Cybrarians Journal

E-ISSN 1687-2215
Home Page: https://journal.cybrarians.info/index.php/cj
Issue 71, April 2024

Managing the preservation and retrieval of research projects data in specialized research centers in Egypt : a literature review

Articles – Full text

Maha Saad Mahmoud Mohamed
Assistant teacher, Information Science Department, Helwan University
Maha.Mohamed@arts.hewlan.edu.eg

Abstract

The study aimed to monitor the most important findings of intellectual production from Arab and foreign contributions closely related to the subject of "management of preserving and retrieving research project data". The study followed the method of objective scientific review. Relying on the bibliographic approach to study the substantive and temporal trends as well as the descriptive approach in its analytical style to clarify the full picture of the intellectual production that dealt with issues of research data management and related topics regarding the preservation and retrieval of research data as well as the role of libraries and data specialists / or information in providing research data services ; In order to be guided by it and to know the purpose of these studies and the methods used, and to stand on the most important findings of these studies, and to stand on the similarities and differences between the current study and previous studies. The study concluded by presenting the most important findings and recommendations for reviewing intellectual production.

Keywords

Research data, research projects, data repositories.

1. The Nature of Research Data

A foundational understanding of research data, its various forms, and its lifecycle is essential for effective management.

Defining Research Data

Research data is broadly defined as the factual records, materials, and observations generated or collected during a research project to validate findings. While definitions vary slightly, a consensus emerges that research data constitutes the primary inputs and evidence of the scientific process.

• Core Concepts: It is described as the "underlying basis for scientific research" (Kim J., 2013; Nielsen & Hjørland, 2014) and the "primary outputs... that constitute the basis of the initial results of this research" (Dora & Kumar, 2015).

• Comprehensive View: Several studies (e.g., Corti et al., 2014; Tripathi et al., 2017; Si et al., 2015) converge on a definition of data as any information created or collected by researchers, whether in digital or non-digital formats. This includes text, documents, images, survey data, interview transcripts, code, and more.

• Validation and Verification: The data serves as the "source and resource necessary for all research results" (Al-Askari, 2018) and the "underpinning of the claim" (Elsayed & Saleh, 2018), enabling the verification and replication of research findings.

Open Research Data

Open research data is a subset of research data that is made freely available for access, use, and redistribution without significant restrictions. This concept is closely tied to the broader open access movement.

• Accessibility: It is defined as "information available to everyone for any purpose at no cost" (Childs et al., 2014).

• Reuse and Innovation: The core principle is that data should be reusable to generate further discoveries and insights. Its management has become a common practice to facilitate the creation of new research outcomes and to preserve them long-term (Tayeh et al., 2018).

• FAIR Principles: Effective data sharing is increasingly guided by the FAIR principles, which state that data should be Findable, Accessible, Interoperable, and Reusable (Zakaria, 2020).

Types and Classifications of Research Data

Research data is highly heterogeneous, varying by discipline, methodology, and format.

• Primary vs. Processed: Data can be categorized as raw, unprocessed data collected directly from sources, intermediate data generated during analysis, and final processed data supporting published results (Badr, 2020).

• Format-Based Classification: The repository registry Re3data.org identifies over a dozen data types, including raw data, images, structured text, databases, source code, and audiovisual materials (Schöpfel et al., 2017).

• A Functional Typology (Spichtinger & Siren, 2018):

◦ Descriptive/Bibliographic Data: Metadata found in indexes and catalogs.

◦ Data Underlying Publications: Data necessary to validate the results presented in a publication.

◦ Curated Data: Data collected and organized in thematic databases or collections.

◦ Raw Data and Datasets: Primary data that is often stored on local hard drives or institutional servers.

The Research Data Lifecycle (RDL)

The RDL is a conceptual model that describes the stages data moves through from its creation to its potential reuse. Understanding this lifecycle is critical for planning RDM activities. Different models exist, including those based on the individual researcher, the organization, or the broader research community (Carlson J., 2014). The key stages and their associated RDM tasks are summarized below.

Lifecycle Stage	Associated Research Data Management Tasks
Creating Data	Research design, creating a data management plan, identifying existing data sources, collecting data (observation, experiments, surveys, simulation), and creating/capturing metadata.
Processing Data	Data entry, transcription, digitization, validation and verification, anonymization of data, data description, and managing/storing the data.
Analyzing Data	Data interpretation and derivation, generating research outputs, and creating permanent versions of the data for preservation.
Preserving Data	Migrating data to best format, selecting an appropriate medium, data storage, creating documentation and metadata, and archiving the data.
Giving Access to Data	Distributing and sharing data, controlling access, establishing copyright, and publishing.
Reusing Data	Follow-up research, reviewing and validating research, new research studies, and for teaching and learning.

Source: Adapted from Perrier et al., 2017; Krahe et al., 2020; Al-Askari, 2018.

2. The Framework of Research Data Management (RDM)

RDM is the active and systematic organization, storage, preservation, and sharing of data created during a research project.

Defining Research Data Management

RDM is a comprehensive term covering all activities and processes related to data throughout its lifecycle.

• Scope: It involves the "organization, documentation, curation, preservation, and access provision for data" (Whyte & Tedds, 2011) and encompasses data creation, processing, analysis, preservation, sharing, and reuse (Borghi et al., 2018; Mushi et al., 2020).

• Objective: The primary goal is to ensure that research data is "well-organized, well-documented, preserved, and accessible" (Corti et al., 2014), thereby maximizing its value and potential for reuse.

• Institutional Framework: Effective RDM requires an institutional infrastructure composed of policies, technological systems, metadata standards, and support services (Qin, 2013).

Importance and Benefits of RDM

Proper RDM yields significant benefits for researchers, institutions, and the wider scientific community.

• Efficiency and Collaboration: It facilitates research activities and enhances collaboration and data sharing among researchers (Jennex & Bartczak, 2015).

• Verification and Integrity: It ensures the long-term preservation of primary research data, allowing for the verification of results and protecting against data loss (Abdul Kadir & Yuns, 2017).

• Increased Impact: Data sharing enabled by RDM can increase the visibility and impact of research (Manu, 2018).

• Compliance: It is often a requirement of funding agencies, who mandate the creation of DMPs and the sharing of data generated from funded projects (Couture et al., 2018).

RDM Systems and Institutional Strategies

Institutions are developing structured approaches to support RDM, often combining technology with policy and services.

• Electronic Laboratory Notebooks (ELNs): These digital platforms are replacing traditional paper notebooks, offering significant advantages for managing, storing, and sharing research data while protecting intellectual property (Baykoucheva, 2015).

• The Institutional Portfolio Model: Pinfield, Cox, & Smith (2014) outline a comprehensive model for institutional RDM support comprising six key components:

◦ Strategies: Defining the overall vision and priorities for RDM within the institution.

◦ Policies: Establishing formal rules for data management, storage, access, and intellectual property.

◦ Guidelines: Providing practical, advisory instructions on best practices for RDM.

◦ Processes: Defining the specific workflows for data handling throughout the RDL.

◦ Technologies: The infrastructure for data storage, preservation, and discovery (e.g., repositories).

◦ Services: The human support offered to researchers, such as training, consultation, and technical assistance.

Data Management Plans (DMPs)

A DMP is a formal document created at the start of a research project that outlines how data will be managed during and after the project.

• Purpose: It serves as a living document that specifies how data will be created, documented, stored, shared, and preserved (Sanjeeva, 2018; Al-Askari, 2018).

• Content: A DMP addresses key questions about data types, metadata standards, storage solutions, access policies, and long-term preservation strategies (Al-Sarraj, 2021).

• Function: It promotes critical thinking about data from the outset, ensures compliance with funder requirements, and lays the groundwork for making data FAIR.

3. Preservation, Sharing, and Retrieval of Data

The ultimate goal of RDM is to ensure that valuable data is preserved, made accessible for sharing, and can be easily retrieved for future use.

Data Preservation and Availability

Data preservation involves processes to ensure the long-term accessibility and usability of digital data.

• Preservation vs. Backup: Digital preservation is distinct from simple backup. While a backup is a copy for short-term disaster recovery, digital preservation is a managed lifecycle process aimed at maintaining the integrity, authenticity, and accessibility of data over time (Higgins, 2008; Kruse & Thestrup, 2014).

• Importance: Preservation is crucial because some research data is irreplaceable. It safeguards the scholarly record, enables future verification, and protects the investment made in research (Tenopir et al., 2011).

• Requirements: Effective preservation requires a robust technical infrastructure, appropriate metadata to describe the data, and clear policies for selecting what data to preserve (Witt, 2008).

Data Sharing and Reuse

Data sharing is the practice of making data used for scholarly research available to others.

• Drivers: The movement is driven by funder mandates, journal policies, and a cultural shift toward greater transparency and collaboration in science (Dora & Kumar, 2015).

• Benefits: Sharing data facilitates the verification of findings, promotes new research by enabling reuse, increases the citation impact of the original study, and accelerates scientific discovery (Pinfield et al., 2014; Piwowar & Vision, 2013).

• Enabling Sharing: Successful sharing depends on proper data documentation, storage in trusted repositories, and the application of principles like FAIR to ensure data is discoverable and usable by others (Childs et al., 2014; Borghi & Van Gulick, 2018).

Data Retrieval and the Role of Metadata

Data retrieval is the process of finding and accessing relevant datasets. This capability is fundamentally dependent on high-quality metadata.

• Discovery and Access: Data retrieval encompasses both "data discovery" (the ability to find the existence of data) and "data access" (the ability to obtain the data itself) (Contaxis et al., 2022).

• Metadata's Critical Function: Metadata—or data about data—is essential for data retrieval. It provides the descriptive, structural, and administrative information needed to discover, understand, and reuse a dataset (Gomez et al., 2016; Quarati & Raffaghelli, 2020).

• Types of Metadata (Kruse & Thestrup, 2014):

◦ Descriptive: Describes the resource for discovery and identification (e.g., creator, title, keywords).

◦ Technical: Provides information about the file format and software needed.

◦ Administrative: Manages the resource, including rights management and preservation information.

4. The Role of Libraries and Information Professionals

Research libraries and their staff are increasingly central to institutional RDM efforts, providing a range of services and expertise to support researchers.

Research Data Services (RDS)

RDS refers to the portfolio of services offered by an institution, typically a library, to assist researchers in managing their data.

• Scope of Services: These services can range from introductory training and consultation on creating DMPs to providing technical infrastructure like data repositories (SANJEEVA, 2018).

• Service Levels: Services are often categorized as:

◦ Educational/Advisory: Training workshops, online guides, and one-on-one consultations.

◦ Technical: Providing access to storage, data repositories, and software tools.

◦ Curation: Active assistance in preparing datasets for deposit, creating metadata, and ensuring long-term preservation.

The Evolving Role of the Research Library

Libraries are leveraging their expertise in information organization and preservation to become key partners in the research process.

• Core Functions: Libraries are well-positioned to lead in developing institutional RDM policies, educating researchers on best practices, and managing data repositories (Cox et al., 2019).

• Support Across the Lifecycle: They can provide support at every stage of the RDL, from helping researchers find existing data for a new project to curating and preserving data for long-term access after a project is complete (Abdel Hameed, 2023).

The Librarian as Data Specialist

The responsibilities of information professionals are expanding to include new data-centric roles.

• New Roles: Librarians are becoming data consultants, trainers, and curators. Their tasks include helping researchers select repositories, create metadata, understand data citation, and navigate intellectual property issues (Si et al., 2015; Latham, 2017).

• Essential Skills: Success in these roles requires a blend of skills:

◦ Technical Skills: Understanding of data formats, repositories, and metadata standards.

◦ Data Literacy: The ability to work with and understand data from various disciplines.

◦ Communication Skills: The capacity to advise and train researchers from diverse backgrounds.

◦ Policy Knowledge: Familiarity with funder mandates, copyright, and institutional policies.

Data Repositories

Data repositories are digital archives designed for the storage, preservation, discovery, and sharing of research data.

• Function: They provide stable, long-term homes for datasets, assign persistent identifiers (like DOIs) to facilitate citation, and create metadata to make data discoverable (Pinfield et al., 2014).

• Examples: Global registries like Re3data.org help researchers find appropriate disciplinary or institutional repositories. Platforms like CKAN are used to build open data portals (Gomez et al., 2016; Quarati & Raffaghelli, 2020).

• Institutional Strategy: Establishing or supporting a data repository is a core component of an institution's RDM strategy, providing the necessary infrastructure for data preservation and sharing.