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

NIST 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.

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.

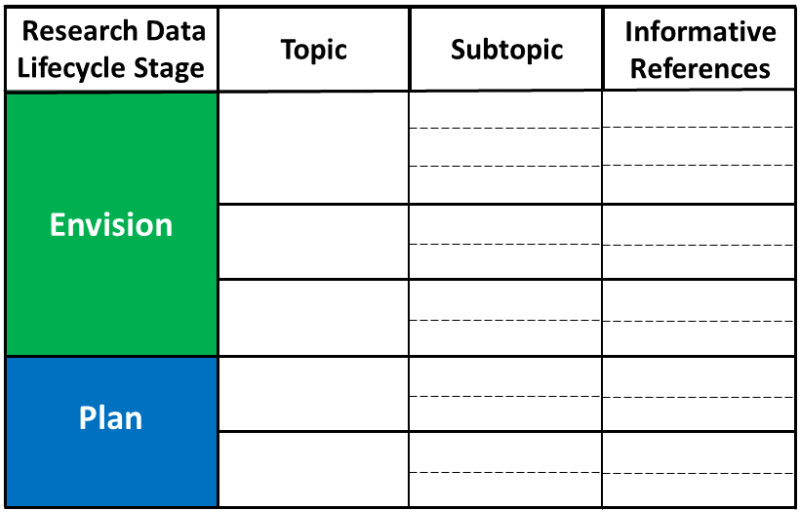

Fig. 1. Partial Organizational Structure of the Framework Core

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.

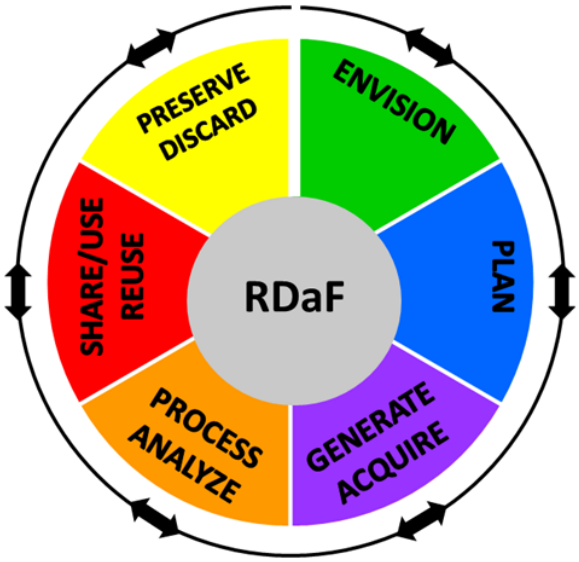

Fig. 2. Research Data Framework Lifecycle Stages

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 Data that is descriptive and concerns phenomena which can be observed but not measured, e.g., language	Nature of objects/subjects	Types and characteristics of entities which are being studied.
	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 Data used in a research study that was not generated by the researchers conducting the study.	From collaborators	Originating from other individuals or other organizations partnering with researchers in an organization.
	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 Numerical data that have undergone rigorous review and critique such that the integrity, reasonableness, and usability are optimized. [90]	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]
	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 involved, tools (e.g., hardware, software), input, and output for each step. Middleware	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]
	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

Share/Use/Reuse: Topic	Subtopic	Definition
		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]
	Publisher agreements and policies	Legal documents that are used to dictate when and how work is published and thereby protect an author's intellectual property from unauthorized use or reproduction. Open access agreements support individual authors to publish open access data at no cost to themselves. Publisher policies are set by the publisher and include, e.g., copyright and licensing, data privacy, and rights and permissions. [191–193]
	Incentives for data publishing	Staff recognition and rewards for widespread dissemination of research data.
	Mitigation of disincentives for data publishing	Practices to remove or reduce barriers that limit dissemination of data (e.g., misinterpretation and misuse of data by others, and lack of recognition and effort for sharing).
Modes of Dissemination Means by which journal articles, datasets, and other data objects are publicly released.	Traditional journal article	A scholarly manuscript submitted to a journal that undergoes a peer review process, an editing and copyediting process, and finally distribution by publishers in the position to print and make high-quality scholarly works available to the world. Such manuscripts typically contain analysis and conclusions, but not digital data objects, e.g., raw data and software. [194]
	Supplementary material	Peer-reviewed material directly relevant to the conclusion of a paper that cannot be included in the printed version for reasons of space or medium (e.g., video clips or sound files) [195]
	On request	Making data available in response to queries typically sent by email. The requester may be required to complete a form, e.g., a data release application agreement. [196]
	Data landing page	A standalone web page that a person [accesses] after clicking on a link from an email, ad, or other digital location. For a dataset, such a web page typically includes a narrative description of the dataset and files or links to files pertaining to the dataset, e.g., the dataset itself and the software used to generate the dataset. [197]
	Workflow	A depiction of a sequence of connected operations or steps that illustrates how data flows through an RDMI. A workflow includes tasks, people involved, tools (e.g., hardware and software), input, and output for each step. [168]
	Mainstream media, e.g., newspapers, radio	Traditional means of communication, such as newspapers, television, and radio that influence large numbers of people. [198]
	Social media, e.g., Twitter, Instagram	A catch-all term for a variety of internet applications that allow users to create content and interact with each other, e.g., Twitter, Facebook, and LinkedIn. [199]
Attribution	Citation metrics	Measures based on the number of times a single entity (e.g., article, data set, or other digital object) published by a

Share/Use/Reuse: Topic	Subtopic	Definition
Acknowledgement of the use of an individual's published articles, data, or other data objects.		researcher is mentioned in the published work of other authors. Indicator of the quality or importance of a published entity. Citation data are available from citation databases, discipline specific databases, and through an emerging range of alternative metrics. [200]
	Citation impact	Quantitative and qualitative tools and methods to measure the impact of an individual's collective work. Quantitative tools, include citation analysis--counting the number of times other authors mention a researcher's published works; the impact factors (IFs) of the journals in which a researcher has published their work (IF is the frequency with which the average article in a journal has been cited in a particular year); and the h-Index for a researcher which is based on the set of the researcher's most cited papers and the number of citations that they have received in other authors' publications. Qualitative methods include anecdotal evidence. [201, 202]
	Dataset citation	The practice of referencing data products used in research (e.g., a DOI or key descriptive information about the data, such as the title, source, and responsible parties). Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse. (see the Joint Declaration of Data Citation Principles) [203–205]
	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]
	Author identity management	Use of a persistent, unique, digital researcher identifier such as ORCID to, e.g., track the scholarly outputs of a researcher, assign appropriate author credit, and eliminate author name ambiguity. [206]
	Use of persistent identifiers	The practice of assigning a long-lasting and unique reference to a digital object. [143, 144]
	Versioning	The process of numbering different release of a software program; the use and management of multiple versions of a document. Version control allows for the ability to revert to a previous revision, which is critical for data traceability, tracking edits, and correcting mistakes. [82, 83, 152]
Modes of Sharing Methods whereby datasets and other digital objects are publicly or privately distributed or are accessible to others upon request.	Standardized formats	The organization of information according to preset specifications that are agreed upon by formal standards bodies or informal community groups.
	Interoperability tools	Methods that provide 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]
	Discovery platform	A software system that uses metadata to identify and recommend sources of data or other digital objects. [207]
	Catalog	A completely organized service that enables any user, from analysts and data scientists to developers, to discover, explore, and use data sources. [208]

Share/Use/Reuse: Topic	Subtopic	Definition
	Registries of repositories	Databases containing information about trusted repositories that are provided by the repository managers and are useful for human and machine users, e.g., the Re3data Repository Registry and the NIST Materials Resource Registry. [209–211]
Access The ability of a user to view and retrieve data [and other digital objects] stored within a database or other repository. Users who have data access can store, retrieve, move or manipulate stored data, which can be stored on a wide range of hard drives and external devices. [212]	Internal access	The ability of individuals in an organization to view and retrieve data and other digital objects that were generated, collected, or processed by an individual or group in the same organization.
	External access	The ability of individuals in organizations other than the organization that generated, collected, or processed the data and other digital objects to view and retrieve such digital resources.
	Programmatic access, i.e., API	The ability of a user to view and retrieve data made possible by an Application Program Interface (API), which is a set of protocols, routines, functions and/or commands that programmers use to facilitate interaction between distinct software services. [154]
	Virtual and physical enclaves	A secure network through which confidential data, such as identifiable information from census data, can be stored and disseminated. In a virtual data enclave, a researcher can access the data from their own computer but cannot download or remove it from the remote server. Higher security data can be accessed through a physical data enclave where a researcher is required to access the data from a monitored room where the data is stored on non-network computers. [213]
	Access vs. visiting	Data visiting is an approach whereby sensitive data stays under the control of the owner and allows the consumers (e.g., analysts or machine learning algorithms) to come to the data to work with it. With data access, users can store, retrieve, move, or manipulate stored data. [214]
	Availability statement	A declaration letting a user know where and how to access data that support the results and analysis. It may include links to publicly accessible datasets that were analyzed or generated during the study, descriptions of what data are available and/or information on how to access data that is not publicly available. [215]
	Mitigation of barriers and economic constraints	Practices that reduce or eliminate the programmatic and administrative constraints and transactional costs of accessing data.
Legal and Licenses Juridical and regulatory issues as pertaining to research data.	Ownership	The act of having legal rights and complete control over a single piece or set of data elements. Ownership defines and provides information about the rightful owner of data assets and the acquisition, use and distribution policy implemented by the data owner. [216]
	Encouragement and support for sharing, use, and reuse	Incentives and human and infrastructural resources that increase the quantity and quality of data assets for access and dissemination.
	Indigenous data rights	Indigenous data sovereignty (IDS) refers to the right of Indigenous peoples to govern the collection, ownership, and application of data about Indigenous communities, peoples, lands, and resources. IDS encompasses data, information,

Share/Use/Reuse: Topic	Subtopic	Definition
		and knowledge about Indigenous individuals, collectives, entities, lifeways, cultures, lands, and resources. [30]
	Intellectual property rights/restrictions	Intellectual property (IP) is something of value (an asset) that is created from an original idea Intellectual property rights (IPR) provide certain exclusive rights to the creators of intellectual property so they can reap commercial benefits from their ideas. [217]
	Usage agreements/terms/licenses and required permissions	Usage agreements: A legally binding contract between the originator of a digital object and the user of the object that spells out the rights and responsibilities of all involved parties. User license: A written contract that gives a user permission to use another party's digital object under a certain set of conditions and typically requires that the user pay a royalty fee. [218, 219]
	Data sharing and licensing agreements	Sharing agreements: Formal contracts that detail what data are being shared and the appropriate use of the data and include provisions concerning access and dissemination. Licensing agreements: Documents that describe what kind of data is being shared with a user and clearly states the purpose and duration of access being provided to the user along with restrictions and security protocols that the user of the data must follow. [22, 23]
	Service-level agreements	A contract between two parties that defines and measures the level of service a data provider will deliver to a user They are an attempt to define expectations of the level of service and quality between data providers and users. [220]
	Terms of service	A legal agreement between a data service provider and a user that details the set of rules and regulations a provider attaches to a software service or Web-delivered product. [221]
	Standardized, machine-actionable license documents	Written contracts in a common, agreed-upon form that can be read, understood, and implemented by a computer. The contract gives a user permission to use a creator's digital object under a certain set of conditions.
	Citation requirements	References to data and other digital objects that are mandated by a data provider, formal agreement, or publishing entity.
Levels of Protection Classification scheme based on potential harm resulting from unauthorized access, disclosure, loss of privacy, compromised integrity, or violation of external obligations. [225]	Unclassified but sensitive information	A term describing data that are not classified for national security reasons, but that warrant or require administrative control and protection from public or other unauthorized disclosure for other reasons, e.g., Personally Identifiable Information (PII) or Business Identifiable Information (BII). The US government uses the term “controlled unclassified information (CUI). [67, 222–224]
	Security classification	A term typically associated with U.S. federal government national security information. NIST has developed a broader document that addresses security controls, defined as the safeguards or countermeasures employed within a system or an organization to protect the confidentiality, integrity, and availability of the system and its information and to manage information security risk. [226, 227]

Share/Use/Reuse: Topic	Subtopic	Definition
	Protection of limited data/secure platforms/enclaves	Limited data: In healthcare, a set of identifiable healthcare information that the HIPAA Privacy Rule permits covered entities to share with certain entities for research purposes if certain conditions are met. Data security platform: aggregates data protection requirements across data types, storage silos, and ecosystems to create an overarching, organization encompassing data security solution. Secure data enclave: a system that allows data owners to control data access and ensure data security while facilitating approved uses of data by other parties. [228–230]
	Constraints and restrictions on data use and sharing	Technical, administrative, or legal limitations on the use and sharing of data.
	Anonymization	The process of preserving private or confidential information by deleting or encoding identifiers that link individuals and the stored data. [231]
Architectures for Application, Use, and Reuse 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. Such a structure should enable a user to capitalize on an organization's data. [56, 57]	Extensibility across communities, including machine-based interactions	A measure of the ability to expand an RDM architecture to enable interactions with a broad group of stakeholders and types of equipment, achieved by adding new functionality or modifying existing functionality. [232]
	Capture of insights from ML and use of these to improve datasets for future AI applications	Recording and retaining information obtained via computer systems that use algorithms and statistical models to enable understanding of complex problems and employing such understanding to develop enhanced datasets for new AI solutions.
	Capture of data performance characteristics	Recording and retaining information concerning the quality attributes of a dataset, e.g., validity, accuracy, completeness, relevance, uniformity and consistency. [233]
	Location of data	Methods whereby, and systems and devices wherein, data are saved and from which data can be retrieved, e.g., on premises, cloud, temporary cache, and removable media.
	Migration strategies and mitigation against data loss	Approaches and practices to eliminate, prevent, or reduce the intentional or unintentional destruction or disappearance of information caused by people, processes, or other means.
	Economic impact of reuse	Monetary benefits of using existing data compared to re-generating identical data.

Table 6. Preserve/Discard Lifecycle Stage

Preserve/Discard: Topic	Subtopic	Definition
Criteria for Preservation Quantitative and qualitative metrics used to assess the need for long-term retention of data. [234]	Use	Instances wherein datasets are utilized for meaningful purposes, e.g., problem-solving and decision-making.
	Impact	Demonstrated, positive outcomes attributed to use of a dataset, e.g., a scientific discovery and a new measurement instrument or product.
	Value	Merit or worth of data in terms of their usefulness and fitness for purpose, e.g., to make sound, fact-based conclusions and decisions.
	Uniqueness	The quality of being unlike any other data in terms of, e.g., type and characteristics. [235]
	Cost	Financial resources required to store and preserve data.
	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]
	Legal and regulatory	Requirements via contract, law, regulation, or other agreement to preserve data.
Sustainability The capacity to maintain or improve the state and availability of [data and an RDM infrastructure] over the long term. [237]	Longevity and support	The amount of time a dataset is retained in an organization and the resources to maintain this retention. [236]
	Funding models	Approaches to build a reliable funding base that will support an organization's core research data projects and services. [38]
	Business models	Approaches to describe how an organization ensures that its research data projects and services provide value. [238]
Storage and Preservation Storage: a process whereby digital data are saved for later uses and access via, e.g., a device or cloud service. Preservation: a series of managed activities necessary to ensure continued stability and access to data for as long as necessary. [178, 243]	Media to store and preserve data	Devices used to retain data in the short-term and long-term. [239]
	File integrity	The process of protecting a file from unauthorized changes or environmental hazards, i.e., validation to determine whether or not a file has been altered after its creation, curation, archiving, or other qualifying event. [240, 241]
	Ability to do advanced searches	Capability to narrow a query through, e.g., the use of filters that eliminate irrelevant information and enable the identification of desired content. [242]
	Backup and recovery	Backup: the process of making copies of data or data files to use in the event the original data or data files are lost or destroyed. Recovery: the process of restoring data that have been lost, accidentally deleted, corrupted or made inaccessible for any reason. [244, 245]
Moving Data from One Service to Another Across Organizations Inter-organizational transit of data.	Roles and responsibilities	The job functions and obligations that enable the movement of data among organizations.
	Registry maintenance and curation	The processes of harvesting, organizing, and handling a collection of data-related resources such as repositories, services, and software, to facilitate ease of user searches and retrieval of information. Examples of registries are re3data and the NIST Materials Resource Registry. [210, 211]

Preserve/Discard: Topic	Subtopic	Definition
	Disciplinary archives	A place to store data from a specific field of study or branch of knowledge that is important but that doesn't need to be accessed or modified frequently (if at all). [69, 246]
Retention and Disposition Schedules A timeline and plan of action based on a policy that addresses which data are important to keep for future use or reference, how that data can be searched and accessed at a later date, and which data are no longer needed and can be destroyed. [248]	Technical decisions, e.g., data archiving	Conclusions regarding retention and disposition of research data that are based on scientific considerations such as merit and future potential usefulness of the data.
	Administrative/policy decisions	Conclusions regarding retention and disposition of research data that are based on logistical or operational considerations, e.g., cost of data archiving.
	Deaccessioning/end-of-life	The formal, documented removal of a data collection or item from its location or custody of an archive service. [247]
	Legal documents	Schedules for retention and disposition of data set by formal contracts or other agreements.
	End-of-life special considerations	Any actions taken before disposition of data that has reached the end of its useful life or will no longer receive continuing support for archiving. An example consideration is adhering to security protocols for sensitive data.
	Recognition of removed data, i.e., tombstone page	Creation of a special type of landing page describing the data item that has been removed that provides a full bibliographic citation, a DOI (if one has been assigned), and a statement on unavailability detailing the circumstances that led to removal of the data item. [249]

4. Overarching Themes

The framework was refined from the published version 1.0 using input from the 2 Opening Plenary Workshops and the 15 Stakeholder Workshops. During this refinement process, 14 themes that spanned the various lifecycle stages were identified. Rather than repeat these themes in each stage, they are listed here with a brief explanation of their meaning in the context of research data. The specific lifecycle stages in which these themes appear are shown in tabular form after a brief narrative for each theme.

In most cases, the overarching themes are supported by explicit references in the framework. In other cases, the themes are implicit. For example, Cost Implications and Sustainability touches on every topic or subtopic, although it is not called out in any lifecycle stage: there is a financial implication to every decision and action that will be considered by those working with research data in any capacity.

Separate tables generated for each overarching theme document the topics and subtopics most closely associated (see Tables 7 through 20 below). There are also two graphics that provide summary information. Figure 3 is a Sankey diagram that provides a visualization of the relationship between each lifecycle stage and each overarching theme. Figure 4 is a matrix table that gives a high-level overview of the relationships between the overarching themes and the topics for each lifecycle stage.

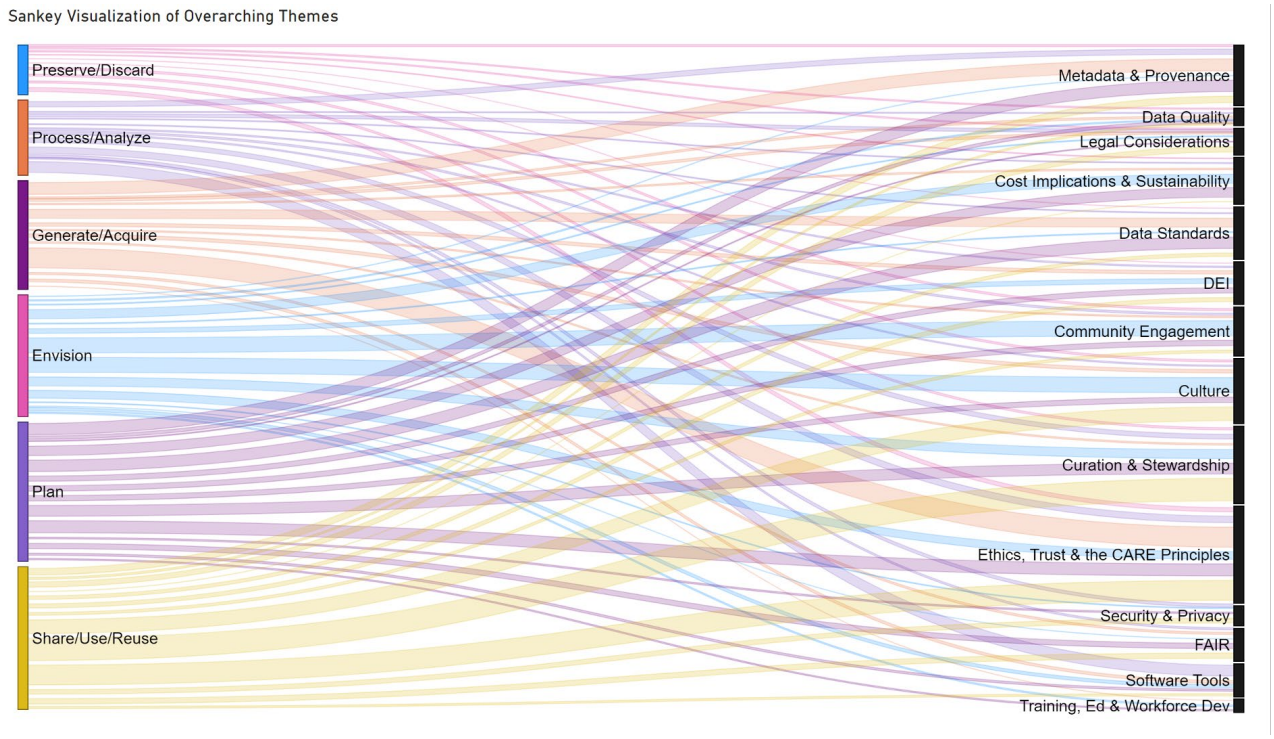

Fig. 3. Sankey Diagram of the Relationships Between Lifecycle Stages and Overarching Themes

Lifecycle Stage	Topic	Ethics, Trust & the CARE Principles	Metadata & Provenance	DEI	Curation & Stewardship	Data Standards	Culture	Community Engagement	Cost Implications & Sustainability	FAIR	Software Tools	Legal Considerations	Data Quality	Security & Privacy	Training, Ed & Workforce Dev
ENVISION	Data Governance—Strategic/Qualitative	x	x	x	x	x	x	x	x	x		x	x		
	Data Governance—Legal and Regulatory Compliance	x	x	x			x	x	x	x		x		x	
	Data Culture and Reward Structure	x	x	x	x	x	x	x	x	x		x	x		x
	Education and Workforce Development		x	x	x	x	x	x	x		x	x	x	x	x
	Resources—Allocation and Sustainability	x			x		x	x	x						x
	Community Engagement	x		x	x	x	x	x	x	x					x
PLAN	Chain of Custody	x	x		x		x					x			
	Financial Aspects of Planning			x	x		x	x	x						x
	Data Management Planning	x	x	x	x	x		x	x	x	x				x
	Data Objects	x				x					x				
	FAIR		x		x	x	x	x		x					x
	Data/Metadata Considerations	x	x	x		x			x	x					
	Data Architecture	x	x			x			x		x			x	
	Hardware and Software Infrastructure	x	x	x		x	x	x	x	x	x	x		x	x
	Research Data Standards	x	x	x	x	x	x	x		x			x		
	Assessment			x	x		x			x					
	Communication and Outreach	x	x	x	x		x	x		x					
	Access Control Associated with Data Sensitivity	x	x	x	x				x	x		x		x	
GENERATE/ACQUIRE	Data Types	x		x		x					x				
	Data Sources	x		x											
	Generated Experimental Data	x	x								x		x		
	Generated Computational Data	x	x						x		x		x		
	Qualitative Data	x	x	x											
	Acquired Data	x	x	x		x			x			x			
	Critically Evaluated (CE) Data	x	x			x						x	x		
	FAIR Principles		x		x	x	x	x		x					
	Community-Based Standards	x	x	x	x	x	x	x		x					x
	Acquisition Software								x		x				
PROCESS/ANALYZE	Types of Processed Data														
	Preparation and Pre-Processing Methods	x	x	x	x		x						x		
	Analysis Methods														
	Modeling	x		x							x				
	Metadata		x	x	x	x		x		x					
	Provenance	x	x	x	x	x		x							
	Software					x	x		x		x	x	x	x	
	Workflows and Middleware	x	x					x	x		x				
	Hardware								x						
SHARE/USE/REUSE	Publishing	x	x	x	x	x	x	x	x	x	x		x		
	Modes of Dissemination	x	x				x	x							
	Attribution	x	x	x	x	x	x								
	Modes of Sharing		x		x	x	x			x			x		
	Access	x		x	x		x			x				x	
	Legal and Licenses	x	x	x	x	x	x	x	x	x	x	x		x	
	Levels of Protection	x		x	x							x		x	
	Architectures for Application, Use, and Reuse	x		x											
PRESERVE/DISCARD	Criteria for Preservation	x	x	x	x	x		x	x			x	x		
	Sustainability						x	x	x						
	Storage and Preservation					x			x						
	Moving Data from One Service to Another across Organizations	x			x	x	x	x							
	Retention and Disposition Schedules	x	x	x	x	x	x	x				x			

Fig. 4. Matrix Diagram of Topics and Overarching Themes

4.1. Community Engagement

Community engagement, typically more broadly for RDM practices and more focused for research data programs, is an intentional set of approaches for both listening to and communicating with stakeholders. Successful research, data management, and data curation come from strong engagement with the community of practice or discipline and the organization in which the research occurs. Engagement is present in all the RDaF lifecycle stages, although there is an emphasis on it within the Envision and Plan stages. Engaging with stakeholders early in the research process may lead to stronger outcomes and uptake of new research. The community engagement that is evident in other lifecycle stages demonstrates that no researcher is an island and that stakeholder engagement throughout the stages is essential for accomplishing the goals set out at the beginning of a research project.

Table 7 lists the topics and subtopics that are most relevant to the overarching theme of Community Engagement.

Table 7. Community Engagement (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance – Strategic/Qualitative	Identification of goals and roles
		Vision and/or policy
		Data management organization
		Organizational values, including DEI
		Data management value proposition
		Needs assessment
		Organization intent regarding FAIR data
		End-use support
		Stewardship
	Data Governance—Legal and Regulatory Compliance	Privacy
		Ethics
	Data Culture and Reward Structure	Roles and responsibilities
		Recognition of data management
		Value of data workers
		Promotion and tenure
		Integrity of research and data
		FAIR data principles
		Incentives and impact for sharing and reuse
		Disincentives for sharing and reuse
		CARE and ethics
	Education and Workforce Development	Workforce skills inventory
		Workforce preparedness in new and advanced technologies

Lifecycle Stage	Topic	Subtopic
		Data management training
		HR's supporting role in workforce development and training
		Promotional paths and career development
	Resources—Allocation and Sustainability	Staffing
	Community Engagement	Stakeholder communities
		Partners/partnerships
		Engagement across knowledge domains and sectors
		Inclusivity in interactions
		Data services and the beneficiaries
Plan	Financial Aspects of Planning	Staffing and training
	Data Management Planning	Purpose/intent of research study and context of anticipated data use
		Specification of data objects, metadata, analysis tools, and workflows throughout the lifecycle
		Data organization, e.g., database, repository, to facilitate future access
		Data management expertise and training
	FAIR	Organizational support for making data more FAIR
	Hardware and Software Infrastructure	Interoperability
		Security and privacy considerations
	Research Data Standards	Sources of standards/guidelines for data/metadata
		Community-based standards/conventions
	Communication and Outreach	Methods to share and reuse data/metadata
		Allocation of credit to project team members
		Promotion of data to communities of interest
		Cross-institution cooperation
		Requests for additional data from community
Generate/Acquire	FAIR Principles	Guidelines/methodologies for each aspect: F, A, I, R
	Community-Based Standards	General vs. domain-specific
		Standards development organizations vs. community consensus
		Vocabulary and ontology
Process/Analyze	Metadata	Responsible parties

Lifecycle Stage	Topic	Subtopic
	Provenance	CRediT taxonomy
	Workflows and Middleware	Collaboration tools
Share/Use/Reuse	Publishing	Repository, i.e., domain, generalist, institutional
		Peer review of datasets and metadata
		Curation
		Publisher agreements and policies
		Incentives for data publishing
		Mitigation of disincentives for data publishing
	Modes of Dissemination	Data landing pages
		Curation
		Publisher agreements and policies
	Legal and Licenses	Indigenous data rights
		Usage agreements/terms/licenses and required permissions
Preserve/Discard	Criteria for Preservation	Use
		Impact
		Value
		Uniqueness
	Sustainability	Longevity and support
		Funding models
	Moving Data from One Service to Another Across Organizations	Roles and responsibilities
		Registry maintenance and curation
		Disciplinary archives
	Retention and Disposition Schedules	End-of-life special considerations

4.2. Cost Implications and Sustainability

The cost of research and research data management is a theme that touches every lifecycle stage and most stakeholders in the research enterprise. From Chief Data Officers and Provosts to bench researchers and grant administrators, cost is a constant focus of the work of individuals in public and private organizations. Administrators and C-suite officers would typically focus their efforts on the stages of Envision and Plan, while researchers, particularly those with curation duties and service provision, have more impact on the cost implications in the Generate/Acquire, Process/Analyze, Share/Use/Reuse, and Preserve/Discard stages.

Sustainability in this usage means sustainable funding, staffing, and preservation models as applied to research data. It is imperative that sustainable plans affecting these three areas are

assessed as they are developed and maintained, or institutions and users risk losing access to valuable datasets.

Table 8 lists the topics and subtopics that are most relevant to the overarching theme of Cost Implications and Sustainability.

Table 8. Cost Implications and Sustainability (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance – Strategic/Qualitative	Data management organization
		Needs assessment
		Organization intent regarding FAIR data
		End-use support
		Stewardship
	Data Governance—Legal and Regulatory Compliance	Risk assessment
		Risk mitigation and management
	Data Culture and Reward Structure	Value of data workers
		Promotion and tenure
		FAIR data principles
		Maintenance of FAIR data
		Incentives and impact for sharing and reuse
		Disincentives for sharing and reuse
	Education and Workforce Development	Workforce preparedness in new and advanced technologies
		Data management training
		Promotional paths and career development
	Resources—Allocation and Sustainability	Sources of funding
		Long-term funding
		Staffing
	Community Engagement	Partners/partnerships
		Data services and the beneficiaries
Planning	Financial Aspects of Planning	Funding models for provisioning resources
		Funding sources
		Decision-making tools to assess costs
		Cost-benefit analysis
		Cost breakdown by lifecycle stage
		Downstream lifecycle costs, e.g., technology refresh, infrastructure maintenance

Lifecycle Stage	Topic	Subtopic
		Staffing and training
	Data Management Planning	Purpose/intent of research study and context of anticipated data use
		Data organization, e.g., database, repository, to facilitate future access
		Data management expertise and training
	Data/Metadata Considerations	Criteria for selection of data/metadata
	Data Architecture	Design
		Hosting and storage, cloud storage
		Security
	Hardware and Software Infrastructure	Organizational research needs
		Sustainability of data vis-à-vis obsolete infrastructure
		Security and privacy considerations
		Staff expertise and support staff
	Access Control Associated with Data Sensitivity	Regulatory compliance
		Sensitive data/PII
		Limited disclosure, e.g., IP
		Licensing for reuse
Generate/Acquire	Generated Computational Data	Hardware
		Parameters and conditions for computation, e.g., OS dependencies, compilers, memory requirements
	Acquired Data	From collaborators
		In repositories
		From the literature
		Aggregated datasets from multiple sources
		Restrictions, fees, and usage agreements
	Acquisition Software	Open source vs. proprietary
		LIMS
Process/Analyze	Software	Commercial vs. custom
		Open-source vs. proprietary
	Workflows and Middleware	LIMS
		Collaboration tools
	Hardware	Compute requirements
		Storage requirements
		Network requirements
		Accelerator requirements
Share/Use/Reuse	Publishing	Repository, i.e., domain, generalist, institutional
		Publisher agreements and policies

Lifecycle Stage	Topic	Subtopic
	Legal and Licenses	Ownership
		Sharing agreements and licensing
		Service-level agreements
	Architectures for Application, Use, and Reuse	Economic impact of reuse
Preserve/Discard	Criteria for Preservation	Cost
	Sustainability	Longevity and support
		Funding models
		Business models
	Storage and Preservation	Media to store and preserve data

4.3. Culture

Culture is the basis for the entirety of a given organization's success in managing data and in nearly every other aspect of running a collective enterprise; culture is what gives an institution or organization its character and consistency over time. Cultures are firmly embedded and stem from both informal practices and formal written policies which can make them difficult to change. Culture shapes norms within an organization and creates glide paths towards ingrained values and behaviors as well as resistance to others. Specifically, culture relates to how research data are valued or supported in an institution.

Table 9 lists the topics and subtopics that are most relevant to the overarching theme of Culture.

Table 9. Culture (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance – Strategic/Qualitative	Identification of goals and roles
		Vision and/or policy
		Data management organization
		Organizational values, including DEI
		Data management value proposition
		Purpose and value of data
		Organization intent regarding FAIR data
		Stewardship
	Data Governance—Legal and Regulatory Compliance	Ethics
		Safety and security assurance
		Risk mitigation and management
		Sharing/licensing

Lifecycle Stage	Topic	Subtopic
	Data Culture and Reward Structure	Roles and responsibilities
		Recognition of data management
		Value of data workers
		Promotion and tenure
		Integrity of research and data
		FAIR data principles
		Maintenance of FAIR data
		Incentives and impact for sharing and reuse
		Disincentives for sharing and reuse
		CARE and ethics
	Education and Workforce Development	Workforce preparedness in new and advanced technologies
		Data management training
		HR's supporting role in workforce development and training
		Promotional paths and career development
	Community Engagement	Stakeholder communities
		Partners/partnerships
		Engagement across knowledge domains and sectors
		Inclusivity in interactions
		Data services and the beneficiaries
Plan	Chain of Custody	Roles and responsibilities
	Financial Aspects of Planning	Funding models for provisioning resources
	FAIR	Organizational support for making data more FAIR
	Hardware and Software Infrastructure	Organizational research needs
		Interoperability
		Security and privacy considerations
		Staff expertise and support staff
	Research Data Standards	Criteria, i.e., general vs. domain-specific standards
		Quality standards
		Community-based standards/conventions
	Communication and Outreach	Methods to share and reuse data/metadata
		Allocation of credit to project team members
		Promotion of data to communities of interest
		Cross-institution cooperation
Generate/Acquire	FAIR Principles	Data born FAIR

Lifecycle Stage	Topic	Subtopic
		Data made FAIR
		FAIR digital objects
		FAIR on a continuous scale
		Guidelines/methodologies for each aspect: F, A, I, R
		Tools to capture FAIR provenance
		FAIR instruments and tools
		Not FAIR data, e.g., legacy
	Community-Based Standards	General vs. domain-specific
		Standards development organizations vs. community consensus
		Metadata format and file structure
		Interoperability
Process/Analyze	Preparation and Pre-Processing Methods	De-identification, anonymization
		Curation
	Software	Commercial vs. custom
		Open-source vs. proprietary
Share/Use/Reuse	Publishing	Repository, i.e., domain, generalist, institutional
		Data papers
		Software
		Updates to datasets and new software versions
		Data linking
		Persistent identifier
		Metadata
		Integrity of data
		Peer review of datasets and metadata
		Reference data/digital objects in journal articles
		Curation
		Incentives for data publishing
		Mitigation of disincentives for data publishing
	Modes of Dissemination	Traditional journal article
		Supplementary material
		On request
		Data landing pages
		Workflows
		Mainstream media, e.g., newspapers, radio
		Social media, e.g., Twitter, Instagram
	Attribution	Dataset citation
	Modes of Sharing	Standardized formats
	Access	Availability statement

Lifecycle Stage	Topic	Subtopic
		Mitigation of barriers and economic constraints
	Legal and Licenses	Ownership
		Encouragement and support for sharing, use, and reuse
		Indigenous data rights
		Sharing agreements and licensing
Preserve/Discard	Criteria for Preservation	Use
		Impact
		Value
		Uniqueness
	Sustainability	Longevity and support
		Funding models
	Moving Data from One Service to Another Across Organizations	Roles and responsibilities
		Registry maintenance and curation
		Disciplinary archives
	Retention and Disposition Schedules	End-of-life special considerations

4.4. Curation and Stewardship

The processes and procedures to make research data shareable and reusable are typically referred to as data curation or stewardship. Both curation and stewardship, and the job roles that are responsible for them, aim to collect, manage, preserve, and promote research data over their lifecycles. Curation is often performed by librarians and others outside of a laboratory or research group, while Data Stewards tend to work with a specific research group, lab, or department (i.e., a specific discipline) to ensure that they are embedded in research projects from the onset of the Plan lifecycle stage. Because curators tend to work outside the lab, they are typically engaged in research projects much later during the Share/Use/Reuse stage, which may introduce complications. The Curation and Stewardship theme implicitly touches each lifecycle stage.

Table 10 lists the topics and subtopics that are most relevant to the overarching theme of Curation and Stewardship.

Table 10. Curation and Stewardship (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance – Strategic/Qualitative	Data management organization

Lifecycle Stage	Topic	Subtopic
		Organization intent regarding FAIR data
		Stewardship
	Data Culture and Reward Structure	Roles and responsibilities
		Recognition of data management
		Value of data workers
		Promotion and tenure
		Integrity of research and data
		FAIR data principles
		Incentives and impact for sharing and reuse
		Disincentives for sharing and reuse
		CARE and ethics
	Education and Workforce Development	Workforce skills inventory
		Data management training
		Promotional paths and career development
	Resources—Allocation and Sustainability	Staffing
	Community Engagement	Stakeholder communities
		Partners/partnerships
		Engagement across knowledge domains and sectors
		Inclusivity in interactions
		Data services and the beneficiaries
Plan	Chain of Custody	Roles and responsibilities
	Financial Aspects of Planning	Staffing and training
	Data Management Planning	Written data management plans (DMPs)
		Specification of data objects, metadata, analysis tools, and workflows throughout the lifecycle
		Machine-readable DMPs
		Data organization, e.g., database, repository, to facilitate future access
		Data management expertise and training
	FAIR	Organizational support for making data more FAIR
		Identification of methods/guidelines vis-à-vis FAIR principles
	Research Data Standards	Criteria, i.e., general vs. domain-specific standards
		Sources of standards/guidelines for data/metadata

Lifecycle Stage	Topic	Subtopic
		Quality standards
		Community-based standards/conventions
	Assessment	Metrics for tracking use and impact measures, including reuse
	Communication and Outreach	Methods to share and reuse data/metadata
		Allocation of credit to project team members
		Promotion of data to communities of interest
		Cross-institution cooperation
		Requests for additional data from community
	Access Control Associated with Data Sensitivity	Identification of responsible parties for access management
		Regulatory compliance
		Sensitive data/PII
		Limited disclosure, e.g., IP
		Licensing for reuse
Generate/Acquire	FAIR Principles	Data made FAIR
		Guidelines/methodologies for each aspect: F, A, I, R
		Not FAIR data, e.g., legacy
	Community-Based Standards	General vs. domain-specific
		Standards development organizations vs. community consensus
		Data format and file structure
		Metadata format and file structure
		Vocabulary and ontology
		Interoperability
Process/Analyze	Preparation and Pre-Processing Methods	Curation
		Normalization of metadata
	Metadata	Types of metadata
		Responsible parties
		Specification of metadata standards
		Linked data structure
		Persistent identifiers (e.g., DOI, ORCID, ARK, ROR, PIDINST, Handles)
	Provenance	Original authoritative copy
		Version identification
		Derivative product
		Aggregation
		Subset
		Timestamp
		CrediT taxonomy

Lifecycle Stage	Topic	Subtopic
Share/Use/Reuse	Publishing	Repository, i.e., domain, generalist, institutional
		Data papers
		Software
		Updates to datasets and new software versions
		Data linking
		Persistent identifier
		Metadata
		Integrity of data
		Quality measures and assessment vis-à-vis fit for purpose
		Peer review of datasets and metadata
		Reference data/digital objects in journal articles
		Curation
		Publisher agreements and policies
		Incentives for data publishing
		Mitigation of disincentives for data publishing
	Attribution	Citation metrics
		Citation impact
		Dataset citation
		Provenance
		Author identity management
		Use of persistent identifiers
		Versioning
	Modes of Sharing	Standardized formats
		Interoperability tools
		Discovery platform
		Catalog
		Registries of repositories
	Access	Internal access
		External access
		Programmatic access, i.e., API
		Virtual and physical enclaves
		Access vs. visiting
		Availability statement
		Mitigation of barriers and economic constraints
	Legal and Licenses	Indigenous data rights
		Usage agreements/terms/licenses and required permissions
		Ownership
		Encouragement and support for sharing, use, and reuse
		Indigenous data rights
		Intellectual property rights/restrictions

Lifecycle Stage	Topic	Subtopic
		Standardized, machine-actionable license documents
		Citation requirements
	Levels of Protection	Constraints and restrictions on data use and sharing
Preserve/Discard	Criteria for Preservation	Use
		Impact
	Moving Data from One Service to Another Across Organizations	Roles and responsibilities
		Registry maintenance and curation
		Disciplinary archives
	Retention and Disposition Schedules	Technical decisions, e.g., data archiving
		Administrative/policy decisions
		Deaccessioning/end-of-life
		End-of-life special considerations
		Recognition of removed data, i.e., tombstone page

4.5. Data Quality

The quality of a given dataset directly impacts its fitness for purpose, usability, and reusability. All parties involved in every stage of a dataset's lifecycle should be cognizant of data quality. The CODATA Research Data Management Terminology [3] definition of data quality includes the following attributes: accuracy, completeness, update status, relevance, consistency across data sources, reliability, appropriate presentation, accessibility. Assessment of data quality is not a single process, but rather a series of actions that, over the lifetime of a dataset, collectively assure the greatest degree of quality.

Table 11 lists the topics and subtopics that are most relevant to the overarching theme of Data Quality.

Table 11. Data Quality (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance – Strategic/Qualitative	Purpose and value of data
		Stewardship
	Data Culture and Reward Structure	Roles and responsibilities

Lifecycle Stage	Topic	Subtopic
	Education and Workforce Development	Data management training
Plan	Research Data Standards	Quality standards
Generate/Acquire	Generated Experimental Data	Verification/validation of output data
	Critically Evaluated (CE) Data	Infrastructure to assure the greatest data integrity, e.g., NIST SRD, KRISS SRD
Process/Analyze	Preparation and Pre-Processing Methods	Data cleaning
		De-identification, anonymization
		Amputation and imputation
		Aggregation
		Validation and verification
		Normalization of metadata
	Software	Testing and validation tools
		Documentation
Share/Use/Reuse	Publishing	Integrity of data
		Quality measures and assessment vis-à-vis fit for purpose
	Modes of Sharing	Standardized formats
Preserve/Discard	Criteria for Preservation	Use
		Impact
		Value
		Uniqueness

4.6. Data Standards

Both disciplinary (e.g., Darwin Core[250] or NeXus[251]) and broad data standards (e.g., PREMIS[252] or schema.org[253]) are implemented by researchers to make their datasets both more FAIR and of higher quality. Researchers may use formal (e.g., ISO[254] or ANSI [255] standards) or de facto (e.g., DataCite [204]) standards for their research community. Use of data standards ensures consistency within a discipline and can reduce cost by decreasing the likelihood that data will have to be created again. Data standards are called out in every lifecycle stage except Envision.

Table 12 lists the topics and subtopics that are most relevant to the overarching theme of Data Standards.

Table 12. Data Standards (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance – Strategic/Qualitative	Stewardship
	Data Culture and Reward Structure	Recognition of data management
		Integrity of research and data
		FAIR data principles
		Maintenance of FAIR data
	Education and Workforce Development	Workforce skills inventory
		Data management training
	Community Engagement	Engagement across knowledge domains and sectors
Plan	Data Management Planning	Written data management plans (DMPs)
		Specification of data objects, metadata, analysis tools, and workflows throughout the lifecycle
		Machine-readable DMPs
		Data organization, e.g., database, repository, to facilitate future access
		Data management expertise and training
	Data Objects	Measurements, including images, audio recordings, and photos/videos
		Observation
		Survey
		Software
		Specimens (physical sample)
	FAIR	Identification of methods/guidelines vis-à-vis FAIR principles
	Data/Metadata Considerations	Criteria for selection of data/metadata
		Nature of data/metadata required
		Methods to capture and store data/metadata
		Metadata schema
	Data Architecture	Model
		LIMS
		Interoperability among different architectures
		Existing standards
	Hardware and Software Infrastructure	Interoperability
		Persistent instrument identifiers
	Research Data Standards	Criteria, i.e., general vs. domain-specific standards
		Sources of standards/guidelines for data/metadata
		Quality standards

Lifecycle Stage	Topic	Subtopic
		Community-based standards/conventions
Generate/Acquire	Data Types	Measurements, including images, audio recordings, and photos/videos
		Text files
		Computations, simulations
		Source code
		Observation
		Survey
		Transaction
		Social media
	Acquired Data	Provenance
	Critically Evaluated (CE) Data	Infrastructure to assure the greatest data integrity, e.g., NIST SRD, KRISS SRD
	FAIR Principles	Data born FAIR
		Data made FAIR
		FAIR digital objects
		Guidelines/methodologies for each aspect: F, A, I, R
		Tools to capture FAIR provenance
		FAIR instruments and tools
	Community-Based Standards	General vs. domain-specific
		Standards development organizations vs. community consensus
		Data format and file structure
		Metadata format and file structure
		Interoperability
Process/Analyze	Metadata	Types of metadata
		Specification of metadata standards
		Linked data structure
		Persistent identifiers (e.g., DOI, ORCID, ARK, ROR, PIDINST, Handles)
	Provenance	Original authoritative copy
		Version identification
		CrediT taxonomy
	Software	Standards, protocols, and interfaces
Share/Use/Reuse	Publishing	Persistent identifier
		Metadata
		Integrity of data
		Curation
	Attribution	Citation metrics
		Dataset citation
		Provenance
		Author identity management
		Use of persistent identifiers
		Versioning
	Modes of Sharing	Standardized formats

Lifecycle Stage	Topic	Subtopic
	Legal and Licenses	Standardized, machine-actionable license documents
Preserve/Discard	Criteria for Preservation	Provenance
	Storage and Preservation	Media to store and preserve data
		File integrity
	Moving Data from One Service to Another across Organizations	Registry maintenance and curation
	Retention and Disposition Schedules	End-of-life special considerations

4.7. Diversity, Equity, and Inclusion

Diversity, Equity, and Inclusion (DEI) is a broad theme covering important social and cultural aspects of the research enterprise. Efforts in DEI center on growing the sense of belonging for everyone in every laboratory, research group, department, or institution. Research data practices are not immune to biases and historical disadvantages must often be addressed through intentional action. DEI is important not just for members of underrepresented and marginalized groups, but for the integrity of the research process as a whole. More inclusive research tends to be more rigorous as it introduces different perspectives that enable more complete and broader interpretations of research data. Given the typical challenges associated with cultural changes within an institution, DEI efforts must be embedded throughout the research data management lifecycle to maximize their effectiveness.

Table 13 lists the topics and subtopics that are most relevant to the overarching theme of Diversity, Equity, and Inclusion.

Table 13. Diversity, Equity, and Inclusion (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance – Strategic/Qualitative	Vision and/or policy
		Organizational values, including DEI
	Data Governance—Legal and Regulatory Compliance	Ethics
		Social license for use and reuse
	Data Culture and Reward Structure	Roles and responsibilities
		Recognition of data management
		Value of data workers

Lifecycle Stage	Topic	Subtopic
		CARE and ethics
	Education and Workforce Development	Promotional paths and career development
	Community Engagement	Stakeholder communities
		Partners/partnerships
		Engagement across knowledge domains and sectors
		Inclusivity in interactions
		Data services and the beneficiaries
Plan	Financial Aspects of Planning	Staffing and training
	Data Management Planning	Purpose/intent of research study and context of anticipated data use
	Data/Metadata Considerations	Nature of data/metadata required
		Methods to capture and store data/metadata
	Hardware and Software Infrastructure	Staff expertise and support staff
	Research Data Standards	Community-based standards/conventions
	Assessment	Goals/definition of success
		Metrics for tracking use and impact measures, including reuse
	Communication and Outreach	Methods to share and reuse data/metadata
		Allocation of credit to project team members
		Promotion of data to communities of interest
		Cross-institution cooperation
		Requests for additional data from community
	Access Control Associated with Data Sensitivity	Identification of responsible parties for access management
		Sensitive data/PII
Generate/Acquire	Data Sources	In-house generation by researchers
		Remote generation by researchers
		In-field generation by researchers
		User facility generation by/for researcher
		Historical
		Human-annotated
	Qualitative Data	Methods and protocols
		Data/metadata/paradata capture methods
	Acquired Data	From collaborators
		From the literature

Lifecycle Stage	Topic	Subtopic
	Community-Based Standards	General vs. domain-specific
		Standards development organizations vs. community consensus
Process/Analyze	Preparation and Pre-Processing Methods	De-identification, anonymization
	Modeling	ML, AI
	Metadata	Responsible parties
	Provenance	CrediT taxonomy
Share/Use/Reuse	Publishing	Curation
		Incentives for data publishing
		Mitigation of disincentives for data publishing
	Attribution	Author identity management
	Access	External Access
		Mitigation of barriers and economic constraints
	Legal and Licenses	Ownership
		Encouragement and support for sharing, use, and reuse
		Indigenous data rights
	Levels of Protection	Unclassified but sensitive information, e.g., de-identification, enclaves
		Protection of limited data/secure platforms/enclaves
		Constraints and restrictions on data use and sharing
	Architectures for Application, Use, and Reuse	Extensibility across communities, including machine-based interactions
Preserve/Discard	Criteria for Preservation	Use
		Impact
		Value
		Uniqueness
	Retention and Disposition Schedules	Deaccessioning/end-of-life
		End-of-life special considerations

4.8. Ethics, Trust, and the CARE Principles

Ethical generation, analysis, use, reuse, sharing, disposal, and preservation of data is a pillar of responsible research and is called out throughout the framework. The phrase “as open as possible, as closed as necessary” [256] comes to mind when working through the ethical implications of sharing data. While ethical choices are often made at the Share/Use/Reuse lifecycle stage, questions and concerns regarding the generation or collection of data that are likely to be examined by an Institutional or Ethics Review Board must be considered in the Plan stage. In the Preserve/Discard stage, it is essential to comply with preservation and disposition

standards. While the subtopics in the framework are a starting point for understanding how ethics touches every aspect of the research data lifecycle, it is also important that a project be securely grounded in the practices of a given discipline; for example, the standards for historical research will differ from those for economic or healthcare research.

Trust is a factor across the Framework and is the basis for relationships between data producers and users, the funding agencies that support projects, and the institutions that host research. Specific populations will also have various ethical considerations, for example, the CARE Principles for Indigenous Data Governance [Rus2020] are quickly becoming the standard for working with indigenous data worldwide.

Table 14 lists the topics and subtopics that are most relevant to the overarching theme of Ethics, Trust, and the CARE Principles.

Table 14. Ethics, Trust, and the CARE Principles (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance – Strategic/Qualitative	Data management value proposition
		Stewardship
	Data Governance—Legal and Regulatory Compliance	Ethics
		Sharing/licensing
	Data Culture and Reward Structure	Roles and responsibilities
		Recognition of data management
		Value of data workers
		Promotion and tenure
		Integrity of research and data
		Incentives and impact for sharing and reuse
		Disincentives for sharing and reuse
		CARE and ethics
	Resources—Allocation and Sustainability	Sources of funding
		Long-term funding
		Staffing
	Community Engagement	Stakeholder communities
		Partners/partnerships
		Engagement across knowledge domains and sectors
		Inclusivity in interactions
Plan	Chain of Custody	Roles and responsibilities
		Implementation authority

Lifecycle Stage	Topic	Subtopic
	Data Management Planning	Written data management plans (DMPs)
		Purpose/intent of research study and context of anticipated data use
		Specification of data objects, metadata, analysis tools, and workflows throughout the lifecycle
		Data organization, e.g., database, repository, to facilitate future access
		Data management expertise and training
	Data Objects	Quantitative and qualitative
	Data/Metadata Considerations	Methods to capture and store data/metadata
	Data Architecture	Design
		Workflow
		Model
		Security
	Hardware and Software Infrastructure	Security and privacy considerations
	Research Data Standards	Criteria, i.e., general vs. domain-specific standards
		Quality standards
		Community-based standards/conventions
	Communication and Outreach	Allocation of credit to project team members
		Promotion of data to communities of interest
		Cross-institution cooperation
		Requests for additional data from community
	Access Control Associated with Data Sensitivity	Identification of responsible parties for access management
		Sensitive data/PII
		Limited disclosure, e.g., IP
		Licensing for reuse
Generate/Acquire	Data Types	Observation
		Survey
		Transaction
		Social media
	Data Sources	In-house generation by researchers
		Remote generation by researchers
		In-field generation by researchers
		User facility generation by/for researcher
		Historical
		Human-annotated
	Generated Experimental Data	Source of object/subjects

Lifecycle Stage	Topic	Subtopic
		Characteristics of object/subjects
		Conditions of research study
		Specification of instruments and tools
		Parameters for instruments and tools
		Methods, protocols, and calibration
		Data/metadata capture methods
		Reproducibility
	Generated Computational Data	Input data/metadata
		Output data/metadata
		Data/metadata capture methods
	Qualitative Data	Nature of object/subjects
		Methods and protocols
		Metadata, e.g., questions in a survey, location of survey
		Paradata, e.g., mode of survey (in person, via phone), responder's response time
		Data/metadata/paradata capture methods
	Acquired Data	From collaborators
		From repositories
		From the literature
		Aggregation of data evaluated by experts
		Restrictions, fees, and usage agreements
	Critically Evaluated (CE) Data	Infrastructure to assure the greatest data integrity, e.g., NIST SRD, KRISS SRD
		Single researcher dataset
		Aggregation of data evaluated by experts
		Reproducibility and uncertainty quantification
		Intellectual property rights
	Community-Based Standards	General vs. domain-specific
		Standards development organizations vs. community consensus
		Metadata format and file structure
		Interoperability
Process/Analyze	Preparation and Pre-Processing Methods	Data cleaning
		De-identification, anonymization
		Curation
		Normalization of metadata
	Modeling	Visualization
		ML, AI
	Metadata	Responsible parties
		Persistent identifiers (e.g., DOI, ORCID, ARK, ROR, PIDINST, Handles)
	Provenance	Original authoritative copy
		Version identification

Lifecycle Stage	Topic	Subtopic
		Derivative product
		Aggregation
		Subset
		Timestamp
		CrediT taxonomy
	Workflows and Middleware	Decisions regarding the need for additional data
		Distributed workflow across sites
Share/Use/Reuse	Publishing	Repository, i.e., domain, generalist, institutional
		Data papers
		Metadata
		Integrity of data
		Peer review of datasets and metadata
		Curation
		Incentives for data publishing
		Mitigation of disincentives for data publishing
	Modes of Dissemination	Traditional journal article
		Supplementary material
		On request
		Data landing pages
		Workflows
		Mainstream media, e.g., newspapers, radio
		Social media, e.g., Twitter, Instagram
	Attribution	Provenance
		Author identity management
	Access	Internal access
		External access
		Programmatic access, i.e., API
		Virtual and physical enclaves
		Access vs. visiting
		Availability statement
		Mitigation of barriers and economic constraints
	Legal and Licenses	Ownership
		Encouragement and support for sharing, use, and reuse
		Indigenous data rights
		Intellectual property rights/restrictions
		Usage agreements/terms/licenses and required permissions
		Sharing agreements and licensing
		Service-level agreements
		Terms of service

Lifecycle Stage	Topic	Subtopic
		Standardized, machine-actionable license documents
		Citation requirements
	Levels of Protection	Unclassified but sensitive information, e.g., de-identification, enclaves
		Protection of limited data/secure platforms/enclaves
		Constraints and restrictions on data use and sharing
		Anonymization
	Architectures for Application, Use, and Reuse	Capture of insights from ML and use of these to improve datasets for future AI applications
Preserve/Discard	Criteria for Preservation	Use
		Impact
		Value
		Uniqueness
		Cost
		Provenance
		Legal and regulatory
	Moving Data from One Service to Another Across Organizations	Roles and responsibilities
		Registry maintenance and curation
		Disciplinary archives
	Retention and Disposition Schedules	Administrative/policy decisions
		Deaccessioning/end-of-life
		End-of-life special considerations

4.9. Legal Considerations

As much as technical realities structure the ways in which data can be gathered, created, published, and preserved, legal considerations constrain and channel the research data lifecycle. Laws form the background rules governing how data can be shared, and institutions that share data often use contracts and agreements that rely upon the legal system to order and enforce the terms. Laws sometimes restrict access, especially for categories of sensitive data such as personally identifiable information, certain types of healthcare information, and business identifiable information but laws can also enable data sharing by providing clear guidelines or directives to provide open data when it is in the public interest. Though legal considerations appear in nearly every one of the six lifecycle stages, meticulous planning and preparation make any constraints less onerous.

Table 15 lists the topics and subtopics that are most relevant to the overarching theme of Legal Considerations.

Table 15. Legal Considerations (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance—Legal and Regulatory Compliance	Privacy
		Safety and security assurance
		Risk assessment
		Risk mitigation and management
		Sharing/licensing
		Jurisdiction for sharing and reuse
	Data Culture and Reward Structure	Disincentives for sharing and reuse
	Education and Workforce Development	HR's supporting role in workforce development and training
Plan	Chain of Custody	Roles and responsibilities
	Hardware and Software Infrastructure	Security and privacy considerations
	Access Control Associated with Data Sensitivity	Identification of responsible parties for access management
		Ease of maintenance and implementation of records
		Regulatory compliance
		Sensitive data/PII
		Limited disclosure, e.g., IP
		Licensing for reuse
Generate/Acquire	Acquired Data	Restrictions, fees, and usage agreements
	Critically Evaluated (CE) Data	Intellectual property rights
Process/Analyze	Software	Open-source vs. proprietary
Share/Use/Reuse	Legal and Licenses	Ownership
		Encouragement and support for sharing, use, and reuse
		Indigenous data rights
		Intellectual property rights/restrictions
		Usage agreements/terms/licenses and required permissions
		Sharing agreements and licensing
		Service-level agreements
		Terms of service
		Standardized, machine-actionable license documents
		Citation requirements
	Levels of Protection	Unclassified but sensitive information, e.g., de-identification, enclaves
		Security classification

Lifecycle Stage	Topic	Subtopic
		Protection of limited data/secure platforms/enclaves
		Constraints and restrictions on data use and sharing
		Anonymization
Preserve/Discard	Criteria for Preservation	Legal and regulatory
	Retention and Disposition Schedules	Administrative/policy decisions
		Deaccessioning/end-of-life
		Legal documents

4.10. Metadata and Provenance

Metadata is the information about a dataset that defines, describes, and links the dataset to other datasets and provides contextualization of the dataset [85]. Metadata are essential to the effective use, reuse, and preservation of research data over time. In the Envision and Plan stages, metadata support legal and regulatory compliance, and are a consideration in planning data outputs and resources.

The table below shows each topic/subtopic that mentions or covers metadata. While the final lifecycle stage (Preserve/Discard) does not explicitly relate to metadata, the existence of descriptive and other metadata is imperative to this stage. The robustness of metadata for a file or dataset determines the level of curation needed for preservation and use: richer metadata allows for better findability, interoperability, and reuse in support of the FAIR data principles, while less robust metadata make all these activities more difficult and time intensive. Poor-quality metadata can render an otherwise important dataset unusable when the creator of the dataset is no longer available.

Included in the Metadata theme is provenance, the historical information concerning the data [37]. Understanding the provenance of a given dataset, including metadata on the experimental conditions used to generate the data, is essential for many disciplines. Without proper provenance documentation, it is difficult to assess the quality and reliability of the data and to publish them with correct metadata. Provenance can be used as a criterion for preservation.

Table 16 lists the topics and subtopics that are most relevant to the overarching theme of Metadata and Provenance.

Table 16. Metadata and Provenance (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance – Strategic/Qualitative	End-use support
		Stewardship

Lifecycle Stage	Topic	Subtopic
	Data Governance – Legal and Regulatory Compliance	Inventory
		Sharing/licensing
	Data Culture and Reward Structure	FAIR data principles
		Maintenance of FAIR data
	Education and Workforce Development	Data management training
Plan	Chain of Custody	Roles and responsibilities
		Implementation authority
		Centralized inventory of services, groups, and resources
		Provenance
	Data Management Planning	Specification of data objects, metadata, analysis tools, and workflows throughout the lifecycle
		Machine-readable DMPs
	FAIR	Identification of methods/guidelines vis-à-vis FAIR principles
	Data/Metadata Considerations	Criteria for selection of data/metadata
		Nature of data/metadata required
		Methods to capture and store data/metadata
		Metadata schema
	Data Architecture	Models
		LIMS
	Hardware and Software Infrastructure	Persistent instrument identifiers
	Research Data Standards	Criteria, i.e., general vs. domain-specific standards
		Sources of standards/guidelines for data/metadata
		Community-based standards/conventions
	Communication and Outreach	Methods to share and reuse data/metadata
		Allocation of credit to project team members
	Access Control Associated with Data Sensitivity	Access Control Associated with Data Sensitivity
Generate/Acquire	Generated Experimental Data	Data/metadata capture methods
		Provenance and capture methods
		Reproducibility
	Generated Computational Data	Versioning
		Data/metadata capture methods

Lifecycle Stage	Topic	Subtopic
		Provenance and capture methods
	Qualitative Data	Metadata, e.g., questions in a survey, location of survey
		Paradata, e.g., mode of survey (in person, via phone), responder's response time
		Data/metadata/paradata capture methods
	Acquired Data	Provenance
		Restrictions, fees, and usage agreements
	Critically Evaluated (CE) Data	Reproducibility and uncertainty quantification
		Intellectual property rights
	FAIR Principles	Data born FAIR
		Data made FAIR
		FAIR digital objects
		Tools to capture FAIR provenance
		FAIR instruments and tools
	Community-Based Standards	Metadata format and file structure
		Vocabulary and ontology
Process/Analyze	Preparation and Pre-Processing Methods	Curation
		Normalization of metadata
	Metadata	Types of metadata
		Responsible parties
		Specification of metadata standards
		Linked data structure
		Persistent identifiers (e.g., DOI, ORCID, ARK, ROR, PIDINST, Handles)
	Provenance	Original authoritative copy
		Version identification
		Derivative products
		Aggregation
		Subsets
		Timestamping
		CrediT taxonomy
Share/Use/Reuse	Publishing	Repository, i.e., domain, generalist, institutional
		Data linking
		Persistent identifier
		Metadata
		Peer review of datasets and metadata
		Curation
		Publisher agreements and policies
	Modes of Dissemination	Modes of Dissemination
	Attribution	Provenance
		Author identity management
		Use of persistent identifiers

Lifecycle Stage	Topic	Subtopic
		Versioning
	Modes of Sharing	Catalog
		Registries of repositories
	Legal and Licenses	Usage agreements/terms/licenses and required permissions
		Sharing agreements and licensing
Preserve/Discard	Criteria for Preservation	Criteria for Preservation
	Retention and Disposition Schedules	Deaccessioning/end-of-life
		Recognition of removed data, i.e., tombstone page

4.11. Reproducibility and the FAIR Data Principles

Touching many of the lifecycle stages are Reproducibility and the FAIR Data Principles. Reproducible research yields data that can be replicated by the author or other researchers using only information provided in the original work [79]. Standards for reproducibility differ by research discipline, but generally the metadata and other contextual information needed for reproducibility are like what is described by the FAIR data principles [29]. These community-based principles have come to define, for many disciplines, what a published dataset should aspire to be. By keeping the principles of findability, accessibility, interoperability, and reusability in mind while planning a project, or as data are collected, the data will be ready for broader reuse when they are publicly released.

Table 17 lists the topics and subtopics that are most relevant to the overarching theme of Reproducibility and the FAIR Data Principles.

Table 17. Reproducibility and the FAIR Data Principles (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance – Strategic/Qualitative	Stewardship
	Data Governance—Legal and Regulatory Compliance	Sharing/licensing
		Social license for use and reuse
	Data Culture and Reward Structure	FAIR data principles
		Maintenance of FAIR data
	Community Engagement	Engagement across knowledge domains and sectors
Plan	Data Management Planning	Data organization, e.g., database, repository, to facilitate future access
	FAIR	Organizational support for making data more FAIR

Lifecycle Stage	Topic	Subtopic
		Identification of methods/guidelines vis-à-vis FAIR principles
	Data/Metadata Considerations	Intended extent of FAIRness
		Metadata schema
	Hardware and Software Infrastructure	Interoperability
		Persistent instrument identifiers
	Research Data Standards	Criteria, i.e., general vs. domain-specific standards
		Community-based standards/conventions
	Assessment	Metrics for tracking use and impact measures, including reuse
	Communication and Outreach	Methods to share and reuse data/metadata
	Access Control Associated with Data Sensitivity	Identification of responsible parties for access management
		Ease of maintenance and implementation of records
		Limited disclosure, e.g., IP
		Licensing for reuse
Generate/Acquire	FAIR Principles	Data born FAIR
		Data made FAIR
		FAIR digital objects
		FAIR on a continuous scale
		Guidelines/methodologies for each aspect: F, A, I, R
		Tools to capture FAIR provenance
		FAIR instruments and tools
		Not FAIR data, e.g., legacy
	Community-Based Standards	Metadata format and file structure
		Interoperability
Process/Analyze	Metadata	Types of metadata
		Specification of metadata standards
		Persistent identifiers (e.g., DOI, ORCID, ARK, ROR, PIDINST, Handles)
Share/Use/Reuse	Publishing	Repository, i.e., domain, generalist, institutional
		Data linking
		Persistent identifier
		Metadata
	Modes of Sharing	Standardized formats
		Interoperability tools
		Discovery platform
		Registries of repositories
	Access	Internal access

Lifecycle Stage	Topic	Subtopic
		External access
		Programmatic access, i.e., API
	Legal and Licenses	Intellectual property rights/restrictions
		Usage agreements/terms/licenses and required permissions
		Sharing agreements and licensing
		Standardized, machine-actionable license documents

4.12. Security and Privacy

Digital data are designed to be easily shared, copied, and transformed, but their mobility can make privacy and security difficult to ensure. Security and privacy issues are fundamentally about trust, both in the institutions and systems that facilitate collection, storage, and transfer of data, as well as the individuals within those institutions. Proper protocols, rationally based on the need to protect vulnerable populations or sensitive information or stemming from common understandings of security needs, promote trust, which can enable greater data mobility. In the European Union, organizations that collect, store, or hold personal data must comply with the General Data Protection Regulation [257] The U.S. does not have such a universal regulation, though various federal laws govern different sectors and types of data, and some states have their own additional regulations. Security and privacy issues arise in the Envision and Plan lifecycle stages, with the results folded into the day-to-day procedures for handling and accessing data and appear again in the Share/Use/Reuse lifecycle stage.

Table 18 lists the topics and subtopics that are most relevant to the overarching theme of Security and Privacy.

Table 18. Security and Privacy (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Governance—Legal and Regulatory Compliance	Privacy
		Safety and security assurance
		Data management organization
		Organizational values, including DEI
	Education and Workforce Development	Workforce skills inventory
Plan	Data Architecture	Hosting and storage, cloud storage
		Security
	Hardware and Software Infrastructure	Security and privacy considerations

Lifecycle Stage	Topic	Subtopic
	Access Control Associated with Data Sensitivity	Identification of responsible parties for access management
		Ease of maintenance and implementation of records
		Regulatory compliance
		Sensitive data/PII
		Limited disclosure, e.g., IP
		Licensing for reuse
Process/Analyze	Software	Security and software updates
Share/Use/Reuse	Access	Internal access
		External access
		Programmatic access, i.e., API
		Virtual and physical enclaves
		Access vs. visiting
		Availability statement
		Mitigation of barriers and economic constraints
	Legal and Licenses	Indigenous data rights
		Intellectual property rights/restrictions
	Levels of Protection	Unclassified but sensitive information, e.g., de-identification, enclaves
		Security classification
		Protection of limited data/secure platforms/enclaves
		Constraints and restrictions on data use and sharing
		Anonymization

4.13. Software Tools

Regarding research data, software tools are programs or utilities for developing applications and analyzing/processing or searching for data. Additionally, software tools are used to generate data from computational and experimental methods, throughout the publication process. An exhaustive list of tools would be ever-changing; more important than a list of tools used in every discipline is the understanding that the tools used during all lifecycle stages can influence later stages.

Table 19 lists the topics and subtopics that are most relevant to the overarching theme of Software Tools.

Table 19. Software Tools (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Education and Workforce Development	Workforce preparedness in new and advanced technologies
Plan	Financial Aspects of Planning	Staffing and training
	Data Management Planning	Machine-readable DMPs
	Data Objects	Software
	Data Architecture	LIMS
		Hosting and storage, cloud storage
	Hardware and Software Infrastructure	Organizational research needs
		Tools to support data-related processes
		Models that connect infrastructure to data processes and workflow
		Interoperability
		Persistent instrument identifiers
		Sustainability of data vis-à-vis obsolete infrastructure
Generate/Acquire	Data Types	Computations, simulations
		Source code
	Generated Experimental Data	Specification of instruments and tools
		Parameters for instruments and tools
		Methods, protocols, and calibration
		Data/metadata capture methods
	Generated Computational Data	Parameters and conditions for computation, e.g., OS dependencies, compilers, memory requirements
	Acquisition Software	Open source vs. proprietary
		LIMS
		Instrument control
		Electronic laboratory notebooks
Process/Analyze	Modeling	Visualization
		Integrated development environments, e.g., Jupyter, Rstudio
	Software	Commercial vs. custom
		Open-source vs. proprietary
		Aggregation tools
		Surveying tools
		Statistical tools
		Calculation and analysis tools
		APIs
		Database management tools
		Testing and validation tools
		Versioning and maintenance
		Source code repository
		Security and software updates

Lifecycle Stage	Topic	Subtopic
		Standards, protocols, and interfaces
	Workflows and Middleware	LIMS
		Laboratory notebooks, e.g., electronic, paper
		Tools for automated metadata capture
		Anomaly detection and correction tools
		Collaboration tools
		Process monitoring and evaluation
		Containerization
		Reusable workflow components
		Microservices
Share/Use/Reuse	Publishing	Software
		Updates to datasets and new software versions
	Legal and Licenses	Usage agreements/terms/licenses and required permissions

4.14. Training, Education, and Workforce Development

Training and education are critical for ensuring that any given organization or individual involved in the research data management process has the necessary skills for research data management. Investment into workforce development is especially important in an area where best practices are still developing. On-the-job training not only helps to promote the standardization that is important in research data management but can also promote equity.

Table 20 lists the topics and subtopics that are most relevant to the overarching theme of Training, Education, and Workforce Development.

Table 20. Training, Education, and Workforce Development (Overarching Theme)

Lifecycle Stage	Topic	Subtopic
Envision	Data Culture and Reward Structure	Value of data workers
		Promotion and tenure
	Education and Workforce Development	Workforce skills inventory
		Workforce preparedness in new and advanced technologies
		Data management training
		HR's supporting role in workforce development and training
		Promotional paths and career development

Lifecycle Stage	Topic	Subtopic
	Resources—Allocation and Sustainability	Staffing
	Community Engagement	Engagement across knowledge domains and sectors
Plan	Financial Aspects of Planning	Staffing and training
	Data Management Planning	Data management expertise and training
	FAIR	Identification of methods/guidelines vis-à-vis FAIR principles
	Hardware and Software Infrastructure	Staff expertise and support staff
Generate/Acquire	Community-Based Standards	General vs. domain-specific
		Standards development organizations vs. community consensus
		Data format and file structure
		Metadata format and file structure
		Vocabulary and ontology
		Interoperability

5. Profiles

Profiles specify those topics and subtopics in the framework that are most relevant for a particular job role in an organization. The framework presents a comprehensive list of the tasks and issues that may arise with respect to research data activities and research data management. However, most organizations or individuals need not be concerned with every subtopic. As described below, NIST has developed a tool that allows individuals and organizations to customize a profile (i.e., select relevant subtopics from the full list) for their specific needs or responsibilities.

To assist with this customization process, the following generic profiles are intended to serve as samples and guides. Users may either modify a generic profile as a starting point for their own profile or build one by selecting relevant subtopics. Profiles may also be used to conduct self-assessments of research data management that indicate areas needing attention and to communicate the results of such self-assessments within an organization or between organizations. Please see the accompanying Appendix C (separate file) which consists of an Excel file with this information as well as a blank template to allow for ease of customization.

Table 21. Generic Profiles

Envision: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	Data/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers
Data Governance – Strategic/Qualitative	Identification of goals and roles	X	X	X	X	X		X	X
	Vision and/or policy	X	X	X	X			X	
	Data management organization		X	X		X		X	
	Organizational values, include DEI	X		X	X			X	X
	Data management value proposition		X	X	X	X		X	
	Data needs assessment	X	X	X				X	X
	Purpose and value of data	X	X	X	X	X		X	
	Organization intent regarding FAIR data			X	X	X		X	
	End-use support		X	X				X	
	Stewardship		X	X	X			X	
Data Governance – Legal and Regulatory Compliance	Privacy								X
	Ethics	X	X	X	X	X		X	X
	Safety and security assurance		X	X		X		X	
	Inventory			X		X			
	Risk assessment		X	X	X				
	Risk mitigation and management		X	X	X			X	
	Sharing/licensing		X	X	X	X		X	
	Social license for use and reuse		X		X	X		X	
	Jurisdiction for sharing and reuse				X			X	

Envision: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	Data/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers
Data Culture and Reward Structure	Roles and responsibilities		X	X	X			X	
	Recognition of data management		X	X	X	X		X	
	Value of data workers		X		X			X	
	Promotion and tenure		X					X	
	Integrity of research and data				X			X	X
	FAIR data principles			X	X	X		X	
	Maintenance of FAIR data		X	X	X	X		X	
	Incentives and impact for sharing and reuse		X	X	X	X		X	
	Disincentives for sharing and reuse			X				X	
	CARE and ethics	X		X	X			X	
Education and Workforce Development	Workforce skills inventory			X	X				
	Workforce preparedness in new and advanced technologies		X	X		X		X	
	Data management training		X	X	X	X			
	HR's supporting role in workforce development and training								
	Promotional paths and career development		X					X	
Resources—Allocation and Sustainability	Sources of funding		X					X	
	Long-term funding		X					X	
	Staffing		X	X				X	
Community Engagement	Stakeholder communities		X	X	X	X		X	

	Modes of communication		X	X	X				
	Partners/partnerships		X	X	X	X			
	Engagement across knowledge domains and sectors			X	X	X		X	
	Inclusivity in interactions		X	X	X				
	Data services and the beneficiaries			X	X			X	

Plan: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	Data/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers
Chain of Custody	Roles and responsibilities		X	X	X			X	
	Implementation authority			X				X	
	Centralized inventory of services, groups, and resources			X			X	X	
	Provenance			X				X	
Financial Aspects of Planning	Funding models for provisioning resources		X					X	
	Funding sources		X		X			X	
	Decision-making tools to assess costs		X						
	Cost-benefit analysis		X	X				X	
	Cost breakdown by lifecycle stage		X					X	
	Downstream lifecycle costs		X		X			X	
	Staffing and training		X	X				X	
Data Management Planning	Written data management plans (DMPs)		X	X	X				X
	Purpose/intent of research study and context of anticipated data use			X	X				

Plan: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	Data/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers
Data Management Planning	Specification of data objects, metadata,			X					

	analysis tools, and workflows throughout the lifecycle								
	Machine-readable DMPs			X					
	Linkage of DMPs to administrative records			X				X	
	Data organization, e.g., database, repository, to facilitate future access		X	X	X	X		X	
	Data management expertise and training		X	X	X				X
Data Object	Quantitative and qualitative			X					
	Measurement, including images, audio recordings, and photos/videos			X					
	Observation			X					
	Survey			X					
	Software			X				X	
	Model	X		X				X	
	Documentation (text)			X	X				
	Specimen (physical sample)								
	Presentation			X	X				
FAIR	Organizational support for making data more FAIR		X	X	X			X	
	Identification of methods/guidelines vis-à-vis FAIR principles			X	X			X	

Plan: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	DATA/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers
Data/Metadata Considerations	Criteria for selection of data/metadata			X	X				X
	Nature of data/metadata required			X					X

	Intended extent of FAIRness			X					X
	Methods to capture and store data/metadata			X					X
	Metadata schema			X					
Data Architecture	Design		X	X	X				
	Processing operations			X					
	Workflow			X					
	Model								
	LIMS								
	Hosting and storage, cloud storage		X	X					
	Configuration management			X					
	Interoperability among different architectures			X	X				
	Security		X	X	X				
	Existing standards			X	X				
Hardware and Software	Organizational research needs		X	X				X	
	Tools to support data-related processes		X	X					
	Models that connect infrastructure to data processes and workflow			X					
	Interoperability			X					
	Persistent instrument identifiers			X					

Plan: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	Data/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers
Hardware and Software	Sustainability of data vis-à-vis obsolete infrastructure		X	X					
	Security and privacy considerations		X	X					
	Staff expertise and support staff		X	X					

Research Data Standards	Criteria, i.e., general vs. domain-specific standards			X				X	
	Sources of standards/guidelines for data/metadata			X	X				
	Quality standards			X	X				
	Community-based standards/conventions			X	X			X	
Assessment	Goals/definition of success	X		X	X				
	Metrics for tracking use and impact measures, including reuse			X	X			X	
Communication and Outreach	Methods to share and reuse data/metadata			X	X			X	X
	Allocation of credit to project team members			X					X
	Promotion of data to communities of interest			X	X				X
	Cross-institution cooperation			X	X	X		X	
	Requests for additional data from community				X			X	
Access Control Associated with Data Sensitivity	Identification of responsible parties for access management			X				X	
	Ease of maintenance and implementation of records							X	
	Regulatory compliance	X		X	X			X	
	Sensitive data/PII			X	X			X	
	Limited disclosure, e.g., IP			X	X			X	
	Licensing for reuse				X			X	

Generate/Acquire: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	Data/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers

Data Types	Measurement, including images, audio recordings, and photos/videos	X				X		X	X
	Text file	X				X		X	X
	Computation, simulation	X				X		X	X
	Source code	X						X	X
	Observation					X		X	X
	Survey							X	
	Transaction							X	
	Social media							X	
Data Sources	In-house generation by researchers	X				X		X	X
	Remote generation by researchers	X						X	
	In-field generation by researchers							X	
	User facility generation by/for researcher	X						X	X
	Historical							X	
	Human-annotated							X	X
Generated Experimental Data	Source of objects/subjects								X
	Characteristics of objects/subjects								X
	Conditions of research study	X							X
	Specification of instruments and tools	X				X			X
Generate/Acquire: Topic	**Subtopic**	**AI Experts**	**Budget/Cost Experts**	**Curators**	**Data/IT Leader**	**Providers of Data Tools**	**Publishers**	**Research Organization Leaders**	**Researchers**
Generated Experimental Data	Parameters for instruments and tools	X				X			X
	Methods, protocols, and calibration	X							X
	Data/metadata capture methods	X							X

	Provenance and capture methods	X							X
	Reproducibility	X			X			X	X
Generated Computational Data	Input data/metadata	X				X		X	X
	Output data/metadata	X						X	X
	Hardware								
	Parameters and conditions for computation, e.g., OS dependencies, compilers, memory requirements					X			X
	Versioning	X				X			X
	Data/metadata capture methods	X				X			X
	Provenance and capture methods	X							X
	Verification/validation of output data	X				X			X
Qualitative Data	Nature of objects/subjects								
	Methods and protocols								
	Metadata, e.g., questions in a survey, location of survey								
	Paradata, e.g., mode of survey (in person, via phone), responder's response time								

Generate/Acquire: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	Data/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers
Qualitative Data	Data/metadata/paradata capture methods								
Acquired Data	From collaborators	X				X		X	X
	From repositories	X				X			X
	From the literature	X				X			X
	Aggregated datasets from multiple sources	X				X			X
	Provenance					X			X

	Restrictions, fees, and usage agreements					X		X	
Critically Evaluated (CE) Data	Infrastructure to assure the greatest data integrity, e.g., NIST SRD, KRISS SRD					X			
	Single researcher dataset	X							
	Aggregation of data evaluated by experts	X			X	X			
	Reproducibility and uncertainty quantification	X							
	Intellectual property rights								
FAIR Principles	Data born FAIR	X			X	X		X	
	Data made FAIR	X		X	X	X		X	
	FAIR digital objects			X		X		X	
	FAIR on a continuous scale			X					
	Guidelines/methodologies for each aspect: F, A, I, R			X	X	X		X	
	Tools to capture FAIR provenance			X		X		X	
	FAIR instruments and tools			X				X	
	Not FAIR data, e.g., legacy	X		X				X	
Generate/Acquire: Topic	**Subtopic**	**AI Experts**	**Budget/Cost Experts**	**Curators**	**Data/IT Leader**	**Providers of Data Tools**	**Publishers**	**Research Organization Leaders**	**Researchers**
Community-Based Standards	General vs. domain-specific	X		X	X	X			
	Standards development organizations vs. community consensus	X			X	X			
	Data format and file structure	X		X		X		X	
	Metadata format and file structure.	X		X		X		X	
	Vocabulary and ontology			X		X		X	

	Interoperability	X		X	X	X		X	
Acquisition Software	Open source vs. proprietary	X							X
	LIMS								X
	Instrument control								X
	Electronic laboratory notebooks	X							X
	Audio and video recordings								

Process/Analyze: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	Data/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers
Types of Processed Data	Tables, spreadsheets	X			X	X	X		X
	Charts, graphs	X			X	X	X		X
	Maps, vectors, images	X			X	X	X		X
	Instrument outputs	X					X		X
	Dynamic data						X		X
	Datasets from models and simulations	X			X	X	X		X
	Structured data, e.g., hierarchical organization	X			X		X		

Process/Analyze: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	Data/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers
Preparation and Pre-Processing Methods	Data cleaning	X		X	X		X		X
	De-identification, anonymization			X			X		
	Amputation and imputation	X		X			X		
	Aggregation	X		X	X		X		X
	Validation and verification	X		X	X		X		X
	Curation	X		X		X	X	X	

	Normalization of metadata	X		X	X	X	X	X	
Analysis	Manual	X			X		X		
	Exploratory	X			X	X	X		X
	Descriptive	X				X	X		
	Diagnostic	X			X		X		X
	Evaluative	X			X		X		X
	Predictive	X					X		X
	Prescriptive						X		
	Correlational	X					X		X
	Statistical	X			X		X		X
	Automated, autonomous	X				X	X		X
Modeling	Visualization	X			X		X	X	X
	ML, AI	X			X	X	X	X	X
	Iterative model fitting	X					X		X
	Integrated development environment, e.g., Jupyter, Rstudio	X			X	X	X		X

Process/Analyze: Topic	**Subtopic**	**AI Experts**	**Budget/Cost Experts**	**Curators**	**Data/IT Leader**	**Providers of Data Tools**	**Publishers**	**Research Organization Leaders**	**Researchers**
Metadata	Types of metadata	X		X	X	X	X		X
	Responsible parties			X	X		X		
	Specification of metadata standards	X		X		X	X		
	Linked data structure			X			X		
	Persistent identifiers (e.g., DOI, ORCID, ARK, ROR, PIDINST, Handles)	X		X	X	X	X		
Provenance	Original authoritative copy	X		X		X	X	X	X
	Version identification	X		X	X	X	X	X	
	Derivative product			X		X	X		

	Aggregation	X		X			X			
	Subset	X		X		X	X			
	Timestamp			X	X		X			
	CRediT taxonomy			X			X			
Software	Commercial vs. custom	X			X	X	X			X
	Open-source vs. proprietary	X			X	X	X	X		X
	Aggregation tools						X			
	Surveying tools						X			
	Statistical tools	X			X		X			X
	Calculation and analysis tools	X					X			X
	APIs	X			X	X	X			X
	Database management tools	X			X	X	X	X		
	Testing and validation tools	X					X			X
	Documentation	X			X	X	X			X
	Reproducibility and uncertainty quantification	X			X	X	X			

Process/Analyze: Topic	**Subtopic**	**AI Experts**	**Budget/Cost Experts**	**Curators**	**Data/IT Leader**	**Providers of Data Tools**	**Publishers**	**Research Organization Leaders**	**Researchers**
Software	Versioning and maintenance	X			X	X	X		
	Systems resilience and adaptability						X		
	Source code repository	X			X		X	X	X
	Security and software updates				X		X		
	Standards, protocols, and interfaces				X		X		
Workflow and Middleware	LIMS								
	Laboratory notebooks, e.g., electronic, paper						X		
	Tools for automated metadata capture	X			X	X	X		X

	Anomaly detection and correction tools	X			X	X			X
	Collaboration tools	X			X	X	X		
	Decisions regarding the need for additional data				X		X		
	Process monitoring and evaluation	X			X				
	Containerization				X				
	Reusable workflow component	X			X				
	Microservices								
	Distributed workflow across sites				X				X
	Comprehensive report generation								
Hardware	Compute requirements	X			X				X
	Storage requirements	X			X			X	X
	Network requirements							X	
	Accelerator requirements								

Share/Use/Reuse: Topic	**Subtopic**	**AI Experts**	**Budget/Cost Experts**	**Curators**	**Data/IT Leader**	**Providers of Data Tools**	**Publishers**	**Research Organization Leaders**	**Researchers**
Publishing	Repository, i.e., domain, generalist, institutional	X	X	X	X	X	X	X	
	Data papers	X		X		X	X	X	
	Software	X	X	X			X	X	X
	Updates to datasets and new software versions	X	X	X		X	X		X
	Data linking			X	X		X	X	
	Persistent identifier			X	X	X	X	X	
	Metadata			X	X	X	X	X	
	Integrity of data			X			X		
	Quality measures and assessment vis-à-vis fit for purpose			X	X		X		X
	Peer review of datasets and metadata	X		X			X		

	Reference data/digital objects in journal articles			X			X	X	
	Curation		X	X		X	X	X	
	Publisher agreements and policies			X	X	X	X	X	
	Incentives for data publishing			X		X	X	X	
	Mitigation of disincentives for data publishing	X		X			X	X	
Modes of Dissemination	Traditional journal article	X		X		X	X	X	X
	Supplementary material	X		X		X	X	X	X
	On request	X		X			X	X	
	Data landing pages	X		X			X		
	Workflows			X			X		
	Mainstream media, e.g., newspapers, radio						X		

Share/Use/Reuse: Topic	Subtopic	AI Experts	Budget/Cost Experts	Curators	Data/IT Leader	Providers of Data Tools	Publishers	Research Organization Leaders	Researchers
Modes of Dissemination	Social media, e.g., Twitter, Instagram						X		
Attribution	Citation metrics	X			X	X	X	X	X
	Citation impact	X		X	X	X	X	X	X
	Dataset citation	X		X	X	X	X	X	X
	Provenance			X		X	X		X
	Author identity management	X		X		X	X	X	
	Use of persistent identifiers	X		X	X	X	X	X	X
	Versioning			X		X	X		X
Modes of Sharing	Standardized formats	X		X	X	X	X	X	
	Interoperability tools	X		X	X	X	X		
	Discovery platform	X		X		X	X		X
	Catalog			X			X		
	Registries of repositories			X		X	X		

Access	Internal access	X	X	X		X	X		X
	External access	X	X	X		X	X		X
	Programmatic access, i.e., API	X	X	X		X	X		
	Virtual and physical enclaves						X		
	Access vs. visiting			X			X		
	Availability statement			X			X		X
	Mitigation of barriers and economic constraints		X	X			X	X	
Legal and Licenses	Ownership		X		X		X	X	X
	Encouragement and support for sharing, use, and reuse			X	X		X	X	X
	Indigenous data rights			X	X		X		X
	Intellectual property rights/restrictions		X	X		X	X	X	

Share/Use/Reuse: Topic	**Subtopic**	**AI Experts**	**Budget/Cost Experts**	**Curators**	**Data/IT Leader**	**Providers of Data Tools**	**Publishers**	**Research Organization Leaders**	**Researchers**
Legal and Licenses	Usage agreements/terms/licenses and required permissions	X	X	X	X		X	X	
	Sharing agreements and licensing	X	X	X			X	X	
	Service-level agreements			X			X	X	X
	Terms of service			X			X	X	
	Standardized, machine-actionable license documents			X		X	X		
	Citation requirements			X		X	X	X	
Levels of Protection	Unclassified but sensitive information, e.g., de-identification, enclaves		X	X			X		
	Security classification		X	X			X		X
	Protection of limited data/secure platforms/enclaves		X	X			X		

	Constraints and restrictions on data use and sharing			X			X	X	
	Anonymization			X			X	X	
Architectures for Application, Use, and Reuse	Extensibility across communities, including machine-based interactions					X	X		X
	Capture of insights from ML and use of these to improve datasets for future AI applications	X					X		X
	Capture of data performance characteristics	X		X			X		X
	Location of data	X	X	X		X	X		X
	Migration strategies and mitigation against data loss			X			X		X
	Economic impact of reuse		X				X		X

Preserve/ Discard: Topic	**Subtopic**	**AI Experts**	**Budget/Cost Experts**	**Curators**	**Data/IT Leader**	**Providers of Data Tools**	**Publishers**	**Research Organization Leaders**	**Researchers**
Criteria for Preservation	Use			X	X			X	X
	Impact			X				X	X
	Value			X				X	X
	Uniqueness			X				X	
	Cost		X	X				X	
	Provenance			X				X	
	Legal and regulatory			X				X	
Sustainability	Longevity and support			X				X	
	Funding models		X	X				X	
	Business models		X	X				X	

Storage and Preservation	Media to store and preserve data	X			
	File integrity	X			
	Ability to do advanced searches	X			
	Backup and recovery	X			
Moving Data from One Service to Another Across Organizations	Roles and responsibilities	X			
	Registry maintenance and curation	X	X		
	Disciplinary archives	X			
Retention and Disposition Schedules	Technical decisions, e.g., data archiving	X			
	Administrative/policy decisions	X			X
	Deaccessioning/end-of-life	X			
	Legal documents				X
	End-of-life special considerations	X			
	Recognition of removed data, i.e., tombstone page	X			

6. Conclusions

This interim version 1.5 of the NIST RDaF has been developed through extensive stakeholder engagement via a total of 17 workshops. Carefully crafted methodologies were used in the development process, which took place over nearly two years. The RDaF is based on a lifecycle model with six stages with a comprehensive list of defined topics, subtopics, and informative references nested under each stage. Version 1.5 also contains full descriptions of 14 pervasive overarching themes. Further, 8 generic profiles detailing the relevant subtopics for 8 common roles in research data and research data management have been created. Finally, the framework contains a list of research data management organizations, with a link to the homepage for each organization. In addition to these features and resources, this effort produced a customizable profile-maker for the management of research data. The following section describes additional web-based tools under development and the process for generating RDaF version 2.0 to be released in Fall 2023.

6.1. Future Work

Given that the research data ecosystem is evolving rapidly, NIST intends to update the RDaF publication on a regular basis (subject to availability of resources) and assist the community, including all organizations[1] and individuals engaged in the use of the framework, in assessing and improving their research data management capacity. We are interested in partnering with organizations with similar aspirations, such as the Australian Research Data Commons, who recently released "Research Data Management Framework for Institutions"[258] and the Research Data Alliance's new working group, the RDA-OfR Mapping the Digital Research Data Infrastructure Landscape."[259] We encourage anyone seeking assistance in using the RDaF or considering developing value-added tools based on the RDaF to contact us at rdaf@nist.gov.

6.1.1. Phase 4: RFI and Feedback to Generate RDaF v. 2.0

The RDaF team developed a Request for Information (RFI) to be posted in the Federal Register to communicate updates to the RDaF and receive additional feedback on Version 1.5. This publication includes the entire list of topics and subtopics for the six lifecycle stages, definitions, informative references corresponding to each subtopic, 14 overarching themes, and 8 generic profiles.

The public will have 30 days after release of the RFI to comment on any aspect of the RDaF. All comments received will be considered in generating version 2.0 of the framework.

6.1.2. Web-based Tools

Given the complexity of the framework, the RDaF team is working on various tools to make it easier to understand and use. A web accessible database will include all the components

[1] It is impossible to provide a comprehensive list of organizations relevant to data management. Such a list would be constantly changing as the research data field is dynamic and new groups and organizations are continually formed. Appendix D catalogs many of the government and non-government (public and private) organizations engaged in research data management.

mentioned, including links between the subtopics and their corresponding informative references. The database will enable searches on any term in the RDaF so a user can navigate the full content. Additionally, the RDaF team hopes to create an automated or guided profile maker.

6.1.3. Interactive Knowledge Graph

An interactive, web-based knowledge graph will accompany version 2.0 of the RDaF. This knowledge graph will allow exploration of the relationships between topics, subtopics, informative references, and job functions within the research data management ecosystem.

6.1.4. Parallel Natural Language Processing

The various workshops held to further develop the RDaF resulted in many transcripts and notes. The methodology subsections above describe a manual, human-driven method of incorporating that feedback to generate version 1.5.

As a supplement and an experimental exercise, the RDaF team is also conducting a parallel natural language processing trial to extract machine learning conclusions. These findings will be compared with the results of the manual process and may be incorporated in RDaF v. 2.0.

References

[1] Hanisch RJ, Kaiser DL, Carroll BC (2021) Research Data Framework (RDaF): Motivation, Development, and a Preliminary Framework Core. (National Institute of Standards and Technology). https://doi.org/10.6028/NIST.SP.1500-18

[2] Data Asset *NIST Computer Security Resource Center Glossary*. Available at https://csrc.nist.gov/glossary/term/data_asset

[3] Research Data Management Terminology *CODATA, The Committee on Data for Science and Technology*. Available at https://codata.org/initiatives/data-science-and-stewardship/rdm-terminology-wg/rdm-terminology/

[4] Techopedia: Educating IT Professionals To Make Smarter Decisions - Techopedia Available at https://www.techopedia.com/

[5] What is the difference between mission, vision and values statements? (2023) *SHRM*. Available at https://www.shrm.org/resourcesandtools/tools-and-samples/hr-qa/pages/mission-vision-values-statements.aspx

[6] Data policy *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/data-policy/

[7] Data management *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/data-management/

[8] Data governance *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/data-governance/

[9] National Institute of Standards and Technology (2018) Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1. (National Institute of Standards and Technology, Gaithersburg, MD), NIST CSWP 04162018, p NIST CSWP 04162018. Available at https://doi.org/10.6028/NIST.CSWP.04162018

[10] What are organizational values? *Workplace from Meta*. Available at https://www.workplace.com/blog/organizational-values

[11] Verlinden N (2021) Organizational Values: Definition, Purpose & Lots of Examples. *AIHR*. Available at https://www.aihr.com/blog/organizational-values/

[12] Briggs LL (2011) Q&A: Solid Value Proposition a Key to MDM Success. *Transforming Data with Intelligence*. Available at https://tdwi.org/articles/2011/02/16/value-proposition-mdm-success.aspx

[13] NOAA Administrative Order 212-15 (National Oceanic and Atmospheric Administration), 212–15, p 4. Available at https://www.noaa.gov/sites/default/files/legacy/document/2020/Mar/212-15.pdf

[14] What is Data Privacy *SNIA*. Available at https://secure.livechatinc.com/

[15] Data ethics *Cognizant Glossary*. Available at https://www.cognizant.com/us/en/glossary/data-ethics

[16] Kengadaran S (2019) Ethics for Data Projects. *Siddarth Kengadaran*. Available at https://siddarth.design/ethics-for-data-projects-5af0af333e71

[17] Bhandari P (2022) Ethical Considerations in Research | Types & Examples. *Scribbr*. Available at https://www.scribbr.com/methodology/research-ethics/

[18] What is Data Security? Data Security Definition and Overview *IBM*. Available at https://www.ibm.com/topics/data-security

[19] K. Molch, R. Cosac (2020) Long Term Preservation of Earth Observation Space Data: Glossary of Acronyms and Terms. Available at https://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/White_Papers/EO-DataStewardshipGlossary.pdf

[20] Karen Scarfone How to Perform a Data Risk Assessment, Step by Step. *Tech Target*. Available at https://www.techtarget.com/searchsecurity/tip/How-to-perform-a-data-risk-assessment-step-by-step

[21] What is Data Risk Management? Why You Should Care? (2022) *The ECM Consultant*. Available at https://theecmconsultant.com/data-risk-management/

[22] Data Sharing Agreements *US Geological Survey*. Available at https://www.usgs.gov/data-management/data-sharing-agreements

[23] Data License Agreement (2021) *Dimewiki*. Available at https://dimewiki.worldbank.org/Data_License_Agreement

[24] Aitken M, Toreini E, Carmichael P, Coopamootoo K, Elliott K, van Moorsel A (2020) Establishing a social licence for Financial Technology: Reflections on the role of the private sector in pursuing ethical data practices. *Big Data & Society* 7(1):2053951720908892. https://doi.org/10.1177/2053951720908892

[25] Sariyar M, Schluender I, Smee C, Suhr S (2015) Sharing and Reuse of Sensitive Data and Samples: Supporting Researchers in Identifying Ethical and Legal Requirements. *Biopreservation and Biobanking* 13(4):263–270. https://doi.org/10.1089/bio.2015.0014

[26] Southekal P (2022) Data Culture: What It Is And How To Make It Work. *Forbes*. Available at https://www.forbes.com/sites/forbestechcouncil/2022/06/27/data-culture-what-it-is-and-how-to-make-it-work/

[27] Scientific Integrity and Research Misconduct Available at https://www.usda.gov/our-agency/staff-offices/office-chief-scientist-ocs/scientific-integrity-and-research-misconduct

[28] What Is Data Integrity and Why Does It Matter? (2021) *Business Insights Blog*. Available at https://online.hbs.edu/blog/post/what-is-data-integrity

[29] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3(1):160018. https://doi.org/10.1038/sdata.2016.18

[30] CARE Principles of Indigenous Data Governance (2023) *Global Indigenous Data Alliance*. Available at https://www.gida-global.org/care

[31] Stakeholder *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/stakeholder/

[32] Numans W, Van Regenmortel T, Schalk R (2019) Partnership Research: A Pathway to Realize Multistakeholder Participation. *International Journal of Qualitative Methods* 18:1609406919884149. https://doi.org/10.1177/1609406919884149

[33] Domain knowledge (2023) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Domain_knowledge&oldid=1136257348

[34] inclusivity (2023) *Cambridge Dictionary* online. Available at https://dictionary.cambridge.org/us/dictionary/english/inclusivity

[35] Data Services (2015) *Techopedia*. Available at https://www.techopedia.com/definition/1005/data-services

[36] CISA Insights: Chain of Custody and Critical Infrastructure Systems Available at https://www.cisa.gov/sites/default/files/publications/cisa-insights_chain-of-custody-and-ci-systems_508.pdf

[37] Provenance *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/provenance/

[38] Perreault G, Kim P, Foster W (2011) Finding Your Funding Model. *Stanford Social Innovation Review* 9:3741. https://doi.org/10.48558/QPQR-QT49

[39] Cost–benefit analysis (2023) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Cost%E2%80%93benefit_analysis&oldid=1136963825

[40] DCC (2013) Checklist for a Data Management Plan. v.4.0. Available at https://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf

[41] Machine readable *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/machine-readable/

[42] What is Data Organization? - Importance & Tips *Sisense*. Available at https://www.sisense.com/glossary/data-organization/

[43] Saul Mcleod (2022) Qualitative vs Quantitative Research: Methods & Data Analysis. *Simply Psychology*. Available at https://simplypsychology.org/qualitative-quantitative.html

[44] Observation Definition & Meaning *Merriam-Webster*. Available at https://www.merriam-webster.com/dictionary/observation

[45] Survey (human research) (2023) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Survey_(human_research)&oldid=1135741584

[46] What is Research Software? *IGI Global*. Available at https://www.igi-global.com/dictionary/knowledge-visualization-for-research-design/69111

[47] Modeling in Scientific Research *Visionlearning Process of Science*. Available at https://www.visionlearning.com/en/library/Process-of-Science/49/Modeling-in-Scientific-Research/153

[48] Documented data *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/documented-data/

[49] Bechhofer S, De Roure D, Gamble M, Goble C, Buchan I (2010) Research Objects: Towards Exchange and Reuse of Digital Knowledge. *Nature Precedings*:1–1. https://doi.org/10.1038/npre.2010.4626.1

[50] Dobreski B, Park J, Leathers A, Qin J (2020) Remodeling Archival Metadata Descriptions for Linked Archives. *International Conference on Dublin Core and Metadata Applications*, pp 1–11. Available at https://dcpapers.dublincore.org/pubs/article/view/4223

[51] Metadata Object Description Schema: MODS (2022) *Library of Congress*. Available at https://www.loc.gov/standards/mods/

[52] What is a Data Workflow? Use Cases & How to Get Started (2023) *Cflow*. Available at https://www.cflowapps.com/data-workflow/

[53] Model *NIST Computer Security Resource Center Glossary*. Available at https://csrc.nist.gov/glossary/term/model

[54] Laboratory Information Management System (LIMS) (2018) *Techopedia*. Available at https://www.techopedia.com/definition/8085/laboratory-information-management-system-lims

[55] Configuration Management (2012) *Techopedia*. Available at https://www.techopedia.com/definition/24822/configuration-controlconfiguration-management-cm

[56] Architecture *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/architecture/

[57] Research Data Architectures in Research Institutions IG (2017) *RDA*. Available at https://www.rd-alliance.org/groups/research-data-architectures-research-institutions-ig

[58] Interoperability *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/interoperability/

[59] Stocker M, Darroch L, Krahl R, Habermann T, Devaraju A, Schwardmann U, D'Onofrio C, Häggström I (2020) Persistent Identification of Instruments. *Data Science Journal* 19(1):18. https://doi.org/10.5334/dsj-2020-018

[60] Data standards *Data.gov*. Available at https://resources.data.gov/standards/concepts/

[61] Data Quality (2022) *Techopedia*. Available at https://www.techopedia.com/definition/14653/data-quality

[62] Standard *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/standard/

[63] Sansone S-A (2016) NIH BD2K workshop report: "Frameworks for Community-based Standards Efforts". Available at https://doi.org/10.6084/m9.figshare.3795816.v2

[64] Ball A, Duke M (2015) How to Track the Impact of Research Data with Metrics. Available at https://www.dcc.ac.uk/guidance/how-guides/track-data-impact-metrics

[65] Alpi KM, Akers KG (2021) CRediT for authors of articles published in the Journal of the Medical Library Association. *Journal of the Medical Library Association* 109(3):362–364. https://doi.org/10.5195/jmla.2021.1294

[66] Regulatory compliance (2023) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Regulatory_compliance&oldid=1147347472

[67] PII *NIST Computer Security Resource Center Glossary*. Available at https://csrc.nist.gov/glossary/term/pii

[68] Intellectual Property Sample Clauses *Law Insider*. Available at https://www.lawinsider.com/clause/intellectual-property

[69] Responsible Conduct in Data Management Glossary Available at https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dmglossary.html#A

[70] Text File (2016) *Techopedia*. Available at https://www.techopedia.com/definition/9707/text-file

[71] Simulation (2019) *Techopedia*. Available at https://www.techopedia.com/definition/5757/simulation

[72] Computation *www.dictionary.com*. Available at https://www.dictionary.com/browse/computation

[73] Source Code (2017) *Techopedia*. Available at https://www.techopedia.com/definition/547/source-code

[74] Transaction Definition & Meaning *Merriam-Webster*. Available at https://www.merriam-webster.com/dictionary/transaction

[75] Social media (2023) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Social_media&oldid=1147905665

[76] User Facility (2014) *Department of Energy OSTI*. Available at https://science.osti.gov/User-Facilities/Policies-and-Processes/Definition

[77] Koch R (2022) Human Annotated Data - All You Need to Know About It. *clickworker.com*. Available at https://www.clickworker.com/customer-blog/human-annotated-data/

[78] Hillemann B (2023) Experimental Data. *Macalester University Dewitt Wallace Library LibGuides*. Available at https://libguides.macalester.edu/c.php?g=527786&p=3608643

[79] Reproducible research *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/reproducible-research/

[80] Hardware (2020) *Techopedia*. Available at https://www.techopedia.com/definition/2210/hardware-hw

[81] System Requirements (2015) *Techopedia*. Available at https://www.techopedia.com/definition/4371/system-requirements

[82] Version control *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/version-control/

[83] Document Versioning (2014) *Techopedia*. Available at https://www.techopedia.com/definition/30702/document-versioning

[84] B.H.Thacker, S.W.Doebling, F.M.Hemez, M.C. Anderson, J.E. Pepin, E.A. Rodriguez (2004) Concepts of Model Verification and Validation., LA-14167, 835920, p LA-14167, 835920. Available at https://doi.org/10.2172/835920

[85] Metadata *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/metadata/

[86] Paradata (2022) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Paradata&oldid=1078821391

[87] Alan F. Karr (2020) Metadata and Paradata: Information Collection and Potential Initiatives. *National Institute of Statistical Sciences*. Available at https://www.niss.org/research/metadata-and-paradata-information-collection-and-potential-initiatives

[88] Repository *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/repository/

[89] Data integrity *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/data-integrity/

[90] Critical Evaluation Criteria (2021) *NIST*. Available at https://www.nist.gov/srd/critical-evaluation-criteria

[91] International vocabulary of metrology – Basic and general concepts and associated terms (VIM), 3rd Edition (2012) Available at https://www.bipm.org/en/search?p_p_id=search_portlet&p_p_lifecycle=2&p_p_state=normal&p_p_mode=view&p_p_resource_id=%2Fdownload%2Fpublication&p_p_cacheability=cacheLevelPage&_search_portlet_dlFileId=41373499&p_p_lifecycle=1&_search_portlet_javax.portlet.action=search&_search_portlet_formDate=1670328688739&_search_portlet_query=VIM&_search_portlet_source=BIPM

[92] Saha CN, Bhattacharya S (2011) Intellectual property rights: An overview and implications in pharmaceutical industry. *Journal of Advanced Pharmaceutical Technology & Research* 2(2):88. https://doi.org/10.4103/2231-4040.82952

[93] FAIR Digital Objects Available at https://fairdo.org/1316-2/

[94] SmartAPI | About (2022) *SmartAPI*. Available at https://smart-api.info/about

[95] What does data format mean? Available at https://www.definitions.net/definition/data+format

[96] File Structure *MIT Communication Lab*. Available at https://mitcommlab.mit.edu/broad/commkit/file-structure/

[97] file structure *SAA Dictionary of Archives Terminology*. Available at https://dictionary.archivists.org/entry/file-structure.html

[98] Bolam M Guides: Metadata & Discovery @ Pitt: Metadata Standards. Available at https://pitt.libguides.com/metadatadiscovery/metadata-standards

[99] Metadata Standards Catalog Available at https://rdamsc.bath.ac.uk/

[100] What is an Ontology? Available at https://www.oxfordsemantic.tech/fundamentals/what-is-an-ontology

[101] Sansone S-A, Rocca-Serra P (2016) Review: Interoperability standards. https://doi.org/10.6084/m9.figshare.4055496.v1

[102] Open-Source Software (2016) *Techopedia*. Available at https://www.techopedia.com/definition/5602/open-source-software-oss

[103] Proprietary Software (2017) *Techopedia*. Available at https://www.techopedia.com/definition/4333/proprietary-software

[104] Electronic Laboratory Notebook (ELN) *NNLM*. Available at https://www.nnlm.gov/guides/data-glossary/electronic-laboratory-notebook-eln

[105] Srivastav AK (2019) Graphs vs Charts. *WallStreetMojo*. Available at https://www.wallstreetmojo.com/graphs-vs-charts/

[106] Instrument output data *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/instrument-output-data/

[107] Dynamic data *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/dynamic-data/

[108] Static Data (2018) *Techopedia*. Available at https://www.techopedia.com/definition/31590/static-data

[109] Dataset *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/dataset/

[110] Banks J ed. (2001) *Discrete-event system simulation* (Prentice Hall, Upper Saddle River, NJ), 3rd ed. Available at https://worldcat.org/title/43945281

[111] Structured data *CODATA, The Committee on Data for Science and Technology*. Available at https://codata.org/rdm-terminology/structured-data/

[112] Data cleaning *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/data-cleaning/

[113] ISO 25237:2017 Health informatics — Pseudonymization, ISO 25237:2017. Available at https://www.iso.org/standard/63553.html

[114] Data Preprocessing (2021) *Techopedia*. Available at https://www.techopedia.com/definition/14650/data-preprocessing

[115] Schouten RM, Lugtig P, Vink G (2018) Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation* 88(15):2909–2930. https://doi.org/10.1080/00949655.2018.1491577

[116] Badr W (2019) 6 Different Ways to Compensate for Missing Data (Data Imputation with examples). *Towards Data Science*. Available at https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779

[117] King T (2018) The Definitive Data Management Glossary. *Solutions Review*. Available at https://solutionsreview.com/data-management/the-definitive-data-management-glossary/

[118] Schwer LE (2007) An overview of the PTC 60/V&V 10: guide for verification and validation in computational solid mechanics. *Engineering with Computers* 23(4):245–252. https://doi.org/10.1007/s00366-007-0072-z

[119] Data Curation *NNLM*. Available at https://www.nnlm.gov/guides/data-glossary/data-curation

[120] Lu M, Zhao Q, Zhang J, Pohl KM, Fei-Fei L, Niebles JC, Adeli E (2021) Metadata Normalization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 10912–10922. https://doi.org/10.1109/CVPR46437.2021.01077

[121] Manual Data Processing: The Secrets of Automation (2021) *Solvexia.com*. Available at https://www.solvexia.com/blog/manual-data-processing-the-secrets-of-automation

[122] Exploratory Data Analysis (2017) *Techopedia*. Available at https://www.techopedia.com/definition/32962/exploratory-data-analysis-eda

[123] Catherine Cote (2021) What Is Descriptive Analytics? 5 Examples. *Business Insights Blog*. Available at https://online.hbs.edu/blog/post/descriptive-analytics

[124] Catherine Cote (2021) What Is Diagnostic Analytics? 4 Examples. *Business Insights Blog*. Available at https://online.hbs.edu/blog/post/diagnostic-analytics

[125] Susan Parker, Gwen Fariss Newman What is evaluation? Available at https://www.eval.org/Portals/0/What%20is%20evaluation%20Document.pdf

[126] Catherine Cote (2021) What Is Predictive Analytics? 5 Examples. *Business Insights Blog*. Available at https://online.hbs.edu/blog/post/predictive-analytics

[127] Catherine Cote (2021) What Is Prescriptive Analytics? 6 Examples. *Business Insights Blog*. Available at https://online.hbs.edu/blog/post/prescriptive-analytics

[128] Rainbow Framework *Rainbow Framework*. Available at https://www.betterevaluation.org/frameworks-guides/rainbow-framework

[129] Positive Correlation: What It Is, How to Measure It, Examples (2022) *Investopedia*. Available at https://www.investopedia.com/terms/p/positive-correlation.asp

[130] Negative Correlation: How it Works, Examples And FAQ *Investopedia*. Available at https://www.investopedia.com/terms/n/negative-correlation.asp

[131] Statistical Analysis (2022) *WallStreetMojo*. Available at https://www.wallstreetmojo.com/statistical-analysis/

[132] statistical data analysis *WhatIs.com*. Available at https://www.techtarget.com/whatis/search/query?q=statistical+data+analysis

[133] Autonomous Things (2019) *Techopedia*. Available at https://www.techopedia.com/definition/33723/autonomous-things

[134] Simulation vs. Visualization - what's the difference? (2017) *Visual Components*. Available at https://www.visualcomponents.com/resources/blog/simulation-vs-visualization-difference/

[135] Machine Learning *Techopedia*. Available at https://www.techopedia.com/topic/318/machine-learning

[136] Artificial Intelligence *Techopedia*. Available at https://www.techopedia.com/topic/87/artificial-intelligence

[137] Priya Pedamkar (2019) Iterative Model. *EDUCBA*. Available at https://www.educba.com/iterative-model/

[138] Integrated Development Environment (2017) *Techopedia*. Available at https://www.techopedia.com/definition/26860/integrated-development-environment-ide

[139] Cofield M (2022) Metadata Basics: Key Concepts. *University of Texas Libraries*. Available at https://guides.lib.utexas.edu/metadata-basics/key-concepts

[140] Dennis AL (2022) The Value of Metadata Governance. *DATAVERSITY*. Available at https://www.dataversity.net/the-value-metadata-governance/

[141] Gilliland AJ (2016) Setting the Stage. *Introduction to Metadata* Available at http://www.getty.edu/publications/intrometadata

[142] Data linkage *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/data-linkage/

[143] Persistent identifier *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/persistent-identifier/

[144] What are Persistent Identifiers (2020) *CERN*. Available at https://sis.web.cern.ch/submit-and-publish/persistent-identifiers/what-are-pids

[145] Authoritative copies *Docusign Developer*. Available at https://developers.docusign.com/docs/esign-rest-api/esign101/concepts/documents/authoritative-copies/

[146] Glossary of data management terms | Research Data Management Service Group (2022) *Cornell University*. Available at https://data.research.cornell.edu/content/glossary

[147] Jeffreys A (2018) Database subsetting. *Redgate*. Available at https://www.red-gate.com/blog/database-devops/database-subsetting-wed-love-hear

[148] Timestamp (2016) *Techopedia*. Available at https://www.techopedia.com/definition/16285/timestamp

[149] CRediT (2011) *CRediT*. Available at https://credit.niso.org/

[150] Commercial Software (2014) *Techopedia*. Available at https://www.techopedia.com/definition/4245/commercial-software

[151] What is Custom Software? Available at https://www.computerhope.com/jargon/c/customso.htm

[152] software *WhatIs.com*. Available at https://www.techtarget.com/whatis/search/query?q=software

[153] Statistics (2023) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Statistics&oldid=1148101750

[154] Application Programming Interface (2022) *Techopedia*. Available at https://www.techopedia.com/definition/24407/application-programming-interface-api

[155] Data Management Software (2013) *Techopedia*. Available at https://www.techopedia.com/definition/11363/data-management-software-dms

[156] Data Validation (2017) *Techopedia*. Available at https://www.techopedia.com/definition/10283/data-validation

[157] What is Software Documentation? Definition, Types and Examples *Tech Target - Software Quality*. Available at https://www.techtarget.com/searchsoftwarequality/definition/documentation

[158] resilience *NIST Computer Security Resource Center Glossary*. Available at https://csrc.nist.gov/glossary/term/resilience

[159] Subramanian N, Chung L (2001) *Metrics for Software Adaptability*. Available at https://personal.utdallas.edu/~chung/ftp/sqm.pdf

[160] What is a Software Repository? (2021) *Full Scale*. Available at https://fullscale.io/blog/software-repository/

[161] Data Management Glossary *National Agriculture Library*. Available at https://www.nal.usda.gov/data/data-management-glossary#W3clib

[162] Update *NIST Computer Security Resource Center Glossary*. Available at https://csrc.nist.gov/glossary/term/update

[163] Resources.data.gov: a Repository of Federal Enterprise Data Resources *Data management & governance resources*. Available at https://resources.data.gov/categories/data-management-governance/

[164] Protocol (2020) *Techopedia*. Available at https://www.techopedia.com/definition/4528/protocol

[165] What is an Interface? (2020) *Computer Hope*. Available at https://www.computerhope.com/jargon/i/interfac.htm

[166] Ryan P Webinar on Keeping a Lab Notebook - Basic Principles and Best Practices. Available at https://www.training.nih.gov/assets/Lab_Notebook_508_(new).pdf

[167] Anomaly Detection (2014) *Techopedia*. Available at https://www.techopedia.com/definition/30297/anomaly-detection

[168] Work Flow (2016) *Techopedia*. Available at https://www.techopedia.com/definition/10072/work-flow

[169] Middleware (2017) *Techopedia*. Available at https://www.techopedia.com/definition/450/middleware

[170] Middleware *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/middleware/

[171] What is Monitoring - Types of Monitoring, Process Monitoring, Validation, Tracking, Performance *Monitoring and Evaluation Studies*. Available at http://www.mnestudies.com/monitoring/what-monitoring

[172] Containerization (computing) (2023) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Containerization_(computing)&oldid=1148088666

[173] Library (Reusable components) (2018) *UiPath Community Forum*. Available at https://forum.uipath.com/t/featureblog-18-3-library-reusable-components/62746

[174] Microservices (2021) *Techopedia*. Available at https://www.techopedia.com/definition/32503/microservices

[175] Workflow Management System *NIST Computer Security Resource Center Glossary*. Available at https://csrc.nist.gov/glossary/term/workflow_management_system

[176] Compute (2016) *Techopedia*. Available at https://www.techopedia.com/definition/6580/compute

[177] Million Instructions per Second (MIPS) *Gartner Information Technology Glossary*. Available at https://www.gartner.com/en/information-technology/glossary/mips-million-instructions-per-second

[178] Storage (2022) *Techopedia*. Available at https://www.techopedia.com/definition/1115/storage

[179] Madden S (2019) Network Speed vs. Bandwidth? *Interconnections - The Equinix Blog*. Available at https://blog.equinix.com/blog/2019/05/09/network-speed-vs-bandwidth/?lang=ja

[180] What is an Accelerator? Available at https://www.computerhope.com/jargon/a/accelera.htm

[181] What Is Hardware Acceleration and When Should You Use It? (2021) *Make Use Of*. Available at https://www.makeuseof.com/what-is-hardware-acceleration/

[182] Stall, Shelley, Martone, Maryann E., Chandramouliswaran, Ishwar, Crosas, Mercè, Federer, Lisa, Gautier, Julian, Hahnel, Mark, Larkin, Jennie, Lowenberg, Daniella, Pfeiffer, Nicole, Sim, Ida, Smith, Tim, Van Gulick, Ana E., Walker, Erin, Wood, Julie, Zaringhalam, Maryam, Zigoni, Alberto (2020) Generalist Repository Comparison Chart. https://doi.org/10.5281/ZENODO.3946720

[183] Data Repository *Egnyte*. Available at https://www.egnyte.com/guides/governance/data-repository

[184] Research data publishing *Springer Nature*. Available at https://www.springernature.com/gp/authors/research-data/research-data-publishing

[185] Support and information Wageningen Data Competence Center Contact form (2015) Why publish research data? *Wageningen University & Research*. Available at https://www.wur.nl/en/value-creation-cooperation/collaborating-with-wur-1/wdcc/research-data-management-wdcc/finishing/why-publish-research-data.htm

[186] DATA UPDATING *Law Insider*. Available at https://www.lawinsider.com/dictionary/data-updating

[187] What is Data Linking? *TIBCO Software*. Available at https://www.tibco.com/reference-center/what-is-data-linking

[188] What is Data Integrity and How Can You Maintain it? *Inside Out Security Blog*. Available at https://www.varonis.com/blog/data-integrity

[189] Sarfin RL (2022) Data Quality Dimensions: How Do You Measure Up? (+ Free Scorecard). *Precisely*. Available at https://www.precisely.com/blog/data-quality/data-quality-dimensions-measure

[190] Research Data Guidelines *Elsevier Author Tools*. Available at https://www.elsevier.com/authors/tools-and-resources/research-data/data-guidelines

[191] Publishing Agreement: Definition & Sample *Contract Counsel*. Available at https://www.contractscounsel.com/t/us/publishing-agreement

[192] OA agreements *Author Services - Taylor & Francis*. Available at https://authorservices.taylorandfrancis.com/choose-open/publishing-open-access/oa-agreements/

[193] Publishing policies | Policies | Springer Nature *Springer Nature*. Available at https://www.springernature.com/gp/policies/publishing-policies

[194] Scholarly Publishing: Traditional and Open Access *Rutgers University Libraries*. Available at https://www.libraries.rutgers.edu/research-tools-and-services/copyright-guidance/copyright-academic-research-and-publication/scholarly-publishing-traditional-and-open-access

[195] Supplementary information | Nature Available at https://www.nature.com/nature/for-authors/supp-info

[196] Submit a Data Request *National Resident Matching Program*. Available at https://www.nrmp.org/match-data-analytics/submit-a-data-request/

[197] What is a Landing Page and Why Should You Use Them? *Mailchimp*. Available at https://mailchimp.com/marketing-glossary/landing-pages/

[198] mainstream media (2023) *Cambridge Dictionary*. Available at https://dictionary.cambridge.org/us/dictionary/english/mainstream-media

[199] Social Media *Techopedia*. Available at https://www.techopedia.com/definition/4837/social-media

[200] Fisher T LibGuides: Research Publishing & Impact: Citation Metrics. *University of Otago Library*. Available at https://otago.libguides.com/research_publishing_impact/citation_metrics

[201] DeGroote S Measuring Your Impact: Impact Factor, Citation Analysis, and other Metrics: Citation Analysis. *UIC Libraries Research Guides*. Available at https://researchguides.uic.edu/c.php?g=252299&p=1683205

[202] Sharma M, Sarin A, Gupta P, Sachdeva S, Desai AV (2014) Journal Impact Factor: Its Use, Significance and Limitations. *World Journal of Nuclear Medicine* 13(2):146. https://doi.org/10.4103/1450-1147.139151

[203] - Data Citation and Policies. Land Processes Distributed Active Archive Center (LP DAAC. *US Geological Survey*. Available at https://lpdaac.usgs.gov/data/data-citation-and-policies/

[204] Cite Your Data *DataCite*. Available at https://datacite.org/cite-your-data.html

[205] Data Citation Synthesis Group (2014) Joint Declaration of Data Citation Principles. (Force11). https://doi.org/10.25490/A97F-EGYK

[206] Research Guides: Author Identity Management: ORCID *Run Run Shaw Library City University of Hong Kong*. Available at https://libguides.library.cityu.edu.hk/aim/orcid

[207] Content discovery platform (2023) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Content_discovery_platform&oldid=1135084424

[208] Data Catalog (2016) *Techopedia*. Available at https://www.techopedia.com/definition/32034/data-catalog

[209] Registry *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/registry/

[210] re3data *re3data*. Available at https://www.re3data.org/

[211] NIST Materials Resource Registry *National Institute of Standards and Technology*. Available at https://www.nist.gov/programs-projects/nist-materials-resource-registry

[212] Data Access (2012) *Techopedia*. Available at https://www.techopedia.com/definition/26929/data-access

[213] Data Enclave *Network of the National Library of Medicine*. Available at https://www.nnlm.gov/guides/data-thesaurus/data-enclave

[214] Weise M, Kovacevic F, Popper N, Rauber A (2022) OSSDIP: Open Source Secure Data Infrastructure and Processes Supporting Data Visiting. *Data Science Journal* 21(1):4. https://doi.org/10.5334/dsj-2022-004

[215] Data Availability Statements - Research Data Policy (2022) *Springer Nature*. Available at https://www.springernature.com/gp/authors/research-data-policy/data-availability-statements

[216] Data Ownership (2012) *Techopedia*. Available at https://www.techopedia.com/definition/29059/data-ownership

[217] Intellectual Property (2022) *Techopedia*. Available at https://www.techopedia.com/definition/5521/intellectual-property-ip

[218] User Agreements 101: What You Need to Know *Ironclad*. Available at https://ironcladapp.com/journal/contracts/user-agreements/

[219] Licensing Agreement: What Is It? 5 Elements To Include Available at https://www.contractscounsel.com/t/us/licensing-agreement

[220] Harper (Michael) (2021) The relationship between data SLAs & data products. *Medium*. Available at https://towardsdatascience.com/the-relationship-between-data-slas-data-products-77207f876072

[221] Terms of Service (2015) *Techopedia*. Available at https://www.techopedia.com/definition/9746/terms-of-service-tos

[222] 12 FAM 540 SENSITIVE BUT UNCLASSIFIED INFORMATION (SBU). *Foreign Affairs Manual* (U.S. Department of State). Available at https://fam.state.gov/fam/12fam/12fam0540.html

[223] De-Identification Guidelines (2018) *Safety and Risk Services - University of Oregon*. Available at https://safety.uoregon.edu/de-identification-guidelines

[224] Controlled Unclassified Information (CUI) (2016) *National Archives*. Available at https://www.archives.gov/cui

[225] Classification Guide: Protection Levels - Information Security & Privacy Office *New School - Information & Privacy Office*. Available at https://ispo.newschool.edu/guidelines/protection-levels/

[226] 5 FAM 480 CLASSIFYING AND DECLASSIFYING NATIONAL SECURITY INFORMATION—EXECUTIVE ORDER 13526. *Foreign Affairs Manual* (U.S. Department of State). Available at https://fam.state.gov/fam/05fam/05fam0480.html

[227] Joint Task Force Interagency Working Group (2020) Security and Privacy Controls for Information Systems and Organizations. (National Institute of Standards and Technology), SP 800-53r5. Available at https://doi.org/10.6028/NIST.SP.800-53r5

[228] 6 Must-Haves in a Data Security Platform *CIO*. Available at https://www.cio.com/article/407778/6-must-haves-in-a-data-security-platform.html

[229] Limited Data Sets and Data Use Agreements (2020) Available at https://www.womans.org/-/media/files/womans/research/policies/limited-data-sets-and-data-use-agreements.pdf?la=en&hash=6772539AC17E04ECE6ECAF00BDA3DB0ED8329F71

[230] Howison M, Angell M, Hicklen MS, Hastings JS (2021) *Protecting Sensitive Data with Secure Data Enclaves* (OSF Preprints). https://doi.org/10.31219/osf.io/jmd7t

[231] Data Anonymization *Corporate Finance Institute*. Available at https://corporatefinanceinstitute.com/resources/business-intelligence/data-anonymization/

[232] Extensibility (2021) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Extensibility&oldid=1008862248

[233] 7 data quality best practices to improve data performance | TechTarget *TechTarget Data Management*. Available at https://www.techtarget.com/searchdatamanagement/tip/Data-quality-best-practices-to-improve-data-performance

[234] Digital Preservation Metrics *Center for Research Libraries: Global Resources Network*. Available at https://www.crl.edu/archiving-preservation/digital-archives/metrics

[235] Definition of uniqueness | Dictionary.com *www.dictionary.com*. Available at https://www.dictionary.com/browse/uniqueness

[236] Data longevity *PCMAG*. Available at https://www.pcmag.com/encyclopedia/term/data-longevity

[237] Harrington LMB (2016) Sustainability Theory and Conceptual Considerations: A Review of Key Ideas for Sustainability, and the Rural Context. *Papers in Applied Geography* 2(4):365–382. https://doi.org/10.1080/23754931.2016.1239222

[238] Business model (2023) *Wikipedia*. Available at https://en.wikipedia.org/w/index.php?title=Business_model&oldid=1145556367

[239] Media (2020) *Techopedia*. Available at https://www.techopedia.com/definition/1098/media

[240] File Integrity (2014) *Techopedia*. Available at https://www.techopedia.com/definition/30616/file-integrity

[241] Integrity *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/integrity/

[242] What Are Advanced Search Options? *Lifewire*. Available at https://www.lifewire.com/what-are-advanced-search-options-3481444

[243] Data Preservation *Network of the National Library of Medicine*. Available at https://www.nnlm.gov/guides/data-glossary/data-preservation

[244] Backup (2022) *Techopedia*. Available at https://www.techopedia.com/definition/1056/backup

[245] Data recovery *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/data-recovery/

[246] Data Archives and Why You Need Them Available at https://cloudian.com/guides/data-backup/data-archive/

[247] Deaccessioning and Disposal: Guidance for Archive Services (2015) Available at https://cdn.nationalarchives.gov.uk/documents/Deaccessioning-and-disposal-guide.pdf

[248] Data retention policy *CODATA Research Data Management Terminology*. Available at https://codata.org/rdm-terminology/data-retention-policy/

[249] DataCite Support Best practices for tombstone pages. Available at https://support.datacite.org/docs/tombstone-pages

[250] Darwin Core Available at https://dwc.tdwg.org/

[251] Taillon JA, Bina TF, Plante RL, Newrock MW, Greene GR, Lau JW (2021) NexusLIMS: A Laboratory Information Management System for Shared-Use Electron Microscopy Facilities. *Microscopy and Microanalysis* 27(3):511–527. https://doi.org/10.1017/s1431927621000222

[252] PREMIS: Preservation Metadata Maintenance Activity (Library of Congress) Available at https://www.loc.gov/standards/premis/

[253] - Schema.org. Available at https://schema.org/

[254] ISO - Standards *ISO*. Available at https://www.iso.org/standards.html

[255] American National Standards Institute - ANSI Home *American National Standards Institute - ANSI*. Available at https://ansi.org/

[256] EUROPEAN COMMISSION Directorate-General for Research & Innovation (2016) H2020 Programme - Guidelines on FAIR Data Management in Horizon 2020. Available at https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[257] General Data Protection Regulation (GDPR) Compliance Guidelines Available at https://gdpr.eu/

[258] Research Data Management Framework for Institutions | ARDC (2023) *https://ardc.edu.au/*. Available at https://ardc.edu.au/resource/research-data-management-framework-for-institutions/

[259] RDA-OfR Mapping the digital research data infrastructure landscape WG Case Statement (2023) *RDA*. Available at https://www.rd-alliance.org/group/rda-ofr-mapping-digital-research-data-infrastructure-landscape-wg/case-statement/rda-ofr

Appendix A. Informative References

Research data management is a complex space. While the RDaF seeks to offer a comprehensive look at the issues and needs, nearly every topic and subtopic listed in the framework could be the subject of an entire project. In addition to definitions for each subtopic and any special terms used therefore, the RDaF also provides Informative References. These provide background information and perhaps best practices for each subtopic so users can obtain a more in-depth understanding as well as plan for their specific job functions. The database included with v. 2.0 will link the Informative References to the corresponding subtopics.

The entire bibliography can be accessed at: https://doi.org/10.6028/NIST.SP.1500-18r1sup1

Appendix B. Acronyms

AAU - Association of American Universities

AGU - American Geophysical Union

AI - artificial intelligence

APARD - Accelerating Public Access to Research Data

API - Application programming interface

APLU - Association of Public and Land-grant Universities

ARK - Archival Resource Key

CARE - Collective benefit, Authority to control, Responsibility, and Ethics

CDO - Chief Data Officer

CE - critically evaluated

CENDI - Commerce, Energy, NASA, Defense Information Managers Group

CEO - Chief Executive Officer

CODATA - Committee on Data of the International Science Council

CRediT - Contributor Roles Taxonomy - used to represent the roles typically played by contributors to scientific scholarly output

DMP - data management plan

DOC - Department of Commerce

DOE - Department of Energy

DOI - digital object identifier

FAIR - Findable, Accessible, Interoperable, and Reusable

HPC - high-performance computing

HR - human resources

IDE - Interactive Development Environment - an application that provides a full suite of features to facilitate software development. Features differ but IDEs typically allow a programmer to code, debug, and preview the effect of their code.

IP - intellectual property

LIMS - Laboratory Information Management System

ML - Machine Learning

NASEM - National Academies of Sciences, Engineering, and Medicine

NIST - National Institute of Standards and Technology

NSF - National Science Foundation

ORCID - Open Researcher and Contributor ID

OSTP - Office of Science and Technology Policy

PIDINST - Persistent Identifiers of Instruments

PII – Personally Identifiable Information

RDA - Research Data Alliance

RDaF - Research Data Framework

ROR - Research Organization Registry - unique identifiers for every research organization in the world

SPARC - Scholarly Publishing and Academic Resources Coalition

Appendix C. Generic Profiles (Separate File)

Generic profiles for eight different research data management roles are included in a separate document. This document reproduces the information in Section 5 and provides a blank template in a format amenable to the generation of customized profiles.

The template can be accessed at: https://doi.org/10.6028/NIST.SP.1500-18r1sup2

Appendix D. Descriptions of Key Organizations

In this Appendix, each key organization is accompanied by a short definition or description to provide some context of their role in research data management.

Academy of Science of South Africa - Officially recognized national science academy that aims to provide evidence-based scientific advice on issues of public interest to government and other stakeholders.
Accelerating Public Access to Research Data (APARD) - A collaboration between the Association of American Universities (AAU) and the Association of Public and Land-grant Universities (APLU) to improve public access to data resulting from federally funded research.
Alfred P. Sloan Foundation - This foundation makes grants primarily to support original research and education related to science, technology, engineering, mathematics, and economics.
American Geophysical Union (AGU) - An association of more than half a million advocates and professionals in Earth and space sciences.
American Library Association (ALA) - The oldest and largest library association in the world which aims to provide leadership for the development, promotion, and improvement of library and information services and the profession of librarianship in order to enhance learning and ensure access to information for all.
Association of American Medical Colleges - A not-for-profit association dedicated to transforming health through medical education, health care, medical research, and community collaborations.
Association of American Universities (AAU) - AAU's 65 research universities transform lives through education, research, and innovation.
Association of Public and Land-grant Universities (APLU) - A membership organization of university leaders collectively working to advance the mission of public research universities. The association's membership consists of more than 250 public research universities, land-grant institutions, state university systems, and affiliated organizations spanning across all 50 states, the District of Columbia, four U.S. territories, Canada, and Mexico.
Association of Research Libraries (ARL) - A nonprofit membership organization of research libraries and archives in major public and private universities, federal government agencies, and large public institutions in Canada and the US.
Australian Research Data Commons (ARDC) - A leading research data infrastructure facility in Australia.
Belmont Forum - A partnership of funding organizations, international science councils, and regional consortia committed to the advancement of transdisciplinary science.
Bill & Melinda Gates Foundation - A foundation that works in various science, technology, and data fields.
Biodiversity Global Information Facility - An international network and data infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth.
BRAIN Initiative - An effort by NIH to better understand how the brain functions.
California Digital Library - DMPTool - A free, open-source, online application that helps researchers create data management plans (DMPs).
CANAIRE - The Canadian Network for the Advancement of Research, Industry, and Education Center for National Research Initiatives (CNRI).

Center for Open Science - A nonprofit which works to ensure that the process, content, and outcomes of research are openly accessible by default.
China Science and Technology Cloud - A national platform to provide scientists with efficient and integrated cloud solutions in the retrieval, access, use, transaction, delivery and other aspects of sharing scientific information and relevant services.
CKAN - An open-source DMS (data management system) for powering data hubs and data portals. CKAN makes it easy to publish, share, and use data. It powers catalog.data.gov, open.canada.ca/data, and data.humdata.org, among many other sites.
Coalition for Publishing Data in the Earth and Space Sciences - A collaboration among research repositories, scholarly publishers, and other stakeholders focused on jointly developing, implementing, and promoting leading practices around the preservation and citation of data, software, and physical samples that lead toward credit and reuse in the Earth, space, and environmental sciences.
Commerce, Energy, NASA, Defense Information Managers Group (CENDI) - CENDI's mission is to increase the impact of federally funded science and technology by improving the management and dissemination of U.S. federal scientific and technical information and data.
Committee on Data of the International Science Council (CODATA) - As the Committee on Data of the International Science Council (ISC), CODATA helps realize ISC's vision of advancing science as a global public good. CODATA does this by promoting international collaboration to advance Open Science and to improve the availability and usability of data for all areas of research.
Commonwealth Scientific and Industrial Research Organisation (Australia) - An Australian Government agency responsible for scientific research.
Data Archiving and Networked Services (DANS, the Netherlands) - The Dutch national center of expertise and repository for research data.
DataCite - A leading global non-profit organization that provides persistent identifiers (DOIs) for research data and other research outputs.
DataONE (Data Observation Network for Earth) - A community driven program providing access to data across multiple member repositories, supporting enhanced search and discovery of Earth and environmental data.
Department of Energy (DOE) - The mission of the Energy Department is to ensure America's security and prosperity by addressing its energy, environmental, and nuclear challenges through transformative science and technology solutions.
Digital Research Alliance of Canada (DRAC) - The Digital Research Alliance of Canada serves Canadian researchers by integrating, championing, and funding the infrastructure and activities required for advanced research computing (ARC), research data management (RDM), and research software (RS).
DKAN - A community-driven, free and open-source open data platform that gives organizations and individuals the ability to publish and consume structured information.
Dryad - Open-source method to enable the open publication and routine reuse of all research data vision.
e-IRG – e-Infrastructure Reflection Group - A strategic body to facilitate integration in the area of European e-Infrastructures and connected services, within and between member states, at the European level and globally.
Earth Science Information Partners (ESIP) - Created by NASA, ESIP supports the networking and data dissemination needs of our members and the global Earth science data community by

linking the functional sectors of observation, research, application, education and use of Earth science.
Economic Commission for Latin America and the Caribbean (ECLAC) - Headquartered in Santiago, Chile, ECLAC is one of the five regional commissions of the United Nations. It was founded with the purpose of contributing to the economic development of Latin America, coordinating actions directed towards this end, and reinforcing economic ties among countries and with other nations of the world.
European Data Infrastructure (EUDAT) - One of the largest infrastructures of integrated data services and resources supporting research in Europe.
European Open Science Cloud (EOSC) - A gateway to information and resources in EOSC, providing updates on its governance and players, the projects contributing to its realization, funding opportunities for EOSC stakeholders, relevant European and national policies, important documents, and recent developments.
European Strategy Forum on Research Infrastructures (ESFRI) - A strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach.
FAIRsharing.org - A community-driven resource with users and collaborators across all disciplines who work together to enable the FAIR Principles by promoting the value and the use of standards, databases and policies.
Fedora Commons - An open-source operating system.
Flatiron Institute - An internal research division of the Simons Foundation, the institute is a community of scientists who are working to use modern computational tools to advance science, both through the analysis of large, rich datasets and through the simulations of physical processes.
Future of Research Communications and e-Scholarship (FORCE11) - A community of scholars, librarians, archivists, publishers and research funders that aims to help facilitate the change toward improved knowledge creation and sharing.
Global Open Findable, Accessible, Interoperable and Reusable (GO FAIR) - A community working towards implementations of the FAIR Guiding Principles. This collective effort has resulted in a three-point framework that formulates the essential steps towards the end goal, a global Internet of FAIR Data and Services.
Harvard Dataverse - A free data repository open to all researchers from any discipline, both inside and outside of the Harvard community, where you can share, archive, cite, access, and explore research data.
Higher Education Leadership Initiative for Open Scholarship (HELIOS) - A cohort of colleges and universities committed to collective action to advance open scholarship within and across their campuses.
Integrated Global Greenhouse Gas Information System - An observation-based information system for determining trends and distributions of greenhouse gasses (GHGs) in the atmosphere and the ways in which they are consistent or not with efforts to reduce GHG emissions.
International Association of Scientific, Technical and Medical Publishers (STM) - An academic publisher.
International Bureau of Weights and Measures (BIPM) - An international organization established by the Metre Convention, through which Member States act together on matters related to measurement science and measurement standards.

International Council for Scientific and Technical Information (ICSTI) - A specialized intergovernmental organization established for ensuring the international exchange of scientific and technical information.
International Development Research Center (Canada) - A Canadian government project that funds research and innovation within and alongside developing regions to drive global change.
International Federation of Library Associations (IFLA) - An international organization that works to represent the interests of the librarian profession and improve services worldwide.
International Science Council (ISC) - Works at the global level to catalyze and convene scientific expertise, advice and influence on issues of major concern to both science and society.
Islandora - A foundation that maintains an extensible, modular, open-source digital repository ecosystem focused on collaborative authorship, management, display, and preservation of digital content at scale.
Kavli Foundation - A foundation that aims to advance science for the benefit of humanity.
Laura and John Arnold Foundation - A foundation that focuses its giving on evidence-based policy solutions.
Materials Genome Initiative - A federal multi-agency initiative for discovering, manufacturing, and deploying advanced materials twice as fast and at a fraction of the cost compared to traditional methods. The initiative creates policy, resources, and infrastructure to support U.S. institutions in the adoption of methods for accelerating materials development.
National Academies of Sciences, Engineering, and Medicine (NASEM) - A nonprofit that provides independent, objective advice to inform policy with evidence, spark progress, and drive innovation.
National Aeronautics and Space Administration (NASA) - The civil space program of the United States.
National Information Standards Organization (NISO) - A non-profit standards organization that develops, maintains, and publishes technical standards related to publishing, bibliographic, and library applications.
National Institute of Standards and Technology (NIST) - A United States federal agency whose mission is to promote innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve quality of life.
National Institutes of Health (NIH) - Part of the U.S. Department of Health and Human Services, NIH is the largest biomedical research agency in the world.
National Library of Medicine (NLM) - The world's largest biomedical library, NLM maintains and makes available a vast print collection and produces electronic information resources on a wide range of topics.
National Optical-Infrared Astronomy Research Laboratory (NOIRLab) - The United States national center for ground-based, nighttime optical astronomy.
National Science and Technology Council (NSTC) - A cabinet-level council of advisers to the President on science and technology including the Subcommittee on Open Science, formerly the Interagency Working Group on Open Science.
ORCID (Open Researcher and Contributor ID) - A global, not-for-profit organization providing a unique, persistent identifier for individuals to use as they engage in research, scholarship, and innovation activities.
Organization for Economic Co-operation and Development (OECD) - An international organization that works with governments, policy makers, and citizens, on establishing evidence-

based international standards and finding solutions to a range of social, economic, and environmental challenges.
Pub Med Central - A free digital repository run by the National Institutes of Health (NIH) that archives open access full-text scholarly articles that have been published in biomedical and life sciences journals.
re3data (Registry of Research Data Repositories) - A global registry of research data repositories from all academic disciplines.
Research Data Alliance (RDA) - Launched as a community-driven initiative in 2013 by the European Commission, the United States Government's National Science Foundation and National Institute of Standards and Technology, and the Australian Government's Department of Innovation, with the goal of building the social and technical infrastructure to enable open sharing and re-use of data.
São Paulo Research Foundation (Brazil) - A public foundation located in São Paulo, Brazil, with the aim of providing grants, funds, and programs to support research, education, and innovation of private and public institutions and companies in the state of São Paulo.
Scholarly Publishing and Academic Resources Coalition (SPARC) - A non-profit advocacy organization that supports systems for research and education that are open by default and equitable by design.
Society for Scholarly Publishing (SSP) - A nonprofit organization formed to promote and advance communication among all sectors of the scholarly publication community through networking, information dissemination, and facilitation of new developments in the field.
Wellcome Trust - A funder of research, policy, and advocacy campaigns, and of building global partnerships.
World Data System (WDS) - An affiliated body of the International Science Council (ISC) that aims to enhance the capabilities, impact and sustainability of member data repositories and data services.
Zenodo - An open repository developed under the European OpenAIRE program and operated by European Organization for Nuclear Research (CERN).

Appendix E. Change Log

In Spring of 2023, the following updates were made to the published RDaF version 1.0 as interim changes pending a full revision to produce and publish a final version 2.0:

- Expanded the topics and subtopics in the lifecycle stages which make up the framework core
- Added 14 Overarching Themes, that apply throughout the lifecycle stages
- Added eight generic Profiles, which identify only those topics and subtopics that are most relevant to specific job functions within the data ecosystem
- Added definitions for the lifecycle stages, topics, and subtopics
- Added Informative References that apply to each subtopic (The database included with version 2.0 will link the Informative References to the corresponding subtopics.)
- Added a methodology section describing how the framework was updated
- Added a section on future work.