DESIRABLE CHARACTERISTICS OF DATA REPOSITORIES FOR FEDERALLY FUNDED RESEARCH

Guidance by the

SUBCOMMITTEE ON OPEN SCIENCE

of the

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

May 2022

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 https://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 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 https://www.whitehouse.gov/ostp.

About the Subcommittee on Open Science

The NSTC Subcommittee on Open Science (SOS) advances Federal efforts to support open science by increasing access to, and use of, the results of Federally funded research and development, including, scholarly publications and digital data. Among its responsibilities, the SOS aims to coordinate and improve implementation of policies to increase access to the results of Federally funded scientific research and to identify additional steps that Federal departments and agencies can take to enhance the preservation, discoverability, accessibility, quality, and utility of research outputs.

About This Document

This document aims to improve consistency across Federal departments and agencies in the instructions they provide to researchers about selecting repositories for data resulting from Federally funded research. It identifies a set of desirable characteristics of online, public access data repositories to help ensure that research data are findable, accessible, interoperable, and reusable (FAIR) to the greatest extent possible, while integrating privacy, security, and other protections. Individual departments and agencies may use the characteristics to guide development of further instructions for the research communities they support.

Acknowledgement

The Subcommittee on Science thanks the Smithsonian Institution for providing copyediting and digital object identifier services for this document.

Copyright Information

Recommended Citation

The National Science and Technology Council, *Desirable Characteristics of Data Repositories for Federally Funded Research,* 2022, DOI: https://doi.org/10.5479/10088/113528

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

Chair

Alondra Nelson, Performing the Duties of Director, Office of Science and Technology Policy

Executive Director

Kei Koizumi, Acting Executive Director, National Science and Technology Council

SUBCOMMITTEE ON OPEN SCIENCE

Co-Chairs

Ryan Donohue, Office of Science and Technology Policy

Alan Tomkins, National Science Foundation

Jerry Sheehan, National Institutes of Health

Skip Lupia, National Science Foundation [through December 2021]

Lyric Jorgenson, National Institutes of Health [through February 2022]

Executive Secretary

Patricia Knezek, National Aeronautics and Space Administration

Senior Advisor

Christopher Steven Marcum, Office of Science and Technology Policy

Additional Contributing Subcommittee Members

Erin Antognoli, United States Department of Agriculture

Louis Barbier, National Aeronautics and Space Administration

Laura Biven, National Institutes of Health

Alasdair Cain, Department of Transportation

Francis Chelsey, Agency for Healthcare Research and Quality

Nino Chkhenkeli, United States Department of Agriculture

Leighton Christiansen, Department of Transportation

Michael Cooke, Department of Energy

Morgan Daniels, United States Agency for International Development

Jason Duley, National Aeronautics and Space Administration

Karen Fallon, National Aeronautics and Space Administration

Terri Geisler, National Aeronautics and Space Administration

Jeff Given, Department of Energy

Martin Halbert*, National Science Foundation

Robert Hanish, National Institute of Standards and Technology

Hope Hongzhu He, Agency for Healthcare Research and Quality

Kirk Keith, United States Geological Survey

Madison Langseth, United States Geological Survey

Hilary Leeds, National Institutes of Health

Alex Montilla, Environmental Protection Agency

Mary Moulton, Department of Transportation

Akshay Narang, Environmental Protection Agency

Tamar Norkin, United States Geological Survey

Jake O'Sullivan, United States Agency for International Development

Dina Paltoo, National Institutes of Health

Corey Portalatin-Berrien, National Aeronautics and Space Administration

Liza Rozen, United States Agency for International Development

Ashley Sands, Institute of Museum and Library Services

Guinevere Shaw, Department of Energy

Gerald Steeman, National Aeronautics and Space Administration

Pimjai Sudsawad*, Administration for Community Living

Robert Swain, Centers for Disease Control

Christopher Thomas, Department of Defense

Mason Thompson, United States Agency for International Development

Ellen Wann, National Institutes of Health

Nancy Weiss, Institute of Museum and Library Services

Bob Williams, Office of the Assistant Secretary for Preparedness and Response, Department of Health and Human Services

Maryam Zaringhalam*, National Institutes of Health

*Denotes subgroup co-chairs

Table of Contents

Abbreviations and Acronyms

DOI	digital object identifier
DPI	digital persistent identifier, synonymous with PID
FAIR	findable, accessible, interoperable, and reusable
NSTC	National Science and Technology Council
OSTP	Office of Science and Technology Policy
PID	persistent identifier, synonymous with DPI
R&D	research and development
RFC	request for comment
SOS	Subcommittee on Open Science

1. Introduction

The *Memorandum on Increasing Access to the Results of Federally Funded Scientific Research* (Memorandum) issued by the White House Office of Science and Technology Policy (OSTP) in February 2013 directs each Federal agency with more than $100 million in annual research and development (R&D) expenditure to require funded researchers to prepare plans describing the proposed approach for managing and sharing digital data resulting from their Federally supported research.[1] To date, more than 20 Federal departments and agencies (herein "agencies") have implemented policies to improve the management and sharing of data resulting from R&D, consistent with the Memorandum.[2, 3] These agencies work together, through the National Science and Technology Council's Subcommittee on Open Science (SOS), to coordinate implementation and disseminate good practices.

A key element of the required data management plans is specification of the digital, online, public access data repository or repositories researchers will use for preserving, maintaining, and providing access to Federally supported research data. While some agencies designate specific repositories to be used for particular types of data (e.g., genomic data, topographical data) or a particular type of research (e.g., Arctic research, social sciences research), for much Federally funded research, the selection of a suitable repository is delegated to the researcher or their institutions. Some agencies provide information to assist researchers in the selection of data repositories. However, this information is inconsistent across agencies, including among those that support research in similar or related disciplines. Until now, agencies had not identified the desirable characteristics of data repositories on which to base their assistance to researchers and their institutions.

To improve the management and sharing of data from Federally funded research, agencies agreed to leverage the SOS to identify a consistent set of desirable characteristics for data repositories that all agencies could incorporate into the instructions they provide to the research community for selecting data repositories. By establishing common expectations, agencies intend to reduce the complexity for the research community–including investigators, program officers, data managers, librarians, and others–in complying with Federal data sharing policies. Federal agencies can also use this set of characteristics to develop or identify suitable repositories for particular types of data.

To carry out this work, agencies within the SOS drew upon existing expertise and experience with data management and sharing. They also reviewed existing criteria promulgated by non-governmental organizations involved in the certification of data repositories (e.g., International Standards Organization, International Science Council). Agencies also took into account input received on a draft set of characteristics issued for public comment in January 2020 (Box 1).

[1] Office of Science and Technology Policy, Memorandum on Increasing Access to the Results of Federally Funded Scientific Research, Feb. 22, 2013. Available at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

[2] 2021 OSTP Congressional Report on Public Access, Nov. 5, 2021. Available at: https://www.whitehouse.gov/wp-content/uploads/2022/02/2021-Public-Access-Congressional-Report_OSTP.pdf

[3] For a compilation of these public access plans, visit https://www.science.gov/publicAccess.html.

Box 1. OSTP Request for Comments on Desirable Characteristics of Data Repositories.

OSTP issued a request for comment (RFC) in January 2020 on a draft set of desired characteristics of repositories for data resulting from Federally funded research. The RFC solicited information on the use and application of the characteristics, their appropriateness to data from Federally funded research, the ability of existing repositories to meet the desired characteristics, whether additional characteristics should be considered, and other considerations.

In total, 119 responses were received from stakeholders in universities, research consortia, data centers, scientific societies, and government agencies. The comments supported the need for a common set of repository characteristics and endorsed the general approach proposed. Comments noted that the proposed characteristics would make data more findable, accessible, interoperable, and reusable (FAIR) and could be met by many existing repositories. Respondents expressed general agreement with most of the proposed characteristics, suggesting some minor reorganization and clarifications.

RFC: https://www.Federalregister.gov/documents/2020/01/17/2020-00689/request-for-public-comment-on-draft-desirable-characteristics-of-repositories-for-managing-and

Responses:https://www.whitehouse.gov/wp-content/uploads/2017/11/Desirable-Characteristics-RFC-Comments.pdf

This guidance document presents the set of desirable characteristics for repositories agreed to by Federal agencies, reflecting the input that OSTP and SOS received and evaluated. It addresses a near-term need to provide greater consistency across agencies, recognizing that future steps will be needed to better coordinate data storage and management to make data from Federally funded research more findable, accessible, interoperable, and reusable (FAIR),[4] as well as more equitable, inclusive, secure, and trustworthy. The endeavor to improve public access to Federally-supported research makes for a more open government, facilitates evidence-based decision making, and yields greater returns on Americans' investments in R&D. This guidance document constitutes one set of tools that agencies can use to advance those goals.

[4] For a discussion of FAIR data, see: https://www.go-fair.org/fair-principles.

2. Desirable Characteristics of Data Repositories

Through the SOS, Federal agencies identified a consensus set of desirable characteristics of repositories for data resulting from Federally funded research. The characteristics are intended to help agencies direct Federally funded researchers toward repositories that enable management and sharing of research data consistent with the principles of making data FAIR and promoting equitable access to research products, and that integrate necessary protections of privacy and security, including human subjects' protections.

The desirable characteristics provided by this guidance document are not intended to be an exhaustive set of features for data repositories; rather they represent general capabilities for researchers, agencies, and institutions to prioritize when selecting repositories to share research data. While the desirable characteristics are intended to be enduring, Federal agencies may update them periodically, in coordination with the SOS, to reflect changing expectations, advances in research and technology, and evolving practices related to data management, privacy, security, and sharing.

While several organizations provide for repository certification standards, Federal agencies have elected not to adopt existing certification criteria, due in part to the cost and complexity of certification processes and of differences in needs and expectations of different agencies and their research communities. Nevertheless, agencies aimed to ensure the desirable characteristics supplied in this guidance document would be consistent with many of the criteria used to certify data repositories. At the same time, agencies expect that many repositories that do not seek certifications will already exhibit these characteristics and will be able to improve their ability to make data FAIR.

Desirable Characteristics for All Repositories

The SOS identified characteristics that are relevant to all repositories that manage and share data resulting from Federally funded research. The characteristics are organized across three themes, similar to those used by certification organizations, to highlight their consistency with certification approaches. The themes are: Organizational Infrastructure, Digital Object Management, and Technology. Table 1 describes the specific desirable characteristics for all repositories from each of the themes in detail.

Table 1. Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded or Supported Research

Organizational Infrastructure	**Free and Easy Access**	The repository provides broad, equitable, and maximally open access to datasets and their metadata free of charge in a timely manner after submission, consistent with legal and policy requirements related to maintaining privacy and confidentiality, Tribal and national data sovereignty, and protection of sensitive data.
	Clear Use Guidance	The repository ensures datasets are accompanied by documentation describing terms of dataset access and use (e.g., reuse licenses and need for approval by a data use committee).
	Risk Management	The repository has documented capabilities for ensuring that administrative, technical, and physical safeguards are employed to comply with applicable confidentiality, risk management, and continuous monitoring requirements for sensitive data.
	Retention Policy	The repository provides documentation on policies for data retention.
	Long-term Organizational Sustainability	The repository has a plan for long-term management of data, including maintaining integrity, authenticity, and availability of datasets; has contingency plans to ensure data are available and maintained during and after unforeseen events.
Digital Object Management	**Unique Persistent Identifiers**	The repository assigns a dataset a citable, unique persistent identifier (PID or DPI), such as a digital object identifier (DOI), to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The unique PID points to a persistent location that remains accessible even if the dataset is de-accessioned or no longer available.
	Metadata	The repository ensures datasets are accompanied by metadata to enable discovery, reuse, and citation of datasets, using schema that are appropriate to, and ideally widely used across, the communities that the repository serves.
	Curation and Quality Assurance	The repository provides or facilitates expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

	Broad and Measured Reuse	The repository ensures datasets are accompanied by metadata that describe terms of reuse and provides the ability to measure attribution, citation, and reuse of data (e.g., through assignment of adequate and openly accessible metadata and unique PIDs).
	Common Format	The repository allows datasets and metadata to be accessed, downloaded, or exported from the repository in widely used, preferably non-proprietary, formats consistent with standards used in the disciplines the repository serves.
	Provenance	The repository has mechanisms in place to record the origin, chain of custody, version control, and any other modifications to submitted datasets and metadata.
Technology	**Authentication**	The repository supports authentication of data submitters. The repository has technical capabilities that facilitate associating submitter PIDs with those assigned to their deposited digital objects, such as datasets.
	Long-term Technical Sustainability	The repository has a plan for long-term management of data, building on a stable technical infrastructure and funding plans.
	Security and Integrity	The repository has documented measures in place to meet well established cybersecurity criteria for preventing unauthorized access to, modification of, or release of data, with levels of security that are appropriate to the sensitivity of data (e.g., the NIST Cybersecurity Framework: https://www.nist.gov/cyberframework).

Additional Considerations for Repositories Storing Human Data

The SOS also identified additional characteristics for repositories storing human data, which must be able to integrate privacy protections, confidentiality, and other capabilities based on considerations. These characteristics are supplemental to those for all repositories in Table 1. These additional considerations are intended to apply to repositories that store de-identified human data. As re-identification of de-identified human data remains a risk for many datasets, additional considerations to protect privacy and security, as described in Table 2, are paramount.

Table 2. Additional Considerations for Repositories Storing Human Data

Fidelity to Consent	The repository employs documented procedures to restrict dataset access and use to those that are consistent with participant consent (such as for use only within the context of research on a specific disease or condition) and changes in consent.
Security	The repository implements and provides documentation of appropriate approaches (e.g., tiered access, credentialing of data users, security safeguards against potential breaches) to protect human subjects' data from inappropriate access.
Limited Use Compliant	The repository employs documented procedures to communicate and enforce data use limitations, such as preventing reidentification or re-distribution to unauthorized users.
Download Control	The repository controls and audits access to and download of datasets.
Request Review	The repository makes use of an established and transparent process for reviewing data access requests.
Plan for Breach	The repository has security measures that include a response plan for detected data breaches.
Accountability	The repository has procedures for addressing violations of terms-of-use and data mismanagement.

3. Applications

Federal agencies intend to use these characteristics as a tool when:

- Assisting Federally funded institutions and investigators, including members of the Federal workforce and the agencies that employ them, in identifying data repositories for storing and providing public access to research data (e.g., when funding agencies do not host the data and/or have not designated specific repositories for use);
- Identifying specific repositories that a Federal agency might designate for use for particular types of data resulting from Federally funded research; and,
- Evaluating data management plans submitted to Federal agencies with requests or applications for research funding.

They may also serve to inform:

- External data repository developers and managers of the characteristics desired by Federal agencies for storing, providing equitable access to, and preserving data resulting from Federally funded research;
- Development of Federal agency repositories to store data resulting from Federally funded research; and,
- Development of best practices, policies, and funding initiatives to create a robust repository infrastructure for data resulting from Federally funded research.

Federal agencies plan to incorporate these characteristics into guidance for their research communities to make decisions around where and how to share data. In addition, the SOS continues to advance activities aimed at improving agency coordination to enhance compliance efforts with public access and data management policies, as well as strengthening open science infrastructure. Future directions for such investment may consider additional desirable characteristics for storing public access data resulting from Federally supported research, such as data that is readily available for use in emerging technologies or modes of science. This broader set of work will enhance accessibility, quality, transparency, and reuse of data resulting from Federally funded research.

Federal agencies do not plan to use the characteristics in this guidance document to assess, evaluate, or certify the acceptability of a specific data repository, unless otherwise required for a particular agency program, initiative, or funding opportunity.

Through dissemination and use of this consensus set of desirable characteristics of data repositories, Federal agencies anticipate improvement in storage, management and reuse of data resulting from Federally funded research. Moreover, increased coordination across agencies is expected to result in greater consistency in data management and access, improved quality and use of data, and enhanced opportunities to replicate findings. The desirable characteristics are also expected to lead to better stewardship of data from human subjects, greater returns on the public's investment in research through data that are more findable, accessible interoperable, and reusable. Ultimately, Federally funded data that are FAIR are also more equitable and can reach a broader, more diverse, and more representative, userbase.