FRAMEWORK FOR CONSIDERING DATA INFRASTRUCTURE AND INTERCONNECTIVITY IN AND AMONG RESEARCH AND DEVELOPMENT INFRASTRUCTURE PROJECTS

A Report by the

SUBCOMMITTEE ON RESEARCH AND DEVELOPMENT INFRASTRUCTURE

of the

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

December 2024

About the National Science and Technology Council

The National Science and Technology Council (NSTC) is the principal means by which the executive branch coordinates science and technology policy across the diverse entities that make up the federal research and development enterprise. A primary objective of the NSTC is to ensure science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares research and development strategies that are coordinated across federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available at http://www.whitehouse.gov/ostp/nstc.

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, homeland security, health, foreign relations, the environment, and the technological recovery and use of resources, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of federal research and development in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the federal government. More information is available at http://www.whitehouse.gov/ostp.

About the NSTC Subcommittee on Research and Development Infrastructure

The Subcommittee on Research and Development Infrastructure (RDI) coordinates federal investments in infrastructure supporting research and development (R&D) across the nation. This coordination ensures that federal R&D infrastructure and the scientific and engineering workforce it supports remain preeminent, relevant, and ready to address the nation’s economic and national security priorities.

About this Document

This report offers a high-level framework for considering data infrastructure and interconnectivity during planning, developing, operating, assessing, and upgrading Research and Development Infrastructures (RDIs). Prevalent challenge areas are also identified. The report aims to be a practical resource for practitioners and federal program managers and to inform partnerships and collaborations on RDI data and infrastructure. Opportunities are identified for collective action to address challenges and disseminate practices.

Disclaimer

References in this document to any specific commercial products, publications, processes, services, manufacturers, companies, trademarks, or other proprietary information are intended to provide clarity and do not constitute an endorsement or recommendation by the U.S. government.

Copyright Information

NATIONAL SCIENCE and TECHNOLOGY COUNCIL

Chair

Arati Prabhakar, Assistant to the President for Science and Technology; Director, OSTP

Executive Director (Acting)

Lisa E. Friedersdorf, OSTP

SUBCOMMITTEE ON RESEARCH AND DEVELOPMENT INFRASTRUCTURE

Co-Chairs

Ann Schwartz, Office of Science and Technology Policy

Jagadeesh Pamulapati, Department of Defense

Harriet Kung, Department of Energy

Linnea Avallone, U.S. National Science Foundation

Executive Secretary

Mariam Elsayed, Department of Energy

Members

Aliya Iftikhar, Office of Science and Technology Policy
Gordon Hengst, Department of Defense
Kartik Sheth, National Aeronautics and Space Administration
Scott Miller, Smithsonian Institution
Alexander Kurien, General Services Administration
Paul Strang, Department of Homeland Security
Joon Park, Department of Agriculture
Mike Villarreal, Department of Agriculture
Geoff Plumlee, U.S. Geological Survey
Dan Hayba, U.S. Geological Survey
Esha Mathew, State Department
Jason Boehm, National Institute for Standards and Technology
John Cortinas, National Oceanic and Atmospheric Administration
Terence Lynch, National Oceanic and Atmospheric Administration
Angela Hadley, Environmental Protection Agency
Elisabeth McClure, Office of Management and Budget

DATA INFRASTRUCTURE WORKING GROUP

Co-Chairs

Michael Cooke, Department of Energy

Bill Miller, U.S. National Science Foundation

Members

Ann Vega, Environmental Protection Agency
Kevin Murphy, National Aeronautics and Space Administration
Andrew Mitchell, National Aeronautics and Space Administration
Elena Steponaitis, National Aeronautics and Space Administration
J.L Galache, National Aeronautics and Space Administration
Robert Hanisch, National Institute for Standards and Technology
Jon O'Neil, National Oceanic and Atmospheric Administration
Lisa Kewley, Smithsonian Institution
Kari Haworth, Smithsonian Institution
Margaret Benoit, U.S. National Science Foundation
Susan Gregurick, National Institutes of Health

Acknowledgements

OSTP and the NSTC thank the following individuals for providing feedback during the development of this report: Laura Biven, Ben Brown, Michelle Buchannan, Ewa Deelman, Jim Fowler, Carol Hawk, Tony LaVoi, Anirban Mandal, Manil Maskey, Lavanya Ramakrishnan, Roland Roberts, Charlotte Roehm, and Scott Thompson. Thanks also go to Scott Miller, liaison to the Interagency Working Group on Scientific Collections, and to the Institute for Defense Analysis, Science and Technology Policy Institute team—Walter Valdivia, Kush Patel, Skandan Ananthasekar, and Matthew Diasio—for their literature review and technical support.

Table of Contents

Abbreviations and Acronyms v

Executive Summary vi

I. Introduction and Aims 1

Definitions 2

II. Framework: Principal Considerations for Data Infrastructure Design and Interconnectivity 4

Science goals and mission priorities for data 4
Users and utilization 5
Data inventory, management, and stewardship 5
Dynamic data ecosystem 6
Project governance and partnering 7
Suggestions for putting the framework into practice 7

III. Cross-cutting Challenge Areas 9

Transferring and managing large-scale data 9
Data integration and data infrastructure interoperability 9
Operating in the commercial cloud 10
Handling sensitive and secure data 10
Workforce development and nurturing 11

IV. Recommendations for Collective Action 12

Dissemination and exchange of practices 12
Coordination 12
Workforce 13

V. References 14

Abbreviations and Acronyms

API	Application Programming Interfaces
AI	Artificial Intelligence
DIWG	Data Infrastructure Working Group
DOE	Department of Energy
DOI	Digital Object Identifier
FAIR	Findable, Accessible, Interoperable, and Reusable
G7	The Group of Seven
G20	Group of 20
GRI	Global Research Infrastructure
GSO	G7 Group of Senior Officials on Global Research Infrastructure
IWGODSP	NSTC Interagency Working Group on Open Data Sharing Policy
ML	Machine Learning
NASA	National Aeronautics and Space Administration
NBER	National Bureau of Economic Research
NIST	National Institute of Standards and Technology
NITRD	Networking and Information Technology Research and Development Program
NSORDI	NSTC RDI National Strategic Overview report
NSF	U.S. National Science Foundation
NSTC	National Science and Technology Council
OECD	The Organization for Economic Co-operation and Development
OSTP	Office of Science and Technology Policy
R&D	Research and Development
RDaF	NIST Research Data Framework
RDI	Research and Development Infrastructure
SDLM	Science Data Lifecycle Model

Executive Summary

Data infrastructure is a central enabler of many large-scale national and international projects and collaborations. Exponential growth in scale and complexity of research data from major experimental and observational research facilities is occurring in tandem with rapid advancements in computing and data technologies that have made it easier to access, interlink, interact with, and analyze research data. Major federal initiatives such as the National Artificial Intelligence Research Resource and National Strategic Computing Reserve, as well as federal goals for open science and public access, necessitate new levels of coordination on scientific data and infrastructure. Research and Development Infrastructure (RDI) data infrastructure practitioners, RDI managers, and federal sponsors thus face challenges when planning, deploying, and operating data systems and services that can keep pace with evolving scientific needs and technology advancements. There are many variables to address, and since RDIs are typically developed independently of one another, the resulting data solutions can be quite different from one another—which can hinder broader interoperability.

This report aims to raise awareness of the importance of comprehensive planning for RDI data infrastructure. This report offers an actionable approach in the form of common framework with key questions that practitioners, managers, and sponsors should ask when planning, developing, operating, and upgrading data infrastructure and interconnectivity (see summary box below). The report further identifies current cross-cutting challenge areas for RDI data infrastructure that would benefit from collective federal and community discussion and action. An ultimate aim is to facilitate partnering and collaboration on sharing and interoperability of RDI data and infrastructure towards maximizing overall scientific impact and benefit to the U.S. scientific enterprise.

Summary of data infrastructure planning framework presented in this report

Science goals and mission priorities for data

- What scientific priorities and objectives drive the data infrastructure investment choices?
- What overall data policies, community norms, and standards will be adopted?
- What is the vision for success related to data, and what quantifiable criteria and metrics will enable evaluation and assessment?

Users and utilization

- Who are the primary users to be served by the data infrastructure to meet the science and mission goals?
- Which science utilization models and workflows will be supported?
- What ways need to be supported for direct user interaction with the data?

Data inventory, management, and stewardship

- What is the data inventory for this effort?
- How will data management be implemented across the data lifecycle?
- How will the data infrastructure implement data governance and stewardship needs?

Dynamic data ecosystem

- How can the data infrastructure be designed to maximize interoperability?
- How will forward-looking development factors be integrated into all stages of the RDI lifecycle?
- Are there processes in place to enable both staff and users to successfully adapt to the changing data resources and services?

Project governance and partnering

- Who are the data stakeholders and how will they be represented in decision-making about the data?
- How will the governance and partnership policies and financial support mechanisms ensure sustainability and resilience of the project?
- In a collaboration or partnership, is there agreement on how the joint effort will be conducted and overseen and how the respective roles and responsibilities apportioned?

I. Introduction and Aims

Research and development infrastructure (RDI) projects and collaborations are large-scale endeavors where data acquisition, curation, sharing and distribution, analysis, and/or archiving are typically in central focus. These data-intensive missions are expanding as new modes of data-driven interactive research—such as automated experimental steering and real-time analysis—and the explosive growth of artificial intelligence (AI) and machine learning (ML) techniques have opened new opportunities for novel science with RDI data. Furthermore, RDIs increasingly seek to interconnect and share data to enable multidisciplinary and geographically distributed collaborations [1]. At the national level, initiatives such as the National Artificial Intelligence Research Resource [2] and the National Strategic Computing Reserve [3], and federal imperatives on open science and public access [4] necessitate radically new levels of coordination and interoperability across experimental and observational facilities, data repositories, and computing resources [5].

RDIs are consequently under continuous pressure to leverage growing data in new and complex ways, and to implement interconnectivity broadly to maximize scientific return in a sustainable way. However, the rapidly evolving landscape of scientific needs for data and connectivity can be difficult to address for new RDI projects and may strain or exceed the capabilities of existing RDIs. Technology solutions are also diverse and change quickly relative to typical RDI development timelines. Plans for data sharing by and among RDIs can encounter roadblocks in many forms, for instance, due to incompatibilities in the formats of the data or metadata, complexity in the data types, as well as governance and use restrictions. Changing user requirements and utilization modes also creates challenges to providing support over time.

Overall, there are a great many variables to address when planning or upgrading RDI data infrastructure and interconnectivity. Since RDIs are typically developed independently of one another, their data solutions are also typically quite different which can hinder broader interoperability. In short, there is a need to foster compatible science-enterprise-level approaches to planning RDI data infrastructure and interconnectivity on common principles and practices to maximize the overall scientific impact and benefit of RDI data.

To address this need, in 2023, the National Science and Technology Council (NSTC) RDI Subcommittee charged a Data Infrastructure Working Group (DIWG) to develop a common high-level framework for considering data infrastructure and interconnectivity during planning, developing, operating, assessing, and upgrading RDIs. This resulting report has the following primary aims:

- Raise awareness of the importance of comprehensive planning for data infrastructure that aligns with sharing and interoperability objectives;
- Offer an actionable approach to inform and assist efforts to share and integrate different types of data and information from and among RDIs; and
- Facilitate and inform interagency and international discussions on sharing and interoperability of both open and protected data generated by U.S. RDIs to successfully achieve shared goals with partners.

The primary intended audience of this report is the community of RDI data infrastructure practitioners, RDI planners and developers, and federal program managers. The report is mainly aimed at a practical level, offering both a planning framework and identification of current challenge areas encountered by RDI data practitioners.

Ultimately, the RDI Subcommittee and DIWG intend this report to advance national goals of maintaining a strong, integrated, and agile research and development enterprise as identified in the NSTC RDI National Strategic Overview report (NSORDI) [1], emphasizing interoperability, openness, transparency, and user-centric approaches. Recommendations for collective action on data infrastructure planning and practice are offered towards achieving these ends.

Definitions

Research and development infrastructure (RDI) is an inclusive term defined in the NSORDI 2021 report as "facilities or systems used by scientific and technical communities to conduct research and development or foster innovation," and comprising three major categories:

- experimental and observational infrastructure;
- knowledge infrastructure (e.g., shared scientific data assets and resources, such as scientific collections, repositories and archives, and related expertise); and
- research cyberinfrastructure (i.e., research computing, data, and networking infrastructure).

RDIs are typically characterized by long planning, implementation and operational life cycles; are focused on well-defined research objectives (or domains); and involve sustained federal support and, possibly, support from other partners. U.S. federal agencies also use other terms for RDIs, such as user facilities, missions, major or mid-scale facilities, and research infrastructure projects and programs. While most RDIs are located and operated domestically, others are part of international partnerships and collaborations (sometimes also termed "global research infrastructures" [6]).

Data infrastructure broadly refers to the array of data systems and services that together enable an RDI's data objectives and supported data lifecycle (see latter definition below). This infrastructure includes both the data-focused elements within an RDI as well as any external independent data resources that an RDI may engage with in a dependent way. Typical examples include:

- User-facing data tools, platforms, and services which are typically used for searching, connecting, accessing, handling, processing, and analyzing.
- Data management and curation systems and protocols, including physical archiving and storage systems, local networking/transfer systems, and curation technologies and protocols (such as digital identifiers for data).
- Software and middleware for operating data infrastructure and executing data workflows.
- Data access and cybersecurity technologies, policies and protocols, including those used for identity management and user authentication.
- Interoperability, sharing and integration resources and services, including interface protocols and standards for user and machine access, as well as semantic services such as ontology, taxonomy, and controlled vocabulary services.

Other large-scale systems such as high-speed networking systems, repositories, workflow management systems, and large-scale computing resources may also provide relevant data resources and services such as caching and storage.

Lifecycles is a term used in two distinct ways in this report, the RDI project lifecycle and the research data lifecycle:

- **RDI lifecycle**: As defined in the NSORDI 2021 report, the RDI lifecycle comprises stages of development, establishment, operation and maintenance, modernization, and repurposing or decommissioning. Different agencies may use alternative terms for each of these stages – for instance, "planning" and/or "design" for development, "construction" or "implementation" for establishment, etc.
- **Research data lifecycle** is the term used to represent the ensemble stages of scientific data that data infrastructure enables and supports, such as acquisition/collection, processing and generation of data products, management, analysis, archiving and curation, delivery and sharing, and deaccession. There are many examples of relevant reference data lifecycles (e.g., [7, 8, 9]).

II. Framework: Principal Considerations for Data Infrastructure Design and Interconnectivity

The following framework identifies and addresses five principal areas for considering data infrastructure and interconnectivity throughout the phases of the RDI lifecycle:

- Science goals and mission priorities for data;
- Users and utilization;
- Data inventory, management, and stewardship;
- Dynamic data ecosystem; and,
- Project governance and partnering.

Each topical area comprises key questions that practitioners, managers, and sponsors should ask to formulate requirements. The framework may be considered a checklist or a roadmap to inform the planning, developing, operating, upgrading and of assessing of data infrastructure and interconnectivity.

The framework is meant to serve as a starting point; additional considerations may be required to meet the specific needs of individual RDI projects or partnerships and collaborations among RDIs. It should also be recognized that the various framework areas are interconnected and decisions in one area may impact considerations in other areas and inform holistic cost/benefit decisions.

Science goals and mission priorities for data

RDIs are typically developed to achieve specific scientific or mission goals, which in turn motivate subsequent decisions about how to structure and manage the data aspects of the project to meet those goals efficiently, effectively, and sustainably.

- What **scientific priorities and objectives** drive the data infrastructure investment choices?
  These priorities and objectives may include central science or mission goals for the RDI or collaboration, user-driven objectives (see **Users and utilization**), and other high-level objectives and requirements.
- What **overall data policies, community norms, and standards** will be adopted?
  Motivating factors may include ensuring data security for sensitive data, promoting interoperability, and maximizing open and equitable access to and use of shared data [10, 11]. Policy examples may include adhering to open science and Findable, Accessible, Interoperable, Reusable (FAIR) [12] data principles and other emerging data regimes.
- What is the **vision for success related to data**, and what quantifiable criteria and metrics will enable evaluation and assessment?

Users and utilization

A user- and utilization-centered design approach is a fundamental driver for successful planning, operations, and upgrading of data infrastructure.

- Who are the **primary users** to be served by the data infrastructure to meet the science and mission goals?
  The primary user base is typically associated with fulfilling the primary goals and objectives of the RDI project or partnership.
    - Are there different priorities in serving various other types or classes of users, such as to support open science goals?
- Which **science utilization models and workflows** will be supported?
  Examples of utilization models and workflows include enabling integration of data from multiple sources; working with time-sensitive and streaming data; facilitating data management, storage, and curation; connecting data to computing and other resources; and supporting *in situ* or remote data processing and analysis (see **Data inventory, management, and stewardship**).
    - What are the exemplar or priority utilization cases and reference workflows that need to be enabled for users to accomplish their research objectives?
- What ways need to be supported for **direct user interaction with the data**?
    - How will data usability requirements be defined?
    - What tools, capabilities, and support need to be provided to users for interactive activities such as data search and discovery, downloading, and *in situ* examination, integration, analysis, and visualization?
    - What kinds of data access do users need to accomplish their research objectives?
    - What types of user support and engagement will be provided, such as documentation, training, assistance, and troubleshooting?
    - Will there be opportunities for users to provide feedback?

Data inventory, management, and stewardship

A thorough and holistic analysis of which data will be supported, the intended usage of the data, and the data governance and stewardship considerations all must be carried out to inform the planning, design, and implementation of data infrastructure.

- What is the **data inventory** for this effort?
    - What are the characteristics of each data source and data product, such as location, size/volume, metadata, identifiers, formats and standards, and static or dynamic nature?
    - What data usage rules and restrictions exist or need to be defined for the data, and how are they determined?
      These rules and restrictions may relate, for example, to scientific, national and economic security, privacy, intellectual property protection, and licensing considerations; ownership, proprietary, legal, liability, and regulatory regimes; and policies, community norms, and standards for access and sharing.

  - How is the data inventory anticipated to evolve over time?
  - Is it important to provide curated or authoritative data sets from trusted sources with fully documented provenance?
- How will **data management** be implemented across the data lifecycle?
  - What data lifecycle model(s) will be utilized to guide the design and operations?
  - What activities will be undertaken or supported in each lifecycle stage and what data management capabilities are required?
  - What systems, services, protocols, and processes will be required for data access, use, and security?
  - What are the data availability requirements?
  - What are the requirements and plans for retention/archiving, preservation, and curation of generated and derived data products after RDI or collaboration end-of-life, and how will these decisions be made?
- How will the data infrastructure implement **data governance and stewardship** needs?
  - How will the required data models, data structures, semantic systems, metadata and other standards, and the control of data quality and integrity be supported by the infrastructure?
  - How will data risks (e.g., privacy, security, integrity protection) be managed?
  - How will compliance with operant policies and standards be implemented and assured?
  - What documentation related to data stewardship is required?

Dynamic data ecosystem

Science objectives for RDI data, and enabling technologies for data and their use, continuously evolve. Proactive consideration of this continuous change during design can ensure that the resulting data infrastructure and interoperability approach are flexible, adaptable, and enable continued access to data.

- How can the data infrastructure be designed to **maximize interoperability**?
  - How does the design process take into consideration integrative approaches to achieve the project's objective?
    Potential models for integrative approaches may include interoperability, integration, and federation with other public and private data infrastructure, systems, and resources.
  - What interfaces and standards should be supported to realize interoperability and streamline future integration with other resources?
- How will **forward-looking development factors** be integrated into all stages of the RDI lifecycle?
  - What processes will be in place to gather anticipated future needs and requirements for the data infrastructure?
    Requirements may be informed, for example, by science goals, user needs, community best practices, anticipated partnerships, and evolving technology.

  - On what cadence will review, assessment, and planning for infrastructure updates and upgrades take place?

- Are there processes in place to enable both staff and users to **successfully adapt** to the changing data resources and services while maintaining necessary operations?
  - What education, training and outreach approaches will be employed to assist users in benefitting from the changes and enhancements?
  - What workforce planning, development, and retraining efforts are necessary to support the evolving data infrastructure technologies and approaches?

Project governance and partnering

Governance and partnering considerations for data infrastructure on RDI projects and collaborations include policy, legal, funding, and oversight mechanisms that ensure engagement of key stakeholders and long-term sustainability and resilience of data infrastructure [13].

- Who are the **data stakeholders** and how will they be represented in decision-making about the data and vision for its use?
- How will the governance and partnership policies and financial support mechanisms ensure **sustainability and resilience** of the project?
  Example considerations may include funding and cost management of systems, operations, and the workforce necessary to support the data infrastructure.
- In a collaboration or partnership, is there agreement on how the joint effort will be conducted and overseen and how the respective **roles and responsibilities** will be apportioned?
  - How will all elements of the framework be jointly or separately addressed?
  - Is there sufficient common understanding regarding the data and data infrastructure (for instance, on terminology, semantics, data structures, metadata schemas, governing laws, standards) and is any alignment/translation effort needed to support the joint effort?
  - How will the joint effort be regularly evaluated and assessed to ensure the desired performance?
  - What governance processes and decision mechanisms are necessary to address differences that may arise among partners regarding the data and data infrastructure?

Suggestions for putting the framework into practice

Following the above framework can inform the development of a concise set of requirements for overall design, operations, and performance of the data infrastructure. Supporting approaches and actions towards that end include:

- Conduct formal requirements gathering and identify reference use cases to drive the design of the data infrastructure and interconnectivity.
- Ensure sufficient effort is placed on cost/benefit and analysis of alternatives during planning; in particular, considering adoption/adaptation of existing solutions versus developing new ones.

- Periodically review current data infrastructure capabilities against future requirements in conjunction with science reviews and continuous assessment of evolving user needs, approaches, and solutions.
- In a collaboration or partnership on data infrastructure and interconnectivity, ensure sufficient time is given in the planning phase to finalize needed formal agreements, particularly those involving sensitive data (see **Handling sensitive and secure data** in Section III).
- Thoroughly assess the workforce needs to support all aspects of the data and RDI lifecycles. Ensuring a sufficient, well-qualified workforce is critical to all areas of the framework and can represent one of the most significant operational costs (see **Workforce development and nurturing** in Section III).

III. Cross-cutting Challenge Areas

In developing this report, the DIWG identified several prevalent challenge areas often faced when developing and upgrading RDI data infrastructure:

- Transferring and managing large-scale data;
- Data integration and data infrastructure interoperability;
- Operating in the commercial cloud;
- Handling sensitive and secure data; and,
- Workforce development and nurturing.

Key open issues are highlighted for each of the challenge areas, which cut across technology, data management and delivery, and socio-technical considerations. These areas and issues would benefit from further concerted cross-agency and community discussion, experience exchange, and action, and thus informed the recommendations in Section IV. Additionally, readers are referred to the RDI Subcommittee report, "U.S. Federal Research and Development Infrastructure," which highlights diverse challenge areas for RDIs [14], and to the NITRD Big Data Strategy update [15] which identifies areas of the scientific data enterprise that need to be addressed strategically.

Transferring and managing large-scale data

Distributed RDIs, large-scale domestic and international collaborative efforts, and complex disciplinary and interdisciplinary research projects increasingly rely on large-scale data transfer and integration from multiple sources to accomplish science objectives and/or facilitate coordination and information sharing among partners and collaborators.

Common challenges include:

- Creating processes to store, share, and transfer large-scale data across the partnering organizations or distributed sites in a large collaboration, while maintaining data security and integrity.
- Facilitating optimal data flow and processing in cases where low latencies and interactivity are important, such as for experiment steering, real-time data processing, and distributed/federated AI/ML learning and inference across geographically dispersed RDIs.
- Addressing situations where limited availability of data transfer capabilities necessitates *in situ* analysis of some or all the data at the source, which in turn requires provisioning of services for local data access, tools for data manipulation and analysis, and computing resources.

Data integration and data infrastructure interoperability

Collaborative and interdisciplinary research relies on the ability to compare and analyze data from different RDI projects and other sources. As RDI data infrastructure projects strive towards interoperability, emergent challenges include how to efficiently manage new modes of operation and better engage with the broader research community to maximize scientific return.

Common challenges include:

- Working towards common schema for generalized interoperability among RDI projects and moving away from point-to-point solutions to facilitate seamless collaboration, scalability, automation, and provide harmonized user experience.
- Identifying governance approaches that facilitate interoperability planning and execution.
- Ways to address incompatibilities in technology, data collection protocols and methods, standards and formats, and reconciling different semantics and metadata interpretations of schema via translation layers.
- Approaches to determining when centralizing shared data infrastructure is most appropriate versus maintaining federated interoperable infrastructure.

Operating in the commercial cloud

Commercial cloud platforms offer a variety of scalable services that can be advantageous for some data-intensive RDI activities and services. Recent examples include migration of some or all of an RDI's data hosting and delivery services from on-premises systems (i.e., physically located at or controlled by the RDI) to the commercial cloud and/or other externally managed platforms; large scale data reprocessing; and training and use of large AI models [16, 17]. The approaches to using the commercial cloud are varied and depend on an RDI's and the sponsoring agency's missions and technical and operational specifics. Extensive planning, analysis of alternatives, and careful implementation efforts are necessary to implement this kind of transition successfully—particularly to avoid scientific service interruption and to understand and control relative costs and benefits as well as risks.

Common challenges include:

- Approaches for defining, bounding, and prioritizing cloud-based data services that meet RDI project, partnering, and user needs. This includes understanding how to support the full spectrum of users and the variety of utilization models through cloud-based services as an alternative to or in conjunction with on-premises services.
- Accounting for anticipated cloud-based data lifecycle activities and associated costs, including for data upload, storage, processing, and analysis, as well as egress or transfer across cloud provider boundaries for necessary data workflows.
- Considering the compatibilities and differences in cloud and on-premises services for issues such as security and compliance regulations, identity management, access controls, optimal data structure and formats, and handling of sensitive/secure data.
- Managing service agreements to ensure operational continuity and maintain flexibility to be able to change cloud providers.

Handling sensitive and secure data

Diverse research and development domains sometimes involve working with sensitive research data that cannot be openly shared and require secure handling, such as healthcare and clinical research data, certain social science data, culturally sensitive data, and data in security and defense research domains. In these cases, dedicated planning, specialized data infrastructure, and additional data

governance and stewardship efforts may be required—such as through secure data storage systems ("data enclaves"), de-identification systems, strict access and utilization protocols, and special governance and legal processes.

Common challenges include:

- Allocating sufficient time and effort to establish appropriate governance mechanisms and address legal rights and terms of use for data sharing and re-use across a domestic or international collaboration or partnership.
- Performing thorough risk assessment and mitigation planning for data access and data integrity.
- Defining a timeline and establishing processes needed to socialize and train users on policies, requirements, and practices necessary to ensure protection and enable access to sensitive data and subsequent data products.
- Establishing the technical and operational requirements and expertise necessary to support secure storage and transfer of data in an ever-evolving ecosystem of security requirements and threats.

Workforce development and nurturing

As data-intensive RDI research pursuits expand rapidly, there is a critical need to nurture and maintain a strong, diverse, capable, and mobile U.S. research data workforce [18, 19, 20]. This need puts great pressure on RDIs to implement effective practices for attracting, training, and retaining data-related staff in a competitive and evolving labor market. Data infrastructure staff represent a wide range of specialized skills spanning infrastructure development and operation, data management, and the scientific use of data—ensuring that end users can maximally benefit from RDI data resources and service. RDIs also play a key role in preparing the next generations of data infrastructure developers and operators.

Common challenges include:

- Identifying the data and data infrastructure skills that are needed for RDI project workforce planning throughout the lifecycle, including the kinds of expertise needed to develop and support cloud-based as compared to on-premises infrastructure.
- Addressing the competitive and evolving nature of the data workforce through hiring and retention practices, outreach to diverse communities, incentives, and training and reskilling opportunities.
- Facilitating recruiting, training, and mobility of the data infrastructure workforce by developing consensus definitions for relevant professional roles across the data ecosystem, such as "data engineer," "data steward," "data manager," and other such positions (e.g., [21, 22]).

IV. Recommendations for Collective Action

This section proposes collective agency and community actions to disseminate practices and approaches for planning data infrastructure (framework in Section II) and addressing common challenge areas (identified in Section III) towards advancing the state of the art.

Dissemination and exchange of practices

Broad dissemination of—and expert exchange on—practices for data infrastructure planning (as embodied in the framework) can accelerate cross-fertilization of existing and novel approaches across the RDI ecosystem, scientific disciplines, and international boundaries.

Recommendation 1: Federal agencies should identify or establish a regular forum for federal managers, RDI leaders, collaboration partners, practitioners, and domain experts to discuss and exchange approaches on data infrastructure planning and implementation.

Coordination

Collaborative community-level exercises have informed strategic planning and requirements gathering efforts for cross-disciplinary RDI projects within and across scientific disciplines (e.g., [23, 24, 25]). A similar approach would benefit multidisciplinary data infrastructure planning, broadly including those who can speak to scientific objectives and utilization, cross-project partnering, and technical, legal, and operational requirements and solutions.

Recommendation 2: Agencies should consider collaborative exercises where appropriate to formulate common forecasts of data infrastructure needs across disciplines to inform respective agency planning.

Interoperability of data infrastructure supported by different agencies needs to be significantly improved to facilitate multidisciplinary integrative research and support geographically distributed workflows spanning data acquisition, computing, and analysis at scale. Harmonizing efforts might include collaborations on discipline-specific elements, such as tools and repositories that serve a research field, and trans-disciplinary elements, such as data caching, workflow systems, and networking, that broadly support all research fields.

Recommendation 3: Agencies should explore opportunities to federate or otherwise harmonize their RDI data systems and services to enable integrative scientific exploration and discovery.

Many agencies have developed tailored arrangements for commercial cloud services to support their respective large-scale science activities. Multi-agency collaboration on cloud services might be beneficial in terms of overall cost, service flexibilities, and facilitating new, broadly useful hybrid scenarios such as connecting cloud-based data to government-supported computing resources (e.g., [2, 5]).

Recommendation 4: Agencies should collectively investigate ways to jointly engage commercial cloud services for government-supported research activities.

International bodies are increasingly focused on data sharing infrastructure and related practices. U.S. participation in these forums shapes the conversations therein, allowing for better synergies between

U.S. RDI policies and practices and those of international partners, which can lead to strengthened collaborations and enhanced scientific impact.

Recommendation 5: Agencies should continue to coordinate to clearly represent U.S. policy and practice considerations for exchange and collaboration of large-scale data in relevant international forums.

Workforce

A diverse and flexible skilled workforce is a critical need that must be thoroughly considered in planning and operating RDI data infrastructure. Nurturing and growing this workforce, drawing on the nation's diversity, will greatly impact and sustain the competitiveness of the critical data infrastructure supporting the U.S. science enterprise.

Recommendation 6: Agencies should collectively explore the types of skills needed for data infrastructure across the science enterprise and identify ways to expand outreach, recruitment, training, career progression, and mobility of data infrastructure practitioners, such as via development of common types of position descriptions that might facilitate hiring processes.

V. References

I. Introduction and Aims

1. NSTC Subcommittee on Research and Development Infrastructure, *National Strategic Overview for Research and Development Infrastructure*, 2021, https://www.whitehouse.gov/wp-content/uploads/2021/10/NSTC-NSO-RDI-_REV_FINAL-10-2021.pdf.
2. National Artificial Intelligence Research Resource (NAIRR) Task Force, *Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource*, 2023, https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf; and multi-partner NAIRR Pilot, https://new.nsf.gov/focus-areas/artificial-intelligence/nairr.
3. Networking and Information Technology Research and Development (NITRD) Committee on Science and Technology Enterprise, *Strategic Computing Reserve: A Blueprint*, 2021, https://www.nitrd.gov/pubs/National-Strategic-Computing-Reserve-Blueprint-Oct2021.pdf.
4. OSTP, *Ensuring Free, Immediate, and Equitable Access to Federally Funded Research*, 2022, https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf.
5. DOE, *Integrated Research Infrastructure Architecture Blueprint Activity*, 2023, https://doi.org/10.2172/1984466.
6. Group of Senior Officials (GSO) on Global Research Infrastructures, https://www.gsogri.org/.
7. Hanisch, R., et al., *NIST Research Data Framework (RDaF), Version 2.0*, National Institute of Standards and Technology, 2024, https://doi.org/10.6028/NIST.SP.1500-18r2.
8. Christopherson, L., *The Major Facilities Data Lifecycle in a Nutshell*, 2021, https://zenodo.org/records/5550224.
9. Faundeen, J., et al., *The United States Geological Survey Science Data Lifecycle Model*, USGS Open-File Report 2013-1265, 2014, https://doi.org/10.3133/ofr20131265.

II. Framework: Principal Considerations for Data Infrastructure Design and Interconnectivity

10. NSTC Interagency Working Group on Open Data Sharing Policy (IWGODSP), *Principles for Promoting Access To Federal Government-Supported Scientific Data and Research Findings Through International Scientific Cooperation*, 2016, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/NSTC/iwgodsp_principles_0.pdf.
11. NSTC Subcommittee on Open Science, *Desirable Characteristics of Data Repositories for Federally Funded Research*, 2022, https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf.
12. Wilkinson, M., Dumontier, M., Aalbersberg, I., et al., *The FAIR Guiding Principles for scientific data management and stewardship*, Sci Data 3, 160018, 2016, https://doi.org/10.1038/sdata.2016.18; GO FAIR, *FAIR Data Principles*, https://www.go-fair.org/fair-principles/.

13. OECD, *Very Large Research Infrastructures, Policy issues and options*, 2023, https://doi.org/10.1787/2b93187f-en.

III. Cross-cutting Challenge Areas

14. NSTC Subcommittee on Research and Development Infrastructure, *U.S. Federal Research and Development Infrastructure: A Foundation of the Nation's Global Scientific Leadership and Economic and National Security*, 2024, https://www.whitehouse.gov/wp-content/uploads/2024/05/NSTC-Report-on-RDI-Global-Competition-and-Modernization.pdf.
15. Networking and Information Technology Research and Development (NITRD) Big Data Interagency Working Group (BD IWG), *Innovating the Data Ecosystem: An Update of the Federal Big Data Research and Development Strategic Plan*, 2024, https://www.nitrd.gov/pubs/Big-Data-Strategic-Plan-2024.pdf.
16. G. B. Berriman, et al., *NSF Major Facilities Cloud Use Cases and Considerations (1.0)*, 2024, https://doi.org/10.5281/zenodo.10481410.
17. T. Cai, et al., *Accelerating Machine Learning Inference with GPUs in ProtoDUNE Data Processing*, Computing and Software for Big Science, 7:11, 2023, https://doi.org/10.1007/s41781-023-00101-0.
18. U.S. Government Accountability Office, *Science and Technology: Strengthening and Sustaining the Federal Science and Technology Workforce*, 2021, https://www.gao.gov/products/gao-21-461t.
19. NSTC Interagency Working Group on Data for the Bioeconomy, *Vision, Needs, and Proposed Actions for Data for the Bioeconomy Initiative*, 2023, https://www.whitehouse.gov/wp-content/uploads/2023/12/FINAL-Data-for-the-Bioeconomy-Initiative-Report.pdf.
20. IEEE Computer Society, *Advancing the Workforce that Supports Computationally and Data Intensive Research*, 2021, https://ieeexplore.ieee.org/document/9492830.
21. USGS, *Data Stewardship and Data Steward*, https://www.usgs.gov/data-management/stewardship.
22. Office of the Director of National Intelligence, *The Intelligence Community Data Management Lexicon*, 2024, https://www.dni.gov/files/ODNI/documents/IC_Data_Management_Lexicon.pdf.

IV. Recommendations for Collective Action

23. National Academies of Sciences, Engineering, and Medicine, *Pathways to Discovery in Astronomy and Astrophysics for the 2020s*, 2023, https://doi.org/10.17226/26141.
24. Particle Physics Project Prioritization Panel (P5) of the High Energy Physics Advisory Panel (HEPAP), *Exploring the Quantum Universe: Pathways to Innovation and Discovery in Particle Physics*, 2023, https://www.usparticlephysics.org/2023-p5-report/.
25. J. Zurawski, et al., *Nuclear Physics Network Requirements Review Final Report*, LBNL-2001602, 2024, https://doi.org/10.2172/2386941.