DLINE JOURNALS

International Journal of Information Studies

Print ISSN: 1911 8408
Online ISSN: 1911 8414

IJIS 2024; 16 (4)
https://doi.org/10.6025/ijis/2024/16/4/126-134

Research Data Management and Curation: What Librarians need to know

[1]Femy Francis
Librarian, Jain University, Kochi campus. India
femyfrancis@jainuniversity.ac.in

[2]Gopakumar V
Head, Knowledge Centre
Digital University, Kerala. India
librarian@duk.ac.in

Received: 20 June 2024
Revised: 6 August 2024
Accepted: 14 August 2024

ABSTRACT

This article provides an in-depth exploration of Research Data Management (RDM) and data curation, emphasising their importance for modern academic research. It defines RDM as a systematic approach to organising, storing, and safeguarding research data, ensuring compliance with legal, ethical, and funding body regulations. Data curation is the ongoing management of data to ensure its accessibility and quality over time. Key components of RDM and curation include data management planning, data collection and documentation, data storage and security, data preservation, and data sharing and access. The article also discusses the benefits of RDM and curation, such as enhanced reproducibility, data reuse, and collaboration, while addressing the challenges, including resource intensity and technical expertise gaps. Best practices, such as early planning, collaboration, and regular review, are recommended to overcome these challenges and ensure the integrity, accessibility, and longevity of research data. The article underscores the critical role of librarians in supporting effective RDM and curation practices, contributing to the advancement of knowledge and scientific discovery.

Keywords: Research Data Management, RDM, Data Curation

1. Introduction

Modern technologies are crucial to many facets of higher education, especially Scholarly communication. Knowledge dissemination in academia and society happens through various methods and channels of scholarly communication. Spichtinger and Siren (2018) define research data as "recorded factual material commonly accepted in the scientific community as necessary to validate research findings, including data sets used to support scholarly publications." It is a well-known fact that researchers are interested in all types of data, regardless of their source or format. Hence, it is also considered as an academic product. Any data collected as part of research can be brought under this category.

Research data includes a wide variety of information formats, including spreadsheets, field notebooks, diaries, audio and video recordings, documents, models, algorithms, scripts, workflows, protocols, software, standard operating procedures, and more.

The advancement of scientific knowledge depends significantly on the thoughtful handling, shielding, storage, and re-utilisation of the research data generated by academic scholars. Research data management (RDM) and data curation are the core elements of the current research process. This process facilitates the collection and management of data in a manner that allows easy access and reuse.

In RDM, research data must be collected, assessed, discarded, documented, conserved, retrieved, reused, and modified. Data curation ensures that data is utilised for its intended purpose, starting from its production.

Definitions

Research Data Management (RDM) is the systematic approach to arranging, classifying, storing, and safeguarding data. It comprises principles and methodologies that ensure data accuracy, accessibility, and sustainability.

Research Data Management (RDM) refers to managing and maintaining data during a project. The process includes planning, collecting, organizing, managing, storing, securing, backing up, preserving, and sharing data. It ensures that research data is handled according to legal, regulatory, ethical, and funding body regulations (Whyte, A. & Tedds, J., 2011).

Data Curation: Data curation refers to the continuous and systematic management of data from generation and storage to obsolescence. This entails ensuring that data remains easily accessible and of superior quality for future use and archiving.

"Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time" (Craigin et al, 2007).

"Data Curation is the encompassing work and actions taken by curators of a data repository to provide meaningful and enduring access to data" (Johnson, L.R. et al., 2018)

Key Components of RDM and Curation

Research data management (RDM) encompasses the procedures for gathering, handling, preserving, sharing, and using the data necessary to support research.

Important components of curation and RDM include:

• Strategies for efficiently storing and backing up data to prevent loss and damage;

• Ways to arrange the information, implementing version control to changed files and employing logical folder structures and appropriate file names;

• How to incorporate quality controls into the data workflow to preserve the data's integrity and reduce the likelihood and consequences of error;

• How to properly describe the data so that scholars and other interested parties can comprehend what the information contains, how it was created, acquired, and handled

• How to handle sensitive information in a way that complies with the Data Protection Act's regulations and your moral responsibility;

• How can data be shared and maintained, usually through appropriate data repositories, so other scholars can consult and re-use them?

1. Data Management Planning

A Data Management Plan (DMP) is a formal document that outlines the methods and tools to manage research data and other research materials throughout the project, from proposal to publication and beyond. The DMP should include details on:

- Data collection and analysis
- Describing the data with metadata and documentation
- Storage, backup, and general management of the data
- Preservation and archiving
- Sharing and reuse
- Ethical and legal compliance

Therefore, the Data Management Plan (DMP) should serve as a dynamic document that clearly outlines the procedures for handling data throughout and beyond the duration of a research project.

Many funding organizations require a data management plan (DMP) to be included in grant applications. It is highly recommended that a DMP be developed for every research project that involves gathering primary data.

2. Data Collection and Documentation

Data collection and documentation are two crucial steps in research data management (RDM).

Data Collection

Data collection refers to gathering information relevant to your research question.

The best-standardised format for the research data depends on the data type that was collected and used. Along with accuracy, efficiency, and participant privacy, the factors to consider include,

- The type of data
- The size of the data
- Interoperability needs
- Long-term preservation requirements

Some examples of standardised formats in RDM include,

- Comma-separated values (CSV)
- JavaScript Object Notation (JSON)
- Portable Document Format (PDF)
- Tag Image File Format (TIFF)
- NetCDF (Network Common Data Form)

Data Documentation

Data documentation refers to creating clear and comprehensive information about the collected data. This information serves as a guide for anyone trying to understand, interpret, and reuse the data.

Steps for Data Documentation include

• **Metadata:** This is the process of capturing key details about the data, such as variable names, units of measurement, and data collection methods. Metadata should be "data about the data."

• **Codebooks:** Creating codebooks to define codes used in surveys or categorise observational data. This guarantees uniformity and accuracy in the interpretation of data.

• **File Naming Conventions:** Utilising precise and standard file naming conventions that accurately represent the content and organisation of the data files. This facilitates the process of locating and categorising the data easily.

• **Methodological Documentation:** Providing a comprehensive explanation of the study methodology adopted to get the data, including any particular protocols or procedures observed.

3. Data Storage and Security

Effective data storage and robust security measures are essential for Research Data Management (RDM).

Data Storage

Data storage pertains to the actual location where the study data is preserved. Storage options include local servers, external hard drives, network-attached storage (NAS), and cloud storage.

When selecting secure data storage, it is important to consider the following factors:

• **Durability and Reliability:** When choosing storage solutions, emphasise durability and reliability by choosing options that include backup features and redundancy measures. This will help minimise the danger of data loss.

• **Scalability:** Select storage options that can handle more research data as it develops.

Accessibility: Ensure the selected storage solution enables authorised users to access the data as required easily.

• **Security Measures:** Choose storage options with strong security features, such as access control, encryption, and data privacy safeguards.

Data Security

Data security includes the methodologies and technological measures to secure research data against unauthorised attacks, modification, or deletion. The primary factors to consider for data security are:

• **Access Control:** Limit data access to authorised users with a valid purpose for it to conduct research. Implement strategies such as user authentication and access rights.

• **Data Encryption:** Employ encryption techniques to hide the data, making it inaccessible without the proper decryption key during storage and transmission. This provides an additional level of security in the event of unauthorised access.

• **Data Backup and Recovery:** Establish a systematic process of creating and storing duplicate copies of data to safeguard against potential loss in the event of a failure in the primary storage system. Conduct periodic testing of the backup and recovery protocols to verify their optimal performance.

• Utilise Data Loss Prevention (DLP) technologies to decrease the risk of unauthorised data exfiltration via email, external drives, or other communication channels.

• **Security Awareness:** Get acquainted with the most effective methods for ensuring data security and integrity. This entails employing robust passwords, exercising caution when opening suspicious emails, and being aware of the information disclosed online.

4. Data Preservation

Data preservation, a core tenet of RDM, focuses on ensuring the long-term accessibility and usability of research data. Data preservation includes storing data and maintaining its quality, integrity, and understandability over time.

Preserving data ensures compliance with requirements like reproducibility, can be used as secondary research data and should comply with the legal requirements.

Data preservation challenges include technological obsolescence, data degradation, metadata loss, etc.

Strategies to be considered for data preservation are as follows:

• **Should be in standardised Formats:** Storing data in well-established, non-proprietary formats like CSV or TIFF increases its long-term readability by different software programs.

• **Metadata Management:** It is crucial for future understanding to create and maintain comprehensive metadata that describes the data, its collection methods, and any relevant context.

• **Migration and Refreshment:** Periodically migrating data to newer storage formats or refreshing storage media can help prevent data degradation and ensure continued accessibility.

• **Data Repositories:** Depositing data in trusted, long-term data repositories provides a secure and well-managed environment for data preservation and sharing. Data repositories offer expertise in data curation, storage, and access, ensuring data is preserved according to best practices. Repositories possess the necessary infrastructure and resources to effectively preserve data for long periods, reducing the potential hazards of technical obsolescence and data degradation. Data repositories frequently offer data discovery and retrieval features, facilitating data sharing and enhancing the exposure of research findings. When choosing a repository, several factors might impact the decision, including the academic discipline, the type of digital content, the file size, institutional regulations and standards, national rules, funding agency guidelines, and access limitations.

Below are a few examples of data repositories:
Two examples of general-purpose repositories are Zenodo and Figshare. The project repositories include Open Science Framework (OSF), Research Equals, Octopus, and CRAN for R-Packages.

Common data repositories include Dryad, Dataverse, etc
Prominent publishers provide several repositories for the storing and retrieval of material. Some examples of recommended repositories are PLOS ONE, Springer Nature Data repository advice, and Elsevier's Public repositories etc.

Therefore, researchers may guarantee the long-term value and accessibility of their research data by using efficient data preservation procedures. This enhances the scientific record and promotes a culture of sharing knowledge and collaboration, accelerating scientific advancement.

5. Data Sharing and Access
There are numerous justifications for openly disseminating scientific data.

• To provide the potential for complete replication of a scientific work.

• The aim is to avoid redundant work and accelerate scientific advancement. Insufficient dissemination of research findings through limited publication can result in significant squandering of research funding and hinder the professional advancement of researchers.

• The purpose is to enhance cooperation and enhance the influence and excellence of scientific research.

• To ensure the widespread accessibility of research findings, given that public funds frequently financed research.

Data should be disseminated in a formal, transparent, and catalogued repository whenever feasible to ensure long-term accessibility. When choosing data repositories, go for ones that provide a DOI (digital object identifier) for your data. This will facilitate the process of others citing your data. Linking the data and other research objects will ensure the researcher receives credit for the work and increase the visibility of the research outputs.

Implementing suitable licenses for data makes it clear to others what actions they can take with the data. The often-utilized licenses include Creative Commons, Open Government License, and Open Data Commons Attribution License. To get maximum value from data sharing, it should be essential to link research data with licences and vice versa, allowing others to understand research work better.

6. Data Curation Activities

Data curation is essential for the efficient management of research data. Data curation is the continuous effort to guarantee the quality, usability, and accessibility of data across the whole research process. The successful implementation of the data curation process necessitates the expertise of data librarians, researchers, and other consumers, including professional skills, education, domain knowledge, and IT abilities. Quality Control, Data integration, Annotation, and Enrichment are all important aspects of data curation in RDM.

Quality Control: Data curation is essential for regularly examining data to ensure its accuracy, consistency, and completeness. This task entails identifying and rectifying flaws, inconsistencies, or omissions in the data. Quality control guarantees the reliability and suitability of the data for analysis.

Data Integration: Data integration is merging data from several sources to build comprehensive datasets, a highly effective method for enhancing research. This may entail integrating data from surveys, experiments, or external databases. Nevertheless, data integration necessitates meticulous planning to guarantee harmonisation and resolve possible disparities among datasets.

Annotation and Enrichment: Data annotation and enrichment include enhancing the value of data by supplementing it with extra information or establishing connections to relevant datasets. This process can be considered a type of data transformation. Annotations can encompass elucidations of particular data points, explications of variables, or contextual details on the data-gathering process. Connecting data to other datasets enables academics to investigate correlations and get a more profound understanding.

7. Benefits of RDM and Curation

Reproducibility: Reproducibility is guaranteed through the use of effective RDM practices and careful data curation operations, which guarantee the quality, correctness, and comprehensiveness of the data. This facilitates the replication of the study by other researchers, enabling them to validate the findings and expand upon the research. Strengthening the reliability of scientific research maintains trust in scientific advancement.

Data Reuse: Data reuse can be facilitated by ensuring that data is thoroughly recorded, kept,

and easily accessible through RDM and curation. This allows researchers to utilise the data for novel research inquiries. This optimises the effectiveness of data collection endeavours and expedites the process of scientific discovery. Therefore, it enables data utilisation for new research inquiries, enhancing the value obtained from the initial data-gathering endeavours.

Collaboration: Efficient research data management promotes collaboration among researchers. Comprehensive and easily obtainable data enables academics from various universities and fields to collaborate, examine merged datasets, and produce novel findings. Therefore, it facilitates cooperation among researchers by offering readily available and thoroughly documented data. Using this collaborative strategy can result in revolutionary breakthroughs that would be unattainable through individual endeavour.

Compliance: Numerous research institutions, funding agencies, and publications enforce RDM criteria. Researchers can assure compliance with regulations and prevent delays or funding restrictions by employing RDM techniques and engaging in data curation activities.

Efficiency: RDM facilitates systematic data management throughout the research process. This implies that researchers allocate less time to locating misplaced data or grappling with incongruities. Convenient access to thoroughly documented data simplifies research processes, enabling researchers to concentrate on analysis and interpretation, resulting in improved research efficiency.

8. Challenges in RDM and Curation

Although RDM and data curation provide significant advantages, they also offer certain obstacles, as outlined below.

· **Resource** -intensive activities that require enormous time, effort, and funding: To effectively implement RDM, one must allocate sufficient time and effort to planning, documentation, organisation, and data curation. Researchers may already be overburdened, and obtaining specific funding for RDM operations can be challenging.

· **Technical Expertise - Skills Gap:** Researchers may lack the necessary technical proficiency to perform intricate data curation tasks, such as data cleansing, transformation, or format migration, which are essential for the long-term preservation of data. Training and support are essential for closing this skills gap.

· **Cultural Barriers in Data Sharing:** Certain researchers exhibit reluctance to publicly share data due to concerns of misinterpretation, misuse, or the potential loss of recognition for their contributions. Establishing a culture that promotes confidence in sharing data and resolving these issues by implementing data use agreements and adhering to correct attribution processes can be beneficial.

· **Intellectual Property (IP) Considerations:** Worries regarding safeguarding intellectual property rights can impede data exchange. Well-defined institutional policies and guidance around data ownership and sharing can assist in addressing these problems.

· **Further difficulties:** Researchers must carefully consider data privacy requirements, particularly when handling sensitive data. Adhering to anonymization and secure storage methods is crucial.

· **Ensuring long-term sustainability** involves continuously dedicating effort and resources to preserving data over an extended period. It is crucial to carefully select dependable storage options and make arrangements for transitioning to different formats.

· **Insufficient standardisation:** The lack of uniformity in data formats and storage methods among various institutions and disciplines might make data sharing and integration difficult. Continuously advocating for the establishment of data standards and the capacity of different systems to work together is a persistent endeavour.

9. Best Practices

The following are some exceptional best practices for data curation and Research Data Management (RDM):

Early Planning: By focusing on data management from the outset of a research project, it is possible to account for storage requirements, allocate funds for data curation activities, and create a data-sharing strategy. This proactive approach prevents the need for a last-minute scramble during the research process. Consequently, it will reduce the risk of data loss or mismanagement. It will ensure that data is collected and documented to enable analysis and future reuse and save time and resources in the long term.

Training and Support: Researchers may lack the requisite skills to manage data effectively. Training gives them the necessary knowledge and resources to manage data cleansing, transformation, curation, and preservation tasks. Consequently, it enhances the overall quality of research data, promotes a culture of data responsibility within the research community, and enables researchers to manage their data effectively.

Collaboration: The expertise of librarians, data scientists, and IT professionals is a valuable asset to the field of RDM. Through collaboration, researchers can capitalise on this expertise and create comprehensive data management strategies. Consequently, collaboration facilitates interdisciplinary collaboration, provides researchers with access to specialised knowledge and resources, and results in more sustainable and robust RDM practices.

Application of Standards: The adoption of common data formats, metadata standards, and documentation practices facilitates the sharing, integration, and repurposing of data across various research projects and disciplines. Consequently, it can mitigate obstacles to collaboration, improve the discoverability and accessibility of data, and guarantee that data remains interpretable over time.

Regular Review: The discipline of RDM is in a state of perpetual evolution as new technologies and methodologies are introduced. Researchers can adapt and integrate these advancements by regularly reviewing data management practices. Consequently, it can guarantee that data management practices remain effective and efficient, enable researchers to capitalise on new resources and tools, and maintain data management by evolving regulations and best practices.

10. Conclusion

Research data management led duplication in research work and helps the researcher improve using current research data. Coincidentally, more than one researcher may follow the same research methods and do duplicate exercises to collect necessary corresponding data for their study. However, they have different individual research objectives and outcomes. Nowadays, most publishers and funding agencies have to make policies on RDM, and authors and researchers must manage the data collected as part of their research work.

Data curation activity of managing and promoting research data from its point of creation ensures its usage for contemporary purposes. It helps in the discovery and re-use of data. The best practices of the data curation process are like a cycle that begins with preparation, understanding the needs and arrangement of infrastructure prerequisite, data receiving, cleaning, documentation, metadata creation, preservation, data publishing, data access, evaluation of the usage and data citations. IT-oriented competencies serve to develop user-friendly data repositories and data cleaning. Academic research knowledge helps to understand the research activities and the need for data curation assistance by data generators. Appropriate use of software and application tools facilitates the development and practice of data curation.

Research data management and curation are essential for ensuring the integrity, accessibility, and longevity of research data. By implementing best practices and overcoming challenges, researchers can enhance the impact and value of their data, contributing to the advancement

of knowledge and scientific discovery.

References

[1] Andrikopoulou, A., Rowley, J., Walton, G. (2021). Research data management (RDM) and the evolving identity of academic libraries and librarians: A literature review. *New Review of Academic Librarianship, 28*(4), 1–13. https://doi.org/10.1080/13614533.2021.1964549

[2] Cox, A. M., Pinfield, S. (2013). Research data management and libraries: Current activities and future priorities. *Journal of Librarianship and Information Science, 46*(4), 299–316. https://doi.org/10.1177/0961000613492542

[3] Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., Stewart, C. (2018). How important is data curation? Gaps and opportunities for academic libraries. *Journal of Librarianship and Scholarly Communication, 6*(1). https://doi.org/10.7710/2162-3309.2198

[4] Pinfield, S., Cox, A. M., Smith, J. (2014). Research data management and libraries: Relationships, activities, drivers, and influences. *PLoS ONE, 9*(12), e114734. https://doi.org/10.1371/journal.pone.0114734

[5] Spichtinger, D., Siren, J. (2017). The development of research data management policies in Horizon 2020. In *Research data management policies* (pp. 11–24). De Gruyter. https://doi.org/10.1515/9783110365634-002

[6] Wiljes, C., Cimiano, P. (2019). Teaching research data management for students. *Data Science Journal, 18*. https://doi.org/10.5334/dsj-2019-038