Research Data Framework (RDaF):

Motivation, Development, and a Preliminary Framework Core

National Institute of Standards and Technology

Preliminary Release

December 1, 2020

Notes to Readers

This publication is the result of an ongoing collaborative effort involving industry, government agencies, universities, institutions, non-profits, and publishers. The National Institute of Standards and Technology (NIST) launched the Research Data Framework (RDaF) project by convening national and international private- and public-sector organizations and individuals in December 2019. This Preliminary Release of the RDaF was published in 2020 as a NIST Special Publication.

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

The research data environment is rapidly changing, and this Framework shall remain a living document. Revisions will be made as we, the stakeholders of the RDaF, gain experience with its application and use.

NIST acknowledges and thanks all of those who have contributed to this Preliminary Framework.

Table of Contents

List of Figures

List of Tables

Executive Summary

In the past decade, research data have become widely recognized as a critical national and global resource, and the risks of losing or mismanaging research data can have severe economic and social consequences. The proliferation of artificial intelligence approaches in all fields has created a huge demand for trustworthy research data in both the natural (e.g., chemistry) and social (e.g., economics) sciences. Further, research data drive innovation and growth in all civilian and military technologies and are essential for advances in human health and other societal concerns. The complexity of research data and the challenges of its management require an organizing framework adaptable to various disciplines, organizations, and job functions.

To address these issues, NIST initiated a program in fall 2019 called the Research Data Framework (RDaF). The overarching goal of the RDaF is to provide the stakeholder community with a structured approach to develop a customizable strategy for various roles in the research data ecosystem. Stakeholder organizations, both US and global, include industry, government agencies, universities, institutions, non-profits, publishers, and the general public. RDaF users, or individuals in an organization, vary from Chief Executive Officers and Chief Data Officers to librarians and researchers. The value of the RDaF to stakeholders and users is multifaceted, and addresses maximizing the value of research data assets, minimizing risks and costs, enabling discovery and innovation, and increasing the productivity and quality of research. In simple terms, the RDaF can be viewed as a map of the research data landscape that can be navigated by stakeholders and users according to their roles and needs.

A Scoping Workshop with 50 invited experts representing government, industry, academia, and other organizations and including international perspectives was held to gauge stakeholder support for the RDaF. The Workshop served to build community consensus on critical aspects of the RDaF and to propose a basic structure. A key recommendation from the Scoping Workshop was to adopt the structure of the widely used NIST Cybersecurity Framework [1] for the RDaF. Provided that funding can be secured, two pilot studies, one on Materials Science and the other involving Research Universities, involving the roles of libraries and publishers, will be carried out in order to evaluate the Preliminary RDaF structure. The first full version of the RDaF is targeted for release in summer 2022. Subsequent versions of the RDaF will be developed as additional disciplines and roles in the research data ecosystem are explored, contingent on funding and interest by the broader research community.

Creating the RDaF

Why: An increasingly complex research data ecosystem with voracious artificial intelligence drivers for trustworthy research data requires systematic, intentional research data asset management.

Purpose: To optimize use and value of strategic research data assets with a coherent research data management strategy.

Scope: Covers management of research data created and/or used by any organization

Status: Confirmed support by government agencies, universities, industry, scholarly publishers, institutions, professional societies, and international stakeholders.

Next Steps: Financial commitment for pilot studies, evaluation of the research data landscape, and community building activities.

1. Introduction

NIST is leading the development of the Research Data Framework (RDaF) with involvement and input from national and international leaders in the broad research data stakeholder community. Research data is defined here as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings [2]." The overarching goal of the RDaF is to provide the stakeholder community with a structured approach to develop a customizable strategy for various roles in the research data ecosystem. The audience for the RDaF is the entire research data community, including all organizations and individuals engaged in any activities concerned with research data management, from CEOs and CDOs to librarians and researchers. This document is organized into four sections: (1) Introductory material; (2) Development of the Preliminary RDaF; (3) A description of the Preliminary RDaF; and (4) Next steps.

1.1 Motivation

As we face the challenges of the 21st Century, research data have become a critical national and global resource, and the risks of losing and mismanaging research data can have severe economic and social consequences. With rapidly advancing information technologies, research data have become ubiquitous and are growing at astronomical rates. Europe and China have recognized this and have moved proactively in developing enterprise approaches to managing research data. China is working aggressively worldwide and has taken a dominant role in Africa. Europe has taken the leadership position in open research with FAIR (Findable, Accessible, Interoperable, Reusable) [3] data and is moving toward implementation with the European Open Science Cloud [4].

The US vies for having the fastest scientific research computers in the world and is applying them to artificial intelligence and data analytics in government and industry. The US continues to be a leader in data-intensive research in many disciplines, with enormously active private, academic, and public sectors. Many innovative projects are being conducted for managing research data in both the natural and social sciences.

There is an increasing variety of stakeholders in the research data ecosystem: government agencies, universities and their research libraries, data repositories, scholarly publishers, professional societies, national and international collaborations, organizations (e.g., CENDI[1], BRDI[2], NASEM[3], CODATA[4], RDA[5], WDS[6], and GO FAIR[7] (see Appendix A, Acronyms and Initialisms), standards bodies, funders (both public and private), industry and the private sector, researchers, and the general public. New job functions such as data stewards and data scientists are emerging, and skilled people are in short supply. How do the roles, responsibilities, and expectations of these diverse stakeholders differ, overlap, or contradict?

[1] https://www.cendi.gov
[2] https://www.nationalacademies.org/brdi/board-on-research-data-and-information
[3] https://www.nationalacademies.org/home
[4] https://codata.org/
[5] https://rd-alliance.org
[6] https://www.worlddatasystem.org
[7] https://www.go-fair.org

With all this, one thing is clear: *better national and international coordination is needed now for both basic and applied research data to ensure we stay competitive and think strategically about the management of such data, arguably one of our strategic national resources*.

1.2 Origin of the Framework

The concept of a Research Data Framework (RDaF) is inspired by the demonstrated success of the *Framework for Improving Critical Infrastructure Cybersecurity* [1], which NIST initially issued in February 2014, and which is hereafter referred to as the NIST Cybersecurity Framework.

The development of the RDaF started with a preliminary scoping study to determine the best approach to get support and uptake from a diverse stakeholder community. The RDaF will focus on the US, but by necessity will include global players and global best practices. Open and FAIR data are essential tenets in the Framework, but it supports the concept of "as open as possible as closed as necessary [5, 6]." The details of the Preliminary RDaF presented herein were informed by a small subset of the research data community; subsequent versions of the RDaF will be informed by the broader community.

The research data space is crowded with well-intentioned and often useful initiatives. However, these initiatives are not well-coordinated efforts focused on a multilateral, ecosystem basis. There are government agencies, of course (e.g., the OSTP/NSTC Subcommittee on Open Science[8] in the US), but also private funders, research data centers and repositories, tool and service providers, research libraries, professional associations and advocacy groups, universities, and the scholarly publishing community (both for- and non-profit). There are integrated efforts such as the Research Data Alliance and CODATA, and topical programs such as the Materials Genome Initiative[9], the Global Biodiversity Information Facility[10], and the BRAIN Initiative.[11] The RDaF will take advantage of this plethora of activities and organizations to facilitate better coordination and thus assure maximum return on the investment in research data infrastructure and interoperability tools. The RDaF will lay the groundwork for an infrastructure to ensure we think strategically about research data as a valuable global resource.

1.3 What is the RDaF?

The research data ecosystem is very complex! There are lots of players, various funding models and sustainability plans. How long should data be kept? How should data quality be assessed? How do we measure the value of research data? The RDaF strives to answer these questions by being:

- A map of the research data space: who, what, where, why, when?
- A dynamic guide for the various stakeholders in research data to understand best practices for research data management and dissemination.
- A resource for understanding costs, benefits, and risks associated with research data management.
- A consensus document based on inputs and conversations amongst the stakeholders in research data
- A tool that may be used to change the research data culture in an organization

[8] See https://www.whitehouse.gov/ostp/nstc, Committee on Science
[9] https://www.mgi.gov
[10] https://www.gbif.org
[11] https://braininitiative.nih.gov

1.4 Legal and Institutional Drivers

The RDaF provides organizations with a structured approach to develop a coherent research data strategy and will provide stakeholders with some common language terms[12] and a basis for coordination. NIST will lead the coordinated effort to develop and maintain a Framework that is useful but voluntary for all sectors of the economy–industry, government, academia, and not-for-profit organizations.

Just as the first version of the NIST Cybersecurity Framework was initially driven by legislation, namely Executive Order 13636: *Improving Critical Infrastructure Cybersecurity* [7], there are federal directives that support the development of the RDaF. These include a series of White House directives, with the most influential being *Increasing Access to the Results of Federally Funded Scientific Research* [8], also known as "the Holden memo," which was issued in February 2013. This memorandum was followed by another memorandum, *Open Data Policy-Managing Information as an Asset* [9] in May 2013, and by Executive Order 13642: *Making Open and Machine Readable the New Default for Government Information* (9 May 1013) [10]." On January 14, 2019, the President signed into law the *Foundations for Evidence-Based Policymaking Act of 2018* [11], which is based on the *OPEN Government Data Act*, House Resolution 1770 [12]. The above-mentioned legislation collectively dictates that US government agencies must make their data publicly available. Complying with these national requirements and considering the massive efforts in the open research/open data world, the US needs to assess and promote the best practices that are emerging in a diverse and complex global ecosystem of research data. The US also needs to coordinate its efforts within an international context. For example, the European Commission, through its European Open Science Cloud [4], aims to create a European research interoperability framework. The RDaF coordination office at NIST intends to keep abreast of these international efforts to achieve a consistent approach across the entire research data lifecycle.

1.5 Value Proposition

The immense value of managing research data is clearly supported by several federal documents. As stated in *Open Data Policy – Managing Information as an Asset* [9],

> *"Managing government information as an asset will increase operational efficiencies, reduce costs, improve services, support mission needs, safeguard personal information, and increase public access to valuable government information. Making information resources accessible, discoverable, and usable by the public can help fuel entrepreneurship, innovation, and scientific discovery—all of which improve Americans' lives and contribute significantly to job creation."*

From 2017 to 2019, the U.S. government released three key documents concerning Federal data: (1) *The Promise of Evidence-Based Policymaking*, which describes improvements on how data are used to generate evidence about policies and programs in the federal government [13]; (2) The *President's Management Agenda: Modernizing Government for the 21st Century*, which sets a priority goal of leveraging data as a strategic asset [14]; and (3) the *President's Management Agenda: Federal Data Strategy 2020 Action Plan* [15], which defines the steps to achieve this goal. The *Foundations for*

[12] Many of the language terms will be dependent on the specific research discipline.

Evidence-Based Policymaking Act of 2018 [11] stipulates the reporting structure for data management shown in the quotation below.

> *"[To] improve Federal data management…The head of each agency shall designate a nonpolitical appointee employee in the agency as the Chief Data Officer of the agency [who] shall be responsible for lifecycle data management…There is established in the Office of Management and Budget a Chief Data Officer Council that shall (1) establish Government wide best practices for the use, protection, dissemination, and generation of data; [and] (2) promote and encourage data sharing agreements between agencies."*

The specific value proposition for the RDaF includes the following benefits:

- **Research Integrity.** The RDaF will enable higher-quality, reproducible, and better-characterized research data, and transparency of the research process.
- **Costs and Efficiency.** The RDaF will aid in establishing and applying best practices to research data management to maximize efficiency and control costs.
- **Risk Management and Reduction.** While risk management and reduction practices are designed to decrease potential negative impacts, they may inadvertently result in missed opportunities. The RDaF will help organizations to assess their current risk positions and to create their own roadmap for improvement, including the management and reduction of risk in business decisions.
- **Scientific Discovery and Innovation.** Scientific discovery and innovation are critical to global competitiveness. The RDaF will embrace the FAIR principles, which promise to increase scientific productivity through better use and reuse of research data.
- **Policy Compliance.** The RDaF will assist organizations to be compliant with research data management and sharing policies from funding organizations and journals/publishers.

1.6 Risk Management

As stated in the NIST Privacy Framework [16], "risk management is a cross-organizational set of processes that helps organizations to understand how their systems, products, and services may create problems for individuals or the organization and how to develop effective solutions to manage such problems…risk assessments produce the information that can help organizations to weigh the benefits of data processing[13] against the risks and to determine the appropriate response—sometimes referred to as proportionality." Further, the NIST Privacy Framework demonstrates an application of risk management to data and privacy, whereby an organization "optimizes beneficial uses of data while minimizing adverse consequences for individuals' privacy and society as a whole [16]." Similar risk management and assessment processes may be applicable to research data in the RDaF.

[13]data processing is a collective set of data actions which include, but are not limited to, collection, retention, logging, generation, transformation, use, disclosure, sharing, transmission, and disposal.

1.7 Relationship to Other NIST Frameworks

As detailed in Section 2, a consensus decision was made to base the RDaF structure on that of the successful Cybersecurity Framework, which NIST initially issued in February 2014 to address the similarly emerging and complex global challenge of cybersecurity. Both the NIST Cybersecurity and the NIST Privacy Frameworks have three basic parts: a Framework Core, Framework Profiles, and Framework Implementation Tiers. In these two Frameworks, a Framework Core consists of four elements: Core Functions (activities), Categories and Subcategories (outcomes), and Informative References (e.g., standards, guidelines, and practices). Framework profiles represent the outcomes based on various factors that an organization has selected from the Categories and Subcategories. As described below in Section 3, the RDaF will adopt the same three basic parts, but with two important differences: (1) For Categories and Subcategories, topics replace outcomes; and (2) Framework profiles for an organization and/or specific role will be developed from the relevant Categories and Subcategories. Completed in October 2019, the nine-volume NIST Big Data Interoperability Framework [17] does not have the three basic parts of the other two completed NIST Frameworks. Future versions of the RDaF will draw upon the Big Data Interoperability Framework as appropriate.

Table 1 compares the Framework Core Functions of the NIST Cybersecurity and Privacy Frameworks with the Core Functions selected for the Preliminary RDaF (see Section 3). The intersections of these three frameworks are evident. For example, Plan, Detect, and Identify all relate to situational awareness; Process/Analyze, Respond, and Control all relate to operational aspects; and Preserve/Discard, Recover, and Protect all relate to final actions.

Table 1: Core Functions of the NIST Frameworks

RDaF	Cybersecurity	Privacy
Envision Plan Generate/Acquire Process/Analyze Share/Use/Reuse Preserve/Discard	Identify Protect Detect Respond Recover	Identify Govern Control Communicate Protect

2. Development of the Preliminary RDaF

Because a framework is only successful if it has buy-in and acceptance from the community, it is important to ensure that a wide range of voices are heard. For research data, the community includes business, academia, government, and other types of stakeholders. It involves roles and players that represent all stages of the research data lifecycle. As noted, the RDaF should be global in scope and reach because the nature and applications of research data are intended for broad adoption. The Preliminary RDaF development process is depicted in the timeline in Figure 1.

2.1 Initial Scoping Study

As a necessary first step, initial research was conducted to characterize the current research data landscape, including:

- Stakeholders and users (see Appendix B);
- Standards and tools already produced and in use;
- Maturity models and indicators (i.e., mechanisms to assess the extent of research data management in organizations); and
- Requirements and gaps in knowledge of best practices, including research data infrastructure.

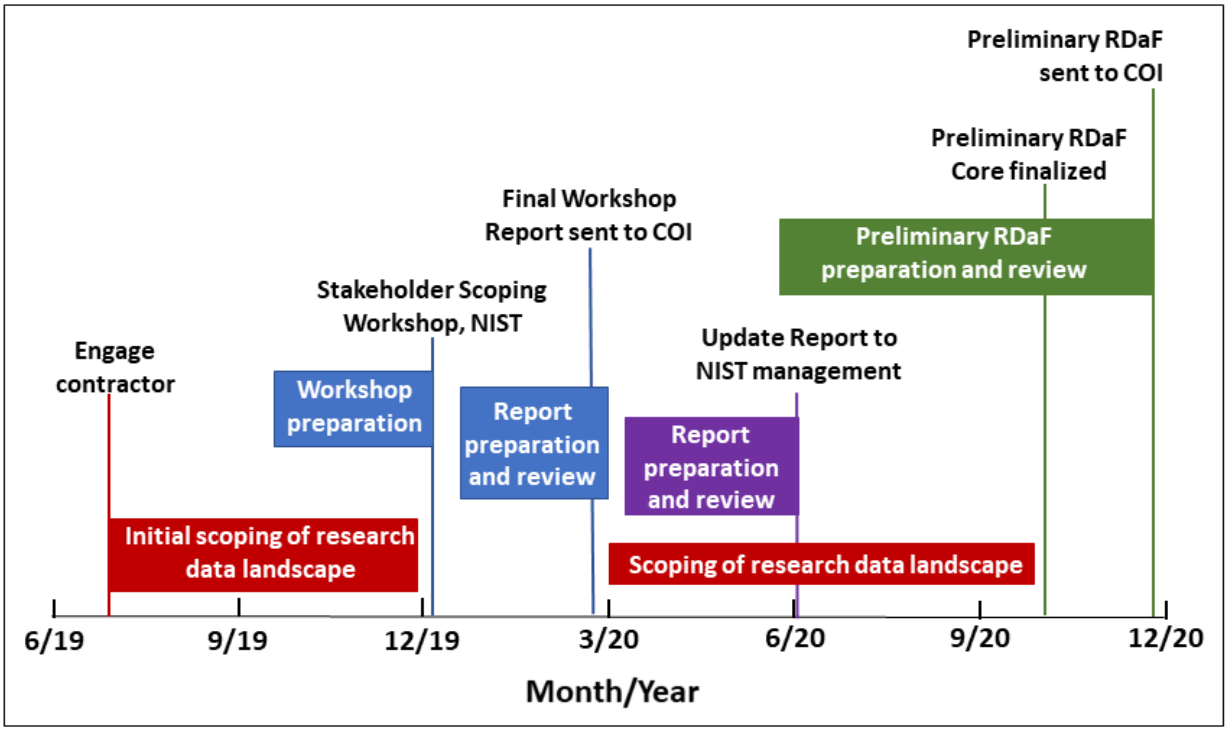

Fig. 1. Timeline for development of the Preliminary RDaF. Contractor: Bonnie Carroll (Information International Associates). COI: community of interest (i.e., the Workshop attendees and others who had expressed interest in following the progress of RDaF.)

The preliminary scoping study gauged stakeholder interest and determined the best approach to creating a framework that would have support from and adoption by a diverse stakeholder community. To this end, a Stakeholder Steering Committee consisting of eight individuals from different parts of the research data ecosystem was recruited to assist and advise in the development of the RDaF. The Steering Committee members are shown in Table 2.

Table 2: Stakeholder Steering Committee members.

Name[14]	**Organization**	**Sector**
Laura Biven	Department of Energy	Government
Mercè Crosas	Harvard University	Academia
Joshua Greenberg	Sloan Foundation	Funder, private foundation
Hilary Hanahoe	Research Data Alliance	International data organization
Heather Joseph	Scholarly Publishing and Academic Resources Coalition	A non-government advocacy organization, libraries
Barend Mons	Leiden Univ., CODATA, GO-FAIR	International data organization
Beth Plale	National Science Foundation	Government, funder
Anita de Waard	Elsevier	Scholarly publisher, private sector

[14] Note that Mark Leggott, Executive Director of Research Data Canada, was added to the Steering Committee in mid-2020.

2.2 Stakeholder Scoping Workshop

To determine the viability and true value of a Research Data Framework as perceived by the community, a Stakeholder Scoping Workshop was held on December 5-6, 2019 at the NIST National Cybersecurity Center of Excellence in Gaithersburg Maryland (see Agenda, Appendix C). The co-chairs of the workshop were Robert Hanisch from NIST and Bonnie Carroll from Information International Associates and CODATA. At the workshop, 51 invited attendees represented a broad spectrum of stakeholders encompassing a variety of job functions within the research data ecosystem. Participants included 19 people from six government agencies, 14 from academia and national laboratories, and seven from six industry segments. Five attendees came from four countries outside the US. See Appendix D for a list of Workshop attendees.

All participants actively and enthusiastically engaged in discussions, break-out sessions, and presentations. Workshop participants resonated with the structure of the NIST Cybersecurity Framework and recommended its basic structure for the RDaF. Two organizing concepts for the Core were considered at the workshop: a research data ecosystem approach or a lifecycle approach, perhaps including a top-level "sphere of responsibility." The lifecycle approach was selected. Each of four break-out groups proposed various lifecycle stages for the Co-Chairs and Steering Committee members to consider in their post-workshop deliberations.

There was consensus that the main target for the RDaF is at an institutional or organizational level such as a Chief Data Officer (CDO), i.e., someone with broad responsibilities for the management of research data across an organization. It was noted that the RDaF also has great value for other roles (i.e., job functions) in organizations such as researchers.

All participants were enthusiastic about remaining involved in the RDaF development and adoption. Since it was unanimously agreed that the RDaF should move forward, it was recommended that NIST move as rapidly as possible to solidify the plan and seek collaborative funding with other government agencies. Continued communication with workshop stakeholders and frequent consultation of the Steering Committee were strong recommendations. In summary, the workshop was effective in building the base for moving ahead and for soliciting support for the Framework's development.

In the following few months, the Workshop co-chairs drafted a report which was vetted by the Steering Committee. The report, *Initial Scoping Study for a NIST-Led Research Data Framework (RDaF)*, was distributed on March 5, 2020 to the RDaF "Community of Interest" (workshop attendees and others who have expressed interest in following the progress of the RDaF). The report contained an initial Framework Core with seven Functions (research data lifecycle stages) and 44 Categories and Subcategories (relevant topics for the seven Functions.)

2.3 Interim Studies and Reports

Following the workshop recommendations, two reports were generated. The first was a brief roadmap document, parts of which are incorporated the present report. The second was a briefing report for NIST upper management and included a budget for continuation of the RDaF project beyond the completion of the preliminary version presented herein. Scoping of the current research data landscape continued in the four bulleted areas given in Section 2.1 and was used to refine the initial Framework Core mentioned above.

2.4 Drafting the Preliminary RDaF

In the five months leading up to the release of this preliminary version of the RDaF, or Preliminary RDaF, the Framework Core was largely finalized with iterative discussions with the Steering Committee. The concepts of Framework Profiles and Framework Implementation Tiers, the latter in terms of data maturity models, were explored. Details of the Preliminary RDaF are presented in the following section. A draft report was vetted by the Steering Committee and released to the RDaF Community of Interest on October 26, 2020.

3. Description of the Preliminary RDaF

Like the NIST Cybersecurity and Privacy Frameworks, the full version of the RDaF will consist of three parts: Framework Core, Framework Profiles, and Framework Implementation Tiers. How each of these parts pertain to research data is described below. To date, collaborative development on the Preliminary RDaF has focused solely on the Framework Core. Descriptions of the Framework Profiles and Implementation Tiers are included only for illustrative purposes.

3.1 Framework Core

The relationship between the four different elements of the Framework Core—Functions, Categories, Subcategories, and Informative References—is shown in Figure 2 for a Core with three Functions. Definitions of the four elements in the context of research data for the RDaF are given below.

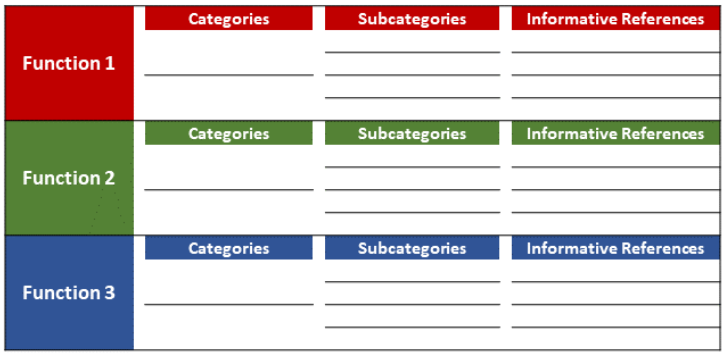

Fig. 2. Four elements of a Framework Core.

(1) **Functions** organize foundational research data-related activities at their highest level. As mentioned in Section 2, a lifecycle approach was selected as the organizing concept of the Framework Core.
(2) **Categories** are topics for a Core Function that are closely tied to programmatic needs and activities, as well as other important factors.
(3) **Subcategories** further divide a Category into more specific topics.
(4) **Informative References** are standards, guidelines, and practices associated with a Subcategory that provide the means to address a topic. Informative References will likely be a combination of resources that are common to all disciplines, organizations, and roles and resources that are specific to the disciplines or organizations, and roles to which the RDaF is being applied.

The Preliminary RDaF Framework Core is presented in Appendix E. The Core contains six Functions, which correspond to stages in the research data lifecycle (see Section 2.4), and Categories and Subcategories for each Function. (Informative References will be identified for the Subcategories in a future version of the RDaF.) The Functions are not intended to form a serial path or lead to a static desired end state. Rather, the Functions should be performed concurrently and continuously to form an operational culture that addresses the research data management needs. The Functions are defined below, as follows:

- ***Envision*** – This Function encompasses the review of the overall strategies and drivers of an organization's research data program. The Envision Function is where choices and decisions are made that together chart a high-level course of action to achieve desired organizational goals. The Categories within this Function are Data Governance Structure, Community Engagement, Data Culture, Reward Structure, Workforce/Career Paths, Data Safety and Security, Strategy, and Risk Management.
- ***Plan*** – This Function encompasses the tactical management positioning in an organization for effective research data management throughout the research data lifecycle. The Categories within this Function are Chain of Control, Economics and Costs in Planning, Funding Planning, Research Data Objects, Hardware/Software Infrastructure, Data Management Planning, Scientific Data Standards, and Assessment and Controls.
- ***Generate/Acquire*** – This Function 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. The Categories within this Function are Sources of Raw Data, Experimental Data Generation, Computational Data Generation, FAIR Principles for Data Generated In-House, External Sources of Data, and Community-Based Standards for Formats.
- ***Process/Analyze*** – This Function concerns the actions performed on generated or acquired research data to yield processed research data, typically using software, from which observations and conclusions can be made. This Function also concerns the data stewardship functions performed by an organization. The Categories within this Function are Data Provenance, Data Architecture, Software Tools, Scientific Workflow Processes and Systems, Data Inventory, Data Modeling and Analytics, Data Representation/Models/Structures, Data Curation, and Metadata.
- ***Share/Use/Reuse*** – This Function 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 an organization. The Categories within this Function are Legal and Licenses, Data Publishing, Data Citation, Internal and External Data Access, Levels of Protection, Applications and Analysis, and Data Architecture for Application and Use.
- ***Preserve/Discard*** – This Function delineates the end-of-use and end-of-life provisions for research data to complete and includes records management, archiving, and safe disposal. The Categories within this Function are Criteria, Data Sustainability, Storage and Preservation of Data, Moving Data from One Service to Another Across Organizations, and Retention and Disposition Schedules.

3.2 Informative References

Informative References are existing standards, guidelines, and practices relevant to a specific Subcategory. Informative References may also include laws, regulations, and other tools. Mappings of Informative References to Subcategories provide implementation support, e.g., help organizations determine which topics to prioritize to attain the desired state of research data management. A gap analysis of such mappings can also be used to identify where revised or additional standards, guidelines, and practices would help an organization to address emerging research data management needs. An

initial list of Informative References is given in Appendix F. These resources can support an organization's use of the RDaF to adopt better research data management practices.

3.3 Framework Profiles

Because the research data world is advancing so rapidly, there are new requirements for research data management as well as new research data-focused professional and managerial roles in many organizations, from high-ranking executives to technicians. Guidelines and checklists to ensure that research data management considerations in the various roles are fully characterized and addressed are now a critical need. The concept of Framework Profiles allows the RDaF to be tailored to different levels of stakeholders/users from a CEO to an individual researcher. To develop a Framework Profile, an organization can review all the Categories and Subcategories and determine which are relevant for an organizational unit and/or job function. Categories and Subcategories can be added as needed to fully adapt the RDaF to a specific stakeholder/user. Framework Profiles can be used to conduct self-assessments of research data management and communicate the results within an organization or between organizations. An example of Framework Profile development using a few Subcategories in the Envision Function/Data Governance Structure Category for various roles in an organization is provided in Table 3.

Table 3. Example of framework profile development.

			Roles				
Function	**Category**	**Subcategory**	CDO	Researcher	Librarian	Funder	Data Steward
ENVISION Review of the overall strategies and drivers of an organization's research data management program.	Data Governance Structure	Data vision and/or data policy	X				
		Legal and regulatory compliance	X				
		Data quality (including Trust and Certification)	X	X	X		X
		Data privacy	X	X	X		X
		Data management value proposition	X			X	X

3.4 Framework Implementation Tiers

Implementation Tiers are not addressed in the Preliminary RDaF but will be included in the next version. For the RDaF, Implementation Tiers will allow an organization to assess its current state of research data management and to develop a roadmap to attain its desired state of research data management. Implementation Tiers can support an organization's decisions regarding research data management and help prioritize areas that would benefit from additional resources. For the RDaF, Implementation Tiers will be described in terms of data maturity, which can be defined as "the extent to which an organisation utilises the data they produce [18]" and "a measurement of the ability of an organization for continuous improvement in [data management] [19]. Maturity indicators, which are mechanisms to assess the extent of research data management in an organization, can be used to determine data maturity.

There are several well-known data management/governance maturity models such as DAMA-DMBOK2 [20], Data Management Capability Assessment Model (DCAM) [21], Data Management Maturity Model [22], IBM Data Governance Council Maturity Model [23], Stanford Data Governance Maturity Model [24], and Gartner's Enterprise Information Management Maturity Model [23]. Maturity models define the fundamental processes of data management and specific capabilities and actions that constitute a path to improvements in data maturity.

4. Next Steps

The objective of the next phase in the development of the RDaF is to test the applicability and usefulness of the Framework Core in Appendix E. To accomplish this objective, two concurrent pilot studies—one in Materials Science and the other in Research Universities, including librarian and publisher roles—will be conducted. A timeline for the next phase is presented in Figure 3. Because continuation of the RDaF effort is contingent on the availability of funding, the timeline begins with month 0. Prior to month zero, funding has been secured and support staff have been identified.

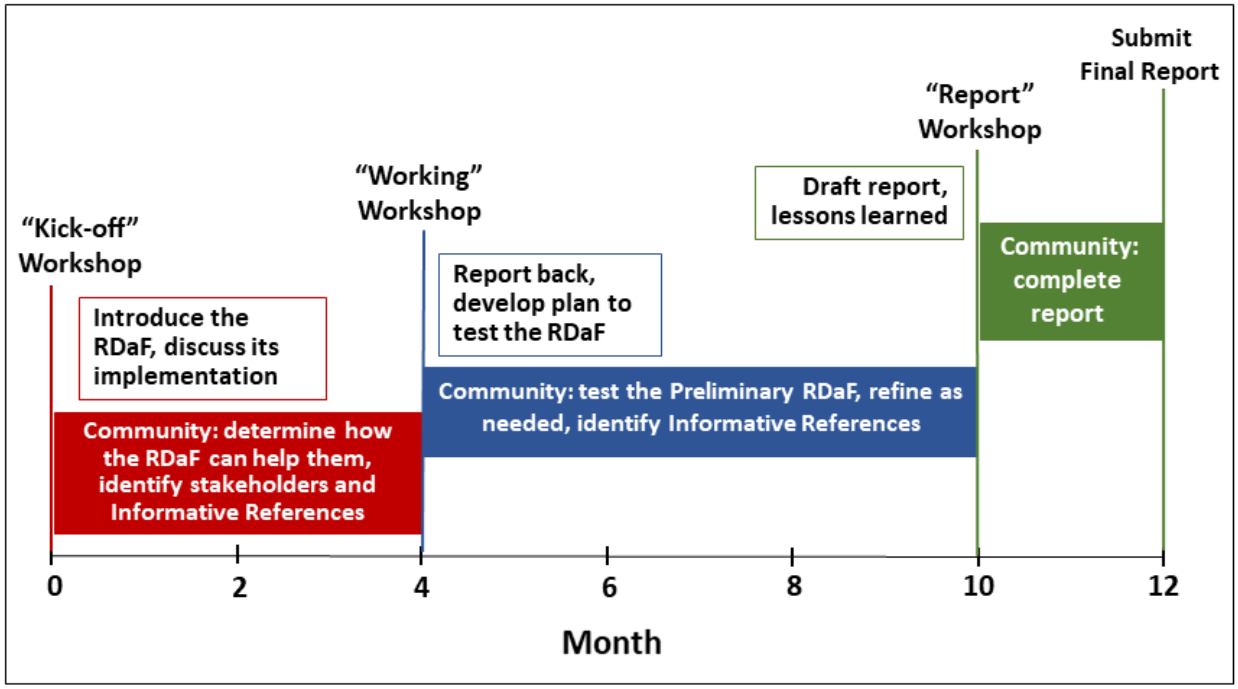

Fig. 3. Timeline for each pilot study.

Each pilot study will have three workshops:

(1) "Kick-Off" Workshop: Attendees will be introduced to the Preliminary RDaF. Implementation of the RDaF will be discussed.
Homework: *Community discusses how the RDaF can help them with research data management and identifies their stakeholders and Informative References for the RDaF Core.*

(2) "Working" Workshop: Attendees will report their findings and plan how to apply the RDaF.
Homework*: Community tests the Preliminary RDaF and identifies refinements to it and Informative References for the RDaF Core.*

(3) "Report" Workshop: Attendees will draft a report on the pilot study findings and discuss lessons learned.
Homework: *Community completes their report.*

The Steering Committee will review the two Pilot Study reports and revise the RDaF as needed. The next version of the RDaF will be released within six months of completion of the Pilot Study reports.

References

[1] National Institute of Standards and Technology (2018) *Framework for Improving Critical Infrastructure Cybersecurity*, version 1.1 (U.S. Department of Commerce, Washington, D.C.). https://doi.org/10.6028/NIST.CSWP.04162018

[2] Intangible property, 2 CFR § 200.315 (2014). Available at https://www.govinfo.gov/app/details/CFR-2014-title2-vol1/CFR-2014-title2-vol1-sec200-315

[3] Wilkinson MD et al. (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* **3**:160018. https://doi.org/10.1038/sdata.2016.18

[4] European Union (2020) *European Open Science Cloud*. Available at https://web.archive.org/web/20201120103536/https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[5] Collins S et al. (2018) *Turning FAIR into Reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data* (European Commission, Brussels, EU). Available at https://web.archive.org/web/20201120103833/https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_0.pdf

[6] Landi A et al (2020) The 'A' of FAIR – As Open as Possible, as Closed as Necessary. *Data Intelligence* 2(1-2), 47-55. https://doi.org/10.1162/dint_a_00027

[7] United States, Executive Office of the President Barack Obama (12 Feb. 2013) Executive Order 13636: *Improving Critical Infrastructure Cybersecurity*. *Federal* Re*gister*, vol. 78, no. 33, 19 Feb. 2013, pp. 11739-11744. Available at https://web.archive.org/web/20201120104102/https://www.govinfo.gov/content/pkg/FR-2013-02-19/pdf/2013-03915.pdf

[8] Holdren JP (2013, Feb. 22) *Increasing Access to the Results of Federally Funded Scientific Research* [Memorandum] Executive Office of the President, Office of Science and Technology Policy. Available at https://web.archive.org/web/20201120104318/https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[9] Burwell SM, VanRoekel S, Park T, Mancini DJ (2013, May 9) *Open Data Policy-Managing Information as an Asset* [Memorandum] Executive Office of the President, Office of Management and Budget. Available at https://web.archive.org/web/20201120104818/https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

[10] United States, Executive Office of the President Barack Obama (12 Feb. 2013) Executive Order 13642: *Making Open and Machine Readable the New Default for Government Information. Federal Register, vol. 78, no. 93, 14 May 2013, pp. 28111-28113.* Available at https://web.archive.org/web/20201120105015/https://www.govinfo.gov/content/pkg/FR-2013-05-14/pdf/2013-11533.pdf

[11] Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. 115-435 §101 132 Stat. 5529 (2019). Available at https://web.archive.org/web/20201120105202/https://www.govinfo.gov/content/pkg/PLAW-115publ435/pdf/PLAW-115publ435.pdf

[12] OPEN Government Data Act, H.R. 1770, 115th Congress (2017-2018). Available at https://www.congress.gov/115/bills/hr1770/BILLS-115hr1770ih.pdf

[13] Commission on Evidence-Based Policymaking (2017*) The Promise of Evidence-Based Policymaking*. Available at

https://web.archive.org/web/20201120105528/https://www.cep.gov/report/cep-final-report.pdf

[14] White House (2018) The President's Management Agenda: Modernizing Government for the 21st Century. Available at https://web.archive.org/web/20201120105737/https://www.whitehouse.gov/wp-content/uploads/2018/04/ThePresidentsManagementAgenda.pdf

[15] White House (2019) The President's Management Agenda: Federal Data Strategy 2020 Action Plan. Available at https://web.archive.org/web/20201120105948/https://strategy.data.gov/assets/docs/2020-federal-data-strategy-action-plan.pdf

[16] National Institute of Standards and Technology (2020) *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management*, version 1.0 (U.S. Department of Commerce, Washington, D.C.). Available at https://web.archive.org/web/20201120110331/https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf

[17] National Institute of Standards and Technology (2019) *NIST Big Data Interoperability Framework*, V3.0 Final Version (U.S. Department of Commerce, Washington, D.C.). Available at https://web.archive.org/web/20201120112757/https://bigdatawg.nist.gov/V3_output_docs.php

[18] Data Orchard (2020) What is data maturity? Available at https://web.archive.org/web/20201120113157/https://www.dataorchard.org.uk/what-is-data-maturity

[19] Steenbeek I (2020 February 10) *Data Management Maturity 101: What is a data management maturity assessment and why does a company need it?* Data Crossroads. Available at https://web.archive.org/web/20201120113359/https://datacrossroads.nl/2020/02/10/data-management-maturity-101-what-is-a-data-management-maturity-assessment-and-why-does-a-company-need-it

[20] DAMA International (2017) *DAMA-DMBOK Data Management Body of Knowledge* (Technics Publishing, Basking Ridge, New Jersey), 2nd Ed.

[21] Enterprise Data Management Council (2018) *Data Management Capability Assessment Model (DCAM) Overview.* Available at *https://web.archive.org/web/20201120130740/https://cdn.ymaws.com/edmcouncil.org/resource/resmgr/featured_documents/EDMC_DCAM_Overview.pdf*

[22] CMMI Institute (2020) *Data Management Maturity (DMM) Model At-A-Glance*. Available at https://web.archive.org/web/20201120142150/https://cmmiinstitute.com/getattachment/cb35800b-720f-4afe-93bf-86ccefb1fb17/attachment.aspx

[23] Taylor K (2020 August 6) *Data Governance Maturity Models Explained*. Available at https://www.hitechnectar.com/blogs/data-governance-maturity-models-explained/

[24] Firican G (2018 August 29) Stanford data governance maturity model. Available at https://web.archive.org/web/20201120132142/https://www.lightsondata.com/data-governance-maturity-models-stanford/

Appendix A: Acronyms and Initialisms

AAU	Association of American Universities
AGU	American Geophysical Union
AI	Artificial Intelligence
ANDS	Australian National Data Service
APARD	Accelerating Public Access to Research Data
API	Application programming interface
APLU	Association of Public and Land-grant Universities
BRAIN Initiative	Brain Research through Advancing Innovative Neurotechnologies® Initiative
BRDI	Board on Research Data and Information
CDO	Chief Data Officer
CENDI	Commerce, Energy, NASA, Defense Information Managers Group
CEO	Chief Executive Officer
CMMI	Capability Maturity Model Integration
CNRI	Center for National Research Initiatives
CODATA	Committee on Data of the International Science Council
DAMA	Data Management Association International
DANS	Data Archiving and Networked Services
DCAM	Data Management Capability Assessment Model
DMBOK	Data Management Body of Knowledge
DMM	Data Management Maturity
DOI	Digital Object Identifier
e-IRG	e-Infrastructure Reflection Group
ESFRI	European Strategy Forum on Research Infrastructures
ESIP	Earth Science Information Partners
EUDAT	European Data Infrastructure
FAIR	Findable, Accessible, Interoperable and Reusable
GEIA	Government Electronics and Information Technology Association
GO FAIR	Global Open Findable, Accessible, Interoperable and Reusable
HPC	High-Performance Computing
HR	Human Resources
ICSTI	International Council for Scientific and Technical Information
IFLA	International Federation of Library Associations
ML	Machine Learning
NASEM	National Academies of Sciences, Engineering, and Medicine
NASA	National Aeronautics and Space Administration
NECTAR	Network for Effective Collaboration Technologies Through Advanced Research
NFAIS	National Federation of Advanced Information Services (now merged with NISO)
NISO	National Information Standards Organization
NIST	National Institute of Standards and Technology

NSTC	National Science and Technology Council
OPEN	Open, Public, Electronic, and Necessary
ORCID	Open Researcher and Contributor ID
OSTP	Office of Science and Technology Policy
RDA	Research Data Alliance
RDaF	Research Data Framework
SPARC	Scholarly Publishing and Academic Resources Coalition
SSP	Society of Scholarly Publishing
STM	International Association of Scientific, Technical and Medical Publishers
WDS	World Data System

Appendix B: Initial List of Stakeholders and Users

Private Funders

- Laura and John Arnold Foundation
- Alfred P. Sloan Foundation
- Bill & Melinda Gates Foundation
- Kavli
- Flat Iron
- Belmont Forum
- Helmsley Charitable Trust
- Wellcome Trust

Data Centers

- World Data System (WDS) and its members (particularly US member centers)

Repositories

- re3data
- DataONE
- Figshare
- Datacite
- Dryad

Tool Providers

- DataCite
- Open Researcher and Contributor ID (ORCID)

Library and Not-for-Profit Organizations

- California Digital Library - DMPTool
- National Information Standards Organization (NISO)
- Association for Research Libraries
- Scholarly Publishing and Academic Resources Coalition (SPARC)
- Center for Open Science
- Center for National Research Initiatives (CNRI)
- International Federation of Library Associations (IFLA)

University Organizations

- Association of American Medical Colleges
- American Association of Universities (AAU)
- Association of Public and Land-grant Universities (APLU)

Publishing Community

- Elsevier
- Nature
- Springer
- Society for Scholarly Publishing (SSP)
- Coalition for Publishing Data in the Earth and Space Sciences
- International Association of Scientific, Technical and Medical Publishers (STM)

Data Organizations

- Committee on Data of the International Science Council (CODATA)
    - NASEM: US National Committee for CODATA, associated with the NASEM Board on Research Data and Information (BRDI)
- Research Data Alliance (RDA)
    - Special focus on RDA-US
- Earth Science Information Partners (ESIP)
- esri, formerly Environmental Systems Research Institute
- International Council for Scientific and Technical Information (ICSTI)
- National Federation of Advanced Information Systems (NFAIS), now merged with NISO
- FORCE11: Future of Research Communication and e-Scholarship
- Commerce, Energy, NASA, Defense Information Managers Group (CENDI)
- Global Open Findable, Accessible, Interoperable and Reusable (GO FAIR)
- World Data System (WDS)

Disciplinary/Topical Initiatives

- Materials Genome Initiative
- Integrated Global Greenhouse Gas Information System
- Biodiversity Global Information Facility
- American Geophysical Union (AGU)
- Accelerating Public Access to Research Data (APARD)

Federal Agencies and Programs

- National Institute of Standards and Technology
- Department of Energy
- National Optical Astronomy Observatory
- National Aeronautics and Space Administration
- National Institutes of Health
- National Library of Medicine

Policy/Studies Organizations

- National Science and Technology Council (NSTC)
    - Interagency Working Group on Open Science
- National Academies of Sciences, Engineering, and Medicine (NASEM)
    - Board on Research Data and Information

International Agencies and Programs

- Organization for Economic Co-operation and Development
- European Data Infrastructure (EUDAT)
- European Open Science Cloud
- International Science Council
- International Bureau of Weights and Measures
- e-IRG – e-infrastructure reflection group (ongoing activity on research data and infrastructure)

Foreign Governments and National Organizations

- Australian Research Data Commons (a merger between ANDS, NECTAR, and Research Data Services, Australia)
- Commonwealth Scientific and Industrial Research Organisation, Australia

- CANAIRE - Advancing Canada's knowledge and innovation infrastructure
- Data Archiving and Networked Services (DANS)
- Academy of Science of South Africa (South Africa)
- International Development Research Center (Canada)
- Economic Commission for Latin America and the Caribbean
- São Paulo Research Foundation (Brazil)
- European Strategy Forum on Research Infrastructures (ESFRI)

Appendix C: RDaF Stakeholder Scoping Workshop Agenda

AGENDA

DAY 1: Dec. 5 ***"Why an RDaF?"***

Time	Topic	Comments
8:30-9:00	**CONTINENTAL BREAKFAST/NETWORKING**	
9:00-9:30	Welcome	**Walter Copan (**NIST Director) **Jim St. Pierre** (ITL Deputy Director; Setting the Framework context at NIST)
9:30-9:45	***Setting the stage***	**Bob Hanisch** *(Director, Office of Data and Informatics, MML, NIST)* What are we trying to achieve at this workshop? What is different about the RDaF from the other NIST Frameworks? Who do we see as stakeholders and beneficiaries? What do we all want to see in Framework and why? How shall we proceed?
9:45-10:30	***Understanding the Ecosystem***: - Stakeholder perspectives	This will be flash talks (5-6 minutes each) from different categories of stakeholders to tee up break-outs. Present who players are; how to engage them; why are they there; what's their role; sustainability/sustainability model. Six minutes each; no slides.
10:30-10:45	**BREAK**	
10:45-11:45	Break out: ***Users and beneficiaries?***	What would be the value of an RDaF Why would you would want a Framework. How could it be of value? (Ideas like RDaF's are expensive. Do we really know that? Realistic estimate of costs – could we explore that? Costs of not having it?) Explore users and beneficiaries; these maybe not the exactly the same.
11:45-12:30	Group Presentations and Discussion	RESULTS: 1) Who are users & beneficiaries; 2) Use cases; 3) What stakeholder groups aren't here that should be included?
12:30-1:30	**LUNCH with table discussions on morning topics**	
1:30-2:15	***Understanding Life Cycle:*** - Stakeholder Perspectives	This will be 5 flash talks (6 minutes) from stakeholders who represent different functions in the data life cycle to tee up break-outs. Objective is to understand how the Framework might address layers or functions in the life cycle.
2:15-3:15	Break-out: ***Applications***	Taking a life cycle perspective What else do you want from an RDaF and identify additional use cases.
3:15-3:30	**BREAK**	
3:30-4:15	Group Presentations and Discussion	RESULTS: 1) What is wanted from a RDaF? 2) Additional User Cases for a Framework
4:15-5:00	***RDaF Contents***	Brief Presentations on what might be in an RDaF to tee up the morning break-outs.
6:00-	Informal Networking/Group Dinners	Informal gathering at the open Bar in the Courtyard by Marriott Rockville 2500 Research Blvd, Rockville, MD 20850. Groups can self-organize for dinner at local restaurants.

DAY 2: Dec. 6 -- *"What should be in it and how to develop it?"*

Time	Topic	Comments
8:30	**GET YOUR COFFEE**	
8:40-9:00	Recap of Day 1	We've gained an understanding of many perspectives and how it can be used and some use cases
9:00-10:00	Break out: ***Content, Components, and Elements*** -What should be in an RDaF?	Pick 2-3 use cases to help guide thinking, but don't be restricted to them. Based on what we've heard and the use cases what should be in an RDaF? What components need to be included to come out supporting use cases and to deliver benefits?
10:00-10:45	Report out & discussion – Developing RDaF content.	Begin to develop a conceptual outline of an RDaF.
10:45-11:00	**BREAK**	
11:00-11:45	Break-out: ***Process.*** -How should we go about it?	How should we proceed from here to complete the scoping study? (Provide the proposed steps?) Assuming a positive outcome of the scoping, what should be the next steps? What is proposed: 1) full scoping 2)Pilot in two disciplines 3)Full development? Who should be involved?
11:45-12:30	Group Presentations and Discussion	RESULTS: Each group should present 1) a path to take the workshop results and complete the scoping. Once the initial scoping is done, 2) what should be the steps to a final Framework and what would it take to get there.
12:30-1:00	Summary and wrap up	What have we heard and where will we go from here? What is the interest in continued involvement of participants?
1:00-3:00	Steering Group Meeting	The Steering Group will convene for a working lunch.

FLASH TALK PRESENTERS

Stakeholder perspectives

Time	Topic	Speakers
9:45-10:30	***Understanding the Ecosystem:*** -	Government: **Laura Biven**, Department of Energy Academia: **Sayeed Choudhury**, Johns Hopkins University Industry: **Vivien Bonazzi**, Deloitte International: **Jean-François Abramatik**, European Open Science Cloud Community: **Hilary Hanahoe**, Research Data Alliance Chief Data Officer: **Ed Kearns**, Department of Commerce
1:30-2:15	***Understanding Life Cycle:***	Funder: **Beth Plale**, National Science Foundation Researcher: **Barend Mons**, Leiden University Medical Center & CODATA Publisher: **Shelley Stall**, American Geophysical Union Infrastructure: **Mark Leggott**, Research Data Canada Library/Archive: **Leah McEvan**, Cornell University
4:15-5:00	***RDaF Contents***	**Anita deWaard**, Elsevier **Merce Crosas**, Harvard **Susan Gregurick**, National Institutes of Health

Appendix D: RDaF Stakeholder Scoping Workshop Attendees

Last Name	First Name	Organization
Abramatic	Jean-François	National Institute for Research in Computer Science and Automation, France
Agarwal	Deborah	Lawrence Berkeley National Laboratory
Allard	Suzanne	University of Tennessee, Knoxville
Ananthakrishnan	Rachana	University of Chicago
Ang	James	Pacific Northwest National Laboratory
Biven	Laura	Department of Energy, Office of Science
Bonazzi	Vivien	Deloitte
Bruce	Elizabeth	Microsoft
Carroll	Bonnie	Information International Associates and CODATA
Choudhury	Golam	Johns Hopkins University
Cragin	Melissa	University of California, San Diego
Crosas	Mercé	Harvard University
Dahlitz	Karen	Australia
de Waard	Anita	Elsevier
Dreisigmeyer	David	US Census Bureau
Erdmann	Christopher	University of North Carolina, Renaissance Computing Institute
Fagnan	Kirsten	Lawrence Berkeley National Laboratory
Federer	Lisa	National Library of Medicine
Govoni	Marco	Argonne National Laboratory
Gregurick	Susan	National Institutes of Health
Hanahoe	Hilary	Research Data Alliance, Italy
Hanisch	Robert	NIST Material Measurement Laboratory
Hanson	Brooks	American Geophysical Union
Honaker	James	Center for Research on Computation and Society
Hudson-Vitale	Cynthia	Association of Research Libraries
Johnston	Lisa	University of Minnesota
Kahn	Scott	LunaDNA
Kaiser	Debra	NIST Material Measurement Laboratory
Kearns	Edward	Department of Commerce
Kitney	Stuart	National Physical Laboratory
Leggott	Mark	Research Data Canada
Lucas	Matthew	Social Sciences and Humanities Research Council of Canada
McEwen	Leah	Cornell University
Medina-Smith	Andrea	NIST Information Services Office
Mons	Barend	CODATA, GO-FAIR, Leiden University
Musen	Mark	Stanford University
Nichols	Lisa	Office of Science and Technology Policy
Plale	Beth	National Science Foundation
Pollard	Tom	Massachusetts Institute of Technology / PhysioNet

Pouchard	Line	Brookhaven National Laboratory
Ricci	James	Department of Energy, Advanced Scientific Computing Research
Robinson	Carly	US Department of Energy, Office of Scientific and Technical Information
Schlenoff	Craig	NIST Program Coordination Office
Sellars	Scott	Department of State
Shyam Sunder	Sivaraj	NIST Acting Chief Data Officer
Stall	Shelley	American Geophysical Union
Strawn	George	National Academies
Uhlir	Paul	Self-employed
Vanderwall	Dana	Bristol-Myers Squibb (Allotrope Foundation)
Woo	Kara	Sage Bionetworks

Appendix E: Preliminary RDaF Framework Core

Notes:
(1) In the Categories and Subcategories, the use of "data" means "research data;"
(2) Bolded words indicate input from the Stakeholder Scoping Workshop; and
(3) A * at the end of a word or group of words indicates that a definition is provided in Appendix G.

FUNCTION (Data Lifecycle Stage)	CATEGORY	SUBCATEGORY
ENVISION Review of the overall strategies and drivers of an organization's research data program.	Data **Governance*** Structure	• Identification of Goals and Roles • Data vision and/or **data policy** • Data management value proposition • Data management organization • Value of data (quantitative or qualitative) • Legal and regulatory **compliance** • **Data quality** (including Trust and Certification) • **Data privacy** • Data ethics
	Community Engagement	• **Stakeholder community(ies)** • Communication with stakeholder community(ies) • Interactions with other organizations • **Cross-community engagement** (across domains and sectors) • Inclusivity in interactions
	Data Culture*	• **FAIR data principles** • **Value of data** • Roles and responsibilities
	Reward Structure	• For data management • Value of **data workers** • **Incentives and institutional credit for data sharing** and reuse • Disincentives for data sharing • Human Resources (HR) involvement
	Workforce/Career Paths	• Workforce skills inventory • HR's role in data **workforce development** • Data management **training** • Workforce preparedness in **new and advancing technologies,** e.g., HPC, AI, ML, and computation services • Promotional paths, continual training, and career development
	Data **Safety and Security**	• Safety and security assurance • Data inventory
	Strategy	• Organizational data management
	Data **Risk** Management*	• Risk assessment • Risk mitigation and management

FUNCTION (Data Lifecycle Stage)	CATEGORY	SUBCATEGORY
PLAN The tactical management positioning in an organization for effective research data management throughout the research data lifecycle.	Chain of Control	• Documentation
	Economics and Costs in Planning	• Decision-making tools on data, including cost-benefit analysis • Cost breakdown, i.e., calculation of costs by data lifecycle stage
	Funding Planning	• Models for **provisioning resources**, i.e., direct, overhead, or mixed
	Data Objects	• Data • Software • Instruments • Publications • Presentations • Other
	Hardware/Software Infrastructure	• Research data
	Data Management Planning	• **Data management plans** (DMPs) • Lifecycle: DMPs as living documents or static proposals
	Scientific Data Standards	• Sources of standards
	Assessment and Controls	• Goals/definition of success • **Metrics** or metrics structure, tracking use and impact measures
GENERATE/ACQUIRE The generation of raw research data and/or the acquisition* of research data by an organization.	Sources of Raw Data*	• Generated In-house experimentally or computationally • Collected from external sources
	Experimental Data Generation	• Specification and recording of instruments and associated metadata • Description and recording of measurement protocols • Methods for data and metadata capture and recording
	Computational Data Generation	• Commercial and/or custom software • Methods for computational variables (metadata) capture and recording
	FAIR Principles for Data Generated In-House	• Data born FAIR
	External Sources of Data	• Data acquired FAIR • **Identification, collection, and recording** • **Metadata harvesting**
	Community-Based Standards for Formats	• Standards development organizations/sources

FUNCTION (Data Lifecycle Stage)	CATEGORY	SUBCATEGORY
PROCESS/ANALYZE The actions performed on generated or acquired research data to yield processed research data, and the research data stewardship* functions performed by an organization.	Data **Provenance**	• Original authoritative copy • Version identification • Provenance of data derived from other data • Provenance of scientific records across all the individual outputs • **Timestamping**
	Data Architecture	• Design • Security • Configuration management • Hosting and storage • Use of cloud
	Software Tools	• Data lifecycle* • Management and analysis • Commercial and/or custom tools • System resilience and adaptability • Maintenance
	Scientific **Workflow** Processes and Systems	• Workflow tools • Laboratory notebooks, i.e., electronic, paper
	Data Inventory	• Formats and standards • Catalogs • Interoperability (across instrument manufacturer file formats)
	Data Modeling and Analytics	• Processes • Tools
	Data Representation, Models, Structures	• Dynamic data
	Data Curation	• Policies and processes • Manpower
	Metadata	• Types of metadata • Responsible parties • Specification of metadata standards • **Linked data** structure • Persistent identification (DOI)

FUNCTION (Data Lifecycle Stage)	CATEGORY	SUBCATEGORY
SHARE/USE/REUSE How research data are disseminated, used, and reused within and outside an organization.	Legal and Licenses	• Ownership of data • Constraints and encouragement for data use • **Intellectual property** rights/restrictions • **Usage agreements/terms/licenses** and required permissions • Terms of service • Data sharing agreements and licensing • Data citation*
	Data Publishing*	• Repositories and referencing data/digital objects from journal articles • Supplementary material • Data linking
	Data Citation*	• Citation Impact
	Internal and External Data Access	• Access internally, e.g., the data generator • Access externally • **Programmatic access**, aka Smart API • Data access vs. data visiting
	Levels of Protection	• Unclassified but sensitive information, e.g., de-identification, enclaves • Security classification • Protecting limited data/secure platforms/enclaves • Data anonymization*
	Applications and Analysis	• **Technologies** for use and analytics, e.g., AI, ML
	Data Architecture for **Application and Use**	• Extensibility across communities, including machine-based interactions • Capturing insights from ML and use of these to improve datasets for future AI applications • Capturing data performance characteristics • Location of data (e.g., relative to instruments or HPC machines, novel computing architectures, in data lakes, in the cloud, transient copies)
PRESERVE/ DISCARD The end-of-use and end-of-life of research data in an organization, including records management and archiving.	Criteria	• Use and impact
	Data Sustainability	• Data longevity and support • Orphan datasets
	Storage and Preservation of Data	• Media to store and preserve data • Data back-up • Data repositories
	Moving Data from One Service to Another across Organizations	• Roles and responsibilities • Moving data from one agency to another, e.g., from a funded support to long-term preservation space • Registration of repositories – roles and responsibilities • Disciplinary archives

FUNCTION (Data Lifecycle Stage)	CATEGORY	SUBCATEGORY

	Retention and Disposition Schedules	• Data **archiving**, i.e., **what is kept and not kept**
		  • Decision processes
		• **End-of-life** issues
		  • Example: Responsible party for keeping raw data* feeds
		  • Example: Store (or not) raw data*, given the large amount of storage needed
		• Deaccessioning/End-of-life
		• Recognition of removed data (gravestone)

Appendix F: Initial List of Informative References

NIST Frameworks

- Cybersecurity Framework Version 1.1 https://www.nist.gov/cyberframework/framework
- Privacy Framework Version 1.0 https://www.nist.gov/privacy-framework/privacy-framework
- Big Data Interoperability Framework: Volume 1, Big Data Definitions [Version 2] https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-1-big-data-definitions-version-2

Other Frameworks to Consider

- Australian National Data Service
    - Creating a Data Management Framework https://www.ands.org.au/guides/creating-a-data-management-framework and https://www.ands.org.au/__data/assets/pdf_file/0005/737276/Creatinga-data-management-framework.pdf
    - An overview of what elements institutions need to consider when planning for an institutional approach to data management. It also has an in-depth analysis of the Capability Maturity Model which can be used to develop an institutional Data Management Framework:
        - **Five elements of data management capability**: Policies and procedures, IT infrastructure; support services, managing metadata, managing research data
        - **Assessed across 5 levels of maturity**: initial, development, defined, managed, optimized
- DAMA (Data Management Association International)
    - Data Management Body of Knowledge Book 2 (DMBOK2) https://www.dama.org/content/what-data-management
    - DAMA-DMBOK2 Framework https://www.datasqlvisionary.com/wp-content/uploads/2018/06/DMBOK-Framework.pdf
- NISO
    - Research Data Management https://groups.niso.org/apps/group_public/download.php/15375/PrimerRDM-2015-0727.pdf
- CMMI Institute
    - Data Management Maturity (DMM) Model https://web.archive.org/web/20201120142150/https://cmmiinstitute.com/getattachment/cb35800b-720f-4afe-93bf-86ccefb1fb17/attachment.aspx

Guidelines and Initiatives

- AAU-APLU Public Access Working Group Report and Recommendations, November 29, 2017, fhttps://www.aau.edu/sites/deault/files/AAU-Files/Key-Issues/Intellectual-Property/Public-Open-Access/AAU-APLU-Public-Access-Working-Group-Report.pdf
- Redd K SteenK., Nusser, S, Smith, T., Walters, T, Chasen, J., Luther, J., and Reecy, J. (2019). Accelerating Public Access to Research Data Workshop. Washington, District of Columbia: Joint

publication by the Association of Public and Land-grant Universities and Association of American Universities. DOI: 10.31219/osf.io/63mxh. https://osf.io/63mxh/

- Accelerating Public Access to Research Data Summit: https://www.aau.edu/national-summit-accelerating-public-access-research-data

Articles, Reports, and Presentations

- Data Management maturity models: a comparative analysis https://datacrossroads.nl/2018/12/16/data-management-maturity-models-a-comparative-analysis/
- Proença D., Borbinha J. (2018) Maturity Models for Data and Information Management. https://www.researchgate.net/publication/327431346_Maturity_Models_for_Data_and_Information_Management
- CMMI Data Management Maturity Model Introduction https://cdn.ymaws.com/www.globalaea.org/resource/collection/68814379-BF7E-41C8-B152-18A617F9C0AA/Data_Management_Maturity_Model_Introduction_-_Dec_12_2014.pdf
- FAIR Data Maturity Model: specifications and guidelines https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines-0
- Data Management Capability Assessment Model (DCAM) Overview https://cdn.ymaws.com/edmcouncil.org/resource/resmgr/featured_documents/EDMC_DCAM_Overview.pdf
- DCAM Working Draft https://dgpo.org/wp-content/uploads/2016/06/EDMC_DCAM_-_WORKING_DRAFT_VERSION_0.7.pdf
- GEIA859A, Data Management Standard, SAE International https://www.sae.org/standards/content/geia859a

Appendix G: Glossary of Terms Used in Appendix E

Data Acquisition The process of acquiring data from some source. For example, data may be acquired by download from a repository, transfer from a data logger, data capture, etc.[15]

Data Anonymization Data anonymization is a type of information sanitization whose intent is privacy protection. It is the process of removing personally identifiable information from data sets so that the people whom the data describe remain anonymous. Data anonymization may enable the transfer of information across a boundary, such as between two departments within an agency or between two agencies while reducing the risk of unintended disclosure, and in certain environments in a manner that enables evaluation and analytics post-anonymization.[16]

Data Citation Data citation is the provision of accurate, consistent, and standardized referencing for datasets just as bibliographic citations are provided for other published sources like research articles or monographs. Typically, the well-established Digital Object Identifier (DOI) approach is used with DOIs taking users to a website that contains the metadata on the dataset and the dataset itself.[17]

Data Culture Data culture is the principle established in the process of social practice in both public and private sectors which requires all staffs and decision-makers to focus on the information conveyed by the existing data and make decisions and changes according to these results instead of leading the development of the company based on experience in the particular field.[18]

Data Governance The policies, procedures, and processes to manage and monitor the organization's regulatory, legal, risk, environmental, and organizational requirements are understood and inform the management of [data] risk.[19]

Data Lifecycle Refers to all the stages in the existence of digital information from creation to destruction. A lifecycle view is used to enable active management of the data objects and resource over time, thus maintaining accessibility and usability.[20]

[15] https://casrai.org/term/data-acquisition
[16] https://en.wikipedia.org/wiki/Data_anonymization
[17] https://en.wikipedia.org/wiki/Data_citation
[18] https://en.wikipedia.org/wiki/Data_culture
[19] Definition of governance taken from the NIST Cybersecurity Framework [1], with the words "data risk" replacing "cybersecurity risk"
[20] https://casrai.org/term/data-lifecycle/

Data Publication The release of research data, associated metadata, accompanying documentation, and software code (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to uniquely and persistently. Data publishing occurs via dedicated data repositories and/or (data) journals which ensure that the published research objects are well documented, curated, archived for the long term, interoperable, citable, quality-assured, and discoverable – all aspects of data publishing that are important for future reuse of data by third-party end-users.[21]

Data Stewardship Data Stewardship is "The most common label to describe accountability and responsibility for data and processes that ensure effective control and use of data assets. Stewardship can be formalized through job titles and descriptions, or it can be a less formal function driven by people trying to help an organization get value from its data."[22]

Raw Data Data that have not been processed for meaningful use. Although raw data have the potential to become "information," they require selective extraction, organization, and sometimes analysis and formatting for presentation. As a result of processing, raw data sometimes end up in a database, which enables the data to become accessible for further processing and analysis in several different ways.[23]

Risk Management Risk management refers to the practice of identifying potential risks in advance, analyzing them and taking precautionary steps to reduce/curb the risk.[24] Data carries tremendous value for organizations while creating new challenges around transparency, accuracy, security, privacy, social expectations, and legal requirements.[25]

[21] https://casrai.org/term/data-publication
[22] Reference [20], pp. 75-76.
[23] https://casrai.org/term/raw-data
[24] Definition of Risk Management (2020 November 20) *The Economic Times.* https://economictimes.indiatimes.com/definition/risk-management
[25] Albinson N, Thomas C, Rohrig M, Chu Y (2019) Future of risk in the digital era, *Deliotte*. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/finance/us-rfa-future-of-risk-in-the-digital-era-report.pdf