Check for updates

NIST Special Publication 1500
NIST SP 1500-18r1

NIST Research Data Framework (RDaF)

Version 1.5

Robert J. Hanisch
Debra L. Kaiser
Alda Yuan
Andrea Medina-Smith
Bonnie C. Carroll
Eva M.Campo

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-18r1

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Special Publication 1500
NIST SP 1500-18r1

NIST Research Data Framework (RDaF)

Version 1.5

Robert J. Hanisch
Debra L. Kaiser
Alda Yuan
Andrea Medina-Smith
Office of Data and Informatics
Material Measurement Laboratory

Bonnie C. Carroll
Consultant

Eva M. Campo
Consultant
Campostella Research and Consulting
Alexandria, VA

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-18r1

May 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

NIST SP 1500-18r1
May 2023

Publications in the SP1500 subseries are intended to capture external perspectives related to NIST standards, measurement, and testing-related efforts. These external perspectives can come from industry, academia, government, and others. These reports are intended to document external perspectives and do not represent official NIST positions. The opinions, recommendations, findings, and conclusions in this publication do not necessarily reflect the views or policies of NIST or the United States Government.

NIST Technical Series Policies
Copyright, Fair Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

Publication History
Approved by the NIST Editorial Review Board on 2023-05-15
Supersedes NIST Series 1500-18 (February 2021) https://doi.org/10.6028/NIST.SP.1500-18

How to Cite this NIST Technical Series Publication
Hanisch, RJ; Kaiser, DL; Yuan, A; Medina-Smith, A; Carroll, BC; Campo, EM (2023) NIST Research Data Framework (RDaF) Version 1.5. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 1500-18r1. https://doi.org/10.6028/NIST.SP.1500-18r1

NIST Author ORCID IDs
Robert Hanisch: 0000-0002-6853-4602
Debra Kaiser: 0000-0001-5114-7588
Alda Yuan: 0000-0001-9619-306X
Andrea Medina-Smith: 0000-0002-1217-701X
Bonnie Carroll: 0000-0001-8924-1000
Eva Campo: 0000-0002-9808-4112

Contact Information
rdaf@nist.gov

Abstract

The NIST Research Data Framework (RDaF) is a multifaceted and customizable tool that aims to help shape the future of open data access and research data management (RDM). The RDaF will allow organizations and individual researchers to customize an RDM strategy. Although NIST is leading the RDaF, most of the content in the current version 1.5 was obtained via engagement with national and international leaders in the research data community. NIST held a series of 3 plenary and 15 stakeholder workshops. Workshop attendees represented many stakeholder sectors: US government agencies, national laboratories, academia, industry, non-profit organizations, publishers, professional societies, trade organizations, and funders (public and private), as well as international organizations. The audience for the RDaF is the entire research data community in all disciplines—the biological, chemical, medical, social, and physical sciences, plus the humanities. The RDaF is applicable from the organization to the project level, encompassing a wide array of job roles engaged in activities concerned with research data management, from Executives and Chief Data Officers to publishers, funders, and researchers.

The RDaF is a map of the research data space that uses a lifecycle approach with six high-level lifecycle stages to organize key information concerning RDM and research data dissemination. Through a community-driven and in-depth process, NIST identified and defined specific, high-priority topics and subtopics for each lifecycle stage. The topics and subtopics are programmatic and operational activities, concepts, and other important factors relevant to RDM and form the framework core. Further, the RDaF team identified Overarching Themes which are pervasive throughout the framework. The core enables organizations and individual researchers to use the RDaF for self-assessment of their RDM status. In addition, each subtopic has several Informative References—resources such as guidelines standards, and policies. As such, the RDaF may be considered a “best practices” document. The RDaF includes sample “profiles” for various job functions or roles, each containing topics and subtopics and individual in the given role needs to consider in fulfilling their RDM responsibilities. Individual researchers and organizations involved in the research data lifecycle will be able to tailor these profiles for their specific job function. The methodologies used to generate the content of this publication are described in detail.

Keywords

Research data; research data ecosystem; research data framework; research data lifecycle; research data management; research data dissemination, use, and reuse; research stewardship; open data.

Table of Contents

List of Tables

List of Figures

Foreword

This interim version 1.5 of the NIST Research Data Framework builds on the Preliminary version 1.0 released in February 2021 and incorporates stakeholder input received since that date. Version 1.5 has more than twice as many topics and subtopics and includes new sections. These new sections include overarching themes, which are terms prevalent in multiple lifecycle stages, profiles, which list the most relevant topics and subtopics for a given job function or role within the research data management ecosystem. A Request for Information (RFI) based on v. 1.5 will be posted in the Federal Register. All comments submitted on this RFI will be considered and the RDaF will be revised as appropriate. This modified RDaF will be presented at a Closing Plenary Workshop to be held in summer 2023 and input will be incorporated into RDaF publication v. 2.0, with a target release date in fall 2023.

Acknowledgments

The success and completeness of the NIST RDaF is wholly dependent on the input and participation of the broad research data community. NIST is grateful to all the workshop participants and others who have provided input to this effort. First and foremost, NIST thanks the members of the RDaF Steering Committee, past and present, who have given sound advice and shared their invaluable expertise since the inception of the RDaF in December 2019: Laura Biven, Cate Brinson, Bonnie Carroll (Chair), Mercè Crosas, Anita de Waard, Chris Erdmann, Joshua Greenberg, Martin Halbert, Hilary Hanahoe, Heather Joseph, Mark Leggott, Barend Mons, Sarah Nusser, Beth Plale, and Carly Strasser.

The team is also grateful to Susan Makar from the NIST Research Library for assistance with the Informative References. Thanks to Eric Lin, James St. Pierre, and Angela Lee for their critical assistance and advice.

Thanks to those who worked on the RDaF team including Breeze Dorsey, Laura Espinal, and Tamae Wong. Thanks as well to Campostella Research and Consulting for providing administrative support for the project and technical support for the natural language processing work. Our appreciation also goes to the NIST Material Measurement Laboratory (MML) leadership for their support and to all participants of the various workshops held to solicit community feedback, particularly those individuals who volunteered to serve as discussion leaders.

And finally, thanks to all involved with the NIST Cybersecurity Framework for forging a path for the RDaF.

Author Contributions

Robert Hanisch: Conceptualization, Methodology, Supervision, Writing- review and editing; **Debra Kaiser**: Formal Analysis, Methodology, Writing- review and editing; **Alda Yuan**: Formal Analysis, Methodology, Project Administration, Writing- original draft, Writing- review and editing, Visualization; **Andrea Medina-Smith**: Data Curation, Formal Analysis, Visualization, Software, Writing- review and editing; **Bonnie Carroll**: Conceptualization, Supervision, Writing- review and editing; **Eva M. Campo**: Data Curation, Visualization, Writing- review and editing.

1. Introduction

NIST's Research Data Framework (RDaF) is designed to help shape the future of research data management (RDM) and open data access. Developed through active involvement and input from national and international leaders in the research data community, the RDaF provides a customizable strategy for the management of research data. The audience for the RDaF is the entire research data community, including all organizations and individuals engaged in any activities concerned with RDM, from Chief Data Officers to publishers to researchers and funders. The RDaF builds upon previous work but distinguishes itself through its emphasis on research data, the community-driven nature of its formulation and its broad applicability to all disciplines, including the social sciences and humanities.

The RDaF is a map of the research data space that uses a lifecycle approach with six high-level lifecycle stages to organize key information concerning RDM and research data dissemination. Through a community-driven and in-depth process, stakeholders identified programmatic and operational activities, concepts, and other important factors relevant to RDM. The framework core lays out these topics and subtopics under the six stages of the research data life cycle, as shown in Fig. 1.

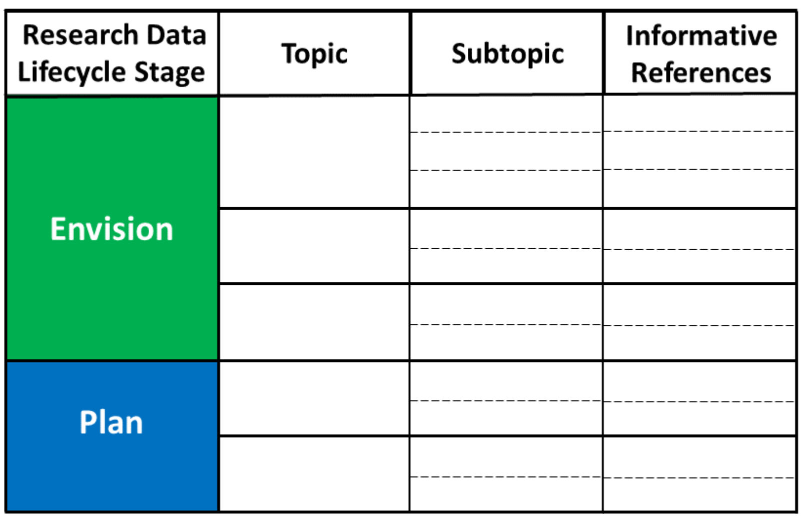

Fig. 1. Partial Organizational Structure of the Framework Core

All elements of the RDaF framework core illustrated in Fig. 1—lifecycle stages and their associated topics and subtopics—are defined. The framework enables organizations and individual researchers to use the RDaF for self-assessment of their RDM status. In addition, each subtopic has several informative references—resources such as guidelines standards, and policies—that assist stakeholders in addressing that subtopic. The RDaF also includes sample profiles, which contain topics and subtopics an individual in a job function or role should consider in fulfilling their RDM responsibilities. Researchers and organizations involved in the research data lifecycle will be able to tailor these profiles.

Since the first RDaF publication in 2021, referred to herein as v. 1.0 [1], NIST has expanded and enriched the framework through community outreach and direct engagement with stakeholders via workshops. These workshops resulted in a more comprehensive list of topics and subtopics relevant to RDM. NIST added definitions and informative references for each subtopic to improve the usability and applicability of the RDaF. In addition to profiles, this document includes overarching themes that appear across multiple lifecycle stages and a list of key organizations in the RDM space. The methodology used to generate the content of this interim publication v. 1.5, is described in detail in the following section. The publication of v. 1.5 will expose a larger group of stakeholders to the RDaF via a Request for Information in the Federal Register. Interested parties in all organizations working with research data are encouraged to give feedback on any part of the framework as this will be incorporated into version 2.0.

As a general note, the terms "data," "datasets," "data assets," "digital objects," and "digital data objects" are used throughout the framework depending on the context. Data is the most general and frequently used term. Dataset means a specific collection of data having related content. A data asset is "any entity that is comprised of data which may be a system or application output file, database, document, or web page."[2] Digital objects and digital data objects typically have a structure such that they can be understood without the need for separate documentation. In addition, the terms "organization" and "institution" used throughout the framework are considered synonymous.

2. Methodology

This section describes the approaches used to further develop the RDaF since the first publication. Throughout all the steps in the methodology, the RDaF Steering Committee was consulted, took leadership roles as discussion leaders, and provided valuable feedback to the process and the results.

2.1. Framework Development Through Stakeholder Input

The RDaF is driven by the research data stakeholder community who can use the framework for multiple purposes, from identifying best practices for RDM and dissemination to changing the research data culture in an organization. To ensure the RDaF is a consensus document, NIST held community engagement workshops as the primary mechanism to gather stakeholder input on the initial framework. Thus far, the workshops have taken place in three phases, each resulting in further examination and refinement of the framework.

2.1.1. Phase 1: Plenary Scoping Workshop and Publication of the Preliminary RDaF

In the Plenary Scoping Workshop held in December 2019, a group of about 50 distinguished research data experts selected the organizing principle of the RDaF—a research data lifecycle—and the six stages of this lifecycle: Envision, Plan, Generate/Acquire, Process/Analyze, Share/Use/Reuse, and Preserve/Discard. Feedback from this workshop contributed to the publication of the RDaF v. 1.0, which provides a structured and customizable approach to developing a strategy for the management of research data. The framework core, modified from v. 1.0, remains the heart of this publication and consists of the six lifecycle stages along with their topics and subtopics.

2.1.2. Phase 2: Opening Plenary Workshops

The second phase of the RDaF development began with two virtual Plenary Workshops held in late 2021. Each had approximately 70 attendees and focused on two cohorts. The University Cohort (UC) Workshop, co-hosted by the Association of American Universities, the Association of Public Land-grant Universities, and the Association of Research Libraries was a horizontal cut across various stakeholder roles in universities (e.g., vice presidents of research, deans, professors, and librarians), publishing organizations, data-based trade organizations, and professional societies. In contrast, the Materials Cohort (MC) Workshop, held in cooperation with the Materials Research Data Alliance, was a vertical cut across stakeholder organizations engaged in materials science, including academia, government agencies, industry, publishers, and professional societies.

Prior to the workshops, the attendees selected, or were assigned to, one of six breakout sessions focusing on a data lifecycle stage. A NIST coordinator sent the attendees a link to the RDaF publication v. 1.0, a list of the session participants and definitions of the topics for the lifecycle stage of the session. The agenda for the two Workshops included an overview talk by Robert Hanisch on the RDaF, a one-hour breakout session, and a plenary session with summaries presented by an attendee of each breakout and closing remarks. During the breakout sessions, a discussion leader, recruited by the RDaF team, solicited input from the 10 to12 participants on the following questions:

1. What are the most important (2 or 3) topics and the least important one?
2. Are there any missing topics?
3. Should any topics be modified or moved to another lifecycle stage?

The identical questions were posed regarding the subtopics for each topic. Attendee input was captured as notes taken by the session rapporteur and the NIST coordinator and an audio recording.

In the few months following the opening plenary workshops, the RDaF team revised the topics and subtopics for the lifecycle stage based on input from the UC and MC workshops. All six of the lifecycle stages were then reviewed side-by-side for consistency and completeness.

The collective review revealed 14 overarching themes which appeared in multiple lifecycle stages. These themes include metadata, data quality, FAIR (Findable, Accessible, Interoperable, and Reusable), provenance, software tools, and cost implications. A separate section of the framework will address all overarching themes in detail.

2.1.3. Phase 3: Stakeholder Workshops

The next step in obtaining community input involved a series of two-hour Stakeholder Workshops focused on specific roles, equivalent to job functions or position titles. Unlike the first two RDaF workshops, these workshops were in the form of smaller focus groups. The goal of holding smaller, job-specific focus groups was to develop profiles to serve as checklists of topics and subtopics important for individuals in a given role with respect to RDM. The target size of these two-hour workshops was 10 to12 participants and, like the opening plenary workshops, had a discussion leader and rapporteur recruited from amongst the workshop participants, as well as a NIST coordinator.

To secure a broad range of feedback, the RDaF team compiled a list of more than 200 invitees, including attendees of previous workshops and additional experts in various research data programs and in RDM, parsing them into fifteen roles. The roles were not perfect fits for all invitees, but the RDaF team assigned people to the role closest to their job function. Those roles were:

- Academic Mid-Level Executives/Heads of Research,
- AI Experts,
- Budget/Cost Experts,
- Curators,
- Data/IT Leaders,
- Data/Research Governance Leaders,
- Funders,
- Institute/Center/Program Directors,
- Open Data Experts,
- Professional Society/Trade Organization Leaders,
- Professors,
- Providers of Data Tools/Services/Infrastructure,
- Publishers,
- Researchers
- Senior Executives.

During the workshops, after a brief refresher about the purpose and structure of the RDaF, participants selected the most relevant of the six lifecycle stages to their assigned role. For each lifecycle stage, participants reviewed the topics and subtopics in more detail. Each of the workshops addressed two to four of the lifecycle stages. In addition to obtaining feedback on the completeness of the framework, the RDaF team asked participants to consider which topics and subtopics had the greatest influence on their role and those over which they had the greatest influence.

2.2. Updating the Framework per Feedback

In parallel with feedback collection, NIST began to make modifications to the framework. This required digesting and categorizing the comments and advice generated through the extensive community engagement described above. This stakeholder input was used to update the core and to help define the profiles, which describe and apply to specific roles within the research data ecosystem.

2.2.1. Stakeholder Workshop Note Aggregation

After the Stakeholder Workshops, the RDaF team designed a common methodology for collecting and analyzing the feedback, using a template to record the input from each workshop. One main goal of these workshops was to obtain information sufficient to produce profiles of the most relevant topics and subtopics for each of the roles represented. This template contained the following:

1. A column for topics and subtopics relevant to, or missing from, the profile for a role
2. A column for topics and subtopics in a lifecycle stage that were missing, misplaced, or unclear
3. A section on feedback that addressed the definition of the role
4. A section on "takeaways" regarding the framework as a whole
5. A section on proposed new overarching themes.

To analyze the feedback from each stakeholder workshop, selected RDaF team members used notes from the rapporteur to familiarize themselves with the discussion. Then these team members viewed the recording of the workshop, read through any written comments provided in the workshop chat, and noted every comment in the appropriate section of the template. After the first draft of the template notes was completed, the team members viewed the recording a second time and added any missing comments. Next, these team members converted each comment and suggestion concerning a topic or subtopic in a lifecycle stage into a potential change for review. Finally, the entire RDaF team considered each potential change and generated an updated version 1.5 of the framework core.

2.2.2. Collecting Feedback for Profile Development

After updating the framework core based on the stakeholder feedback, the next step involved generating generic profiles for each role addressed by a workshop. As the feedback from the stakeholder workshops concerning profiles varied in form and specificity, more data was needed to develop these profiles.

The RDaF Team used the updated framework to develop blank checklists of topics and subtopics for the lifecycle stages discussed at each Stakeholder Workshop. Respondents returned a total of about 60 checklists.

The responses were analyzed for similarities and several roles were modified. For example, Professors and Researchers were grouped together to form one role because of the involvement Professors have in their groups' research. After consideration of the response data, the RDaF Team decided to generate 8 generic profiles: AI Experts, Curators, Budget/Cost Experts, Data and IT Experts, Providers of Data Tools, Publishers, Research Organization Leaders, and Researchers.

For each generic profile, the RDaF team first calculated the percentage of responses that labeled a subtopic as relevant. When 50% or more of the respondents considered a subtopic to be relevant, it was presumptively deemed relevant for the generic profile. Next, the team considered all comments received with the profile responses as well as all the notes from the Stakeholder Workshop to further flesh out the generic profile. Lastly, the RDaF team consulted with experts in these roles to finalize the generic profiles.

3. Framework Core – Lifecycle Stages, Topics, and Subtopics

The following tables represent the framework core, developed primarily through stakeholder feedback. Each research data lifecycle and its topics and subtopics is a separate table. As shown in Fig. 2, the research data lifecycle is cyclical rather than linear. Each of the stages can lead into any of the others, and each organization or individual may approach the stages in different ways. It is likely that an organization will be involved in all lifecycle stages simultaneously, though with different levels of intensity or capacity. The goal of the framework is to be comprehensive while remaining flexible. Not every researcher or organization that works with research data will find every topic and subtopic relevant. See the profiles section for more detail on how to tailor the framework to suit individual circumstances.

The six lifecycle stages as depicted in Fig. 2 are defined below.

Envision – This lifecycle stage encompasses the review of the overall strategies and drivers of an organization's research data program. The Envision lifecycle stage is where choices and decisions are made that together chart a high-level course of action to achieve desired organizational goals.

Plan – This lifecycle stage encompasses the activities associated with preparing for data acquisition, selection of data formats and storage solutions, and anticipation of data sharing and dissemination strategies and policies.

Generate/Acquire – This lifecycle stage covers the generation of raw research data, both experimentally and computationally, within an organization, and the collection or acquisition of research data produced outside of an organization.

Process/Analyze – This lifecycle stage concerns the actions performed on generated or externally acquired research data to yield processed research data, typically using software, from which observations and conclusions can be made.

Share/Use/Reuse – This lifecycle stage outlines how raw and processed research data are disseminated, used, and reused within an organization and any constraints or encouragements to use/reuse. It also includes the dissemination, use, and reuse of raw and processed research data outside of an organization.

Preserve/Discard – This lifecycle stage delineates the end-of-use and end-of-life provisions for research data in an organization and includes records management, archiving, and safe disposal.

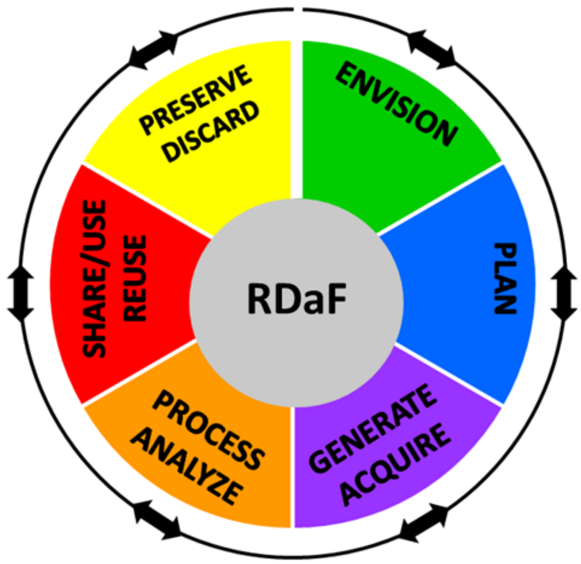

Fig. 2. Research Data Framework Lifecycle Stages

Many lexicons (i.e., vocabularies) exist in the research data management space. Though the RDaF does not intend to introduce an entirely new lexicon, it is important to be precise with the use of key terms. For each topic and subtopic, the RDaF therefore provides definitions to serve as a reference point. These definitions allow users to better understand what tasks and responsibilities are associated with each topic and subtopic. To produce these definitions, the RDaF team performed a search of common data lexicons such as CODATA's Research Data Management Terminology and Techopedia [3, 4]. Additionally, the RDaF team searched more broadly for common definitions and literature-specific definitions, including ones for the informative references that provide guidance in the implementation of the RDaF. The final definitions are also formulated to be consistent with stakeholder feedback. Individual researchers and organizations should keep in mind that these definitions are not prescriptive and consider their own context when determining whether the definitions provided are the best fit.

Table 1. Envision Lifecycle Stage

Envision: Topic	Subtopic	Definition
Data Governance – Strategic/Qualitative The policies, procedures, and processes pertaining to authority, control, and shared decision-making (planning, monitoring, and enforcement) over the management of data assets. [8, 9]	Identification of goals and roles	An exercise to define the objectives of, and responsible individuals for, various aspects of research data management (RDM).
	Vision and/or policy	Vision: An aspirational state an organization wishes to achieve with respect to RDM. Policy: A set of recommended and sometimes mandatory high-level principles that establish a guiding framework for RDM. [5, 6]
	Data management organization	An RDM infrastructure (RDMI) of human and capital resources that supports data-related activities, e.g., policies, planning, and sharing, as well as practices and projects, e.g., data acquisition, control, and protection. [7]
	Organizational values, include DEI	A set of core beliefs that function as guides to what is seen as good and important in an organization and the guiding principles that provide an organization with purpose and direction. (Values ideally include diversity, equity, and inclusivity.) [10, 11]
	Data management value proposition	A clear statement that indicates exactly what benefits an organization will derive from an RDM program. [12]
	Data needs assessment	An evaluation of the types of data that are critical to accomplish the stated goals of an organization.
	Purpose and value of data	A clear statement of the need for, use of, and benefit derived from, research data.
	Organization intent regarding FAIR data	The extent to which an organization supports the internal adoption and use of the FAIR data principles.
	End-use support	Components of the RDMI within an organization that enable data to be prepared and processed for its ultimate application, including reuse.
	Stewardship	The application of rigorous analyses and oversight to ensure that data assets meet the needs of users. [13]
Data Governance – Legal and Regulatory Compliance The policies, procedures, and processes to manage and monitor an organization's regulatory [and] legal responsibilities and risks pertaining to data. [9]	Privacy	The practice of protecting and properly handling sensitive data, including personal, proprietary, and confidential data. [14]
	Ethics	Moral principles pertaining to data practices, e.g., analysis and dissemination, that have the potential to adversely impact people and society (e.g., to minimize bias and maintain the privacy of personal data). See also the Global Data Ethics Project. [15–17]
	Safety and security assurance	The practice of protecting digital information from unauthorized access, denial of access, theft, or corruption throughout its entire lifecycle. [18]
	Inventory	A function that provides organization capabilities for archiving management such that data products can be grouped, searched and identified for retrieval, statistics and reorganization. Also, a list of available items stored and/or controlled in a storage warehouse system. [15, pg 19]
	Risk assessment	A systematic process for the identification and evaluation of potential threats to and vulnerabilities of an organization's data assets, e.g., unauthorized access to sensitive data. [20]
	Risk mitigation and management	A process for the development and implementation of appropriate strategies to control, reduce, or eliminate

Envision: Topic	Subtopic	Definition
		potential threats to and vulnerabilities of an organization's data assets as identified by a risk assessment. [21]
	Sharing/licensing	Data sharing agreement: A formal contract that details what data are being shared and the appropriate use for the data. Licensing agreement: A formal contract that states the purpose and duration of access being provided to the recipient licensee along with restriction and security protocols the recipient licensee of the data must follow. [22, 23]
	Social license for use and reuse	An unwritten agreement whereby a group of public stakeholders accept that certain datasets may be applied for purposes other than those for which the data were originally intended, e.g., healthcare data. [24]
	Jurisdiction for sharing and reuse	Legal requirements as set by an authoritative entity (e.g., local and national regions) concerning the dissemination of data by an organization and subsequent use of the data by other organizations. [25]
Data Culture and Reward Structure The collective beliefs and behaviors of the people in an organization concerning the value and management of research data. Practices designed to recognize the advantages and accomplishments of sharing data. [26]	Roles and responsibilities	The job functions and obligations that enable the establishment of a desired data culture and reward structure.
	Recognition of data management	Processes and practices that provide acknowledgement and rewards for good RDM at all levels in an organization.
	Value of data workers	Recognition of the benefits that staff performing data-centric jobs or functions provide to an organization.
	Promotion and tenure	Career advancements that are linked to good research processes, practices, and outcomes.
	Integrity of research and data	For research: The condition resulting from adherence to professional values and practices when conducting, reporting, applying, and disseminating results of the work.[27] For data: The accuracy, completeness, and quality of data as they are maintained over time and across formats.[28]
	FAIR data principles	Guidelines that allow digital objects (e.g., data, algorithms, and workflows) to be Findable, Accessible, Interoperable, and Reusable. [29]
	Maintenance of FAIR data	Ongoing infrastructural support to sustain FAIR data principles and practices.
	Incentives and impact for sharing and reuse	Staff recognition and rewards for widespread dissemination and application of research data and the beneficial effects of such dissemination.
	Disincentives for sharing and reuse	Barriers that limit dissemination of data, e.g., misinterpretation and misuse of data by others, lack of recognition, and the effort required for sharing.
	CARE and ethics	The CARE (Collective benefit, Authority to control, Responsibility, and Ethics) Principles for Indigenous Data Governance are people and purpose-oriented, reflecting the crucial role of data in advancing Indigenous innovation and self-determination. (These principles complement the existing FAIR principles for indigenous data governance.) Ethics: Moral principles pertaining to data practices, e.g., analysis and dissemination, that have the potential to adversely impact people and society (e.g., to minimize bias and maintain the privacy of personal data). [15, 30]

Envision: Topic	Subtopic	Definition
Education and Workforce Development Training to provide staff with the necessary skills and expertise for data-related activities and RDM. Includes leadership support and formal and informal training.	Workforce skills inventory	A catalog of an organization's capabilities in essential data processes.
	Workforce preparedness in new and advanced technologies	Assessment of needs for and provision of training in the skills and expertise of an organization's staff pertinent to novel and leading-edge areas of research, e.g., AI.
	Data management training	In-classroom, on-line, and/or hands-on instruction for staff to attain the skills and expertise required to manage data across all lifecycles.
	HR's supporting role in workforce development and training	Involvement of an organization's HR department in establishing and implementing instructional courses for staff to expand their skill sets and expertise in research data programs and RDM.
	Promotional paths and career development	Documented approaches for recruitment, advancement, and retention of staff in data-centric jobs in an organization and expansion of data-related skills and expertise for all technical jobs.
Resources—Allocation and Sustainability The distribution and longevity of funding to attain and maintain robust research data programs and RDM infrastructure.	Sources of funding	Entities that provide financial support for research data programs and RDM infrastructure (e.g., capital and human resources).
	Long-term funding	Sustained financial support for research data activities and RDM infrastructure.
	Staffing	Provision of sufficient resources to support RDM staff and researchers engaged in RDM activities.
Community Engagement Outreach and interactions among organizations or individuals with shared goals or interests concerning research data activities or RDM.	Stakeholder communities	Individuals, groups, and organizations that have an interest or stake in RDM or research data activities in general and in particular domains. [31]
	Modes of communication	Ways by which information about data and data management are shared and discussed.
	Partners/partnerships	Partner: Two or more organizations or individuals that share responsibility and control of ideas, processes, and outcomes of research data activities. Partnership: An agreement between organizations and individuals to collaborate on such activities.. [32]
	Engagement across knowledge domains and sectors	Interactions among groups or individuals having expertise in different specific, specialized disciplines or fields, or expertise in different technology areas. [33]
	Inclusivity in interactions	The practice of including all types of people or ideas and treating them all fairly and equally. [34]
	Data services and the beneficiaries	Solutions for data tasks (e.g., data transfer, storage, and analytics) and the organizations or individuals deriving value from such solutions. [35]

Table 2. Plan Lifecycle Stage

Plan: Topic	Subtopic	Definition
Chain of Custody A complete, fully documented step-by-step history of a data asset in an organization, i.e., who has possession of a data asset, at what time, and for what purpose, at all times throughout the lifecycle of the data asset. [36]	Roles and responsibilities	The job functions and obligations for tracking data assets.
	Implementation authority	Person empowered to grant access to data assets, e.g., a Chief Data Officer.
	Centralized inventory of services, groups, and resources	An organization-wide catalog of elements supporting data-related activities at various levels of an organization, including capital (e.g., HPC), virtual (e.g., domain repositories), and human (e.g., Data Steward and AI interest group) components.
	Provenance	The historical, documented record of a data asset that contains details on its origin—where, when, how, and by whom it was generated/acquired/processed—and on all alterations to the data asset. [15, pg 24, 31]
Financial Aspects of Planning Factors to consider in estimating or assessing the costs associated with all research data and RDM activities over the data lifecycle.	Funding models for provisioning resources	Approaches for providing financial support for data-related activities and infrastructure, including direct, (e.g., grants, contracts, and institutional), overhead, or mixed. [38]
	Funding sources	Entities that provide financial support for research data activities and infrastructure, e.g., capital and human resources.
	Decision-making tools to assess costs	Methods to determine the financial requirements of various data activities and infrastructure (e.g., cost-benefit analysis, market analysis, and decision trees).
	Cost-benefit analysis	A systematic approach to estimating the strengths and weaknesses of alternative actions to determine options which provide the best approach to achieving benefits while preserving savings. [39]
	Cost breakdown by lifecycle stage	Identification of funds required for each data activity in a project (e.g., hardware, software, and staffing for data generation), or for an RDM infrastructure (e.g., centralized data services).
	Downstream lifecycle costs	Funds required after establishment of an RDM infrastructure (e.g., technology refresh and maintenance) or for later-stage data activities (e.g., long-term preservation).
	Staffing and training	Costs incurred in assuring that new staff with appropriate skills and expertise are hired for specific data activities and that existing staff attain new and advanced skills through instructional courses.
Data Management Planning The process of organizing and specifying objectives and activities throughout the research data lifecycle.	Written data management plans (DMPs)	Documents that provide sufficient detail on the following topics: Administrative Data, Data Collection, Documentation and Metadata, Ethics and Legal Compliance, Storage and Backup, Selection and Preservation, Data Sharing, and Responsibilities and Resources. DMPs are referred to herein as a research data "roadmaps." [40]
	Purpose/intent of research study and context of anticipated data use	Clear articulation of research objectives in terms of data products that are essential to address specific research and/or technical requirements.
	Specification of data objects, metadata, analysis	Detailed descriptions of all information, processes, software, and hardware required from conception to completion of a research data project.

Plan: Topic	Subtopic	Definition
	tools, and workflows throughout the lifecycle	
	Machine-readable DMPs	Research data roadmap documents in a form that can be used and understood by a computer. [41]
	Linkage of DMPs to administrative records	Interconnection of a research data roadmaps to operational data, e.g., agreements, transactions.
	Data organization e.g., database, repository, to facilitate future access	The practice of categorizing, classifying, and storing data with sufficient detail and specificity such that the data are readily discoverable and usable by others. [42]
	Data management expertise and training	In-class, on-line, and/or hands-on instruction for staff to attain the skills and knowledge required to manage data in a research study.
Data Object An entity that, together with associated Representative Information (i.e., metadata), is produced or used in a research study. [15, pg 13]	Quantitative and qualitative	Quantitative data are numerical data, e.g., measurements and some controlled observations and questionnaires. Qualitative data are defined as non-numerical data, e.g., text, video, photographs, or audio recordings. [43]
	Measurement, including images, audio recordings, and photos/videos	A quantity in various formats, including numerical, visual, and auditory.
	Observation	A fact or occurrence often involving measurement with instruments. [44]
	Survey	A list of questions aimed for extracting specific data from a particular group of people. [45]
	Software	A computer-based application that converts inputs into outputs to support the user in one or more research tasks. [46]
	Model	A representation, pattern, or mathematical description that can help scientists replicate a system, process, or research result. [47]
	Documentation (text)	Comprehensive information that accompanies a dataset, including all associated metadata, data dictionary, description of methods and instruments [and software used to generate/collect and] process the data, and other supporting data (e.g., duplicate sample results, replicate analyses). [48]
	Specimen (physical sample)	A tangible object that may observed or tested to determine its properties or characteristics.
	Presentation	Material assembled to explain and describe research results or processes to an audience.
FAIR Findability, Accessibility, Interoperability, Reusability: a set of guiding principles to support the reusability of data that are beneficial to all scholarly digital research objects. [29, 49]	Organizational support for making data more FAIR	Institutional resources to improve the extent of "FAIRness" of data. (FAIRness is used herein to denote a continuum state ranging from no FAIR aspects to fully FAIR.)
	Identification of methods/guidelines vis-à-vis FAIR principles	An exercise to locate techniques and recommended procedures related to FAIRness.

Plan: Topic	Subtopic	Definition
Data/Metadata Considerations Factors to take into account prior to conducting a research study.	Criteria for selection of data/metadata	Requirements and needs by which decisions are made regarding what information to generate, collect, and document in a research study.
	Nature of data/metadata required	Specification of the requisite types and characteristics of selected information.
	Intended extent of FAIRness	The degree to which data and metadata are meant to comply with the FAIR data principles.
	Methods to capture and store data/metadata	Techniques or means by which data/metadata are collected, recorded, and preserved.
	Metadata schema	The overall structure of data about the data. Two examples of general-purpose metadata schema are Dublin Core and MODS (Metadata Object Description Schema). [50, 51]
Data Architecture The fundamental structure of an organization's research data management (RDM) system embodied in its components, their relationships to each other and to the environment, and the principles guiding its design and evolution. Includes, for example, the underlying storage technologies, networking, hardware, system interfaces, authentication mechanisms, data brokers, monitoring platforms, semantic interoperability tools, long-term preservation services, and HPC. [56, 57]	Design	A set of principles that are formulated from specific strategies, rules, models, and guidelines for the management and flow of a dataset throughout its lifecycle.
	Processing operations	Methodology for translating raw data into useable information. Specific methods include, e.g., data preparation, validation, sorting, aggregation, analysis, and reporting.
	Workflow	The process of managing data in a structured manner. It involves collecting, organizing, and processing data so that it can be used for various purposes. [52]
	Model	A detailed description or scaled representation of the relationships and data flow between different components of an RDM system, typically in the form of a diagram or flowchart. [53]
	LIMS	A laboratory information management system (LIMS) is a software system developed to support laboratory operations (e.g., track specimens and workflows and aggregate data). [54]
	Hosting and storage, cloud storage	Methods whereby, and locations wherein, data are saved and from which data can be retrieved.
	Configuration management	The actions of tracking and controlling changes in the hardware and software components, e.g., updates and version control. [55]
	Interoperability among different architectures	The capability to communicate, execute programs, or transfer data among different RDM systems in a useful and meaningful manner that requires the user to have little or no knowledge of the unique characteristics of those systems. [58]
	Security	Features of the architecture that protect data from unauthorized access, denial of access, corruption, or theft throughout their entire lifecycles. [18]
	Existing standards	Standards relevant to data architecture, including schema (e.g., based on SQL, JSON), format (e.g., JSON, XML), and APIs (e.g., Google Search for the web).
Hardware and Software Infrastructure The physical and	Organizational research needs	Essential resources required to accomplish the objectives of research projects and RDM (e.g., centralized infrastructure, appropriate training, and support staff).
	Tools to support data-related processes	Items, e.g., instruments, methods, utility software, and APIs, that enable research.

Plan: Topic	Subtopic	Definition
non-physical functional components that collectively form a foundation for conducting research and RDM.	Models that connect infrastructure to data processes and workflow	A detailed description or scaled representation of the relationships between data tasks and movement and the hardware and software components in an RDMI. [53]
	Interoperability	The capability to seamlessly communicate, execute programs, or transfer data among various functional components .that requires the user to have little or no knowledge of the unique characteristics of those components. [58]
	Persistent instrument identifiers	Globally unique, persistent, and resolvable identifiers of operational scientific instruments enable research data to be persistently associated with such crucial metadata, helping to set data into context. The Research Data Alliance Working Group Persistent Identification of Instruments (PIDINST) developed a metadata schema and prototyped schema implementation and demonstrated the viability of the proposed solution in practice. [59]
	Sustainability of data vis-à-vis obsolete infrastructure	Concerns regarding the ability to reproduce and reuse data if the hardware and software components become outdated or non-functional.
	Security and privacy considerations	The degree of protection of data from unauthorized access, denial of access, corruption, or theft provided by the hardware and software. [18]
	Staff expertise and support staff	Personnel with the appropriate skills and knowledge to maintain and update the hardware and software infrastructure as needed, and personnel to interface with researchers using the infrastructure.
Research Data Standards Documents, including codes, specifications, recommended practices, classifications, test methods and guides, that describe how data should be stored or exchanged for the consistent collection and interoperability of that data across different systems, sources, and users. [60, 62]	Criteria, i.e., general vs. domain-specific standards	Requirements and needs by which decisions are made regarding the type of research standard, i.e., broadly applicable or limited to a particular field of research.
	Sources of standards/guidelines for data/metadata	Origins of accepted practices consisting of discrete, reusable components, e.g., data types, identifiers, schemas, and formats. Examples include the Dublin Core Metadata Initiative and Schema.org. [60]
	Quality standards	Guidelines that provide sufficient information to allow all users to readily evaluate the degree of "fitness for purpose" of the data. Key data quality components include completeness, accuracy, integrity, consistency, and timeliness. [15, pg 26, 57]
	Community-based standards/conventions	Community-based data and metadata standards are typically long-term endeavors with many different players and types of efforts. Such standards facilitate reuse of data integrative analysis and comparison to other datasets and linkage of data with other research products, such as scholarly material, algorithms and software. [63]
Assessment Evaluation of the success of a research project against expectations set before the project has started.	Goals/definition of success	Statement of project objectives; list of accomplishments demonstrating that these objectives were met.
	Metrics for tracking use and impact measures, including reuse	Quantitative and qualitative indicators of positive influence or outcomes, (e.g., number of citations of a dataset and anecdotal evidence of reuse of a dataset). [64]
Communication and Outreach	Methods to share and reuse data/metadata	Approaches to disseminate data/metadata and to facilitate reusability of data/metadata, (e.g., use of open repositories and maximizing the FAIRness of data).

Plan: Topic	Subtopic	Definition
Engagement and interactions among groups and individuals working in similar research areas.	Allocation of credit to project team members	Properly documenting and recognizing each team member's contributions to a project. [65]
	Promotion of data to communities of interest	Modes to communicate the existence and location of datasets to targeted groups, e.g., special-topic data publications and presentations at topical workshops.
	Cross-institution cooperation	The process of working with other institutions or organizations on a shared activity (e.g., informal collaborations, formal partnerships, and agreements).
	Requests for additional data from community	Soliciting data contributions from partners and stakeholders on areas of mutual interest.
Access Control Associated with Data Sensitivity Methods and requirements to limit the individuals or groups permitted to view or use protected data.	Identification of responsible parties for access management	A determination of those individuals authorized to both prohibit and permit access to sensitive data.
	Ease of maintenance and implementation of records	The extent to which sensitive data can be kept up to date and made accessible to authorized individuals and groups.
	Regulatory compliance	Efforts by organizations to ensure that they are aware of and take steps to conform to relevant laws, policies, and regulations concerning sensitive data (e.g., medical records). [66]
	Sensitive data/PII	Data that needs to be controlled due to certain risks. Personally Identifiable Information (PII) is any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means. [67]
	Limited disclosure, e.g., IP	Restricting release of data to specific legal circumstances and often requiring notification to the data provider. Intellectual Property (IP) refers to certain exclusive rights granted by law to the owner of, e.g., a novel data product. [68]
	Licensing for reuse	Legal agreement that allows one party to use another party's data subject to certain conditions.

Table 3. Generate/Acquire Lifecycle Stage

Generate/Acquire: Topic	Subtopic	Definition
Data Types Classifications or categories of data. [69]	Measurement, including images, audio recordings, and photos/videos	Quantity that can be in various formats, including numerical, visual, and auditory.
	Text file	A type of digital, non-executable file that contains letters, numbers, symbols and/or a combination without any special formatting (e.g., ASCII, EBCDIC). [70]
	Computation, simulation	Computation: an act, process, or method of computing. Simulation: any research or development project where researchers or developers create a model of some authentic phenomenon to mimic the outcomes that happen in the natural world. [71, 72]
	Source code	The set of instructions and statements written by a programmer using a computer programming language. This code is later translated into machine language by a compiler. [73]
	Observation	A fact or occurrence often involving measurement with instruments. [44]
	Survey	A list of questions aimed for extracting specific data from a particular group of people. [45]
	Transaction	Data that describe an exchange or transfer of goods, services, or funds. [74]
	Social media	Interactive technologies that facilitate the creation and sharing of data through virtual communities and networks. [75]
Data Sources Description of circumstances whereby data are produced. Origin of data.	In-house generation by researchers	Data created by researchers within an organization and at a physical location internal to the organization.
	Remote generation by researchers	Data created by researchers within an organization through control of an instrument or device at a location other than the organization.
	In-field generation by researchers	Data created by researchers within an organization at a physical location external to the organization, which may be a natural environment.
	User facility generation by/for researcher	Data created by researchers or facility staff at a federally sponsored research facility available for external use to advance scientific or technical knowledge. [76]
	Historical	Data generated or collected in the past, which may have uncertainties due to, e.g., age and loss of metadata.
	Human-annotated	The process of adding metadata or other information [in different formats] to data by a person such as labels or tags to describe the content or context of images, and labels or tags to classify or extract relevant information from text. Such annotation allows the AI and ML models to categorize the data and approve the execution of relevant tasks. [77]
Generated Experimental Data Data produced by automation or active intervention by a researcher to induce	Source of objects/subjects	Origin of items used in an experiment.
	Characteristics of objects/subjects	Distinct features of items used in an experiment, e.g., appearance and properties.
	Conditions of research study	Description of the external physical environment in which data were collected (e.g., temperature, atmosphere). Such conditions are elements of metadata.

Generate/Acquire: Topic	Subtopic	Definition
and measure changes or to create differences when a variable is altered. [78]	Specification of instruments and tools	Identification and documentation of measurement equipment and other items, e.g., software, methods, and materials, used in an experimental research study. Descriptions of the technical details and requirements of each item.
	Parameters for instruments and tools	Variables or settings on an instrument that are maintained and controlled during an experiment (e.g., laser intensity, gas flow rate, and rate of data collection).
	Methods, protocols, and calibration	Techniques and procedures used in the generation of data.
	Data/metadata capture methods	Techniques and procedures for collecting and recording information, both short-term and long-term.
	Provenance and capture methods	Techniques and procedures for collecting and recording the historical, documented record of a data asset that contains details on its origin—where, when, how, and by whom it was generated/acquired/processed—and on all alterations to the data asset. [19, 37]
	Reproducibility	The ability to replicate data using identical tools (e.g., documented metadata, code, methods, and instruments) employed previously by the original researchers or by other researchers, without the need for any additional information or communication with the original researchers. [79]
Generated Computational Data Data produced by using calculations, models, simulations, or other methods. Can be produced manually or using a computer or other device. [71, 72]	Input data/metadata	Information of any type that is entered manually or via an automated process into an instrument, computer, or other device.
	Output data/metadata	Electronic data produced by an instrument, processor, computer, or other device.
	Hardware	The physical elements that make up a computer or electronic system and everything else involved that is physically tangible, including the monitor, hard drive, memory and the CPU. [80]
	Parameters and conditions for computation	Hardware or software system requirements, or configurations, that are necessary for a hardware or software application to run smoothly and efficiently, e.g., operating system dependencies, compilers, and memory requirements. [81]
	Versioning	The process of numbering different releases of entities, e.g., software, hardware, and documents, for the purposes of tracking and recording changes. This provides the ability to revert to a previous revision, which is critical for data traceability and data re-creation, tracking edits, and correcting errors. [82, 83]
	Data/metadata capture methods	Techniques and procedures by which information is collected and recorded.
	Provenance and capture methods	Techniques and procedures for collecting and recording the historical, documented record of a data asset that contains details on its origin—where, when, how, and by whom it was generated/acquired/processed—and on all alterations to the data asset. [15 pg 24, 31]
	Verification/validation of output data	Verification: The process of determining that a computational model accurately represents the underlying mathematical model and its solution Validation: The

Generate/Acquire: Topic	Subtopic	Definition
		process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model. [84]
Qualitative Data	Nature of objects/subjects	Types and characteristics of entities which are being studied.
Data that is descriptive and concerns phenomena which can be observed but not measured, e.g., language	Methods and protocols	Techniques; standard operating procedures; sets of rules, and guidelines.
	Metadata	Data about data, i.e., data that define and describe the characteristics of other data. Using a survey as an example, metadata include the questions in and location of the survey. [85]
	Paradata	Data about the process by which data were collected. Formalized data on methodologies, processes and quality associated with the production and assembly of statistical data. (Using a survey as an example, paradata include the mode of the survey and responders' response times. Note that the term paradata is typically associated with social science disciplines; in physical and medical science disciplines, paradata would be included in metadata.) [86, 87]
	Data/metadata/paradata capture methods	Techniques and procedures for collecting and recording any type of data, either manually or via an automated process into an instrument, computer, or other device.
Acquired Data	From collaborators	Originating from other individuals or other organizations partnering with researchers in an organization.
Data used in a research study that was not generated by the researchers conducting the study.	From repositories	Originating from a destination designated for data storage. Operations of a repository include preservation, management, and provision of access for digital materials that may have different types and formats. [88]
	From the literature	Originating from a publication.
	Aggregated datasets from multiple sources	Data compiled from disparate studies, organized, and summarized so that conclusions can be drawn, and decisions made from such data-rich collections.
	Provenance	The historical, documented record of a data asset that contains details on its origin—where, when, how, and by whom it was generated/acquired/processed—and on all alterations to the data asset. [19, 37]
	Restrictions, fees, and usage agreements	Mechanisms that may limit the use of acquired data.
Critically Evaluated (CE) Data	Infrastructure to assure the greatest data integrity	A foundation composed of practices, processes, and procedures designed to produce data that are clean, traceable, and fit for purpose. NIST and KRISS are two institutions that produce critically evaluated data named Standard Reference Data. [89]
Numerical data that have undergone rigorous review and critique such that the integrity, reasonableness, and usability are optimized. [90]	Single researcher dataset	A group of data that originates from an individual researcher.
	Aggregation of data evaluated by experts	The process by which data from disparate sources are compiled, reviewed, critiqued, and summarized by subject matter experts.
	Reproducibility and uncertainty quantification	Reproducibility: The ability to replicate data using identical tools (e.g., documented metadata, code, methods, and instruments) employed previously by the original researchers or by other researchers without the need for any

Generate/Acquire: Topic	Subtopic	Definition
		additional information or communication with the original researchers. Uncertainty quantification: assignment of a numerical value to a non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand. Critically evaluated data have great reproducibility and small uncertainty. [79, 91]
	Intellectual property rights	Legally enforceable claims for owners of original ideas, inventions, and creative expressions. [92]
FAIR Principles Findability, Accessibility, Interoperability, Reusability: four concise and measurable guidelines designed and broadly endorsed to support the reusability of data.	Data born FAIR	Data objects that comply with the FAIR principles when first generated or produced.
	Data made FAIR	Data objects that are transformed or changed in some manner so that they comply with the FAIR principles.
	FAIR digital objects	Standardized, autonomous and persistent entities, which contain the information needed about different kinds of digital objects (e.g., data, metadata, documents, software, semantic assertions, etc.), to enable both humans and machines to Find, Access, Interoperate, and Reuse (FAIR) these digital objects in highly efficient and cost-effective ways. [93]
	FAIR on a continuous scale	Recognition that there is a degree of FAIRness that ranges from fully FAIR to not FAIR, perhaps represented on a numerical scale.
	Guidelines/methodologies for each aspect: F, A, I, R	Means, e.g., standards, best practices, and protocols by which the findability, accessibility, interoperability, and reusability of data may be improved.
	Tools to capture FAIR provenance	Techniques and procedures for collecting and recording the collective information on the FAIRness of a data asset, from its origin to the present.
	FAIR instruments and tools	Equipment, devices, methods, standards, and other tools that enable the findability, accessibility, interoperability, and reusability of data (e.g., SmartAPI). [94]
	Not FAIR data (e.g., legacy data)	Data that are not findable, accessible, interoperable, and reusable to any degree for various reasons, e.g., obtained using old or obsolete instruments or software.
Community-Based Standards Documents, including codes, specifications, recommended practices, classifications, test methods and guides, that are developed by a group with common interests.	General vs. domain-specific	Broadly applicable as opposed to limited to a particular field or area.
	Standards development organizations vs. community consensus	Formal, recognized, standards bodies (e.g., ISO and ASTM International), as opposed to informal, self-assembled groups of individuals or institutions with shared interests (e.g., professional societies).
	Data format and file structure	Data format: the organization of data according to preset specifications. File structure: the way data and code are organized within a file with the goal of reusability. In the context of standards, the syntax, encoding, and file format or media type for storing or transmitting your data (e.g., CSV and JSON). [60, 95–97]
	Metadata format and file structure.	Metadata format: The organization of information metadata according to preset specifications. File structure: The way metadata are organized within a file. In the context of standards, a metadata standard is a high-level document which establishes a common way of structuring and understanding data and includes principles

Generate/Acquire: Topic	Subtopic	Definition
		and implementation issues for utilizing the standard. See the RDA Metadata Standards Catalog. [95, 96, 98, 99]
	Vocabulary and ontology	Vocabulary: A compendium of standardized terms with consistent semantic definitions. Ontology: a description of data structure–of classes, properties, and relationships in a domain of knowledge. [60, 100]
	Interoperability	The capability to seamlessly communicate, execute programs, or transfer data among various functional components that requires the user to have little or no knowledge of the unique characteristics of those components. In the context of standards, Interoperability standards enable the operational processes underlying exchange and sharing of information between different systems to ensure all digital research outputs are Findable, Accessible, Interoperable and Reusable, according to the FAIR principles. [58, 101]
Acquisition Software Computer programs that enable the collection and procurement of data.	Open source vs. proprietary	Programs freely distributed with the source code that anyone must be able to modify and subsequently redistribute modified versions thereof vs. programs that are copyrighted and bear limits against use, distribution and modification that are imposed by their publisher, vendor, or developer. Such programs remain the property of their owner/creator and are used by end-users/organizations under predefined conditions. [102, 103]
	LIMS	A laboratory information management system (LIMS) is a software system developed to support laboratory operations (e.g., track specimens and workflows, and collect, annotate, and aggregate data). [54]
	Instrument control	Software for configuring the operating parameters of an instrument.
	Electronic laboratory notebooks	A software tool that digitally replicates paper lab notebooks traditionally used in the sciences to record experimental results. [104]
	Audio and video recordings	Digital records used to store and preserve the audible or both audible and visual components of events.

Table 4. Process/Analyze Lifecycle Stage

Process/Analyze: Topic	Subtopic	Definition
Types of Processed Data Classifications or categories of data. [69]	Tables, spreadsheets	Table: Numerical and textual information arranged in rows and columns. Spreadsheets: A computer program that can capture, display and manipulate data arranged in rows and columns.
	Charts, graphs	Visual representations of datasets. Charts can take the form of a diagram, picture, or graph. Graphs charts that show the mathematical relationship between varied groups of data. [105]
	Maps, vectors, images	Representations of the relationships between variables, i.e., quantities, phenomena, or entities. Map: a diagrammatic depiction of the association of two or three variables. Vector: a linear depiction of two independent variables; Image: a visual representation of an object.
	Instrument outputs	Raw electronic data generated by a piece of equipment, device, or other tool before any human action on the data and before any processing of the data. [106]
	Dynamic data	Data which are changing frequently and at asynchronous moments. Data that may change after it is recorded and has to be continually updated. [107, 108]
	Datasets from models and simulations	Organized collections of data generated by models (Representations, patterns, or mathematical descriptions that can help scientists replicate a system, process, or research result) and simulations (projects where researchers or developers create a model of some authentic phenomenon to mimic the outcomes that happen in the natural world.) [47, 71, 109, 110]
	Structured data, e.g., hierarchical organization	Data whose elements have been organized into a consistent format and data structure within a defined data model such that the elements can be easily addressed, organized, and accessed in various combinations to make better use of the information (e.g., a relational database). [111]
Preparation and Pre-Processing Methods Techniques by which raw data are transformed into complete datasets with consistent formatting such that data analysis can subsequently be performed. [114]	Data cleaning	The process of detecting and correcting corrupt or inaccurate records from a dataset. This process involves identifying, replacing, modifying, or deleting complete, incorrect, inaccurate, inconsistent, irrelevant, and improperly formatted data. [112]
	De-identification, anonymization	A process by which personal data is irreversibly altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party. [113]
	Amputation and imputation	Amputation: A process whereby some valid data points are selectively deleted from a complete dataset. Imputation: A process used to determine and assign replacement values for missing, invalid, or inconsistent data. [115, 116]
	Aggregation	A process used to combine datasets, usually taken collectively or in the form of a summary.[117]
	Validation and verification	Validation: The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model. Verification: The process of determining that a

Process/Analyze: Topic	Subtopic	Definition
		computational model accurately represents the underlying mathematical model and its solution. [84, 118]
	Curation	The ongoing processing and maintenance of data throughout its lifecycle to ensure long term accessibility, sharing, and preservation. Data curation is composed of research data management and digital preservation and involves processes such as adding metadata to make data more findable and understandable, ingesting data into a repository, validating file checksums and file fixity checks, and other tasks for organizing, cleaning, describing, enhancing, storing, and preserving data. [119]
	Normalization of metadata	The adjustment of metadata elements into standard formats. [120]
Analysis Statistical and/or logical techniques that are systematically applied to describe and illustrate, condense and recap, and evaluate and interpret data, with the goal of producing new, meaningful information. [69]	Manual	Collection, organization, and transformation of data by a human without using a machine or any other tool. [121]
	Exploratory	Techniques that typically use visual tools to, e.g., determine the main characteristics of datasets, find relationships among datasets or variables that may have been unknown or overlooked, and discern trends or differences among datasets. [121, 122]
	Descriptive	Techniques for answering the question, "What happened?", e.g., identifying trends and relationships using current and historical (past) data. [123]
	Diagnostic	Techniques for answering the question, "Why did this happen?", e.g., determining the causes of trends and correlations among datasets or variables. [124]
	Evaluative	Techniques for the systematic determination of merit, worth, value, or significance of datasets, e.g., relevance to the project objectives. [125]
	Predictive	Techniques for answering the question, "What might happen in the future?", e.g., making assumptions about the future using historical data, either manually or with machine-learning algorithms. [126]
	Prescriptive	Techniques for answering the question, "What should we do next?", e.g., informing an optimal course of action or decisions and strategies, often via machine learning. [127]
	Correlational	Techniques that provide a statistical measure indicating how strongly two variables are related and whether that relationship is positive (e.g., when one variable increases, the other also increases) or negative (e.g., when one variable increases, the other decreases). [128–130]
	Statistical	Techniques whereby data are interpreted to uncover patterns and trends. The five basic techniques are mean, standard deviation, regression, hypothesis testing, and sample size determination. [131, 132]
	Automated, autonomous	Techniques that require no human guidance or direct intervention and are based solely on machines, e.g., self-driving vehicles. [133]
Modeling A class of computational	Visualization	Techniques for the representation of data (e.g., graphs, images, and diagrams). Transforming numerical data into a visual or pictorial context in order to assist users in better understanding what the data is telling them. [117, 134]

Process/Analyze: Topic	Subtopic	Definition
methods whereby a representation, pattern, or mathematical description is used to replicate a system, process, or research result. [47]	ML, AI	Machine learning (ML) is a methodology that uses statistics and mathematical models to detect patterns in historical data and learning algorithms to make predictions about new data. Artificial intelligence (AI) is a field of study in which computerized systems can learn, solve problems, and autonomously achieve goals under varying (and sometimes uncertain) conditions. ML is a subset of AI strategies. [135, 136]
	Iterative model fitting	A technique whereby the parameters of a model are adjusted in repeated cycles to improve accuracy of the computation. [137]
	Integrated development environment, e.g., Jupyter, RStudio	An application that facilitates application development, typically a graphical user interface (GUI)-based workbench designed to aid a developer in building software applications combined with all the required tools at hand. Common features include, e.g., debugging, version control, and data structure browsing. [138]
Metadata Data about data, [i.e.], data that define and describe the characteristics of other data. [85]	Types of metadata	The three main categories or classifications of metadata are descriptive, structural, and administrative. [139]
	Responsible parties	Those Individuals whose duties or job functions include the management of metadata, e.g., data owner or metadata steward. [140]
	Specification of metadata standards	Identification and description of those metadata standards categorized as four types: format/technical interchange, structure, content, and value. Standards include recommended practices, classifications, test methods, and guides. [141]
	Linked data structure	A deliberate design for the organization of data (structure) wherein information (metadata) is brought together from different sources (linked) to create a new, richer dataset. [142]
	Persistent identifiers (e.g., DOI, ORCID, ARK, ROR, PIDINST, Handles)	A unique and long-lasting reference that allows for continued access to an entity (e.g., document, dataset, instrument, webpage, contributor, and organization). A persistent identifier (PID) may be connected to a set of metadata describing an object rather than to the object itself. [143, 144]
Provenance The historical, documented record of a data asset that contains details on its origin – where, when, how, and by whom it was generated/acquired /processed – and on all alterations to the data asset. [19, 37]	Original authoritative copy	The single, distinct, absolute version of a dataset from the originating source that is unique, identifiable, and unalterable without detection. It should be sufficient to allow a third party to reproduce the results of the research. [145]
	Version identification	Determination, for a specific time, data, computer code, software, and documents, that allows for the ability to revert to a previous revision, which is critical for data traceability, tracking edits, and correcting mistakes. [82]
	Derivative product	Any data, publication, illustration or visualization, or other work that rearranges, presents, or otherwise makes use of an existing data set. [146]
	Aggregation	A process used to combine datasets, usually taken collectively or in the form of a summary. [117]

Process/Analyze: Topic	Subtopic	Definition
	Subset	A portion of a dataset that is referentially intact. [147]
	Timestamp	Temporal information regarding an event that is recorded by a computer and then stored as a log or metadata. [148]
	CRediT taxonomy	Contributor Roles Taxonomy (CRediT) consists of a high-level taxonomy, including 14 roles, that can be used to represent the roles typically played by contributors to research outputs. [149]
Software A set of instructions, data, or programs used to operate computers and execute specific tasks. [152]	Commercial vs. custom	Commercial software is any software or program designed and developed for licensing or sale to end-users or for serving a commercial purpose (e.g., off-the-shelf programs and games). Custom software is made for an individual or organization and performs tasks specific to their needs. [150, 151]
	Open-source vs. proprietary	Open-source typically refers to software that is freely distributed with source code that can modified by users and modified versions may be redistributed. Proprietary typically refers to software that is copyrighted and bears limits against use, distribution, and modification that are imposed by its publisher, vendor or developer. The software remains the property of its owner/creator and is used by end-users/organizations under predefined conditions. [102, 103]
	Aggregation tools	Software or programs that enable the combination of datasets. [117]
	Surveying tools	Software or programs that aid in the gathering of responses to questions aimed at extracting specific data from a particular group. [45]
	Statistical tools	Software or programs used in statistics, i.e., the collection, organization, analysis, interpretation, and presentation of masses of data. [153]
	Calculation and analysis tools	Software or programs that produce knowledge from organized data to draw conclusions, highlight useful information, and support decision-making.
	APIs	An Application Program Interface (API) is a set of protocols, routines, functions and/or commands that programmers use to facilitate interaction between distinct software services. [154]
	Database management tools	Software or programs that aggregate diverse data into a database or other consistent resource, handle different types of queries, provide security, and perform other functions. [155]
	Testing and validation tools	Methods to determine if software performs the functions for which it was designed. Software to help ensure that the data sent to connected applications is complete, accurate, secure, and consistent. [156]
	Documentation	Written information that describes the software product to the people who develop, deploy and use it, including technical manuals and online material, such as online versions of manuals and help capabilities. The term is sometimes used to refer to source information about the product discussed in design documentation, code comments, white papers and session notes. [157]

Process/Analyze: Topic	Subtopic	Definition
	Reproducibility and uncertainty quantification	Reproducibility: The ability to replicate data using identical tools (e.g., documented metadata, code, methods, and instruments) employed previously by the original researchers or by other researchers without the need for any additional information or communication with the original researchers. Uncertainty quantification: assignment of a numerical value to a non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand. [79, 91]
	Versioning and maintenance	The process of numbering different releases of a particular software program for both internal use and release designation. It allows programmers to know when changes have been made and track changes enforced in the software. At the same time, it enables potential customers to be acquainted with new releases and recognize the updated versions. [83]
	Systems resilience and adaptability	Resilience: The ability of a software system to continue to operate under adverse conditions while maintaining essential operational capabilities; and to recover to an effective operational state in an acceptable time frame. Adaptability: The ability of a software system to tolerate changes in its environment without external intervention. [158, 159]
	Source code repository	A storage location for source code (the fundamental component of a computer program) that holds code, makes code available for use, and organizes code in a logical manner. [160, 161]
	Security and software updates	Patch, upgrade, or other modification to code that corrects security and/or functionality problems in software. [162]
	Standards, protocols, and interfaces	Standards: Codes, programs, and associated documentation that describe how data should be stored or exchanged for the consistent collection and interoperability of that data across different systems, sources, and users. Protocol: set of rules and guidelines. Interface: A program that allows a user to interact with computers in person or over a network or the controls used in a program that allow the user to interact with the program. [163–165]
Workflow and Middleware Workflow is a depiction of a sequence of connected operations or "steps" that illustrates how data flows through an RDMI. A workflow includes tasks, people	LIMS	A laboratory information management system (LIMS) is a software system developed to support laboratory operation (e.g., track specimens, collect and annotate data and workflows, and aggregate data). [54]
	Laboratory notebooks, e.g., electronic, paper	A complete record of the hardware, software, procedures, materials, observations, and relevant thought processes that would enable the work and resulting data to be reproducible. This generally includes an explanation of why the research was done, including any necessary background and references, how the research was performed, the actual data (raw and processed), and where the data are stored. [166]
involved, tools (e.g., hardware, software), input, and output for each step. Middleware	Tools for automated metadata capture	Software, hardware, and methods used to collect and record data about data without the need for manual instruction.
	Anomaly detection and correction tools	Software, hardware, and methods used to identify items (e.g., operations, observations, events, and results), that do

Process/Analyze: Topic	Subtopic	Definition
is a software layer or "glue" situated between applications and operating systems that makes it easier for software developers to perform communication and input/output, so they can focus on the specific purpose of their application. **[168–170]**		not conform to the expected pattern or result (anomaly detection) and to restore such items to the expected pattern or result (anomaly correction). [167]
	Collaboration tools	Software and/or software systems that enable communication and sharing of documents, data, analysis, and/or visualization amongst individuals who are not co-located.
	Decisions regarding the need for additional data	Conclusions by researchers that more data are needed to accomplish project goals.
	Process monitoring and evaluation	Periodic tracking of the operation and results of a workflow component by systematically gathering and analyzing data to assure that the component is functioning properly. [171]
	Containerization	Operating system-level virtualization or application-level virtualization over multiple network resources so that software applications can run in isolated user spaces called containers in any cloud or non-cloud environment, regardless of type or vendor. [172]
	Reusable workflow component	A discrete piece of software that can be compiled and packaged as an activity and reused in multiple processes, thereby reducing duplication and enabling sharing of the software with others. [173]
	Microservices	An approach to software development in which a large application is built from modular software components (i.e., microservices), each of which does one defined job, e.g., messaging. [174]
	Distributed workflow across sites	Computerized information system that is responsible for scheduling and synchronizing the various tasks within the workflow across physical or virtual locations, in accordance with specified task dependencies, and for sending each task to the respective processing entity. [175]
	Comprehensive report generation	The production of a single document which includes all the information needed to reproduce a dataset, including, e.g., methods, format standards, and software versions.
Hardware The physical elements that make up a computer or electronic system and everything else involved that is physically tangible (e.g., processor, memory, storage, communication ports, and peripheral devices). [80]	Compute requirements	Specifications of the raw processing power of a computer to meet the needs for activities, applications, or workloads. (Such power may be characterized as the rate at which operations are performed, e.g., million instructions per second (MIPS).) [176, 177]
	Storage requirements	Specifications and needs for devices and components that store data on a long-term basis for later uses and access (e.g., hard disks and network-attached storage devices). In contrast to storage, memory is the short-term location for temporary data storage. [178]
	Network requirements	Network capability is characterized by stability of the signal, throughput (transfer rate of data from a source system to a destination system), and bandwidth (the amount of data that can be transferred per second, in megabits/sec). [179]
	Accelerator requirements	Specifications and needs for hardware devices designed to improve the overall performance of the computer. Hardware acceleration is a process where applications

Process/Analyze: Topic	Subtopic	Definition
		offload certain computing tasks to specialized hardware components within the system, enabling greater performance and efficiency. [180, 181]

Table 5. Share/Use/Reuse Lifecycle Stage

Share/Use/Reuse: Topic	Subtopic	Definition
Publishing Public disclosure of research datasets and supporting data objects, e.g., associated metadata and software code, in a manner such that the datasets are findable and reusable for others for future research and researchers. Published datasets ideally have a persistent identifier (PID). [185]	Repository, i.e., domain, generalist, institutional	A broad term that refers to a designated location where a collection of digital objects is stored in an organized manner such that the collection is findable, searchable, accessible, and reusable. Types of repositories include domain-specific (e.g., discipline or subject matter); generalist (a variety of data types, format, and content); and institutional (i.e., within an organization). [88, 182, 183]
	Data papers	Publications that contain datasets, without having to be at the stage of presenting further analysis and conclusions, as in a traditional research paper. [184]
	Software	A set of instructions, data, or programs used to operate computers and execute specific tasks. [152]
	Updates to datasets and new software versions	Datasets: The functional process of renewing information already contained in a data base or stored elsewhere that results in the creation of a new record and may result in storage of existing data as history. Software: Patch, upgrade, or other modification to code that corrects functionality problems in software. [162, 186]
	Data linking	The process of collating and cross-referencing data from different sources in order to create a more valuable and meaningful dataset. [187]
	Persistent identifier	A long-lasting and unique reference to a digital object of various types (e.g., document, dataset, and webpage). Persistent identifiers (PIDs) are labels that locate, identify and share information about digital objects. A PID may be connected to a set of metadata describing an object rather than to the object itself. [143, 144]
	Metadata	Data about data, i.e., data that define and describe the characteristics of other data. [85]
	Integrity of data	The reliability and trustworthiness of data throughout its lifecycle. The assurance that a digital object is uncorrupted and can only be accessed or modified by those authorized to do so. [69, 188]
	Quality measures and assessment vis-à-vis fit for purpose	The degree to which a dataset meets the requirements for its planned usage as determined by an evaluation of quality metrics (e.g., accuracy, completeness, consistency, and timeliness). [189]
	Peer review of datasets and metadata	An editorial process prior to publication of a dataset whereby people with a similar degree of expertise and experience as the author review and provide input on the integrity and quality of the dataset.
	Reference data/digital objects in journal articles	Journals have different guidelines concerning the publication of digital objects, e.g., raw data and software, that accompany a traditional article. Examples of these guidelines are depositing data in a relevant repository, citing a dataset by its PID, and linking the dataset to the article. [190]
	Curation	The ongoing processing and maintenance of data throughout its lifecycle to ensure long term accessibility, sharing, and preservation. Data curation is composed of