October 2023

ROLE OF HARVARD DATAVERSE PROJECT IN RESEARCH DATA MANAGEMENT SERVICES

Surbhi Arora
arora.surbhi03@gmail.com

Rupak Chakravarty
rupak@pu.ac.in

ROLE OF HARVARD DATAVERSE PROJECT IN RESEARCH DATA MANAGEMENT SERVICES

Surbhi Arora* Research Scholar (ORCID: 0000-0002-4593-4195) (surbhi03@pu.ac.in,) Library and Information Science Department, Panjab University, Chandigarh, PIN: 160014

Rupak Chakravarty, Professor (ORCID: 0000-0001-5046-1663) (rupak@pu.ac.in), Library and Information Science Department, Panjab University, Chandigarh, PIN: 160014

ROLE OF HARVARD DATAVERSE PROJECT IN RESEARCH DATA MANAGEMENT SERVICES

Abstract

Purpose: To simplify working with and sharing research data, researchers want infrastructures that provide the highest level of accessibility, stability, and reliability. The Harvard Dataverse Project (HDP) (https://dataverse.org/) is compiling a growing list of such infrastructures. In this regard, the objective of this paper is to provide an overview through an analysis of the activities of the Dataverse website in managing research data.

Design/ Methodology/ Approach: The study examines the statistics systems and other critical resources concerning upload and use the dataverse/ datasets/ files upto October 2022. This includes the creation of dataverse, category of dataverse, uploaded total datasets, file downloads trend, publication of dataverse or datasets, most approachable subject to share and browse data, the most recommended file type of research data and access level of research data. The basic resources include top metadata sources and data citation standards of dataverse project.

Findings: It is noted that behaviours associated with structured study outcomes are more evident in developed countries as opposed to developing countries according to top author affiliation which is from the USA. The findings also show that research data in Medicine, Health, and Life Sciences is more uploaded and structured, whereas data in Social Sciences is more browsed and structured. Overall, the generation of dataverses, datasets, files, their downloads, and publication dataset is on the rise. The maximum contribution of data developers is found as Master, Daniel M. and Stager, Lawrence E whereas research project category and data and image file format are seen as highly used to organize data. On the dataverse website, good citation standards are being noticed, as well as the fact that 97 percent of data are available to reuse because contributors waive their copyright licences under CC0.

Originality/ Value: This study presented an overall picture of the growing research data practices throughout the investigation on the Harvard Dataverse platform. The research proposed best practices focused on RDM operations to improve the amount of Research Data activities.

Keywords: Research Data Management; Dataverse; Datasets; Data Citation Standard; Harvard Dataverse Project

1. Introduction

The Dataverse Project is an open source web application that allows users to share, store, reference, browse, and examine research data (Harvard Dataverse[1]). It allows easier to share data with others and makes it simpler to copy other people's work (King and King[2] [3]). The Dataverse Project is housed and developed at Harvard's Institute for Quantitative Social Science (IQSS), along with many collaborators and contributors worldwide. The Dataverse Project was built with an earlier Virtual Data Center (VDC) project, which spanned 1997-2006 as a collaboration between the Harvard-MIT Data Center (now part of IQSS) and the Harvard University Library. The Dataverse is now an open-source web tool that allows researchers to exchange, save, cite, explore, and analyse data. Researchers, data authors, publishers, data distributors, and affiliated institutions all receive appropriate credit via a data citation with a persistent identifier (e.g., DOI, or Handle). Multiple dataverses are stored in a Dataverse repository. Each dataset contains descriptive metadata and data files, and each dataverse contains datasets or additional dataverses (including documentation and code that accompany the data). Features of the Harvard Dataverse Project (HDP) are listed below:

Table 1: Features

Support for FAIR Data Principles	Versioning (History of changes to datasets and files are preserved)
Data citation for datasets and files (EndNote XML, RIS Format, or BibTeX Format)	Custom Terms of Use (CC0 waiver by default)
OAI-PMH (Harvesting) (using standardized metadata formats: Dublin Core, Data Document Initiative (DDI), OpenAIRE, etc.)	Guestbook (Optionally collect data about who is downloading the files from datasets.)
APIs for interoperability and custom integrations	File hierarchy
Login via Shibboleth	Faceted search
Login via ORCID, Google, or GitHub	Restricted files
DataCite integration	Customization of Dataverse Collections
Usage statistics and metrics	Dropbox integration
Schema.org JSON-LD	Notifications
Preview and analysis of tabular files	Widgets
External Tools	User management
Fixity checks for files	Mapping of geospatial files
Publishing workflow support	Handling large data
File download in R and TSV format	Pull header metadata from Astronomy

	(FITS) files

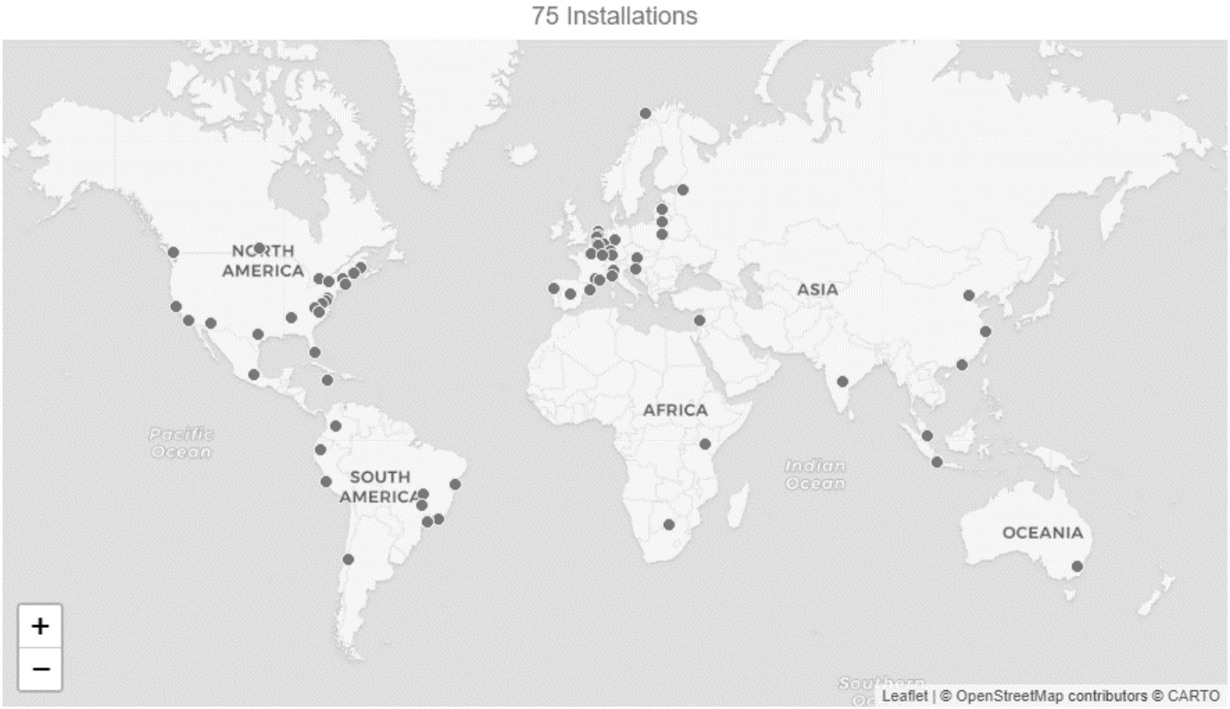

Figure 1: Installations

This study presents an overall picture of the research data practices throughout the investigation of the Harvard Dataverse Project website. The research elaborates the best practices focused on Research Data Management (RDM) operations to improve the amount of research data activities.

2. Study scope

The study approached the The Harvard Dataverse website (https://dataverse.harvard.edu/) to analysis the data. This website is chosen for a variety of reasons i.e. it is a free data repository open to all researchers from any discipline, both inside and outside the Harvard community, where they can share, archive, cite, access, and explore research data. Each Dataverse collection (or virtual repository) is a personalised collection of datasets that may be used to organise, manage, and show data. Researchers can choose to make their data available to the general public, restrict access, and set unique terms of use. When researchers submit their data, they instantly receive a standard data citation with a DOI, and their metadata is accessible and searchable via search engines, even if the data is limited or restricted.

3. Study objectives

3.1 To understand the adoption and growth trends with regard to creating dataverse repository using HDP.

3.2 To identify out the top category of dataverse and further their rate of share.

3.3 To examine the total datasets uploaded and determine which are the most in each month of the year.

3.4 To investigate the dataverse usage trends.

3.5 To identify the prominent subject, contribution of top authors and author affiliations.

3.6 To identify the top metadata sources, research data file type.

3.7 To analyse the access level of research dataverse/ datasets/ files

3.8 To examine the data citation standards of dataverse project

4. Methodology

The Harvard Dataverse website is chosen for the compilation, presentation and analysis of the findings. The researcher examines the statistics systems and other critical resources up to October 2022. The approaches to statistics include the involvement of contributors to upload and use of dataverse/ datasets/ files. These statistics of contribution include the creation of dataverse, category of dataverse, uploaded total datasets, file downloads trend, publication of dataverse or datasets, most approachable subject to share and browse data, the most recommended file type of research data and access level of research data. The basic resources include top metadata sources and data citation standards of dataverse project.

5. Results and Discussion:

5.1 Total Dataverses

Dataverse is a virtual container that may be configured and managed by its owner to store research data studies (including datasets and other dataverses). The dataverse project has experienced an increase in the number of dataverse repository creations in recent years, with a peak of 12800 creations expected by the end of October, 2022. Out of the total created repositories, 6058 dataverse repositories are searchable.

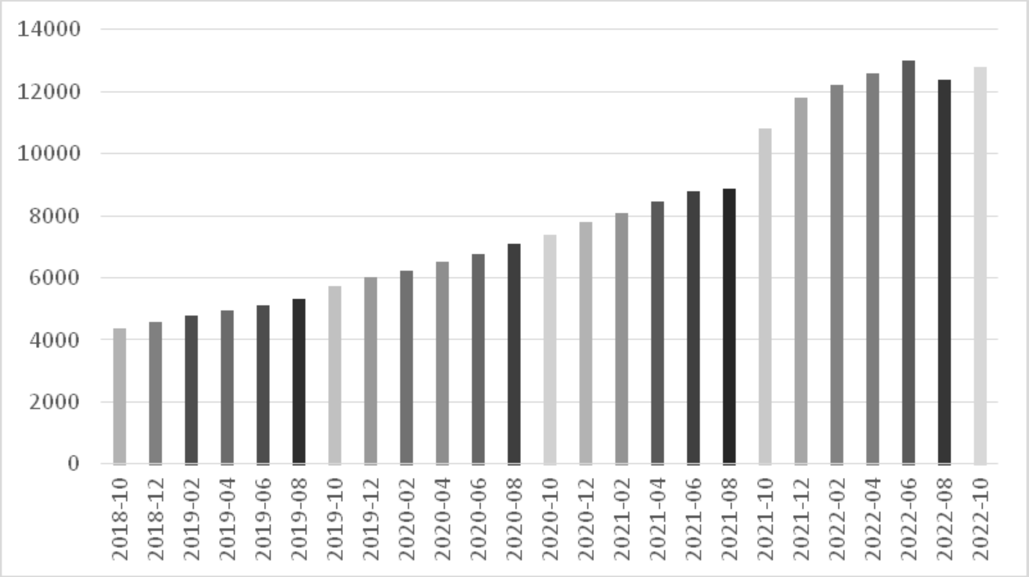

Figure 2: Total Dataverses

Dataverse is a global project that encourages researchers, publishers, and organisations to use research data management systems. To arrange their data, authors can build a dataverse community. It is observed that awareness and adoption is increasing as the creation of dataverses are increasing.

5.2 Dataverse category

A total of 30 categories are registered in the Harvard dataverse project. These 30 categories are Research Project followed by researchers, uncategorized organization or institution, research group, department, laboratory, journal, teaching course, research institute, project de pesquisa, organizacao ou instituicao, cerca centres, universities, faculty, Proyecto de investigacion, organizacion O institucion, pesquisador, sem categoria, grupo de investigacion, grupo de pesquisa, curso de ensinee. It is observed that maximum research data is organized in the category of the research project (32%).

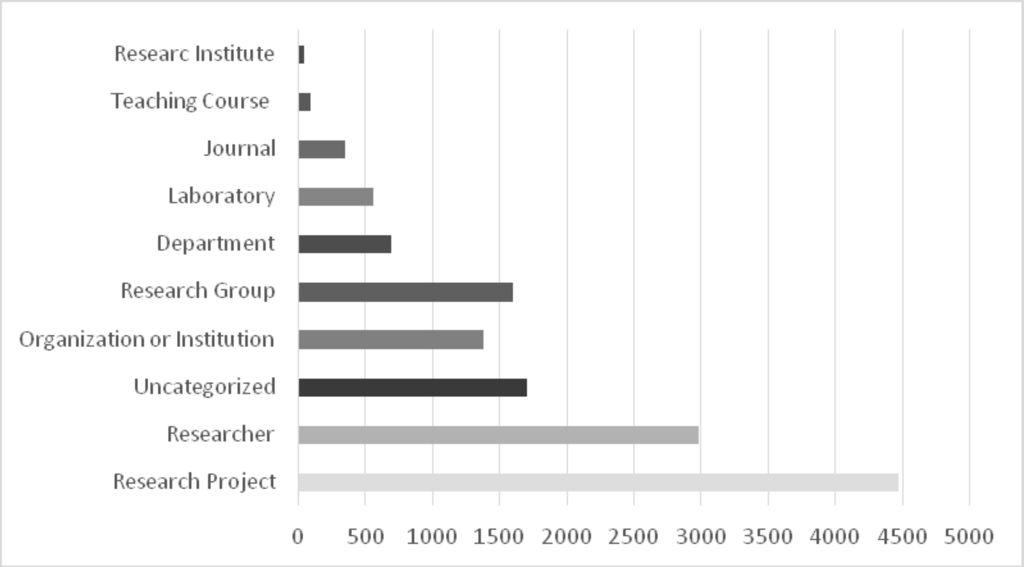

Figure 3: Category of Dataverse

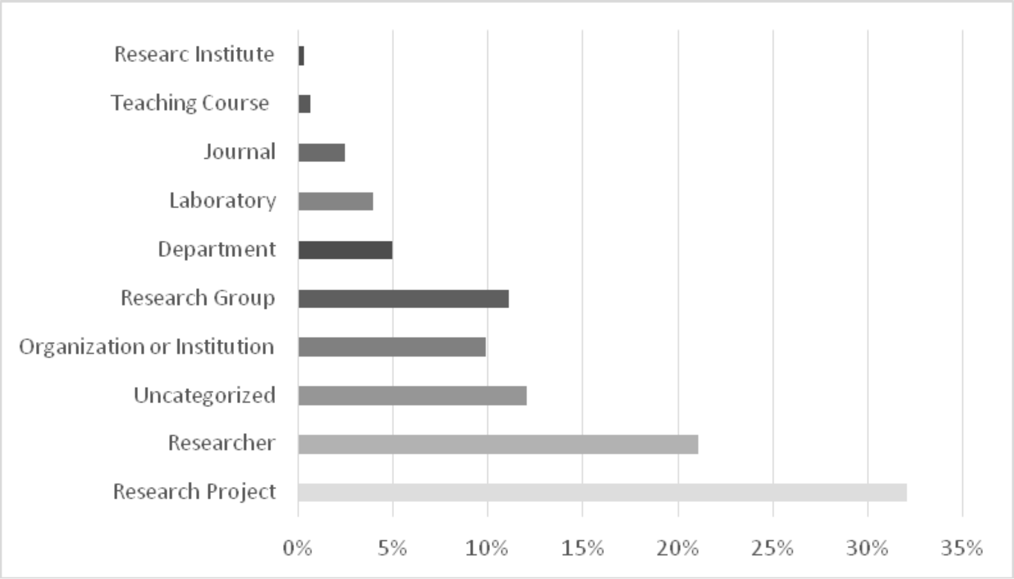

Figure 4: Share of Dataverse category

All categories are important where research data are produced whether these are research project, researcher, organization or institution, department, laboratory or teaching courses (Arora and Chakravarty[4]). Throughout this sense, these all categories should upload and share their datasets on the Harvard dataverse project. In this present finding, research institutes and teaching courses enjoy fewer contributions as a comparison to research projects and researchers.

5.3 Total Datasets

Datasets are a study, experiment, set of observations, or publication that is uploaded by a User. A dataset can be made up of a single or numerous files. In the Harvard dataverse repository, a total of 152,004 datasets have been stored or uploaded, with 637 datasets belonging to the last 30 days' activity. From total datasets uploaded, 80866 datasets have deposited and 71138 datasets have harvested. In comparison to the previous years, the highest number of datasets are uploaded in June, 2022.

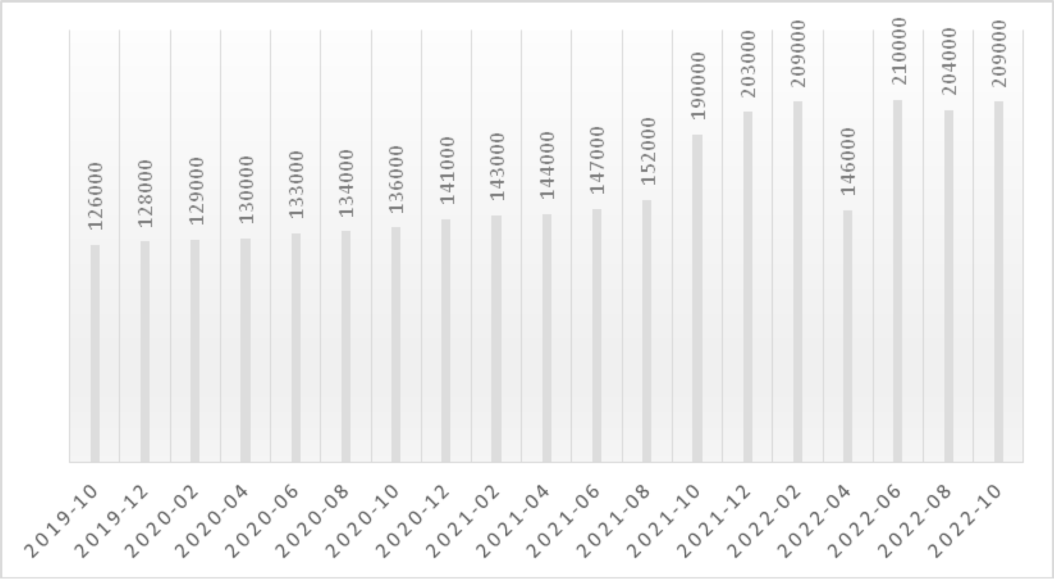

Figure 5: Total Datasets

Dataverse is a global level project that allows researchers to upload their research data through the datasets or files option. And it is interesting to know that the dataset uploading trend is increasing towards 2022.

5.4 Total files

A total of 1,821,674 files are deposited in the Harvard dataverse repository which can be find, share, and archive across all research fiels. It is observed that the trend of deposited files is increasing in comparison to previous years. It's also been discovered that the highest number of 2290000 files are deposited in June month of 2022.

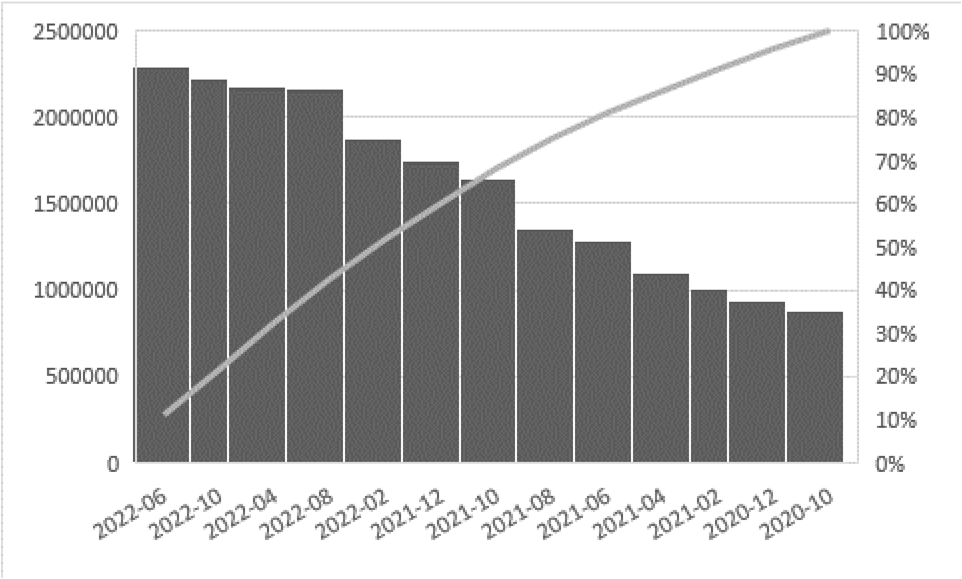

Figure 6: Total files

The Harvard dataverse project allows authors to upload their data in three terms i.e. Dataverse, Dataset and Files. In this current finding, the Authors' interest in uploading their research data files is higher in 2022 than in prior years.

5.5 Total file downloads

A total of 42,322,117 files are downloaded in the Harvard dataverse project, from which 972,926 files were from past 30 day's activity. From total downloads, 1,131,803 files are deposited in the Harvard dataverse website. It is observed that a maximum no. of files are downloaded in June, 2022.

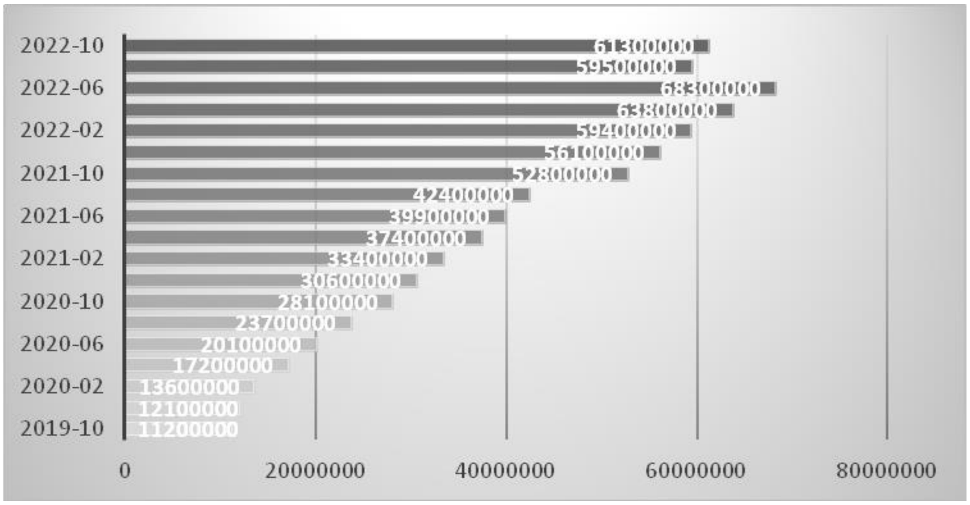

Figure 7: File downloads

The majority of researchers or data developers permits to access their data to a third party. Throughout this sense, maximum type of users should enjoy the benefits to access that research data for observing their further work of research. In this present finding, the number of downloads are increasing.

5.6 Publication year

A function is available for User Submissions by which users can make their dataverse and/or their dataset containing data files publicly available and publicly searchable on the Harvard Dataverse application through search engine and third-party search engines (e.g., Bing Search or Google Search). It is observed that maximum data were made published and available to the public in the year 2021, whereas the trend of this activity is down towards 2022.

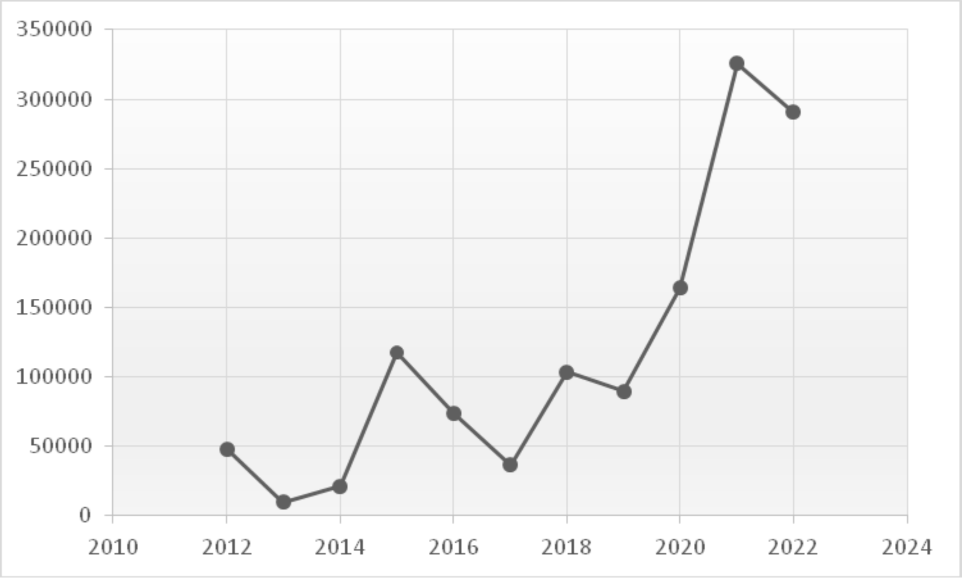

Figure 8: Publication year

The Harvard Dataverse project allows depositors to upload and make public their datasets. In this sense, all data developers should make their data public to make it more visible and to gain more credit in terms of citations. In this study, it was discovered that the level of awareness is low, peaking in 2021 but then dwindling by 2022.

5.7 Subject

There are a total of 21 subjects that have signed up to deposit data in the Harvard dataverse project. These include medicine, health and life science, social sciences, arts and humanities, earth and environmental sciences, agricultural sciences, law, computer and information science, engineering, physics, chemistry, business and management, astronomy and astrophysics, mathematical sciences, forest management & restoration (fmr), forest and human well-being (Hwb), sustainable landscapes & food (SLF), architecture, arts and humanities (ex: English, history, foreign language), social sciences (ex: education, politics, sociology, economics, psychology) and Value chain, Finance & investments (VFi). It is observed that maximum research data are deposited in the Harvard dataverse is belongs to medicine, health and life sciences disciplines (54%) followed by social sciences (18%) and Arts and Humanities (10%). Whereas least in Value Chain, Finance & investment (VFi) (0.012%). It's also worth noting that Social Sciences have the most browsing data (47%) compared to Mathematical Sciences, which has the least record (0 %).

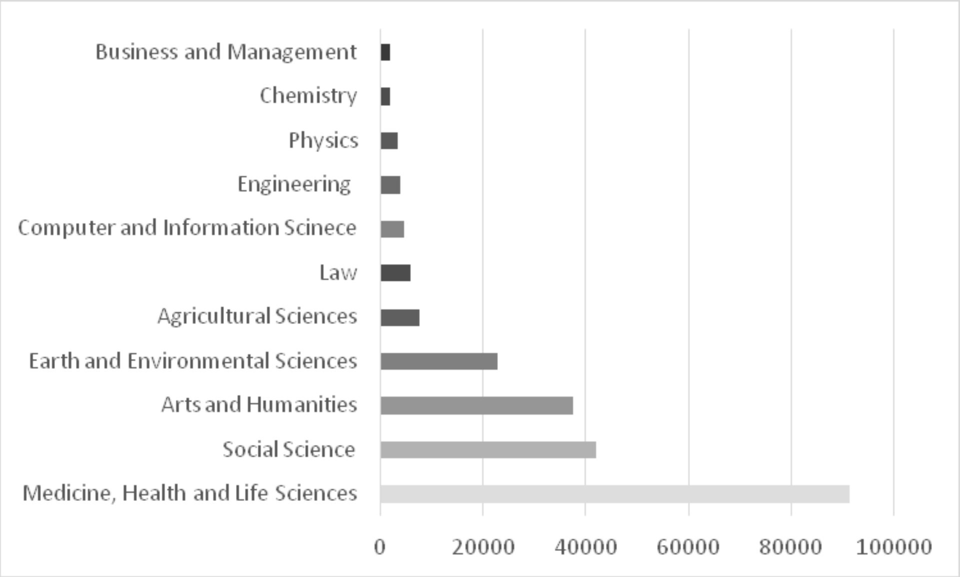

Figure 9: Share of datasets by most common subjects

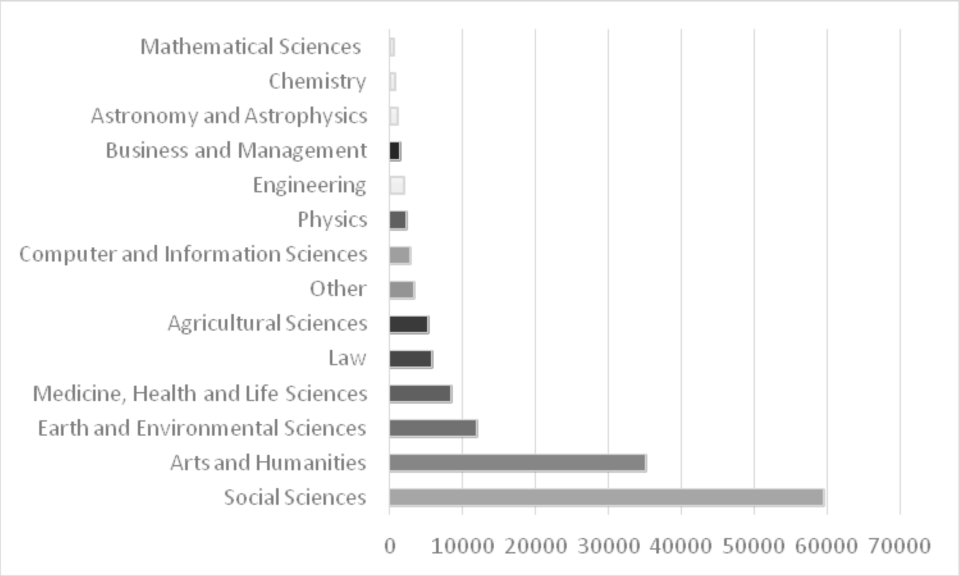

Figure 10: Browse of datasets by most common subjects

Research is not solely associated with only one discipline, but it is also associated with all the disciplines whether it is STEM or AHSS. In this perspective, ensuring the preservation and availability of research data for posterity is one of the primary responsibilities of all disciplines engaged in current research. However, in comparison to medicine, health and life sciences, and social sciences, business and management and chemistry make a smaller contribution in this scenario. In comparison to Mathematical Sciences, the contribution of Social Sciences is higher, according to the browse and search options.

5.8 Authors' contribution

The author is the person(s) who collected the data in the dataset, as well as the person(s) who conducted the research that led to the dataset's production. This individual can but does not have to, be the same as the Depositor (Gupta, Arora and Chakravarty[5]). In the Harvard dataverse project, a total of ten authors contributed. These authors include Master, Daniel M., Stager, Lawrence E., Curtis A. Bradley, Oona A. Hathaway, Digital archive of Massachusetts anti-slavery and anti-segregation petitions, Massachusetts archives, Boston MA, Jack L. Goldsmith, US department of commerce, bureau of the census, geography divisions, GnpIS, Gallup organization and Government of Canada. It is observed that the contribution of Master, Daniel M and Stager, Lawrence E. are maximum (23%) and least contribution is seen of Government of Canada (5%) and Gallup Organization (5%).

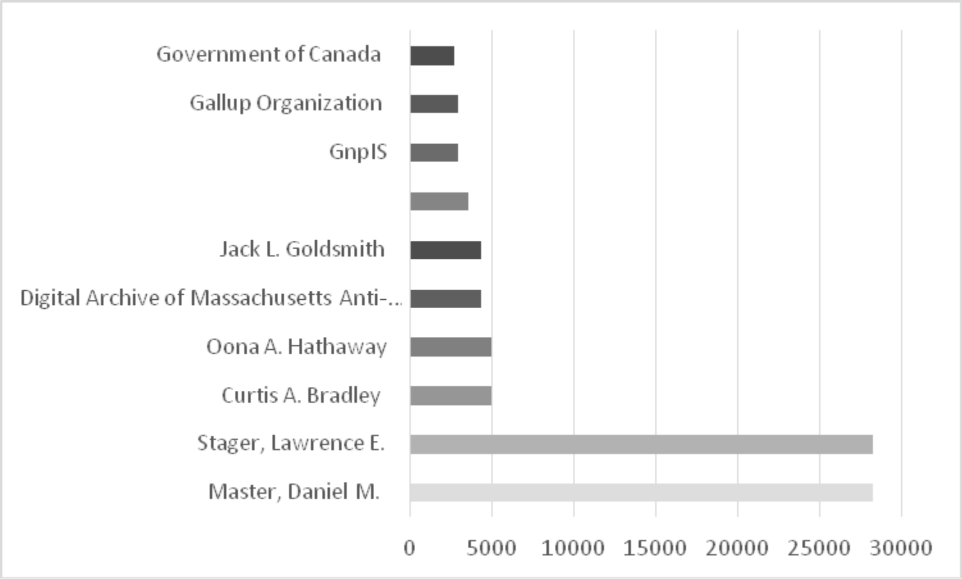

Figure 11: Contribution of Author

Many contests receive dataverse services, of which only a few authors are active and from that few authors only some are continuing contributing. In this regard, it is observed that the Government of Canada and the Gallup Organization contributed less, while Master, Daniel M., and Stager, Lawrence E. contributed more.

5.9 Author affiliation

The "affiliation" in scientific articles refers to the institute to which each author belongs. The affiliation of the author who is the owner of the deposited data on the Harvard dataverse project is found to be from ten different institutions. These 10 author affiliation includes Harvard University, Wheaton College, Harvard law school, yale law school, duke law school, department of national defence, stiching RING, unknown, statistics Canada and Statistique Canada. It is also being seen that the maximum authors or owners of data belong to Harvard University (29%) followed by Wheaton college (26%) and Harvard law school (9%). The least affiliation is being seen from statistics Canada (2%) and Statistique Canada (2%).

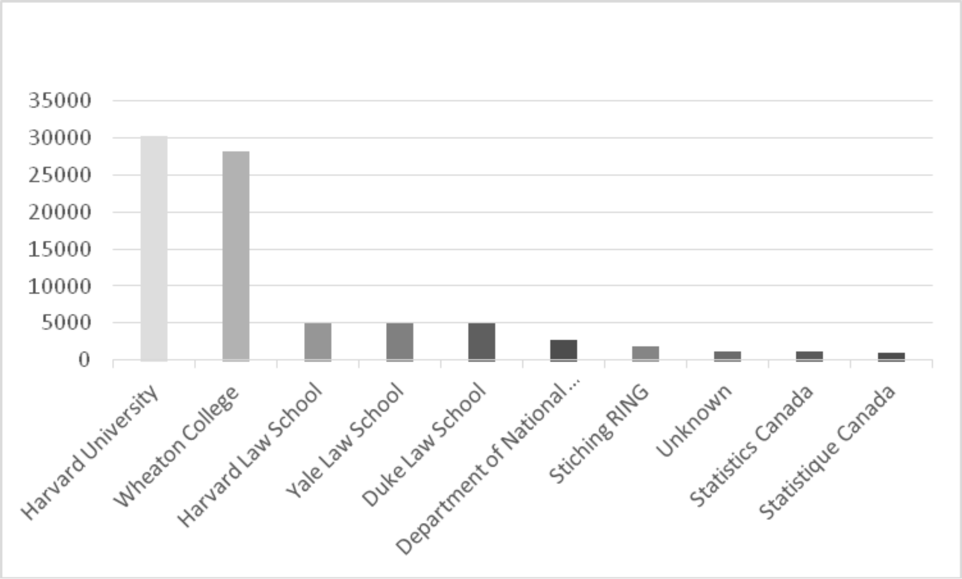

Figure 12. Author affiliation

Harvard Dataverse project is not only allowed to contribute towards research data activities to Harvard university but also allow the same to all organizations and institutions around the globe. In this finding, it is observed that only a few author affiliations are enjoying the benefits of research data management services. From the few, Statistique Canada and statistics Canada are seen as lower contribution as a comparison to highest contributed one i.e. Harvard University followed by Wheaton College and Harvard Law School.

5.10 Metadata sources

Metadata is a piece of accompanying information, either in a separate file or otherwise included in the dataset materials, about a particular dataset or dataverse, including but not limited to the Author's name, publishing date, the title of data contents, description of contents, and other such related information. There are two ways to accompany the information of data i.e. Harvard dataverse and harvested option. Harvard dataverse, accounts for 61% the most common source for metadata while 39% has been as the result of harvesting external data sources.

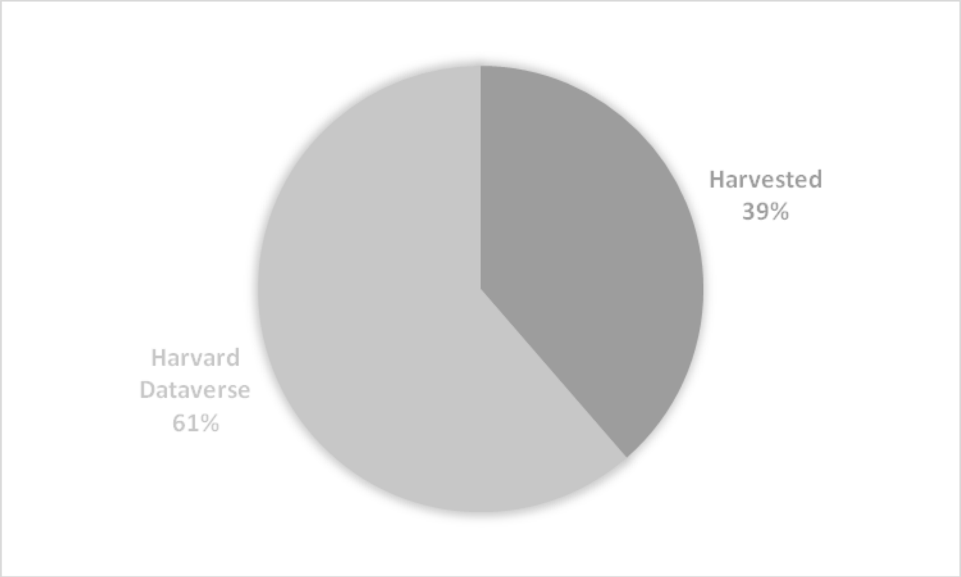

Figure 13. Metadata sources

The two sources are available to extract the information from datasets. Therefore, much of the data management personnel's essential duty is to handle any element of data through uploading their datasets or files. For this case, there is a higher contribution of a source of metadata is Harvard dataverse as opposed to harvested way.

5.11 File type

A total of 20 file types of research data are uploaded on Harvard dataverse website. These file types include data, image, text, unknown, document, tabular data, archive, code, FITS, audio, shape, video, other, model, chemical, binary, test, biosequence, message and multipart. It is observed that the maximum kind of file type is used as data (21%) followed by image (19%) and text (17%). Least is regarded as a multipart file (0%).

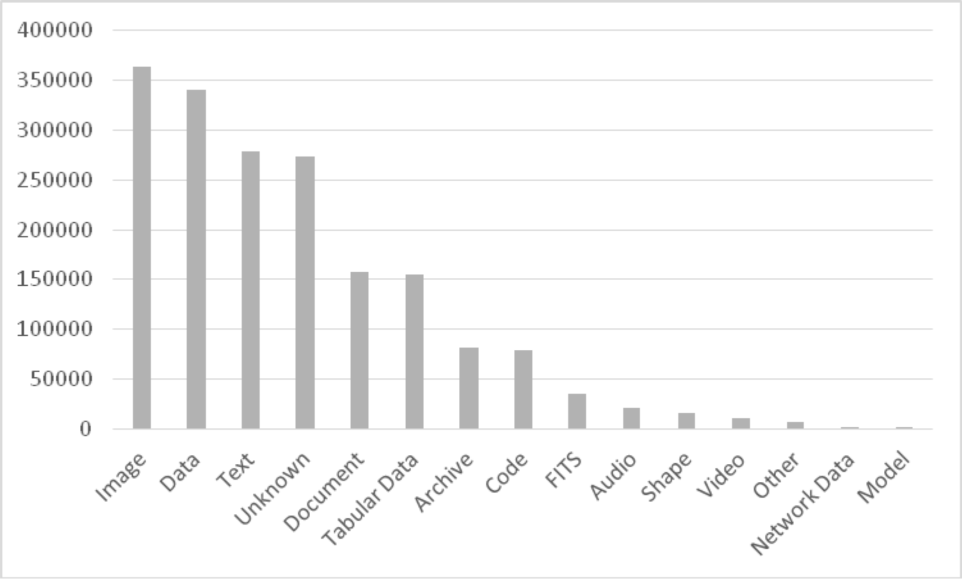

Figure 14. File type

All sort of research data files are essential whether these are images, code, audio or video. Throughout this sense, the organizations, librarians, researchers and publishers are largely responsible for handling all sorts of data files. In this present finding, multipart, message and biosequence enjoy fewer contributions as a comparison to Data and Image file formats.

5.12 Access

Access refers to Data Usage License Agreement that is between a Depositor and a Downloader that governing the limits and restrictions (or lack thereof) of how the downloaded User Submissions can be used. Data Use Agreement is the restricted data usage license agreement option that Harvard Dataverse offers. If applied to a User Submission, the Data Use Agreement will be a legally-binding license contract between the Depositor and any Downloader of that User Submission. Creative Commons licences are made easier with the help of Harvard dataverse. The CC0 Public Domain Dedication allows authors to unambiguously waive all copyright control over their data in all jurisdictions worldwide in the context of a Dataverse installation. Without breaching copyright, data released under CC0 can be freely copied, edited, and disseminated (even for commercial purposes). And it is observed that 97% of data are in open access, can be downloaded whereas 3% of data are restricted that cannot be downloaded but their Differentially Private (DP) Metadata can be accessed for restricted tabular files if the data depositor has created a DP Metadata Release. A Dataverse project also use embargoes to keep file content unavailable until the embargo end date after a dataset version is published. As a result, it won't be possible to download files or view file previews. The result is the same as when a file is restricted, with the exception that no further action is required to end the embargo at the designated date, and requests for file access are not permitted during the embargo. And it observed that only few datasets available through Embargoed then Public followed by embargoed then restricted option.

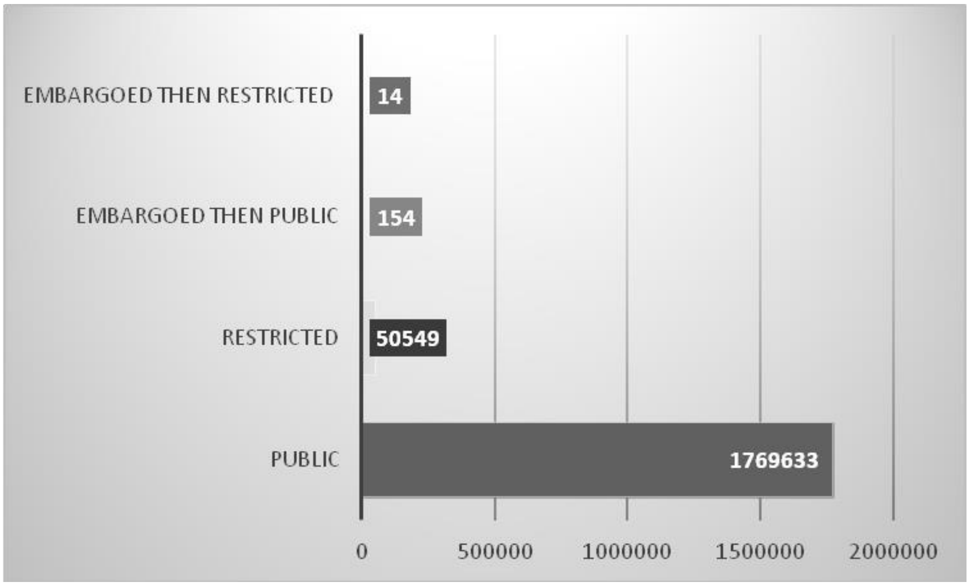

Figure 15: Access

Research datasets can be reusable and interoperable with other datasets if the access of that data is not restricted. In this way, maximum users should get benefits to access datasets for their further research. And it's worth noting that the vast majority of datasets are open to the public or have only a few restrictions.

5.13 Data citation

The Dataverse Project standardises dataset citations to make it easy for academics to publish their findings and receive credit and acknowledgement. The citation is generated and published automatically when researchers create a dataset in a Dataverse repository. The Dataverse Project, as an open-source platform and research data repository, is dedicated to assisting researchers, journals, and organisations in making scientific data accessible, reusable, and open (where possible), which includes applying community-accepted data publication standards (Altman and Crosas[6]). A data citation in a Dataverse repository has seven components i.e. author name(s), year (date published in the Dataverse repository, title, global persistent identifier: DOI or Handle, publisher (repository that published the dataset), version number, universal numerical fingerprint (UNF): for tabular data. The Joint Declaration of Data Citation Principles (2014), a synthesis of all previously existing principles and activities on data citation, is illustrated in the picture below as an example of how the data citation is expressed in a Dataverse repository.

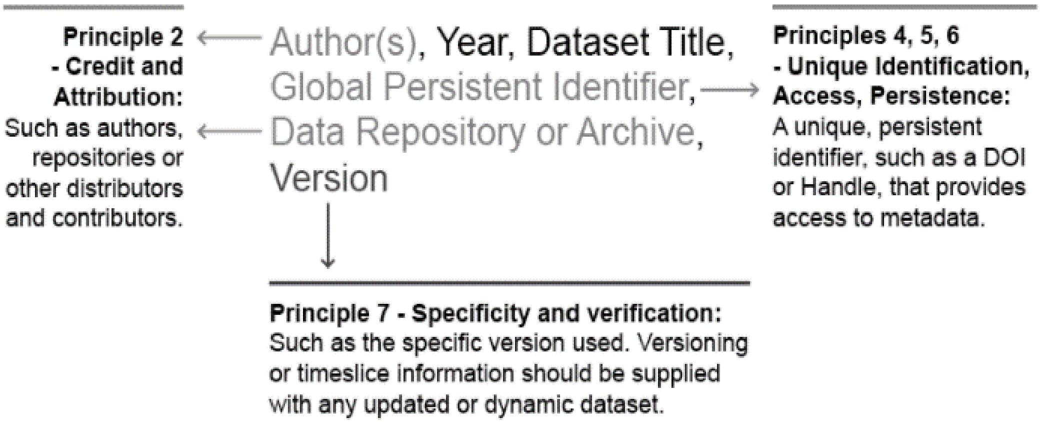
Principle 2
- Credit and
Attribution:
Such as authors,
repositories or
other distributors
and contributors.
Author(s), Year, Dataset Title,
Global Persistent Identifier,
Data Repository or Archive,
Version
Principles 4, 5, 6
- Unique Identification,
Access, Persistence:
A unique, persistent
identifier, such as a DOI
or Handle, that provides
access to metadata.
Principle 7 - Specificity and verification:
Such as the specific version used. Versioning
or timeslice information should be supplied
with any updated or dynamic dataset.

Figure 16: Data Citation Standards

Example of a data citation based on the Joint Declaration of Data Citation Principles (2014).

The citation standard outlined here provide proper acknowledgement to authors as well as permanent identification through the use of global, persistent identifiers in place of URLs, which can change frequently (Arora and Chakravarty[7]). The use of Universal Numerical Fingerprints (UNFs) assures the scientific community that data recovered is identical to that published a decade ago, even though storage media, operating systems, hardware, and statistical programme format have changed.

6. Discussion

According to the findings of this study's analysis of the Harvard Dataverse website, developing nations' contributions to RDM upload and proper support are minimal. The study raises the bar for librarians and other stakeholders when it comes to describing the nuances of RDM operations. The findings suggest that developing countries have few contributions and representatives to work with the dataverse project due to a lack of awareness and fear of data loss. The current study will fill these gaps, allowing librarians and researchers to better manage their data while adhering to the terms of the data sharing agreement. Future research should consider librarians' active participation in informing academics and other stakeholders about the benefits of sharing their data on the dataverse website in particular. The effort should also consider how librarians might upload data from scholars and organisations to the dataverse website within the terms of proper agreements.

7. Conclusions and Recommendations

The Harvard dataverse initiative performed a vital contribution in enhancing and organising research data, according to the report. In particular, it identifies how research data are organized in the dataverse platform in terms of contribution from number of creation of dataverse repository, category of dataverse, upload of datasets and file, number of file downloads, publication of datasets/ dataverse, most approachable subject to share of datasets, most approachable subject according to browse, contribution of authors, contribution of author affiliations, metadata sources, recommended file type, access level of datasets and data citation standards.

The study concludes that maximum research data are organized and uploaded by the authors in Medicine, Health and Life Sciences whereas according to browse/ searchable option, mostly data are organized in the discipline of Social Sciences. The Harvard dataverse seen in upward trend in terms of creation and upload of dataverse, datasets, file and its downloads. The trend of publish datasets are not seen as much high. The contribution of Master, Daniel M. and Stager, Lawrence E. are seen as higher as their affiliation is found from Chicago (USA). According to author affiliation entity, the maximum contribution to organize the data are from Harvard university (USA). So it can also be interpreted that the highly contribution of data developer and then uploader are from USA or developed country. The research project category is very active when it comes to managing research data. The majority of authors waive their copyright licences to make their data more visible and reusable, whereas data and image file formats are largely organised. The Dataverse project also maintains citation requirements to ensure that contributors are properly credited. The report offers the following actions to reinforce and develop RDM practises in a sustainable manner based on the research findings.

1. In order to bridge the divide between Science, engineering and Social Science in particular, study data needs to be exchanged and coordinated across all disciplines
2. Study academics, librarians and other stakeholders need to be aware of RDM
3. It is needed to develop research data repositories on institutions, center then international level.
4. There should be a unique role of librarians to exchange the researchers' and organizations' data on dataverse website within proper legal agreements

8. References and Bibliograpies

[1] Harvard Dataverse. *Dataverse.harvard.edu.* https://dataverse.harvard.edu/ . (*Retrieved on Nov. 08, 2022*).

[2] KING (G). Overview of a Proposed Standard for the Scholarly Citation of Quantitative Data. *IASSIST Quarterly*, *30*(2), 18. 2007.

[3] KING (G). Replication, Replication. *PS: Political Science and Politics*, *28*, 1995. p.444–452. https://j.mp/2oSOXJL. *(Retrieved on Nov. 06, 2022).*

[4] ARORA (S) and CHAKRAVARTY (R). Preserving Global Research Data: Role and Status of Re3data in RDM. *Library Philosophy and Practice (LPP) (ISSN 1522-0222) (E-Journal), 5550*, 2021. p.1–22.

[5] GUPTA (N), ARORA (S) and CHAKRAVARTY (R). Science Mapping and Visualization of Research Data Management (RDM): Bibliometric and Scientometric Study. *Library Philosophy and Practice (LPP) (ISSN 1522-0222) (E-Journal), 6096*. 2021. 1–24.

[6] ALTMAN (M) and CROSAS (M). The Evolution of Data Citation: From Principles to Implementation. https://iassistquarterly.com/public/pdfs/iqvol371_4_altman.pdf. *(Retrieved on Nov. 02, 2022).*

[7] ARORA (S) and CHAKRAVARTY (R). Making research data discoverable: an outreach activity of Datacite. *Library Philosophy and Practice (E-Journal), 5199*. 2021b. https://digitalcommons.unl.edu/libphilprac/5199/ *(Retrieved on Oct. 02, 2022).*

About Authors

Surbhi Arora, JRF (surbhi03@pu.ac.in) is based at the Library and Information Science department, Panjab University, Chandigarh.

Professor Rupak Chakravarty (rupak@pu.ac.in) is based at the Library and Information Science department, Panjab University, Chandigarh.