challenge to data sharing. Later on, Borgman[11] , *et al.* presented experience of a large archive, 'Data Archiving and Networked Services Institute of the Netherlands'. It mentions that the archive claims to manage more than 50 years of data from the social sciences, humanities, and other domains. Their study revealed interesting facts about infrequent

submission of datasets by academic researchers and the general trend of restricting access to their files. It also revealed that contributors and consumers of datasets are diverse groups that overlap minimally.

Cousijn[12], *et al.* asserted that data cannot be made openly available straightaway for facilitating data reuse. It used acronym FAIR and stated that the curated data needs to be made available in a FAIR way, where FAIR stands for Findable, Accessible, Interoperable and Reusable. Thus, there should be rich metadata that also includes identifiers to associated data resources. It was mentioned that additional efforts are needed to accomplish this. Stakeholders in the data repository have their defined role to make available the curated datasets to researchers, who can reuse data and associated outputs for their research pursuits. The requirement of fairness was highlighted by Wikinson[13], *et al.* They argued that by implementing FAIR principles, the submitted datasets will be managed more rigorously. This will lead to the benefit of the academic community for pursuits of knowledge discovery and innovation.

Faniel[14], *et al.* highlighted an important aspect of researcher's confidence on other's data before its reuse. They argued that before reusing datasets created by other researchers certain checks are need to be made. Researchers need to assess the data's relevance and ensure the understandability of data. Apart from this, the researcher should evaluate whether the data can be trusted. Therefore, the supply and making the research data accessible does not guarantee its reuse. In their work the authors attempted to examine how the reusability of other's experimental data are assessed for model validation.

Literature focusing on the distinct nature of academic data curation work and its associated challenges could not be found. Considering the distinct purpose and that substantial research is contributed by academic institutions and their doctorate research scholars toil hard and ensures continuous generation of datasets, though heterogeneous, the challenges of research data curation for academic institutions needs to be explored separately.

3. INVESTIGATION

Preliminary information has confirmed that there is dearth of efforts in data curation field by the academic institutions in India. In order to harness the vast production of research data in academic institutions, the factors behind the stark difference of popularity of research data curation between research institutions and academic institutions need to be identified. For identification of these factors, an investigation was done by the method of personal interviews with 30 academicians and 15 practicing library professionals of distinct academic institutions in India, as interview subjects. The investigation was confined to public funded academic institutions of India. The selection of the subjects was done on the criteria of having substantial research contributions for the academicians and involvement in ICT activities for library professionals. Before interview, the subjects were briefed about the study and the present status of research data curation. The subjects were also demonstrated the use of some open as well as commercial research data repositories. During the interviews, the subjects were presented sets of questions in two stages, in which the first stage questions were common and intended to know about the present status of reuse of research data and curation activities done at their institution, if any. The second stage questions were designed based on the outcome of the first interview to drill down in the negative causes. The descriptive answers by these subjects were grouped by clustering to arrive at conclusive answers for this study.

Figure 1. Ratings of challenges.

The subjects were asked to rate the challenges from 1 to 5, where number 1 assigned to most important challenge and 5 assigned to the least important one, for ease in understanding. These ratings were reversed while tabulating the identified challenges in order to assign weights for comparing them. The comparison of these challenges is given in Fig. 1, wherein sums of tabulated ratings are averaged on all the subjects.

The identified challenges having average ratings above 1 are considered as significant. Accordingly, six of the eleven identified challenges are considered important after analysis. This result is put to a test of counting highest rating of 5, which is represented in Fig. 2. In this test, three out of six significant challenges were identified as prominent. These are: (i) dependability of sourced data, (ii) low reuse, and (iii) contribution rationale for researchers. These three challenges may need special attention to overcome the low proliferation of the research data curation practice in academic institutions.

4. CHALLENGES FOR RESEARCH DATA CURATION IN ACADEMIC INSTITUTIONS

Some critical observations, related to challenges before academic institutions, came out after analysis and descriptive response of the subjects. They are presented below.

4.1 Dependability of Sourced Data

Subjects identified and expressed their concerns regarding the reliability of the curated research dataset. Their use in research requires a dependable dataset for the reuse of which the present researcher will be answerable. The primary concern regarding it was that the datasets were generated and deposited by researchers or research groups and do not bear an institutional authentication. In most of the cases the documentation associated with datasets were incomplete or failed to satisfactorily describe the process involved in generating dataset. For datasets linked with publications, it was observed that dataset support the analysis in the publication, but such is not the case otherwise. Subjects, who are researchers, were vehemently in favour of running their own process to derive a reliable dataset for their own research.

4.2 Contribution Rationale for Researchers

It was a common observation that, in academic institutions, the rationale for contribution is not defined in the system of research data curation. In an academic institution, datasets are generated as a result of research activities, mainly as part of doctoral research and short-term research projects. By and large there are no institutional programme to promote and support a particular research after completion of its term, which is comparatively quite short, and submission of report. Further, citations for datasets not linked with publication are not included in academic recognition in the country. So, there is lack of perceptible rationale for the researchers to contribute the datasets generated, as part of their research, to a data repository. Though data reuse may create chances of corresponding paper's citation, the data citation of

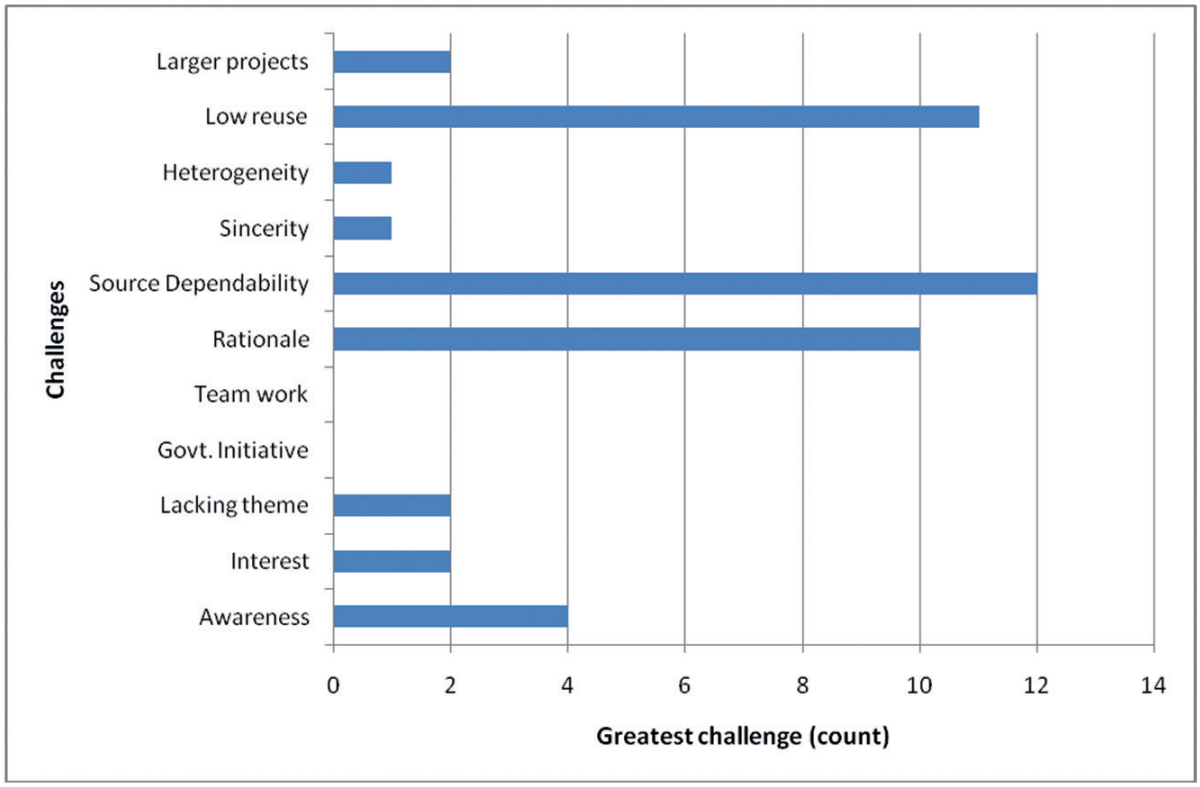

Figure 2. Highest rated challenge.

the datasets, not linked with paper, do not increase author's citation counts.

4.3 Low Reuse

Another general concern was the re-usability and the possible low reuse of the curated datasets. The proposition of curating datasets may fall out due to low expectancy of its future reuse or data citation. Awareness about dataset citations was also found scarce. All the institutions interviewed have informed that research works are generally planned to be substantially different from previous researches in the respective academic departments, making chances of in-house reuse even less.

4.4 Awareness

The subjects informed about little know-how for data citations and research data curation. Organised data curation activities for research data have never happened in their academic institutions. Thus, the required expertise is also a concern. An argument also emerged against the need of citing the source for data reuse, with claim that the anti-plagiarism tools are not able to check data reuse. The existence of open repositories for free archiving of research data was also little known. The use of the most popular directory of data repositories, re3data (Data Cite) was also having little awareness.

4.5 Lack of Theme

An important and primary observation of the professionals was regarding the difference in characteristics of research conducted in a research institution and an academic institution. Most of the research institutions are themselves thematic and they conduct thematic research projects over a very long period. This prolonged existence creates huge and unique collection of datasets in the theme of the research. The uniqueness and the size of the collection create an opportunity for the research institution to organize the collection of datasets into a data repository, which also happens to be a thematic data repository, for wider consumption by the theme researchers around the world. Such an opportunity for thematic research data repository is elusive for the academic institutions whose activities are dominated by imparting of instruction.

4.6 Heterogeneity

Research institutions that run their research data repositories have their theme, around which they curate their research data. Such is not the case of academic institutions that will have heterogeneous small collections of datasets. An academic institution conducts research activities in numerous subject areas. So, it remains small in numbers for individual subject areas. Running own data repository by academic institutions becomes infeasible considering the reduced chances of reuse due to heterogeneity, unless harvested by another popular repository.

5. CONCLUSIONS

People in academic institutions are agreeing with the idea of curating the research data sets for their reuse by some other researchers with due citations to the creator of the datasets. Along with promoting the research cause, the data sharing through repositories may also benefit the researcher by way of increased citations to their work. Simultaneously, they identify huge barriers that may turn the running of such a programme in their academic institution infeasible. Data curation in academic institutions are not picking up due to various challenges including rationale, characteristics of datasets, dependability, disinterest of academicians along with lack of awareness about the tools &processes and requirement of citations for data.

For a possible solution of the challenges, some expectations that are derived from the answers of the subjects are laid down. Firstly, standalone data repository will not serve to promote the reuse. Either the datasets be deposited in other data repositories or the popular data repositoriesare made to harvest institutional data repository, in order to overcome the challenges of heterogeneity and low reuse. An idea was to build several thematic data repositories and they be managed by respective thematicacademic consortiums. Secondly, there is an acute need of examining the documentation of deposited datasets to ensure its completeness and enable other researchers to rely on the authenticity of the process used for generating the dataset. Thirdly, awareness about the data citations and research data curation is essential for proliferation of the best practices and promote the research data deposition and its reuse.

REFERENCES

1. Atkins, D.E.; Droegemeier, K.K.; Feldman, S.I.; Garcia-Molina, H.; Klein, M.L.; Messerschmitt, D.G.; Messina, P.; Ostriker, J.P. & Wright, M.H. Revolutionizing science and engineering through cyber infrastructure. Report of the National Science Foundation Blue-Ribbon Advisory Panel on cyber infrastructure, USA. 2003. https://www.nsf.gov/cise/sci/reports/atkins.pdf (Accessed on 04 August 2022).
2. DST. National Data Sharing and Accessibility Policy (NDSAP). Department of Science and Technology, Government of India, New Delhi, 2012. https://dst.gov.in/sites/default/files/gazetteNotificationNDSAP.pdf
3. NSF. Sustainable Digital Data Preservation and Access Network Partners (DataNet). National Science Foundation, USA, 2007. https://www.nsf.gov/pubs/2007/nsf07601/nsf07601.pdf
4. National Science Board. Long-lived digital data collections: enabling research and education in the 21st century. National Science Board, USA, 2005. www.nsf.gov/pubs/2005/nsb0540/ (Accessed on 08 August 2021).
5. Gold, A. Cyber infrastructure, Data, and Libraries, Part 1: A Cyber infrastructure Primer for Librarians.

D-Lib Magazine, 2007, **13**(9/10), www.dlib.org/dlib/september07/gold/09gold-pt1.html (Accessed on 09 August 2022).

6. Brase, Jan, [*et. al.*]. Approach for a joint global registration agency for research data. *Infor. Services Use*, 2009, **29**(1), 13-27. doi: 10.3233/ISU-2009-0595
7. Lewis, M. Libraries and the management of research data. *In* Envisioning Future Academic Library Services, edited by McKnight, S. Facet Publishing, London, 2010, 145-168.
8. Heller, A.; Blaabjerg, N. J.; Clausen, N. F.; Christensen-Dalsgaard, B. & Dorch, B. Forskningsdataog Open Access: Et deff-projekt (Research Data and Open Access: A DEFF Project). DEFF, DTU, Denmark, 2011, https://backend.orbit.dtu.dk/ws/portalfiles/portal/102016658/DEFF_Forskningsdata_og_Open_Access_Final_2_.pdf (accessed on 18 August 2021).
9. Weber, N.M.; Baker, K.S.; Thomer, A.K.; Chao, T.C. & Palmer, C.L. Value and context in data use: Domain analysis revisited. *In* Proceedings of the American Society for Information Science and Technology, 2012, **49**, 1-10. doi: 10.1002/meet.14504901168.
10. Borgman, C.L. The conundrum of sharing research data. J. Am. Soc. Infor. Sci. & Technol. 2012, **63**, 1059-1078. doi: 10.1002/asi.22634.
11. Borgman, C.L.; Scharnhorst, A. & Golshan, M.S. Digital data archives as knowledge infrastructures: Mediating data sharing and reuse. *J. Assoc. Infor. Sci. Technol.* 2019, 70, 888-904. doi: 10.1002/asi.24172.
12. Cousijn, H.; Habermann, T.; Krznarich, E. & Meadows, A. Beyond data: Sharing related research outputs to make data reusable. *Learned Publishing,* 2022, **35**, 75-80. doi: 10.1002/leap.1429.
13. Wilkinson, M.; Dumontier, M.; Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data (Article no. 160018), 2016. doi:10.1038/sdata.2016.18.
14. Faniel, I.M. & Jacobsen, T.E. Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data. *Computer Supported Cooperative Work,* 2010, **19**(3/4), 355-375. doi:10.1007/s10606-010-9117-8.
15. DataCite. Finding a repository. https://www.datacite.org/services/find-repository.html (Accessed on 02 August 2022).

CONTRIBUTORS

Dr Manish Kumar Singh, is an Information Scientist in Central Library of Banaras Hindu University, Varanasi with prior teaching experience of the subjects of Computer Applications for six years. He has his research interest in Database indexing, Data repository and Digital preservation.
He has contributed to conceptualization of the study, design of analytical framework and writing & review of the article.

Dr Gireesh Kumar T.K., is an academician with Department of Library and Information Science in Banaras Hindu University, having his research interests in Digital information systems, Cultural heritage, Open source software, Library automation, ERM, Academic writing, Scientometrics, Life skills.
He contributed to the computational analysis, writing and editing of the article.