

**Public Responses Received for Request for  
Information 85 FR 3085: *Draft Desirable  
Characteristics of Repositories for Managing and  
Sharing Data Resulting From Federally Funded  
Research***

**January 17, 2020 – March 6, 2020**

**Federal Register Notice:**

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

**IMPORTANT DISCLAIMER:**

**This document is a compilation of comments provided in response to a public Request for Information issued by the Office of Science and Technology Policy (OSTP). The information contained herein does not represent and is not intended to represent any position, recommendation, or views of the White House, OSTP, or any U.S. Government organization.**

**From:** Kathryn Reynolds <[K.Reynolds@cabi.org](mailto:K.Reynolds@cabi.org)>  
**Sent:** Monday, January 20, 2020 7:18 AM  
**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>  
**Subject:** [EXTERNAL] RFC Response: Desirable Repository Characteristics

Dear Whom it May Concern,

I saw with interest your compilation of a series of desirable characteristics to aid in the selection of repositories. I wonder whether you are aware of [a similar effort from the FAIRSharing community](#), who are also taking comments on a set of FAIR specific repository criteria they have compiled. If you were interested in combining your efforts with this group (and haven't done so already) I am sure they will be extremely receptive to learning from your findings and vice versa.

I am a data analyst at the Centre for Agriculture and Biosciences International (CABI). We are a non-profit with significant amounts of research data, and also work helping to enable data access within the agricultural space. We are currently working with the Gates foundation to help their data outputs from their soil and agronomy projects more FAIR, and are concurrently developing a CKAN repository to publish some of our own data assets. As such, I look forward to reading the finalized output of your project, as it may help to advise my own work of facilitating data sharing by producing (and pointing to) FAIR data repositories. I think even if researchers are willing to share data, they are often not sure where or how best to do so, and documentation helping in this process is well overdue!

Thank you for your time,

**Kathryn Reynolds**  
Junior Data Analyst

CABI Head Office  
Nosworthy Way  
Wallingford  
Oxfordshire  
OX10 8DE  
United Kingdom

Telephone: [+44 \(0\)1491 829358](tel:+44%201491829358)  
Email: [K.Reynolds@cabi.org](mailto:K.Reynolds@cabi.org)  
Visit us at [www.cabi.org](http://www.cabi.org)

January 23, 2020

Office of Science and Technology Policy  
Email: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

Re: RFC Response: Desirable Repository Characteristics

Dear Reviewers,

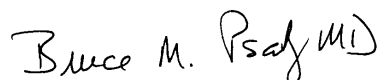
I would like to suggest an additional characteristic that should be included as a "desirable characteristic."

As the leaders of the CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) Consortium funded by HL105756, my colleagues and I are accustomed to sharing data for among NIH-funded cohort studies for genetic analyses. Both for pooled analyses and for meta-analyses, it is essential to harmonize the phenotypes of interest for each analysis. The current data-sharing permission structures and mechanisms make it difficult for us to save, preserve, and share the harmonized data. Our harmonization efforts, useful for one publication, frequently go wasted.

Eric Boerwinkle, Steve Rich, and I described this problem in a Perspective Piece entitled "Innovation in Data Sharing at the NIH," *New England Journal of Medicine* 2019; 380: 2192-5. As we concluded there, "Major advances leveraging these large-scale genomic and phenotype data require not only contemporary analytics based on deep learning and artificial intelligence, but also administrative and regulatory innovation.... Administrative innovations in data sharing to promote big-data science will not emerge on their own. The NIH can devise a new set of data-access policies and regulations that would be fit for the purpose and appropriate for current and future forms of biomedical data."

Please see the publication for a more complete discussion and explanation for the need for "administrative and regulatory innovation" to promote data sharing.

Cordially,



Bruce M. Psaty, MD, PhD  
Professor, Medicine and Epidemiology  
Co-director, Cardiovascular Health Research Unit

**From:** HELLMAN, BARRY M CIV USAF AFMC AFRL/RQHV <[barry.hellman@us.af.mil](mailto:barry.hellman@us.af.mil)>  
**Sent:** Friday, January 24, 2020 9:54 AM  
**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>  
**Subject:** RFC Response: Desirable Repository Characteristics

Two suggestions:

1. DoD currently publishes reports in DTIC. However, there is no place to properly archive finalized data files that go with the report (e.g. Excel spreadsheets, specialized source code, input and output data files that are used with an analysis tool like Finite Element Analysis). There should be a cloud based server (with appropriate distribution limitations enforced) to archive finalized data files that correspond to technical reports.
2. DoD uses the Technology Readiness Level (TRL) a great deal in development roadmaps. While the definitions of TRL are universally accepted, there is no standard way to document that a certain technology has reached a TRL. I recommend coming up with a short standardize template (2-3 pages) for a program manager or principal investigator to document when a technology has reached a TRL an include citations for appropriate references. The concepts of Manufacturing Readiness Level (MRL) and Integration Readiness Level (IRL) are also sometimes used and would also benefit from a standardized documentation method.

Barry Hellman  
AFRL/RQHV  
937-255-3088

**From:** Bruce M Psaty <[psaty@uw.edu](mailto:psaty@uw.edu)>  
**Sent:** Monday, January 27, 2020 8:58 AM  
**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>  
**Subject:** [EXTERNAL] RE: RFC Response: Desirable Repository Characteristics

I can send a copy of the pdf of the perspective from the N Engl J Med if that would help. Let me know. Thanks

=====

Bruce M. Psaty, MD, PhD  
Professor, Medicine & Epidemiology  
University of Washington  
Cardiovascular Health Research Unit  
1730 Minor Avenue, Suite #1360  
Seattle, WA 98101-1466  
Phone: 206/221-7775  
Fax: 206/221-2662  
Email: [psaty@u.washington.edu](mailto:psaty@u.washington.edu)

# Institutional Health Data Repository

# Table of Contents

- 1. Mission ..... 3
- 2. Vision ..... 4
- 3. Data Governance ..... 5
- 4. Governance Bodies ..... 5
- 5. Data Governance Committee ..... 5
- 6. Data Warehouse Architecture ..... 6
- 7. Purpose of the IHDR ..... 7
- 8. Security and Data Access ..... 8
- 9. Intellectual Property ..... 9

## **Mission:**

To create and maintain a data warehouse of health data for the region that conforms to FAIR data principles (Findable, Accessable, Interoperable, Reusable) and provides curated health data for the UB and Buffalo Translational Consortium (BTC) research community and our partners. This will be accomplished through a comprehensive data governance plan and Master Patient Index that takes in data from multiple clinical partners and payers and provides interoperable data to researchers across our campuses and extends to our research partners. This requires a comprehensive extract, transfer and load protocol with data cleaning and imputation. We will add to this framework a set of data analysis and predictive analytic (Machine Learning) tools to ensure that our research community has access to the best computational and data analytic resources to further their research goals in a secure environment.

## **Vision:**

To securely provide FAIR and Interoperable health data and tools to our UB and BTC research community and to our partners.

## **Operations and Management:**

The Institutional Health Data Repository (IHDR) will be operated as a division of the University at Buffalo under the Vice President for Health Sciences with governance as described in the following section. All current resources allocated to the current Institute for Healthcare Informatics will be transitioned to this new unit.

A Director position will be searched and filled once adequate funding is identified. This position will be PhD-level with experience in health data research management. In the interim, the Chair of the JSMBS Bioinformatics Department will serve as Director of the IHDR.

A strategic operations plan will be developed by the Director or Interim Director, staff, and supporting UB constituents during the Spring 2020 semester and submitted to the Executive Oversight Council by July 1, 2020.



## Data Governance:

### Governance Bodies

- Executive Oversight Council – Provide executive guidance and funding decisions to support operation and growth of the IHDR.
  - Membership: VP Health Science, VP Research, UB CIO, Great Lakes Health (GLH) Representative (one from Kaleida Health and one from Erie County Medical Center (ECMC)), Chair of the Department of Biomedical informatics, a Basic Science Representative, Chair of Computer Science and the Director IHDR
  - This body will give the overall direction of the IHDR
  - This body will have budgetary oversight
  - The EOC will meet quarterly
- Advisory Committee – Provide operational guidance including methods for access to data, data policies, regulatory compliance, and technology architecture.
  - Membership: Director IHDR, JSMBS Dept Chairs, CTSA Director, past Director of IHI, VPCIO Director of Enterprise Infrastructure Services, Dean of Libraries, Chair of the Department of Biomedical Informatics, Basic Science Representative, Computer Science Representative and GLH Representatives, outside Advisors from other Universities and / or industry
  - This body will provide strategic direction as to content and set the policies regarding data access and cost of usage of the IHDR
  - The Advisory Committee will meet bi-annually

### Data Governance Committee

For each datatype in the IHDR we will develop both systematic and formal definitions. These will be vetted with the data contributors. All merged data will be evaluated to ensure equivalence in meaning. Representations will be using formal Ontology and will have an Ontology advisor on the committee.

- Membership: UB CIO, their data governance representative, an Ontology expert from at least two Decanal Units, the Chair of the Department of Biomedical Informatics, Ontology staff, Chair of Pathology, director of the Center of Excellence in Bioinformatics and the IHDR director, and a representative from Kaleida and one from ECMC and a representative from other data providers..
- This committee will meet monthly
- This committee will work on the data standardization and governance of new variables throughout GLH
- They will handle Material Transfer Agreements
- The Goal is to lead to semantic interoperability across our Research and Clinical environment

## UB Cloud based – Data Warehouse Architecture:

### Importing Data into the Data warehouse

- From each GLH partner, corporate partners and UB, we will import and merge data from individuals across the health system
- This will be accomplished through establishing a Master Patient Index where patients are matched using:
  - o Name
  - o SSN
  - o Address
  - o DOB
  - o Cell Phone Number
  - o Problem Lists (to identify people whose Medications or Problems are vastly different)
  - o Date of Death (to ensure that we have accurate dates and ages)
- Data will be stored in multiple formats including
  - o OMOP (Relational Data Model)
  - o I2b2 (Relational Data Model in a STAR Schema)
  - o PCORNet (Relational Data Model)
  - o Elastic Search (Fast Indexing and Retrieval)
  - o NOSQL database (Hashtables serialized to disk)
  - o GRAPH DB (triple store)
  - o Neo4J
  - o This is necessary as each of these systems have associated with them different tooling and properties that serve individual purposes
- Artificial Intelligence and Machine Learning tools will preprocess the data and provide situational awareness of shifts in the data over time.
- We will Extract the data from its source (Cerner, MediTech, Allscripts, etc.)
- We will Transfer the data to the IHDR
- We will Load the data into each of the Models that we will maintain
- We will perform data Cleaning
- We will develop and utilize rules for excluding records with beyond a threshold of missing or conflicting data to avoid injecting errors into the data warehouse
- We will use several Imputation methods to handle missing data
- We will allow pathways for Great Lakes Health personnel, Students, Residents, Fellows, and Faculty to gain access to the data utilizing a secure cloud infrastructure maintained in UB on premises data centers.

## **Purpose of the IHDR:**

In addition to the mission and vision described earlier, the IHDR will benefit UB, GLH, and the greater WNY community in the following ways:

1. Recruitment to Clinical Trials
2. Automated Retrospective Research
3. Improved Clinical Practice (clinical decision support, population health, biosurveillance, etc.)
4. Improved Patient Safety
5. Improved ability to administrate the practice
6. Improved education of Residents and Medical Students
7. Clinical Decision Support
8. Precision Medicine & Personalized Medicine
9. New Drug Development
10. New Laboratory Test Development
11. Regenerative Medicine
12. Gene Therapy
13. General Scientific Advances
14. Population Health Impact
15. Administrative business intelligence for our hospital partners
  1. Data Aggregation
  2. Machine Learning / Artificial Intelligence
  3. Data Indexing
  4. Data Visualization
  5. Learning Health System
  6. Improved Quality and Patient Safety
16. Provision of FAIR Data for Practice and Population Management and for Research
17. Linkage to our Biobank

## **Security, Confidentiality and Access to Data:**

Recognizing the central role and importance of Security and Confidentiality in making the IHDR work for our community and partners we have put together a strong HIPAA secure plan.

The data will be held on HIPAA compliant secure data infrastructure in UB on premises data centers within the VPCIO area. IHDR data will be stored in a secure private cloud resource on separate encrypted physical hardware, on a separate subnet, firewall protected in a locked room that will be accessed physically only by HIPAA trained and vetted personnel from IT or Facilities. Some GPU enabled devices will be part of the IHDR cloud resource. Connections between the IHDR cloud resource and the Center for Computational Research (CCR) will be provided to a set of hardware reserved only for that purpose and will be accessed via a Virtual Private Network (vpn). All data will be encrypted in flight and at rest. Encryption technologies employed will be Federal Information Processing Standard (FIPS) 140-2 compliant.

Authorized and dual authenticated users will access the IHDR via a Virtual Machine (VM) set up for them based on their IRB approvals and will require two factor authentication (using Duo Mobile). Users doing preparatory to research investigation will have access to a specialized environment which is only capable of providing aggregate results (numbers). All data uploaded to the IHDR cloud will be run through antivirus software and executables will be tagged to the authorized sender.

The setup for the IHDR will include a Virtual Machine (VM) with a set of preloaded software:

### **Databases**

- Structured Data in SQL (i2b2, OMOP & PCORNet)
- Unstructured data in NOSQL and Elastic Search and Splunk
- Claims Data
- Image Data

### **Data Analytics**

- R Software
- Microsoft tools
- SAS
- Machine Learning
- Python
- Java VM
- Docker
- Uploaded Data Sets
- Uploaded Screened Code

### **Infrastructure**

- VDI
- GPU compute
- Identity Management (accounts)

- Storage
- Security and Privacy Controls

As all needed analysis should be able to be performed on the VMs provided, there is no need to remove any identified or line level data from the IHDR environment. Any data merging should be able to be done on the IHDR environment. Requests for increased data storage can be handled by the IHDR administrative team and can be increased to handle large data such as image or genomic datasets. Software for Genomic Assembly, Metagenomic analysis and pathway analysis for systems biology will also be provided.

Publications using the IHDR should include an acknowledgement of the resource. Grants intending to use the IHDR should reference the resource using standardized Facilities language provided by the IHDR team and approved by the executive committee and should include 3% direct costs to the resource to defray its cost to the administration.

### Intellectual Property

Recognizing the value of the data in the IHDR, we realize that responsibility to share with the data providers the intellectual property and companies that are made possible through the use of that data.

**From:** [dredie@verizon.net](mailto:dredie@verizon.net) <[dredie@verizon.net](mailto:dredie@verizon.net)>

**Sent:** Monday, February 3, 2020 3:43 PM

**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>

**Subject:** [EXTERNAL] OFFICE OF SCIENCE AND TECHNOLOGY POLICY Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

In response to the OFFICE OF SCIENCE AND TECHNOLOGY POLICY Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research, I submit two comments for your consideration.

Comment 1: There is no mention of Government funded IR&D by contractors. Like the direct funded research, IR&D is paid for with public tax dollars and therefore should be discoverable. If the contractors do not want to share the information, they should not be reimbursed.

Comment 2: The veracity of this data repository will be dependent on the agencies making the data available. Coming from a DoD background for nearly 40 years, I can attest to the reticence of DoD employees (and their contractors) to take any extra steps to report data. Your repository, if it is to be robust and complete, must be populated seamlessly.

Thank you for the opportunity to comment – Dr. Edie Williams

**From:** Anna Greene <[a.greene@alexslemonade.org](mailto:a.greene@alexslemonade.org)>  
**Sent:** Monday, February 24, 2020 10:27 AM  
**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>  
**Subject:** [EXTERNAL] RFC Response: Desirable Repository Characteristics

Hello,

My comment is related to the use of standard NIH repositories for non-NIH funded studies. It does appear that NIH will allow deposition of non-NIH-funded data into NIH repositories such as SRA or dbGaP, but that it must go through an approval process first: <https://datascience.cancer.gov/data-sharing/genomic-data-sharing/non-nih-investigators>. It's not clear to me how often these data are rejected or if they are in general are accepted, but my comment is to strongly encourage NIH to accept non-NIH-funded data without asking investigators to go through a more rigorous process than NIH-funded investigators. At Alex's Lemonade Stand Foundation, we require that our funded researchers share all unique resources, including data, openly with the research community. It's much more difficult for them to do so if NIH rejects their submissions to what are considered the standard in the field repositories available for genomic and other large-scale data. NIH should embrace that these repositories such as SRA and dbGaP are single source of truth repositories which should accept appropriate data submissions from non-NIH-funded work.

Thanks!  
Anna

Anna Greene, PhD  
Director of Science  
*Alex's Lemonade Stand Foundation*  
*Fighting childhood cancer, one cup at a time*  
[a.greene@alexslemonade.org](mailto:a.greene@alexslemonade.org) | 610-649-3034 | [www.alexslemonade.org](http://www.alexslemonade.org)

Comments in response to Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research.

Name: Tim Whiteaker

Organizational affiliation: The University of Texas at Austin

Primary scientific discipline: Physical science (water resources engineering)

Roles: researcher, data manager

I have comments on Part I. Desirable Characteristics for All Data Repositories. Overall it looks great, and is in line with the repositories I typically use. Paragraphs J and K may be hard to implement, unless the repository only allows a certain set of formats to be archived (J). For K, the repository would need something like diff capabilities on a Git repo, which is related to J since you need a file format that can be easily diff'd. I still think J and K are desirable, but the responsibility may fall more upon the data submitter than the data archive for the reasons I mentioned.



To: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

From: Margaret C. Levenstein, Director, Inter-university Consortium for Social and Political Research (ICPSR)

Subject: RFC Response: Desirable Repository Characteristics

Date: March 1, 2020

Thank you for this opportunity to comment on the Office of Science and Technology Policy (OSTP) draft set of desirable characteristics of data repositories used to locate, manage, share, and use data resulting from Federally-funded research. As the director of the Inter-university Consortium for Political and Social Research (ICPSR), the largest archive of digital social and behavioral science data in the world, I am familiar with the range of characteristics of effective repositories. I am also a research professor at the University of Michigan.

ICPSR supports the existing desirable characteristics for all data repositories, but would like to highlight the tension between "Free & Easy to Access and Reuse" (Desirable Characteristic F) and "Long-term sustainability" (Desirable Characteristic B). Infrastructure to manage, preserve, and disseminate data is costly, especially when the data are large and complex. Likewise, preparing data for reuse requires significant investment -- often by domain or specialty repositories. In the ecosystem of repositories that exist, "free" data often do not include the necessary metadata for reuse and long-term preservation. ICPSR advocates for the federal government to "commit to sustaining institutions that assure the long-term preservation and viability of research data. Agencies supporting research must back up the new open-access requirements with funding to ensure their success....These are modest costs to assure a strong return on public investments in research and to enable uses of data unanticipated by the original investigators" ([Sustaining Domain Repositories for Digital Data: A White Paper](#)).

ICPSR particularly supports the attention to data on human subjects, even if "deidentified." Protecting the privacy of human subject data requires technological, social, and regulatory dimensions. Perfect and permanent anonymization is essentially impossible for many important use cases. The amount of data already available about individuals and the low cost of computational capacity make re-identification easier than at any previous time. In order to balance the utility of data with privacy protection, repositories need to manage and provide tiered access to data of different levels of sensitivity and the credentialing of data users to create a culture of responsible data management and privacy protection. Repositories can be characterized by their ability to ensure differential and effective consequences for breaching responsible data use and to deploy different technologies for both making data safe and/or making safe the technological platforms where the data are analyzed. Tiered access should balance safe people, safe places, and safe data.

In addition to the existing desirable characteristics for all data repositories, we suggest including the following characteristics, many of which are adapted from the recent draft paper, [Data Repository Selection: Criteria That Matter](#).

- Collection Development Policy - This criterion is whether a repository has a transparent policy detailing the range of data that are considered in scope for the repository and is useful to the audience of users (including data contributors and data users) accessing the services of the repository.
- Data Deposition Conditions - This repository characteristic details any restrictions the repository places on who it will accept data from.
- Dataset Usage Information - Repositories differ in the extent to which they allows researchers insight in data reuse by systematically collecting and sharing this information (e.g. number of views, downloads).
- Data Preservation Policy - It is important that a repository provides to the user community documentation about how long-term preservation of the data is ensured. Repositories can be characterized by various aspects of their approach to digital preservation.
- Certification - Whether a repository has been certified for its compliance with standards for trusted digital repositories is an important characteristic. There is growing community support around the value of the CoreTrustSeal certification for repositories.

Thank you again for this opportunity to comment on the draft set of desirable characteristics of data repositories.

Margaret C. Levenstein, Ph.D.

Director, [Inter-university Consortium for Political and Social Research](#)

Research Professor, [Institute for Social Research](#) and [School of Information](#)

Adjunct Professor of Business Economics and Public Policy, [Ross School of Business](#)

University of Michigan

Ann Arbor, MI 48106-1248

[MaggieL@umich.edu](mailto:MaggieL@umich.edu)

**To:** White House Office of Science and Technology Policy (OSTP)

**From:** Wesley Stites, Associate Vice Chancellor for Research and Innovation; Steve Krogull, Associate Chief Information Officer; and Melody Herr, Head, Office of Scholarly Communications on behalf of the University of Arkansas, Fayetteville (UAF)

**Date:** 27 February 2020

Thank you for the opportunity to review and comment on the desirable characteristics of repositories for managing and sharing data resulting from federally funded research. The list aligns well with emerging standards for data repositories. On behalf of UAF, I ask OSTP to consider the following additions.

*Quality Assurance:* a designated administrator oversees the deposit of data (including metadata) to ensure that it meets FAIR standards

*Supervision of Use:* a designated administrator oversees the use of human data

*Assistance and Training:* the repository should provide assistance and training for all aspects of sharing and using data

*Intellectual Property:* data is made available under an open license, analogous to [Creative Commons Licenses](#) or the [GNU General Public License](#), which specifies the terms of use and requires that a proper citation/attribution and the license visibly accompany all products resulting from use of the data

Of course, data repositories come with costs and it is important that federal funding agencies are prepared to help with both the direct and indirect costs of establishing and maintaining them.



PUBLIC RESPONSIBILITY IN  
MEDICINE AND RESEARCH

**Chair**

*Natalie L. Mays,  
BA, LATG, CPIA*

**Vice Chair**

*Suzanne Rivera, PhD, MSW*

**Secretary**

*Martha Jones, MA, CIP*

**Treasurer**

*Owen Garrick, MD, MBA*

**Board of Directors**

*Albert J. Allen, MD, PhD*

*Elizabeth A. Buchanan, PhD*

*Holly Fernandez Lynch, JD, MBE*

*Bruce Gordon, MD*

*Mary L. Gray, PhD*

*F. Claire Hankenson,  
DVM, MS, DACLAM*

*Karen M. Hansen*

*Megan Kasimatis Singleton,  
JD, MBE, CIP*

*Jori Leszczynski, DVM, DACLAM*

*Vickie M. Mays, PhD, MSPH*

*Gianna McMillan, DBe*

*Robert Nobles, DrPH, MPH, CIP*

*Stephen Rosenfeld, MD, MBA*

**Ex Officio**

*Elisa A. Hurley, PhD  
Executive Director*

March 2, 2020

Comments submitted online to: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

Sean C. Bonyun,  
Chief of Staff  
Office of Science and Technology Policy  
Executive Office of the President  
Eisenhower Executive Office Building  
1650 Pennsylvania Avenue  
Washington, DC 20504

RE: Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research (85 *Federal Register* 3085)

Dear Mr. Bonyun:

Public Responsibility in Medicine and Research (PRIM&R) appreciates the opportunity to comment on the Office of Science and Technology Policy's Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research, published January 17, 2020.

PRIM&R is a nonprofit organization dedicated to advancing the highest ethical standards in the conduct of research. Since 1974, PRIM&R has served as a professional home and trusted thought leader for the research protections community, including members and staff of human research protection programs and institutional review boards (IRBs), investigators, and their institutions. Through educational programming, professional development opportunities, and public policy initiatives, PRIM&R seeks to ensure that all stakeholders in the research enterprise understand the central importance of ethics to the advancement of science.

PRIM&R endorses the White House Office of Science and Technology Policy's (OSTP) efforts to improve the consistency of guidelines that federal R&D-funding agencies provide to their grantees and other stakeholders about best practices in long-term storage of data from federally funded research. We especially appreciate the current step of

developing a proposed, common set of desirable characteristics of data repositories that agencies can use to support their current Public Access and data sharing efforts. As the request for public comment notes, this kind of forward thinking has the potential not only to improve government-operated repositories, but also to lead to better and more consistent practices across repositories run by non-governmental entities.

PRIM&R has long believed that harmonization of federal policies around research can be an important and effective means of supporting the conduct of responsible research, as long as it does not negatively affect the interests and welfare of research subjects. Harmonization can reduce policy redundancies that do little to add to research oversight and drain limited research resources, and can foster the consistent adoption of best practices. Harmonization of policies is clearly desirable in the data sharing and management space.

In 2018, we submitted comments in response to the National Institutes of Health (NIH)'s RFI on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research,<sup>1</sup> in which we expressed concerns about the proliferation of data repositories that follow a variety of rules and procedures. We pointed out that this has the potential to weaken the overall value of the data sharing enterprise. More recently, in January 2020, we submitted comments to the NIH on their Draft Policy for Data Management and Sharing,<sup>2</sup> in which we requested the NIH itself play a role in vetting grantees' proposed data repositories and sharing platforms to ensure they support the secure and ethical sharing of data.

The OSTP's proposed recommendations on repository governance issues are a welcome step in the right direction in terms of promoting harmonization of policies that both reduce burden and enhance responsible research. To that end, we hope the final document will include a strong recommendation that the Subcommittee on Open Science member agencies put language in their grants and contracts *explicitly* requesting adherence to this common set of desirable characteristics in data repositories. Such a move will amplify the benefits of harmonization, and, likely, the utility of the data sharing enterprise.

PRIM&R also appreciates that the draft acknowledges that there are important additional human subject protections considerations when the data repository involves human data, and that these considerations are relevant even if that data is deidentified. To that end, we support the draft's general language on privacy, but urge that as the OSTP further develops its common set of characteristics and considerations, or provides further guidance in this area, it include language about the need for repositories themselves to have in place mechanisms for preventing or discouraging reidentification of deidentified data, in addition to enforcing submitters' data use restrictions. PRIM&R has publicly commented on

---

<sup>1</sup> [Response to the National Institutes of Health's RFI on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research](#). (2018). Public Responsibility in Medicine and Research (PRIM&R).

<sup>2</sup> [Response to the National Institutes of Health's Draft NIH Policy for Data Management and Sharing and Supplement Draft Guidance](#). (2020). Public Responsibility in Medicine and Research (PRIM&R).

reidentification issues extensively and we would be happy to serve as a resource on this important topic if that is of interest.

PRIM&R for the most part endorses the OSTP's current list of "Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded or Supported Research" as appropriately comprehensive and flexible. Below we offer a few additional considerations we think might improve the two sets of desirable characteristics outlined in the draft:

Desirable Characteristics for All Data Repositories:

- We strongly urge the OSTP to add to the list of desirable characteristics that data repositories have a mechanism for ensuring credit for data generators. Giving those who generate data credit for their contributions to the scientific enterprise will incentivize researchers to share their data in the spirit of open science. We direct the OSTP to recently released expert recommendations on how data repositories can play a role in ensuring data generators receive credit for making their data available for future reuse.<sup>3</sup>

Additional Considerations for Repositories Storing Human Data (Even if De-Identified):

- We agree there should be "plans for addressing violations of terms-of-use by users and data mismanagement by the repository." These plans should construe "terms-of-use" as broadly as possible and explicitly include research service agreements. We would like to also note the government as a whole needs to reconsider what penalties should be levied if research subjects' rights are violated during the course of data sharing. It also needs to assess how to determine who should be held responsible for such violations. We believe limiting penalties to just a rescission of funding is likely to be insufficient and an inadequate deterrent to future bad actors.
- We believe the "Fidelity to Consent" consideration as written is likely to be inadequate as a guide for repository developers or those who are evaluating data management plans. We agree that researchers have an obligation to use data in a manner consistent with original consent, as a matter of respect for persons, and data repositories, as gatekeepers for such uses, should do their part to limit dataset access to uses consistent with consent. To that end, we urge OSTP to make clear that repositories that store human data have a responsibility to establish mechanisms for attaching permissions granted in the original consent, as machine-readable metadata, to the data itself.

Furthermore, we note that ensuring that future uses of data are consistent with consent may not always be straightforward. It is not clear what it means to be faithful to consent when, for example, (1) the original consent was silent regarding

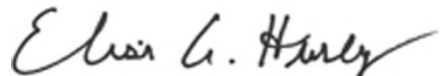
---

<sup>3</sup> [Credit Data Generators for Data Reuse](#), Pierce, H., Dev, A., Statham, E., & Bierer, B. (2019). *Nature*.

whether and in what ways data would be shared or used in the future, or (2) the original consent promised that the data that is stored and shared would remain deidentified, when today's technologies and methodologies, including the aggregation of data sets, make permanent deidentification impossible. Given these complexities, we suggest future policies on this important topic provide additional guidance, perhaps including examples, about what fidelity to consent means or entails in these sorts of circumstances.

Thank you again for the opportunity to comment and for the OSTP's work on this important issue. We hope our comments on the current draft will be useful in your next stage of policymaking in this area. PRIM&R stands ready to provide any further assistance or input that might be useful. Please feel free to contact me at 617.303.1872 or [ehurley@primr.org](mailto:ehurley@primr.org).

Respectfully submitted,

A handwritten signature in cursive script that reads "Elisa A. Hurley".

Elisa A. Hurley, PhD  
Executive Director

cc: PRIM&R Public Policy Committee, PRIM&R Board of Directors



Arizona State University

Tempe, AZ 85281

[www.asu.edu](http://www.asu.edu)

---

Date: March 2 2020

Lisa Nichols, Ph.D.  
Assistant Director for Academic Engagement  
Office of Science and Technology Policy  
Executive Office of the President  
(202) 881-9943  
[Lisa.M.Nichols@ostp.eop.gov](mailto:Lisa.M.Nichols@ostp.eop.gov)

Submitted online to: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

Dear Dr. Nichols,

Arizona State University appreciates the opportunity to respond to Request for Comment (RFC) 85 FR 3085 seeking comments on the Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research. We are pleased to submit these comments for your consideration.

Sincerely,

Philip Tarrant  
Research Data Management Officer  
Arizona State University, Tempe, AZ 85281  
(480) 727-7860  
[philip.tarrant@asu.edu](mailto:philip.tarrant@asu.edu)

## 1. DESIRABLE CHARACTERISTICS FOR ALL DATA REPOSITORIES

**A. Persistent Unique Identifiers:** Assigns datasets a citable, persistent unique identifier (PUIID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The PUIID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.

**ASU supports this position and has no additional comments.**

**B. Long-term sustainability:** Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.

**ASU supports this position and has no additional comments.**



**C. Metadata:** Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.

**ASU Comment:** While we commend the goal of many data repositories to reduce the metadata load on researchers submitting data, the reality is that a stellar dataset without adequate metadata may be unusable by researchers not involved in the original research. Therefore, we recommend that the language of any policy should clearly define what “sufficient” means. For example, a data table containing growth data collected on 5/12/2016 may be interpreted differently (December 5<sup>th</sup> or May 12<sup>th</sup>) depending on the consumer’s location unless the date format MM/DD/YYYY is included in the metadata. Currently, this granularity of metadata is rarely expected for datasets submitted to many repositories.

**D. Curation & Quality Assurance:** Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

**ASU supports this position and has no additional comments.**

**E. Access:** Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.

**ASU Comment:** The desire to provide open access with minimal restriction is understandable. However, if we interpret open access to mean anonymous downloads then we lose the link between the dataset and the consumer. At a minimum a dataset downloader should be encouraged to provide some contact information. We are not suggesting that account creation be required, but a download form could include a name and email address. At this point the repository should also request permission to follow up later to seek a small amount of usage information. This contact information would enrich repository metrics and could be used for communications. Retaining this link with the data consumer permits several actions: 1) follow up to see if/how the data were used in the consumer’s research, 2) request feedback regarding any quality issues noted with the data, and 3) follow up to remind the consumer of their responsibilities with respect to citation of datasets in any publications.

**F. Free & Easy to Access and Reuse:** Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

**ASU Comment:** Data re-use comes with responsibilities to interpret the data as intended and cite it appropriately. Data repositories should encourage data consumers to properly consider the context of the data they are re-using and ensure it is congruent with their usage. Recommended citation text should be provided at point of download or within the dataset metadata.

**G. Reuse:** Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).

**ASU Comment:** Metrics will be important for managing usage, quality, customer satisfaction and return on investment, but only if the correct measurement data are collected. Retaining a complete transaction record from submission to download to usage will be the only way to ensure the repository has end-to-end metrics. Metadata and PUIDs alone will not ensure that datasets are tracked and correctly cited.

---

Tracking where datasets go (see Section E comment) will provide a better opportunity for tracking their re-use.

**H. Secure:** Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>) or the National Institute of Standards and Technology's 800-53 controls (<https://nvd.nist.gov/800-53>).

**ASU supports this position and has no additional comments.**

**I. Privacy:** Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.

**ASU supports this position and has no additional comments.**

**J. Common Format:** Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.

**ASU supports this position and has no additional comments.**

**K. Provenance:** Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.

**ASU supports this position and has no additional comments.**

## II. ADDITIONAL CONSIDERATIONS FOR REPOSITORIES STORING HUMAN DATA (EVEN IF DE-IDENTIFIED)

**A. Fidelity to Consent:** Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).

**ASU supports this position and has no additional comments.**

**B. Restricted Use Compliant:** Enforces submitters' data use restrictions, such as preventing reidentification or redistribution to unauthorized users.

**ASU supports this position and has no additional comments.**

**C. Privacy:** Implements and provides documentation of security techniques appropriate for human subjects' data to protect from inappropriate access.

**ASU supports this position and has no additional comments.**

**D. Plan for Breach:** Has security measures that include a data breach response plan.

**ASU supports this position and has no additional comments.**

**E. Download Control:** Controls and audits access to and download of datasets.

**ASU supports this position and has no additional comments.**

**F. Clear Use Guidance:** Provides accompanying documentation describing restrictions on dataset access and use.

**ASU supports this position and has no additional comments.**

**G. Retention Guidelines:** Provides documentation on its guidelines for data retention.

**ASU supports this position and has no additional comments.**

**H. Violations:** Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository.

**ASU supports this position and has no additional comments.**

**I. Request Review:** Has an established data access review or oversight group responsible for reviewing data use requests.

**ASU supports this position and has no additional comments.**

Comments on USA “Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded or Supported Research”

D Carlson

Senior Chief Editor (and co-founder)  
Earth System Science Data journal

Overall, a useful if somewhat confused set of guidelines without practical impact. Without enforcement via funding, and if trying to accommodate all possible variants and objections, the thus-neutered recommendations will have nearly zero impact. Fundamentally, unless data providers find easy, free and useful services, recommendations to data centers will prove ineffective. In the USA for example, while NSF has for many decades encouraged university researchers to NOT apply 2-year proprietary periods, until or unless funded researchers find easy alternatives with clear benefits, researchers will always revert to known (protectionist) patterns of behavior.

- A. Persistent Identifiers - essential. Not possible to maintain data sharing or data tracking without, e.g. DOI. Providers and data archive centers must adhere to full DOI requirements for clarity, anonymity (e.g. a DOI should not include journal or institutional name although most do), version control etc. DOI currently used as a convenient label to avoid serious data archive responsibilities. Mandatory federal requirements for permanent data identification as a condition of all funding could go a long way.
- B. Long-term sustainability. Define ‘long-term’! At ESSD we suggest 10 years. Climate records would require 50 years or longer? Related to permanent identifier issue above: if a data center goes out of business (as happens too often for US-based data centers), data protected by a DOI can move easily and transparently to a replacement data archive.
- C. Metadata. Some communities develop and support elaborate useful schemes, other communities have not a clue. Skill and capabilities inversely related to frequency of and need for real-time exchange and access. Any requirements need to link to data distribution patterns.
- D. Curation and Quality Assurance. Although data centers like to claim this function, even very specialized (e.g. serving a narrow discipline) data centers fail. Solution lies in data publication where the journal itself works with multiple data centers and authors/data provider will choose generalist vs. specialist vs. completely agnostic data archive services based on ease of use, speed of service, registration requirements, etc. Quality derives from peer review, not from data service. Curation capabilities need to compete for customers based on ease and usefulness of services. Mandatory data repository requirements, e.g. to date center formerly called NODC for most past oceanographic data, have largely failed.
- E. Access. Remains the most pervasive, most persistent barrier to free and open exchange. Providers often want to ‘protect’ their data for a variety of reasons. Data centers often hide behind registration steps and user-ID tracking systems, ostensibly to meet funding requirements. Without an independent third party - data journals, for example - to provide initial access checks and access follow-up (as both ESSD and Nature’s Scientific Data perform), authors and data centers will maintain limits and barriers forever.
- F. Free and easy access and reuse. Not currently honored or tracked by most data centers. Free perhaps, but with a serious list of conditions. Embargoes, proprietary exclusion periods, share-alike license requirements. Unless some (again) third party entity promulgates and enforces true free unrestricted access, data centers will protect as often and as much as possible. Relates to licenses, data provider expectations, and data center (and national and funding agency) policies.
- G. Reuse - tracking of reuse as currently practiced in most cases violates user anonymity. Data centers and tracking organizations, which could build much better tracking algorithms based on permanent identifiers (as some have) continue instead to rely on user emails.

When reuse serves as a qualifying metric for continued data center funding, as too often happens, reuse tracking based on user information becomes accepted and expected cost of 'doing business'.

- H. Secure as described here "unauthorized access or release of data" violates every free open access standard above. Free and open access to data means access sans authorization and unrestricted release. Does not, can not and should not apply to any data except in the case of pre-agreed confidential human data.
- I. Privacy. Not relevant to most earth system science data. Too often offered as an excuse for not developing truly-anonymous identification and use services while simultaneously promoted (e.g. via EU GDPR) as a window-dressing solution to actual serious privacy issues. Institutions often proclaim GDPR adherence will simultaneously requiring user ID for product access. Essentially: give us your email address but we promise we will not share it onward. How does that build trust?
- J. Common format. Easy to request, almost impossible to implement. .csv but not Excel? R codes but not MatLab libraries? netCDF provided useful services for a time period but now big (TB) data sources present new challenges. Google Earth Engine or other competing access services leave the old days of common formats far behind. The concept as written does not reflect current reality nor keep up with present data trends.
- K. Provenance - should rather form a subset of permanent identifiers? A DOI, for example, if properly applied and adhered to, provides excellent version control. As the CDIAC example (prominent data center at ORNL that closed) shows, a valid DOI can ensure provenance, but not vice versa. For 'living' data (data updated on regular schedule) the issue becomes one of simultaneous backward and forward compatibility/traceability. Links to all prior versions should lead users seamlessly to current version, while current version should provide adequate links to all prior versions? Provenance = abstract term without practical application.

This response to the White House Office of Science and Technology Policy’s “Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research” is submitted on behalf of the Open Research Funders Group. The Open Research Funders Group (ORFG) is a partnership of 16 philanthropic organizations committed to the open sharing of research outputs. We believe this benefits society by accelerating the pace of discovery, reducing information-sharing gaps, encouraging innovation, and promoting reproducibility. The ORFG engages a range of stakeholders to develop actionable principles and policies that promote greater dissemination, transparency, replicability, and reuse of papers, data, and a range of other research types. Our current roster of member organizations includes the Alfred P. Sloan Foundation, the American Heart Association, the Arcadia Fund, the Bill & Melinda Gates Foundation, the Eric & Wendy Schmidt Fund for Strategic Innovation, the Gordon and Betty Moore Foundation, Howard Hughes Medical Institute, the James S. McDonnell Foundation, the John Templeton Foundation, Arnold Ventures, the Leona M. and Harry B. Helmsley Charitable Trust, the Lumina Foundation, Open Society Foundations, Templeton World Charity Foundation, the Robert Wood Johnson Foundation, and the Wellcome Trust. Collectively, the ORFG members hold assets in excess of \$100 billion, with total annual giving in the \$10 billion range. Members’ interests range the entirety of the disciplinary spectrum, including life sciences, physical sciences, social sciences, and the humanities. This response has been prepared by Greg Tananbaum, the chief administrator of the Open Research Funders Group, in conjunction with representatives of the ORFG membership.

The Open Research Funders Group is supportive of the White House Office of Science and Technology Policy’s commitment to advance open science and foster implementation of agency Public Access Plans. Identifying best practices for the long-term preservation of data from Federally funded research is a critical component of these efforts. The ORFG is pleased to provide succinct input to the OSTP regarding desirable characteristics of data repositories. These recommendations are drawn from both the direct experience of our members, many of whom have open data policies for the research they fund, and our engagement with the broader scientific community.

Federal grant recipients should, first and foremost, be expected to deposit their data in a data environment that supports the FAIR data sharing principles - findable, accessible, interoperable, and reusable. The FAIR principles are at the core of the open data and reproducibility movement. Any repository housing Federally supported data should clearly and publicly articulate how it conforms to the core components of FAIR:

### **Findable**

- (Meta)data are assigned a globally unique and persistent identifier
- Data are described with rich metadata (defined by R1 below)
- Metadata clearly and explicitly include the identifier of the data they describe
- (Meta)data are registered or indexed in a searchable resource

### **Accessible**

- (Meta)data are retrievable by their identifier using a standardized communications protocol
  - The protocol is open, free, and universally implementable
  - The protocol allows for an authentication and authorization procedure, where necessary
- Metadata are accessible, even when the data are no longer available

### **Interoperable**

- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- (Meta)data use vocabularies that follow FAIR principles
- (Meta)data include qualified references to other (meta)data

### **Reusable**

- (Meta)data are richly described with a plurality of accurate and relevant attributes
  - (Meta)data are released with a clear and accessible data usage license
  - (Meta)data are associated with detailed provenance
  - (Meta)data meet domain-relevant community standards

Given the wide range of projects funded by Federal agencies, no single repository will be universally applicable to house all funded datasets. Instead, the ORFG recommends that Federal agencies should provide grant recipients with a degree of latitude in selecting the most appropriate repository to house their research data. In order for Federally funded research to reach their widest audience and have their deepest impact, these data should be deposited in repositories with clear and explicit guidance along the following dimensions, over and above the FAIR components articulated above:

- **Re-Use.** The repository must allow any interested party to freely access the data without restriction on research reuse, using a CC0 or similar license. This should be codified in the repository's terms of service.

- **Security.** The repository must describe how datasets are stored and protected from vulnerabilities such as credentials theft or hacking. For any data that require gatekeeping on human subject protection or similar grounds, the repository must describe how this information is accessed and protected.
- **Stability.** The repository must have a clearly articulated funding mechanism or business plan to provide reasonable assurances that the data will be available for the indefinite future. It should also have a continuity plan addressing what will happen to the data in the event the repository is discontinued.
- **Fee Structure.** Any costs associated with data deposit and data maintenance must be clearly articulated. This includes details about whether fees are one-time or recurring, as well as how the size of the dataset may impact the cost. The repository must make these costs structures publicly available without restriction.
- **Subject Focus.** There are hundreds of domain-specific repositories in operation at this writing. In general, grant recipients should be encouraged to deposit their data in a repository that is appropriate for the subject matter in question. Further, if a repository consistent with the considerations articulated in this document has emerged within a specific research community as the default resource in that field (e.g., GenBank for DNA sequences), grant recipients should, as a general rule, be encouraged utilize that repository. This optimizes the ability of others to discover and build upon the data.
- **Metadata.** The repository must require a depositor to provide sufficient metadata provided to enable the dataset to be used by others. These metadata should be searchable so that repository visitors can easily discover appropriate datasets.
- **File Formats.** The repository should be able to accommodate all aspects of the grant recipients' dataset, regardless of file type and size.
- **Machine Extraction.** The data stored in the repository should be available in a machine-readable and machine-interpretable format, preferably via API (Application Programming Interface). This will encourage text and data mining, meta-analysis, and information extraction, and additional knowledge discovery.

The Open Research Funders Group appreciates the opportunity to comment on this project, and we are eager to assist in its eventual rollout.



Comment on the Draft Desirable Characteristics of Repositories to Consider for Managing and Sharing Data Resulting from Federally Funded or Supported Research

From: Ben Heavner,

Affiliations: Member of the TOPMed Data Coordinating Center, University of Washington

Department of Biostatistics

Primary Scientific Disciplines: Life Sciences

Role: Researcher, Data Coordinator

Comments:

I am offering these comments on behalf of the members of the TOPMed Data Coordinating Center.

With regard to Section I of the Draft Desirable Characteristics of Repositories to Consider for Managing and Sharing Data Resulting from Federally Funded or Supported Research, we suggest that an additional desirable characteristic should be added addressing the desirability of tools to facilitate data deposition in any repository. Such tools could include software APIs, documentation, standardized submittal methods or portals, or other tools aimed to make it easier to submit data to a repository.

With regard to Section II of the Draft Desirable Characteristics of Repositories to Consider for Managing and Sharing Data Resulting from Federally Funded or Supported Research, we note that the “Fidelity to Consent” guidance of “consistent with original consent” is insufficient since research participants may update their consent. Therefore, it would be desirable for a repository to have capabilities for data providers to revise consent (and the associated dataset access controls).

# Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research

Scout Calvert and Shawn Nicholson  
MSU Libraries

## Preliminary

The Background section appropriately describes the current research data repository context. It is helpful that this set of characteristics is delimited; this allows for a variety of worthy solutions to be developed and tested. It references key developments in research data sharing, including community developed standards (FAIR principles). It anticipates that periodic change may be needed. My feedback here is that this is a quickly developing area that may be energized by the identification of desirable characteristics, so it would be helpful to plan for the first review and update in the relatively near future (five years maximum).

## I. Desirable Characteristics for All Data Repositories

### A. *Persistent Unique Identifiers:*

This is essential. The suggestion for a landing page is very helpful and easy to implement; it should be standard. I'd add, that forward-looking repositories will implement unique identifier systems for researchers (e.g., ORCID), organizations (e.g., <https://ror.org/>), and data types (e.g., <http://www.typeregistry.org/registrar/#>). These may count as reasonable and achievable "desirable characteristics for all data repositories" by the time these characteristics are implemented.

### B. *Long-term sustainability:*

This is admirable and desirable, but should be bounded for now with anticipation of further developments in the automated management of data that will make sustainability more consistently achievable. Perhaps for this category, and others, transparency of the plan (availability for review) will be a necessary component for long-term sustainability. A couple questions confound implementing this characteristic:

What is meant by long-term? This varies by discipline and data type. Perhaps this can be handled by a data repository declaring its definition of "long-term" for its disciplinary context. Ten years is probably a good minimum for most kinds of data; some repositories (e.g., social sciences data, ecological data) can be expected to plan for a much longer sustainability horizon.

What is meant by integrity? Authenticity and availability of datasets can be achieved via machine processes. Integrity is more difficult, depending on the definition. If this means bit-level integrity and checksum processes, that's a reasonable standard. But some data

requires forward migration or export from proprietary formats to maintain long-term integrity. Datasets that live in a custom database may be difficult to sustain. One way a repository might handle this is to not accept data in proprietary formats or in databases, which would jeopardize those data.

A contingency plan should include a succession plan: a named, trustworthy organization that has agreed to accept the data should the repository be decommissioned.

#### *C. Metadata:*

I would include more metadata than this: metadata about the repository, researchers and their affiliations (see under item A), provenance, as well as any metadata automatically generated at ingest that can assist the dataset in being discoverable and usable by machines.

#### *D. Curation & Quality Assurance:*

This is a genuinely sticky requirement. It is desirable and possible, but represents potentially very high labor costs. If a repository provides such a mechanism (such as external review or curation), but the mechanism is not required for use of the repository, that could help mitigate those costs (but would get around the spirit of the guideline). Such a mechanism could potentially allow or encourage other ways of providing peer review and curation of datasets, providing that other efforts at treating data as a first class research object (primarily incentive structures) continue to build support.

If, in the future, repository infrastructure is developed that allows machine review of heterogeneous datasets in the repository, that could assist in curation and quality assurance, but until then, this could be a cost-prohibitive proposition for repositories that could block their development.

#### *E. Access:*

This is desirable and achievable. Major general repositories (e.g., Dataverse, Zenodo) already provide open access to datasets with clear licensing regimes. Privacy and confidentiality are more challenging, and presently rely on the good intentions of the uploader, with expectations varying across disciplines and national contexts. Perhaps some computational review of datasets (as suggested for curation and quality assurance) as part of ingest could detect information likely to be identifying, but so far as I know this is not implemented anywhere. One straightforward solution would be to ensure the development of one or more repositories that specialize in light touch curation of human subjects datasets; reducing labor costs would reduce the cost of use and decrease the temptation to upload poorly de-identified data to a free general repository. If machine actionable DMPs continue their development trajectory, IRB

could trigger the selection of an appropriate repository with the expertise and affordances to handle this data.

*F. Free & Easy to Access and Reuse:*

Desirable to an extent, and many examples of this exist. It depends on what “broadest possible” implies. Some data may not be ethically sharable with a “public domain” designation; researchers may feel uncomfortable sharing their data for commercial purposes. Perhaps some nuance about how a repository can provide and support a variety of “open” licenses to encourage sharing and reuse, not just public domain.

*G. Reuse:*

Common repository frameworks already enable this and the collection of associated metrics. This should be encouraged at every opportunity, and processes for more reliable tracking data reuse should continue to be developed.

*H. Secure:*

Adherence to security criteria is desirable. I don’t know enough about either of these standards to know if they are the right ones for a typical data repository. There should be a security protocol, but it’s not likely to be one size fits all. A standard for climate science (which may be targeted purposefully and maliciously) and human subjects data might not be the right standard for other disciplines or data types and could discourage repository development.

If a specific standard is adopted, I would encourage consultation with the repository community before specifying.

*I. Privacy:*

This is reasonable. Is the expectation that this should be available on the website of the repository for inspection by potential depositors? How will a typical depositor be able to assess whether these are the right safeguards? Or will this be left for specific funding agencies to determine?

*J. Common Format:*

This is desirable but difficult to achieve without researcher participation or additional labor costs. At some point, this may become common computationally, through curation-at-rest processes. It is desirable for repositories to encourage non-proprietary formats but mandating them may mean some data are never deposited.

*K. Provenance:*

This is desirable and automatable. Perhaps extremely large, continuously changing datasets could present a problem, depending on the granularity required in the logfile.

## **II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

### *A. Fidelity to Consent:*

This is desirable and will require an additional layer of human labor. Perhaps in the future some aspects can be automated through credentialing, though that displaces the labor to other places in the hopes of reducing error and labor costs down the road. But this is not an onerous recommendation for repositories charged with storing human subjects data. Making such repositories common and affordable will increase compliance and reduce the temptation to improperly store and share human data.

### *B. Restricted Use Compliant:*

I am not sure how this can be consistently implemented. Perhaps through analysis-in-place or only accepting data that doesn't have these parameters. Data use agreements can aid in this, but it's unclear if they have genuine power to "enforce" or "prevent."

### *C. Privacy:*

Desirable. However, for human subjects data with identifiers, additional specifications may be necessary.

### *D. Plan for Breach:*

### *E. Download Control:*

### *F. Clear Use Guidance:*

### *G. Retention Guidelines:*

All desirable and achievable, and in some cases necessary.

*H. Violations:* Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository.

The first I understand; the second sounds very much like conflict of interest.

*I. Request Review:* Has an established data access review or oversight group responsible for reviewing data use requests.

Desirable, achievable, and necessary.

## **Additional considerations.**

The absence of “succession plan” under Long-term sustainability is a major omission. This is a desirable characteristic that should include some description of what would count.

With specifications and solutions for machine-actional DMPs in development, it would be helpful to anticipate desirable characteristics for repositories that would allow them to be part of maDMP implementation. This is partly a problem of metadata and persistent identifiers described above (ORCID, <https://ror.org/>, <http://www.typeregistry.org/registrar/#>) and partly a problem of APIs, which aren't mentioned. Perhaps a desirable characteristic of all data repositories is to have a general expressed intention to work toward FAIR data principles and toward machine-actionable repository features.

Unique data types. These desirable characteristics appear to leave enough leeway to accommodate unique datasets.

It would also be helpful if this provided some clarity on both the term “archive” and the term “data repository.” By “archive” researchers often mean some place to put data that they are no longer using in order that it should be able to be consulted in case of any questions about the research, or data that they are obligated to keep even though it has no research value to them. Data archivists mean almost exactly the opposite: data that is so valuable and irreplaceable that it must be carefully curated, described, forward migrated, preserved, and protected, even given the substantial costs of curation (e.g., NACJD). Repository may be used by researchers to mean a place just to store data where sharing is incidental (e.g., a department server). For data stewards, a repository is where to put data to ensure it is available for sharing that is as frictionless as possible given the data itself and other constraints, and where it can be counted on to be available for referencing in the future, for some undefined though not necessarily extremely long period.

There's a lot of mileage to be had out of simple transparency. The information these characteristics describe should be available on repository websites to enable and encourage researchers to make informed choices about repositories.

## COAR / SPARC Response to the OSTP Draft Desirable Characteristics of Repositories Managing Data

March 3, 2020

COAR and SPARC thank OSTP for the opportunity to provide feedback to the [Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research](#). Good data management is critical for ensuring validation, transparency of research findings, as well as to maximize impact and value of publicly-funded research through data reuse.

Repositories provide crucial services that manage and provide access to data, articles, and a wide array of other types of scholarly content and are essential community tools for good data management. As we seek to expand national and international capacities to support research data management, we need to make sure that repositories are using best practices for managing data, while at the same time ensuring that requirements are not so overly onerous that they result in excluding a large number of repositories.

Our general comments related to the current draft characteristics are as follows:

- In general, we agree with many of the proposed characteristics, but suggest that they be reorganized in order to distinguish between (1) the objectives of the policy (access, integrity, etc.) followed by (2) the specific practices (metadata, licenses, etc.) that support each objective. In addition, it would be useful if the policy could include a core set of the most essential characteristics, while also pointing to desirable characteristics, that could assist repositories in improving their practices over time.
- In order to support the international nature of research, it is important to ensure that data are interoperable across jurisdictions. We strongly encourage the OSTP to align policy requirements where possible with other countries and regions.
- The current repository landscape includes both domain and general purpose repositories. An implicit assumption in the current OSTP draft seems to be that all data repositories are domain repositories. General repositories (most often managed by university libraries) play a critical role by providing sustainable and long lived services for data management for those researchers who do not have access to an appropriate domain repository, and we would encourage OSTP to explicitly support both types of repositories.
- In some cases, the characteristics proposed in the draft would fall under the responsibility of the data creators/providers (access and reuse rights, data format), making it difficult, if not impossible, for repositories to enforce these in the context of the repository.
- And finally, because this is a rapidly evolving landscape, and technology and standards for data management will surely change over time, it will be important for OSTP to review and update these characteristics regularly. Providing guidance on an update schedule and process would be useful.

With these comments in mind, we propose the following framework for the most essential characteristics of data repositories. Our proposal is based on input from the repository community in the US and elsewhere, and with consideration to the current recommended characteristics outlined in a number of other contexts: [Data Citation Roadmap for scholarly data repositories](#), [Core Trust Seal](#), [FAIR data principles](#), [PLOS “Criteria that Matter”](#), [TRUST](#), and [COAR Next Generation Repositories Technologies](#).

We have not included “highly desirable” or “nice to have” criteria in this submission. However, COAR is in the process of developing an internationally-vetted assessment framework for repositories with several levels of compliance in the coming months and would be happy to share this with OSTP once it is developed.

Following the framework, we also provide specific comments related to the current draft characteristics published by OSTP.

## **About COAR and SPARC**

[COAR](#) is an international association with over 150 members and partners from around the world representing libraries, universities, research institutions, government funders and others. COAR brings together individual repositories and repository networks in order to build capacity, align policies and practices, and act as a global voice for the repository community.

[SPARC](#) is a coalition of 240+ libraries in the U.S. and Canada that works to enable the open sharing of research outputs and educational materials in order to democratize access to knowledge, accelerate discovery, and increase the return on our investment in research and education.

## **For more information, please contact:**

Kathleen Shearer, Executive Director, COAR: [kathleen.shearer@coar-repositories.org](mailto:kathleen.shearer@coar-repositories.org)  
Heather Joseph, Executive Director, SPARC: [heather@sparcopen.org](mailto:heather@sparcopen.org)



## Essential Characteristics for Repositories Managing Research Data Framework

Objective	Essential Characteristics
<b>Discoverability of data</b>	<ul style="list-style-type: none"> <li>● High quality metadata (discipline-based or general metadata schema, e.g. Datacite or Dublin Core metadata) with an OAI-PMH feed</li> <li>● Repository has well documented APIs</li> <li>● Repository assigns a citable, persistent unique and universal identifier (PUID) that points to the landing page of the dataset<sup>1</sup> (even in cases where data is no longer available or data is not available for security purposes)</li> </ul>
<b>Equitable, free and ongoing access to data</b>	<ul style="list-style-type: none"> <li>● There is no cost to the user for accessing data once it is published</li> <li>● Repository ensures ongoing access to data for a publicly stated time frame</li> <li>● Repository has a contingency plan to ensure data are available and maintained during and after unforeseen events</li> </ul>
<b>Reuse of data</b>	<ul style="list-style-type: none"> <li>● Repository supports the use of machine readable licenses (e.g. Creative Commons Licenses)</li> <li>● Repository provides citable PUIDs<sup>2</sup></li> </ul>
<b>Data integrity and authenticity</b>	<ul style="list-style-type: none"> <li>● Repository provides information about data provider(s) including contact information of the person(s) responsible for the data</li> <li>● Repository provides a record of all changes to metadata and data in the repository</li> <li>● Repository provides documentation of its practices that prevent unauthorized access/manipulation of data</li> </ul>
<b>Quality assurance</b>	<ul style="list-style-type: none"> <li>● Repository undertakes basic curation of metadata and data<sup>3</sup></li> <li>● Repository provides documentation about what curation processes are applied to the data and metadata</li> </ul>

<sup>1</sup> Many existing repositories use Handles as persistent identifiers, so these should be admissible.

<sup>2</sup> A citable PUID would involve the persistent identifier expressed as an URL resolving to a landing page specific for that dataset, and that landing page must contain machine readable metadata describing the dataset. We recommend the use of [signposting](#) protocol to support this.

<sup>3</sup> As defined by the CORE Seal of Approval, basic level of curation involves brief checking and addition of basic metadata or documentation where needed.

<p><b>Privacy of sensitive data (e.g. human subjects, etc.)</b></p>	<ul style="list-style-type: none"> <li>● In cases where the repository is collecting sensitive research data, the repository provides tiered access based on the different levels of security requirements of data</li> <li>● In cases where the repository is collecting sensitive research data, the repository has mechanisms that allow data owners to limit access to authorized users only</li> </ul>
<p><b>Sustainability and preservation</b></p>	<ul style="list-style-type: none"> <li>● Repository (or organization that manages repository) has a long term plan for managing and funding the data repository</li> <li>● Repository has a public data retention policy that defines the duration of time the data will be preserved and documentation about preservation practices</li> </ul>
<p><b>Other</b></p>	<ul style="list-style-type: none"> <li>● Repository has a contact point or helpdesk to assist data depositors and data users</li> <li>● Repository provides documentation about the scope of data accepted into the repository</li> </ul>

# I. Desirable Characteristics for All Data Repositories

*Our specific responses/comments to each element are provided in the blue text below.*

**A. Persistent Unique Identifiers:** Assigns datasets a citable, persistent unique identifier (PUID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The PUID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.

*We agree with this requirement, which should be agnostic in terms of type of PUID used.*

**B. Long-term sustainability:** Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.

*This section is currently a mix of requirements, (preservation practices, sustainability of operations, emergency planning). We suggest these be disambiguated into two objectives: (1) sustainability and preservation, and (2) ongoing access.*

**C. Metadata:** Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.

*We agree that quality and comprehensive metadata is required to support a number of objectives (discovery, citation, reuse, and preservation). Metadata requirements may be different for each of these objectives, and it would be valuable to outline the distinct requirements for each objective. In addition, while some domains already have well developed standards for metadata, others do not. Therefore, we suggest a reference to general purpose metadata standards is also acceptable (e.g. DataCite Metadata Schema or Dublin Core)*

**D. Curation & Quality Assurance:** Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

*We agree that a basic level of curation for both metadata and data should be a requirement, but more extensive curation to data will often need to be undertaken by the data creators and/or data curator(s). We suggest a requirement of basic curation at the repository, and a recommendation for the repository to support more extensive data curation by the creators and/or curators.*

**E. Access:** Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.

*This is an objective; we suggest that you update this to include specific requirements related to this including open free access, continuous availability, and open APIs.*

**F. Free & Easy to Access and Reuse:** Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

*There may be cases when researchers wish to deposit and share their data within the research team, and some repositories can support this requirement. Therefore, we suggest this is reworded to, "There is no cost for the user to access the data once it is published."*

**G. Reuse:** Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).

*There are three main requirements needed to support reuse: citation metadata, permanent unique identifiers, and the use of machine readable, standardized licenses. We suggest that you include all of these as requirements to support data reuse.*

**H. Secure:** Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>) or the National Institute of Standards and Technology's 800-53 controls (<https://nvd.nist.gov/800-53>).

*This issue is really related to data integrity, as non-sensitive data will be freely accessible. We suggest that this is reworded as follows, "Repository provides documentation of its practices that prevent unauthorized access/manipulation of data". In addition, there are several other requirements needed for data integrity: documentation of provenance, and versioning/changes to data. We suggest you also list these elements.*

**I. Privacy:** Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.

*There are repositories that collect exclusively data that will be made openly available. This requirement should be clarified, "In cases where the repository is collecting sensitive data, it will provide documentation related to the safeguards in place to protect data from access breaches."*

**J. Common Format:** Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.

*Although repositories can recommend formats, it is the data creators that determine the format of the data they collect. We suggest that this is a responsibility of the researchers and data creators and that this should be a requirement included in a data management plan.*

**K. Provenance:** Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.

*Provenance of data is important for data integrity and assurance, and we agree that this is an important requirement. However, we suggest the terminology be changed from "logfile" to "record of changes."*

## **II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

*In terms of storing human data (or other sensitive data), it is the responsibility of the researcher to ensure that access conditions reflect consent and ensure that human data is appropriately de-identified. The role of the repository may be to support a variety of access levels (including restricting access to authorized users) and adopt practices that ensure secure management of data. It should be noted that not all repositories collect sensitive data.*

*Additionally, not all restricted/sensitive data need to be treated the same way by the repository, and in some cases, it is important that they are not treated the same. Therefore, tiered access to data is something that should be supported by repositories collecting sensitive data.*

- A. *Fidelity to Consent*: Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).
- B. *Restricted Use Compliant*: Enforces submitters' data use restrictions, such as preventing reidentification or redistribution to unauthorized users.
- C. *Privacy*: Implements and provides documentation of security techniques appropriate for human subjects' data to protect from inappropriate access.
- D. *Plan for Breach*: Has security measures that include a data breach response plan.
- E. *Download Control*: Controls and audits access to and download of datasets.
- F. *Clear Use Guidance*: Provides accompanying documentation describing restrictions on dataset access and use.
- G. *Retention Guidelines*: Provides documentation on its guidelines for data retention.
- H. *Violations*: Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository.
- I. *Request Review*: Has an established data access review or oversight group responsible for reviewing data use requests.

## RFC Response: Desirable Repository Characteristics

### ORCID as a Desirable Characteristic of Repositories

As a member of the persistent identifier (PID) community, [ORCID](#) strongly supports characteristic [1A: Persistent Unique Identifiers](#). In order to optimize public access and realize FAIR principles, we suggest expanding use of PIDs to include person identifiers, specifically:

- Minimally, metadata for each dataset include an open persistent unique person identifier for the primary creator of the dataset.
- Ideally, metadata should include open persistent unique person identifiers for all contributors to the dataset.

Person identifiers enhance visibility of and access to datasets by enabling machine-readable connections between datasets and researchers who contributed to their creation. This is particularly valuable in a networked repository ecosystem, where a dataset may physically reside in a location separate from a public web interface that provides access to it.

Person identifiers also play a role in operationalizing FAIR data principles, particularly:

- [I3: \(Meta\)data include qualified references to other \(meta\)data](#) Person identifiers allow establishing machine-readable connections between datasets whose metadata contain the same person identifiers.
- [R1.2: \(Meta\)data are associated with detailed provenance](#) Person identifiers allow authoritatively attributing datasets (or actions taken on datasets) to individuals, regardless of name duplication, variation or change over time.

While several person identifier systems exist, we recommend a non-proprietary system such as ORCID. A non-profit with a community of over 8 million users and 1,000 organizational members (including 7 US government agencies), ORCID has become a de facto global standard. The National Academies of Sciences, Engineering, and Medicine characterize ORCID as an “enabling technology” in [Open Science by Design: Realizing a Vision for 21st Century Research](#). [COAR](#) includes ORCID in its [Recommendations for Next Generation Repositories](#). Finally, ORCID is an active participant in the repository community; ORCID recently convened a [task force](#) of global leaders in the repository community, which published its [Recommendations for supporting ORCID in repositories](#) in 2019.

Thank you for considering our feedback.

Contact: Liz Krznarich, Tech Lead, New Projects, ORCID [e.krznarich@orcid.org](mailto:e.krznarich@orcid.org)

# Comments for Office of Science and Technology Policy (OSTP)

## Contributors:

Megan Potterbusch, George Washington University, Libraries and Information Science, Data Services Librarian, email: mpotterbusch@gwu.edu

Ann Myatt James, George Washington University, Human Geography, Data Services Librarian, email: ajames31@gwu.edu

We submit for consideration our comments, recommendations, and suggestions regarding the draft set of desirable characteristics of data repositories used to locate, manage, share, and use data resulting from Federally funded research [FR Doc. 2020-00689 Filed 1-16-20; 8:45am].

We've organized our remarks in accordance with the sections numbered I and II and alphabetically listed subsections. Each of our comments are outlined as follows:

## Section I

- A. Would recommend a PUID landing page to also contain metadata.
- B. This subsection suggests a lot of implied effort but without many specific details. For example, we would suggest definitions be provided for terms like "integrity" and "stable technical infrastructure" and what counts as an "unforeseen event". We're also curious to know if the concept of long-term sustainability would suggest the need for an emergency preparedness strategy or if this is primarily a financially-oriented use of the term.
- C. Does this requirement for metadata include mandatory inclusion of data dictionaries/codebooks along-side data that are deposited? If so, this bares specifying.
- D. This subsection looks great!
- E. Consider including language in this subsection to ensure repositories are providing open access to datasets in ways that respect cultural integrity or a similar concept. Ensuring datasets and research are handled in culturally appropriate ways will be especially important aspects for consideration in projects that have been created in collaboration with native peoples and/or historically marginalized communities.
- F. It is unclear to us what the concept of "timely manner" means when processes of curation is involved, access can be delayed.
- G. It is unclear to us what kind of tracking is being recommended in this guidance about reuse. Would these recommendations include citations, downloads, bibliographies, and/or all the above? Providing clarity on these points would be helpful.
- H. Seems fine

- I. Seems very beneficial for researchers to have this privacy information easily accessible and clearly outlined as this section recommends.
- J. This subsection does not seem specific enough. Although it is clear that many metadata standards would be acceptable, which makes sense, it is not clear to what degree they should be standardized. For example must repositories structure metadata in a standard metadata format such as Qualified Dublin Core as opposed to a locally modified Dublin Core? Additionally, must the repositories require that certain data formats be used by the depositors? Could the repositories rely on the honor system for submitters or would the digital repositories need to have a system for screening the type of data submitted?
- K. Looks good

## Section II

Our primary concern with this section is that the language used is a bit too high level or non-specific to be really practical for several of the user types mentioned in the previous section.

- A. It is unclear to us how one would go about assessing fidelity to consent. Would it be possible this process would be undertaken by a human, computer, or either? Would suggest including some clarifying language to add clarity for those looking to implement the final guidance.
- B. It is unclear to us how digital repositories would be expected to enforce submitters' data use restrictions and what this should look like when this guidance is operationalized. We would recommend adding clarifying language to this subsection.
- C. We appreciate that this subsection includes language that refers specifically to the type of data that should be planned for by the repository.
- D. It is unclear to us if the repository needs to have a general response plan or if the response plan needs to meet some kind of criteria outlining or scaffolding an appropriate or reasonable response. We would recommend additional, clarifying language be added to provide easier to operationalize guidance.
- E. Looks good
- F. Looks good
- G. It is unclear to us what type of documentation this guidance is referencing. For example, is the guidance referring to documentation that outlines guidelines for data retention by the digital repository (i.e. we will provide discoverability and access to this dataset for ten years)? Or, is the documentation referenced in this section calling for guidelines for the retention of data by the recipients of said data (i.e. we will destroy the data after 2 years in accordance with our data use agreement)? Additional language in this section would help to clarify this guidance.
- H. It is unclear to us what is meant in this subsection as a reasonable plan. Will such a plan include standards or guidance that will help the user navigate the system? Our concern is that even if a plan exists that doesn't make it a good and/or logical.
- I. Looks good.



Request for Public Comment on Draft Desirable Characteristics of Repositories for  
Managing and Sharing Data Resulting From Federally Funded Research  
FR Doc. [2020-00689](#)

Response from the RCSB Protein Data Bank

Filing Name: Stephen K. Burley

Filing Organization: RCSB Protein Data Bank

Date: March 5, 2020

The Protein Data Bank (PDB) was established by the scientific community in 1971 as the 1<sup>st</sup> open access digital data repository in biology and medicine. In its 49<sup>th</sup> year of operations, the PDB is central to research and education in fundamental biology, biomedicine, bioenergy, and bioengineering/biotechnology. The PDB data repository currently houses >160,000 atomic level biomolecular structures determined by crystallography, NMR spectroscopy, and 3D electron microscopy. It is managed by the Worldwide Protein Data Bank partnership (wwPDB; [www.pdb.org](http://www.pdb.org)) according to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability.

Through an internet information portal and downloadable data archive, many millions of researchers and educators freely access 3D structure data for large biological molecules (protein, DNA, and RNA). These are the molecules of life, found in all organisms on the planet. Knowing the 3D structure or shape of a biological macromolecule is essential for understanding the role the molecule plays in health and disease of humans, animals, and plants, food and energy production, and other topics of concern to global prosperity and sustainability.

The RCSB PDB (RCSB.org) operates the US data center for PDB, serves as Archive Keeper for the global PDB archive, and delivers PDB data at no charge to millions of Data Consumers without limitations on usage. Studies of website usage, bibliometrics, and economic benefits document the enormous impact of the PDB data on basic and applied research, clinical medicine, education, and the United States economy.

Access to PDB data and services contribute to patent applications, US Food and Drug Administration approvals of new medical entities, publication of scientific studies, innovations that can lead to new product development and company formation, and STEM education.

RCSB PDB is funded by the National Science Foundation (DBI-1832184), the US Department of Energy (DE-SC0019749), and the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences of the National Institutes of Health under grant R01GM133198.

RCSB PDB appreciates the opportunity to provide comments in response to Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research (FR Doc. [2020-00689](#)).

The RCSB PDB strongly supports the proposed characteristics listed under section “**I. Desirable Characteristics for All Data Repositories**” and notes that they follow previously published standards, including the 2016-11 Core Trustworthy Data Repositories Requirements v01.00; and Wilkinson, M.D. et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3 (160018), 1-9; and van der Aalst W.M.P. et al. (2017) Responsible Data Science. *Business & Information Systems Engineering* 59, 311-3, and those discussed at the 2019 NIH Workshop on Trustworthy Data Repositories for Biomedical Sciences (<https://datascience.nih.gov/data-ecosystem/trustworthy-data-repositories-workshop>).

The proposed characteristics would be strengthened by inclusion a **clear definition of primary data repositories as “stores of experimental data and metadata produced by researchers.”** These data, the work product of federally funded or supported research, need to be curated by domain experts and validated, preserved, and freely distributed. A bright-line distinction should be made between these primary data repositories and derived data resources (a.k.a. knowledgebases) that aggregate information and results of value-added computations and analyses with primary experimental data and metadata stored in the primary data resources.

In reviewing the proposed characteristics, the RCSB PDB found that they were all valuable descriptions.

The RCSB PDB recommends that characteristics *B. Long-term sustainability*, *D. Curation & Quality Assurance*, and *F. Free & Easy to Access and Reuse* be strengthened to ensure robust and enduring public availability of federally-funded research data.

Most importantly, from the standpoint of the RCSB PDB, is inclusion of language that makes federal research funders explicitly responsible for covering the costs of long-term FAIR-compliant storage, maintenance, periodic remediation, and delivery of experimental data and metadata produced by the researchers they fund (perhaps by mandating a modest set aside of total research expenditures to ensure that the research data are made freely available in perpetuity).

The RCSB PDB also recommends inclusion of the following additional Characteristics:

**Transparency:** This would include detailed documentation and clear, public disclosure of all characteristics listed in a way that clearly indicates how the overall goals of the data repository are being met.

**Community Engagement:** It is essential that data repositories know their user communities and meet their needs, and that appropriate oversight and expert advisory review are utilized.

**Technology:** Data repositories must provide a technology platform capable of supporting the secure, persistent, and reliable services enumerated in Sections I and II.

**Life-cycle Management:** A robust and cost-effective data ecosystem depends critically on anticipating community needs for new data repositories. Proactive mechanisms should be put in place to establish new data repositories that reflect rapid evolution of the experimental tools used by federally funded and supported researchers. By the same token, mechanisms need to be put in place to periodically evaluate existing data repositories and provide for orderly transition of those no longer required to meet user needs.

**From:** Ge Peng <[gpeng@ncsu.edu](mailto:gpeng@ncsu.edu)>  
**Sent:** Thursday, March 5, 2020 10:46 AM  
**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>  
**Cc:** Ge Peng <[gpeng@ncsu.edu](mailto:gpeng@ncsu.edu)>  
**Subject:** [EXTERNAL] RFC Response: Desirable Repository Characteristics

**Name:** Ge Peng, PhD  
**Affiliation:** North Carolina Institute for Climate Studies (NCICS), North Carolina State University (NCSU)  
**Primary scientific discipline:** physical sciences  
**Role:** researcher

The draft was nicely put together – it is timely and will be very useful.

Below are my comments **in red** for your consideration. Please feel free to contact me at [gpeng@ncsu.edu](mailto:gpeng@ncsu.edu) if you have any questions or need any additional information.

Hope it helps.

Best regards,

Ge Peng, PhD

-----  
Section I.

**G. Reuse:** Provides information about consent for reuse. A machine-understandable reuse license should be included in the metadata, even if the federal research data by default are open, to maximize the values of federal data. Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).

**General Comments:**

- 1) There is a redundancy among E, F, and G. Rewording may be helpful.
- 2) For transparency, reproducibility, and improved usability, it may be helpful to require repositories to provide documentation on data processing steps and error sources, preferable using a consistent template. Perhaps an item on Documentation after C. Metadata, e.g.,  
**D. Documentation:** Ensures datasets are accompanied by documentation sufficient to enable use and transparency including data processing steps and error sources, preferably using a consistent document template that is standard to the community the repository serves.

--

Ge Peng, Ph.D.  
Research Scholar  
[North Carolina State University](http://www.ncsu.edu)  
North Carolina Institute for Climate Studies (NCICS)  
151 Patton Ave, Asheville, NC 28801 USA  
[gpeng@ncsu.edu](mailto:gpeng@ncsu.edu)  
o: +1 828 257 3009  
f: +1 828 257 3002  
ORCID: <http://orcid.org/0000-0002-1986-9115>

March 4, 2020

Dear Dr. Droegemeier:

The American Physiological Society (APS) appreciates the opportunity to submit remarks in response to the request for comments on draft desirable characteristics for repositories for managing and sharing data resulting from federally funded research. As a publisher of 15 scientific journals, the society's publications policies<sup>1</sup> already encourage authors to "make data that underlie the conclusions reported in the article freely available via public repositories or available to readers upon request."

As a general comment on the implementation of data deposition policies for federally-funded research, the government should consider the costs and administrative burdens associated with data deposition and should seek to harmonize requirements across federal agencies to the greatest extent possible.

With respect to the specific characteristics detailed in the federal register notice, APS offers the following comments on selected provisions.

- I. A. The use of Persistent Unique Identifiers (PUID) for data submissions is absolutely necessary for locating deposited data. Only in rare instances should data become unavailable once it has been deposited. As noted in (B.), long-term sustainability of data repositories is important and each repository should have back-up plans to preserve and transfer data if there was a need to shut down. Federal agencies will need to determine how to fund long-term data storage that extends beyond the end of each award period and preferably for a much more extended period of time.
  
- C. Standard terminology should be used as much as possible to describe data sets. This should include clear annotation, and definitions should be provided as needed. Important questions about metadata include: What metadata will be required? To what extent will an accompanying description of methods be required along with the data for the purposes of replicating experimental results?
  
- D. Curation and quality assurance are highly desirable for data repositories, but it is not clear how this expertise will be provided. How will submitted data be evaluated for quality? Current costs for data storage are sometimes significant depending on the volume of data, and the addition of curation and quality assurance will add to those costs, which must be considered. Where shared data has not undergone peer review in the context of publication, how will the quality of the data be assessed? Will it be evaluated before, or after it is made public? As data from all federally-funded projects begins to accumulate, the sheer volume of the data available will limit the ability of the scientific community to examine and provide meaningful review via informal crowdsourcing.



E, F. Repositories should be designed to provide ease of access both for scientists depositing the data and for users accessing it.

G. Tracking data citation through the use of PUIDs is straightforward, but more details are needed about how repositories might track data usage in order to understand how that would be accomplished. Will users be required to create unique sign in profiles?

H. Repositories should be able to provide access to data in a manner that is automatically consistent with any necessary restrictions on access and reuse such as intellectual property concerns.

J. Research generates an enormous range of data types. Therefore, it will be difficult and perhaps impossible to develop a common format for depositing data into databases. In some cases, specialized software may be required to access and view the data – for example imaging data from different sources. How to make the necessary software available and ensuring long-term compatibility between the software and the data should be considered in the development of repositories. A critical question is also what constitutes “data”. Many labs generate thousands of individual data points or sets each day – do they all need an individual PUID? Are they treated individually or as a data collective for each experiment or set of experiments?

K. Repositories should maintain information about any changes made to data or metadata deposited in them. In addition, they should have security measures in place to ensure that information is not changed in an inappropriate or fraudulent manner after deposition.

As OSTP works to increase access to the results of federally-funded research, APS appreciates the opportunity to provide input. We hope we will have the opportunity for continued conversations on these complex and important topics.

Sincerely,

**A**

Meredith Hay, Ph.D.  
President  
American Physiological Society

<sup>1</sup><https://journals.physiology.org/author-info.data-repositories>

RFC Response: Desirable Repository Characteristics  
2020-03-03

Submitter Information

Karen Stocks, Director, Geological Data Center of Scripps Institution of Oceanography  
Primary scientific disciplines: Oceanography  
Role: Data Facility Manager

I thank the OSTP for the opportunity to comment on the proposed Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research.

**1. Proposed use and application of the desirable characteristics**

The overall goal of supporting open science in general, and the OSTP memorandum on “Increasing Access to the Results of Federally Funded Scientific Research” in particular, is important, timely and worthwhile. In addition, providing consistent guidance across agencies on the characteristics desired in data repositories is useful. We are supportive of the approach, but we have two overarching comments:

- *Few repositories currently meet these characteristics.* While all are desirable, the implementation of any OSTP guidelines cannot be framed such that repositories not meeting these guidelines are deemed inappropriate for use or insufficient. Further, larger generic repositories with institutional support are more likely to meet more of these characteristics, but smaller domain-specific repositories often better meet the critical and heterogeneous needs of their scientific communities. These guidelines should not have the unintended effect of discouraging the use of these valuable specialized facilities.
- *Mappings between these desired characteristics and Core Trust Seal and ISO 16363 are needed to reduce the burden on repositories.* Demonstrating compliance with requirements is effort intensive; if a repository has produced documentation to address one of the existing standards, this should be sufficient to describe compliance with OSTP desired characteristics. While the RFC states “Federal agencies would not plan to use these characteristics to assess, evaluate, or certify the acceptability of a specific data repository” it is inevitable that, if adopted, data facilities will be asked to demonstrate their degree of compliance.

**2. Comments on specific draft characteristics**

Overall, the specific characteristics are appropriate and inclusive, with the caveat that they are currently aspirational for the large majority of earth sciences data facilities. Below are specific comments on individual draft characteristics.

**Long-term sustainability:** We recommend that “long-term” be defined. Further, it is critical that sustainability expectations are consistent with funding mechanisms. Many earth sciences domain data repositories are funded by the National Science Foundation, NOAA, and other agencies on 3-5 year grant cycles. If this is the funding commitment that the federal agencies make, then this should be considered sufficient (though it is reasonable to request contingency plans for transferring the data should funding not be sustained). As mentioned above, many domain specific repositories provide critical services that general repositories cannot, but are more likely to be funded by shorter term awards (even though many have a long term record of sustained funding). OSTP should be aware of the consequences their guidelines may have on changing the landscape of generic vs domain, and agency-funded vs institutionally supported, repositories

**Curation & Quality Assurance:** We recommend that the expected level of Quality Assurance is defined. While it is appropriate to ask repositories to demonstrate that they can ensure data and metadata meet content and format standards, and that the repository is not introducing errors, it is not reasonable to expect all data facilities to undertake scientific quality assurance. Terminology varies among disciplines - e.g. terms like Level 0, level 1, Level 2 QA are not universal - so this is often difficult to communicate generically. Checking, for example, that data are provided with the right parameter name, in the correct units, in a standard format, with standard quality flags is an appropriate level of QA to expect; checking if the measured value appears high given past similar measurements is a level of scientific QA that few repositories can meet, and one can argue should fall to the expert scientist submitting the data.

**Free & Easy to Access and Reuse:** While the “broadest possible terms of reuse or documented in the public domain” is appealing, this is inconsistent with Goal G “Enables tracking for data reuse”. Data use can only be tracked if citation/attribution is requested, and a public domain statement does not support citation. Licenses such as “CC 4.0 BY” that allow wide use while still requiring attribution are not “the broadest possible” but would better meet the stated OSTP goals.

**Privacy:** It is not clear why Privacy is separate from Goal H “Secure”. Privacy is generally one element addressed by security, and considerations such as monitoring and risk management around PII and other privacy concerns are generally part of a security plan.

**Provenance:** tracking all metadata changes is an expectation that few earth sciences data repositories can currently meet.

We also recommend that OSTP consider an **additional characteristic around Financial Transparency**. As the selling of user data in various forms becomes more common, and more problematic, it would be valuable for data facilities to make a clear statement about their funding model.





## Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

March 5 2020

The American Statistical Association (ASA) is pleased to provide comments in response to OSTP's [Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research](#), as invited in the *Federal Register* of January 17, 2020 (85 FR 3085).

ASA's comments were written by members of the ASA Committee on Privacy and Confidentiality and are found on the following pages.

Thank you for your consideration.

*Questions on this document can be directed to the ASA Director of Science Policy Steve Pierson, [pierson@amstat.org](mailto:pierson@amstat.org).*

Authors (from the American Statistical Association, Committee on Privacy and Confidentiality):  
Lars Vilhuber, Cornell University, Member  
Stefan Bender, Research Data and Service Center of the Deutsche Bundesbank, Member  
Frauke Kreuter, University of Maryland, Member  
Stephanie Shipp, Biocomplexity Institute, University of Virginia, Member  
Aleksandra Slavkovic, Penn State University, Member  
Tom Krenzke, Westat, Chair

The authors are responding in their capacity as members of the American Statistical Association's Committee on Privacy and Confidentiality, and are not representing their respective home institutions. They serve voluntarily and without remuneration on this Committee. The Committee's role is described on its website <https://community.amstat.org/cpc/home>. Relevant for our response to the RFC, the Committee has the charge:

- To monitor and encourage new technical developments related to privacy and confidentiality of data collected or used for statistical purposes.
- To develop appropriate liaison with Congressional Committees and Federal agencies on matters relating to privacy and confidentiality.

The authors come from a variety of disciplines in addition to statistics. They have degrees in sociology, economics, and in their various positions, have experience in creating, managing, and expanding research data centers holding confidential research data, and providing secure, unbiased, controlled access to these research data.

In our response, we will focus on the privacy and confidentiality aspects of the proposed repository characteristics. We draw on examples from the United States, Canada, Germany, and the United Kingdom.

In particular, we will respond primarily to questions of access (I.E.) ease of access (I.F.), fidelity to consent (II.A.). We consider that II.B-F. are not fundamentally different from the overarching question of access (I.E.), and that II.I. (request review) is a variant of I.F. We have additional comments on documentation of privacy (I.I.), and on the availability of metadata (I.C.).

*I.E. Access.* The suggested criteria require “broad, equitable, and maximally open access to datasets,” moderated by privacy and confidentiality considerations. We note that there are many considerations why privacy and confidentiality considerations might apply, not just fidelity to consent for human data (II.A.) and compliance with restricted use conditions for human data (II.B.). Additional confidentiality considerations include financial, company, biogenetic, and national security considerations in the domains of biology, nuclear physics, engineering, to name a few. When federal funds are used to support research that use, analyze, generate, or produce such products, safeguards and access restrictions also need to be imposed. These are not fundamentally different from those for human data. To reprise (Desai, Ritchie, and Welpton 2016)<sup>1</sup>, in all cases, repositories must need to assess whether access satisfies appropriate criteria

---

<sup>1</sup> Desai, T., Ritchie, F., and Welpton, R. (2016). ["Five Safes: designing data access for research"](#). *Bristol Business School Working Papers in Economics*. All URLs in this document were last consulted on March 4, 2020.

along five dimensions (the “Five Safes”): *Safe projects* (Is this use of the data appropriate?), *safe people* (Can the researchers be trusted to use it in an appropriate manner?), *safe data* (is the disclosure risk in the data appropriate for the purpose?), *safe settings* (from where and how is the researcher accessing the data?) and *safe outputs* (are the published outputs appropriately protected?). These five dimensions can be usefully applied to data ranging from full public use data (freely downloadable without need for any controls) via medium-security data (released to researchers under enforceable data use agreements) to highly classified data. They should thus be criteria applied by and for all federal funded repositories.

## I.F. Ease of access

Where necessary, access restrictions must be imposed. At the same time, repositories should leverage and implement the broadest possible set of tools to make access as easy as possible. The gold standard in terms of ease of use remains public-use data in the public domain, available for direct download, and with few if any use restrictions.

Clearly, when access is subject to some level of control, ease of use must necessarily be reduced. For instance, in the simple case where registration is required to ensure that users agree to terms of use, various access mechanisms can be implemented. Repositories should strive to allow for seamless access using both human and machine-initiated tools. The UK Digital Economy Act of 2017 enshrines a principle of proportionality.<sup>2</sup>

For instance, users could register once, agree to terms of use, and then obtain an access token which allows them to initiate future downloads from the same provider via an API using machine-initiated (automatic) downloads, while still complying with all terms of use. This is standard in many other common situations in the private industry, but is less frequent amongst current repositories.

Similarly, current restrict-access research data centers – a form of repository with access controls – require users to go through user vetting (“safe users”) for every repository afresh, without reference to prior vetting at other repositories with similar or identical criteria. For a given repository, project vetting (“safe projects”) for a user’s multiple projects happens independently every time, without reference to prior projects. Furthermore, current repositories are often separated into distinct “data silos”, where data sits in distinct repositories, and data that is primarily hosted at one repository cannot be also accessed at a separate repository. This is still generically true at the federal level, despite progress under CIPSEA (Title V of the E-Government Act of 2002, PL 107–347<sup>3</sup> and Title III of the Evidence Act of 2018, PL 115-435<sup>4</sup>). Impediments are also the norm for federal-state data sharing, and for government-private or government-academic data sharing. Though such data sharing across repositories occurs on a regular basis, each one is subject to laborious ad-hoc re-negotiations.

---

<sup>2</sup> Principle 5, <https://www.gov.uk/government/publications/digital-economy-act-2017-part-5-codes-of-practice/research-code-of-practice-and-accreditation-criteria>

<sup>3</sup> <https://www.govinfo.gov/link/plaw/107/public/347?link-type=pdf>

<sup>4</sup> <https://www.congress.gov/bill/115th-congress/house-bill/4174>

Repositories for federally funded data should be held to implement efficient mechanisms that allow for user and project vetting to be streamlined, and that repositories be allowed to share data or be accredited by multiple data owners, thus greatly increasing ease of access. In what follows, we illustrate three examples that have taken first steps, or even successfully implemented such streamlined processes.

## **Example 1: Researcher accreditation**

ICPSR at the University of Michigan has been developing a “researcher passport” (Levenstein, Tyler, and Davidson Bleckman 2018). Key element is “a credential that identifies a trusted researcher to multiple repositories and other data custodians, [...] durable and transferable digital identifier issued by a central, community-recognized data steward.” One possible steward might be a federally mandated entity. A portable digital credential is being considered by the European Union. In the UK, the “Digital Economy Act of 2017” went further, and implemented a legal status of “accredited researcher,” with criteria laid out in the law itself, and a government panel to consider and vet requests for accreditation.<sup>5</sup>

Such a credential or accreditation would allow for efficiencies in the vetting process, and greatly ease access to data subject to access controls. We note that these must be “standard procedures”, ideally initiated or controlled by federal government entity. They are unlikely to work if not mandated, as the current situation suggests.

## **Example 2: Streamlining of project vetting**

One of the costliest steps in providing secure and ethical access to restricted-access data is the per-project vetting process. While efforts are underway in the US to streamline the application process for federal data in support of the Evidence Act of 2018, less emphasis has been put on the approval process for applications. Currently, even where there is a streamlined application process, each application is evaluated individually, an often lengthy process. For other federally funded repositories, no single application process is envisioned that we know of.

Canada may serve as an example of a system that has attempted to streamline and accelerate such a system, reducing the barriers to restricted-access federal data.<sup>6</sup> Since 2019, certain classes of applicants for access are automatically pre-approved, meaning that they no longer have to go through a review process (they must still satisfy all security clearance criteria). Such applicants include any tenured professor at an accredited Canadian university, or recipients of peer-reviewed funding.

---

<sup>5</sup> <https://www.statisticsauthority.gov.uk/about-the-authority/better-useofdata-statistics-and-research/betterdataaccess-research/better-use-of-data/>

<sup>6</sup> <https://www.statcan.gc.ca/eng/microdata/data-centres/guide>

## Example 3: Coordination among networks of research centers

For better transportability and transferability of sensitive research data, coordination or mutual accreditation of secure repositories should be encouraged. The Federal Statistical Research Data Centers are a successful example in the context of data held by federal agencies, but have been slow in expanding the range of agencies and data. Loose coordination among NIH-funded repositories is an issue for the sharing of biomedical data.

Examples of stronger coordination exist in Germany and the UK. Administrative Data Research UK (ADR UK) plays an important role in bridging the gap between government and academia in the realm of administrative data, and in partnership with the Office of National Statistics (ONS).<sup>7</sup> Multiple “hubs” coordinate and implement access. In Germany, the German Data Forum has successfully established a decentralized network of accredited research data centers (RDCs) as a model solution for scientific data access.<sup>8</sup> A total of 31 research data centers are currently accredited and coordinated by the German Data Forum. Research data centers are annually evaluated. This infrastructure enables researchers to gain flexible access to a wide range of data. The UK and German networks also have an important additional component: outreach. The ADR UK Strategic Hub coordinates public engagement activities, helps to gauge public opinion regarding the use of the administrative data. The German Data Forum advises the German federal government and the governments of the Länder (states) on expanding and improving the research data infrastructure. It facilitates a continuous exchange between data producers and the data users in science and research with the aim of improving access to high-quality and scientifically potent data.

While these examples are primarily focused on data held and made available by the federal government, similar examples in the US are emerging. The Administrative Data Research Network (ADRN) is such an example, bringing together research projects that use data provided by various state and local levels. Many university-based secure computing environments exist, serving an important role, but must be authorized by data providers for each new project. A stronger coordination, for instance an accreditation mechanism for secure repositories for any source of data, has yet to emerge.

## I.C. Metadata

Finally, we point out that effective repositories of confidential data urgently need high-quality metadata (I.C.) on their data holdings, so that researchers can find, assess the utility of, and request access to research data that is pertinent for their scientific endeavors. Metadata on confidential data, when available, is currently scattered throughout various disconnected sites, often in disregard of widely available metadata standards. In general, there are few confidentiality concerns regarding the availability of metadata, and where these arise, for instance in the statistical metadata on extreme values, there are well-established measures to

---

<sup>7</sup> <https://www.adruk.org/our-mission/our-mission/>

<sup>8</sup> <https://www.ratswd.de/en>

handle these. We note that a critical element of the metadata needs to be the documentation of privacy-protecting measures applied to the microdata or the outputs (I.I.). Analyses that do not take full account of the statistical properties of the protection mechanisms are at risk of bias and other statistical problems. Analysts need to know exactly how to take into account these legitimate manipulations of the data. This can only be achieved through detailed information on those manipulations as part of the metadata.

Metadata (and the “connected” microdata) need to be findable, accessible, interoperable, and reusable (FAIR). The best implementations emerging in France and Germany are central metadata catalogs. Data.gov and efforts at various US universities (for instance, the Census Bureau data portal at ICPSR<sup>9</sup>) are a step in the right direction. Repositories that are subject to any future rules that may come out of this consultation should be instructed to provide metadata in such standards, and to provide metadata through standard API that can be queried and crawled by aggregating sites.

---

<sup>9</sup> <https://census.icpsr.umich.edu/census/>

The FAIRsharing Community welcomes the opportunity to respond to this White House Office of Science and Technology Policy’s [RFI on the “Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research”](#). We note a close similarity with the work we are doing, and we would like to bring this to your attention in this response.

FAIRsharing (<https://fairsharing.org>) and DataCite (<https://datacite.org>) have joined forces with a group of publisher representatives (signatories of this response) who are actively implementing data policies and recommending data repositories to researchers. The result of our work is a set of **proposed criteria** that journals and publishers believe are important for the **identification** and **selection** of **data repositories**, which can be recommended to researchers when they are preparing to publish the data underlying their findings.

A table summarizing the proposed criteria, their definitions, and ideal values for each criterion is available in our pre-print at <https://osf.io/m2bce/>.

The article also provides more background information on the rationale for our work, which began in January 2018 and has also been presented at a number of sessions during the 12th, 13th and 14th Research Data Alliance Plenaries. This year, we also opened the work for community feedback and received almost 60 responses, 70% of which are from repository managers (the majority in the life sciences), and many of which are on behalf of organizations such as ELIXIR, Core Trust Seal and CSIRO Australia. We are currently reviewing this extensive feedback in order to refine the proposed criteria.

Evidently there is an overlap between our criteria and yours. For example, both lists feature criteria on Persistent Unique Identifiers, Metadata, Model and Format Standardization, Accessibility, Licensing, Reuse, as well as other FAIR-related criteria. We therefore would welcome a discussion on how we could potentially align and/or collaborate, particularly as some funders have expressed an interest in joining the next phase of our work.

## SIGNATURES

Name	Organization	Primary scientific discipline	Role
Susanna-Assunta Sansone (0000-0001-5306-5690)	University of Oxford, FAIRsharing Founder	Life sciences	Researcher
Peter McQuilton (0000-0003-2687-1982)	University of Oxford, FAIRsharing Coordinator	Life sciences	Researcher
Helena Cousijn (0000-0001-6660-6214)	DataCite	<i>generic</i>	Service provider
Matthew Cannon (0000-0002-1496-8392),	Taylor & Francis	<i>generic</i>	Publisher
Wei Mun Chan (0000-0002-9971-813X)	eLife Sciences Publications	Life sciences	Publisher

Sarah Callaghan (0000-0002-0517-1031)	Elsevier	<i>generic</i>	Editor
Ilaria Carnevale (0000-0001-8509-0495)	Elsevier	Life sciences	Editor
Imogen Cranston (0000-0002-7134-499X),	F1000 Research	<i>generic</i>	Publisher
Scott Edmunds (0000-0001-6444-1436)	GigaScience, BGI Hong Kong Tech Ltd.	Life sciences	Editor
Nicholas Everitt (0000-0001-8343-8910)	Taylor & Francis	<i>generic</i>	Publisher
Emma Ganley (0000-0002-2557-6204)	Procols.io	<i>generic</i>	Service provider
Chris Graf (0000-0002-4699-4333)	Wiley	<i>generic</i>	Publisher
Iain Hrynaszkiewicz (0000-0002-9673-5559)	PLOS	<i>generic</i>	Publisher
Varsha K. Khodiyar (0000-0002-2743-6918)	Springer Nature	<i>generic</i>	Service provider
Thomas Lemberger (0000-0002-2499-4025)	EMBO Press	Life sciences	Publisher
Catriona J. MacCallum (0000-0001-9623-2225)	Hindawi Ltd	<i>generic</i>	Publisher
Hollydawn Murray (0000-0002-8243-2493)	F1000 Research	<i>generic</i>	Publisher
Kiera McNeice (0000-0003-2839-4067)	Cambridge University Press	<i>generic</i>	Publisher
Philippe Rocca-Serra (0000-0001-9853-5668)	University of Oxford, FAIRsharing co-Founder	Life sciences	Researcher
Kathryn Sharples (0000-0003-2809-6828)	Wiley	<i>generic</i>	Publisher
Marina Soares E Silva (0000-0001-9530-627X)	Elsevier	<i>generic</i>	Product Manager
Jonathan Threlfall (0000-0001-8599-4320)	F1000 Research	<i>generic</i>	Publisher



Comments on Desirable Characteristics for Data Repositories  
Eric Lancon, elancon@bnl.gov

FAIR metrics should be defined and values computed for data repository (and catalogue) to measure the FAIRNESS w.r.t. Go FAIR

<https://www.go-fair.org/fair-principles/fairification-process/>

Access and availability of data should be guaranteed (through SLA?)

Capability to process the data should also be addressed, a repository is of little usefulness if data cannot be analysed and processed.

The list of publications or scientific results linked to given used datasets / data repository should be available in the repository.

The software (version, architecture, code repositories) used to generate (or analyse) the datasets is not mentioned in the RFC

Data loss is not addressed (this happens) what is the mitigation plan?

How are data management plans and repositories related? Can they be linked through templates and semantics?

## **Comment on : Desirable Repository Characteristics**

Arcot Rajasekar  
Professor,  
School of Information and Library Science  
University of North Carolina, Chapel Hill, NC  
[rajasekar@unc.edu](mailto:rajasekar@unc.edu)

This comment is based on more than twenty years of experience in designing, developing and deploying large-scale data grids - the Storage Resource Broker (SRB) and integrated Rule Oriented Data Systems (IRODS) - for scientific and business communities. Based on my experience and working closely with large-scale projects (CyVerse, HydroShare, Bioinformatics Research Network (BIRN)) I would suggest using the iRODS as a vehicle for achieving a data repository which meets almost all the demands as outlined in the call for comments and beyond. I give a short synopsis for each item that were emphasized in the CFC as desired characteristics for a data repository to help identify how iRODS provides the functionality. A short blurb about the iRODS is added at the end of my comments. Further information can be found at the iRODS Consortium website ([irods.org](http://irods.org)).

Desired Characteristics:

- A. *Persistent Unique Identifiers*: iRODS defines what are called zones which provide a way of defining a domain name service. Datasets and Collections (similar to a folder hierarchy) provide a virtual name to the datasets stored under the domain. Apart from that each dataset also is given a unique identifier (unique at the zone level) and provision is made for adding arbitrary number of external GUIs to be identified as metadata for each data item. Access can be based on zone-collection-dataname triplets or by unique GUIs associated with the datasets. Similar to DNS services, one can easily deploy a Zone Name Service that can identify the physical address of the resources in that zone (A zone can have a number of distributed resources – but because of peer-to-peer networking, one can access any file by connecting to any resource in the zone).
- B. *Long-term sustainability*: iRODS federates multiple levels of resources -including cloud and tape system access. The virtual naming and metadata support provide long-term sustainability for datasets stored in iRODS. Moreover, because of the virtualization of resource names as well as user names (apart from data name virtualization) the need for physical names is obviated and thus provide ease of solving technological obsolescence through transfer from an old

storage system to a new one without applications being aware of the move. Replication is an inherent property of the iRODS system and one can write policies about how, when and where datasets are replicated for improving access and sustainability, and providing fault tolerance.

C. *Metadata*: iRODS has a built-in metadata catalog (iCAT) which natively provides storing attribute-value-unit triplets for any dataset/collection, users and resources. Moreover, iRODS supports access to external metadata catalogs including triple stores, elastic search engines and SQL and NoSQL databases which can use unique identifiers to associate metadata for all objects stored in iRODS. Because of these any kind of metadata (including cross-references and external references) can easily be associated with datasets in iRODS.

D. *Curation & Quality Assurance*: iRODS is policy oriented and one can write rules and policies as needed to manage and automate the full data life cycle. Integrity checking, fidelity and fixity checks can be done on events (ingest/modification), periodically or by user request and through replication of objects automatic recovery can be done via machine-executable policies. Inbuilt support for multiple checksums provides a way to create digital signatures which can easily assure quality of the data as well as recovery from any bit rot or malicious degradations.

E. *Access*. iRODS provides authentication and authorization on a very fine scale. Third-party authorization and authentication, multi-level authorizations and challenge-response checks, are all easily possible through policy implementation. Moreover, iRODS provide faster access and ingestion through parallel data transfers mechanisms.

F. *Free & Easy to Access and Reuse*: Concepts of authenticated access, public data, anonymous access and ticket-based access all provide ease of access to datasets stored in iRODS.

G. *Reuse*: iRODS provides a way to ingest new data and other digital artifacts (including containers) into the system and cross-reference them. Hence provenance can be captured very easily.

H. *Secure*: Multi-level authentication and authorization (including external services) make it very flexible to create a highly secure data repository. With periodic checking one can easily verify for any security breaches.

I. *Privacy*: Multi-level authentication can easily provide compliance to all levels as needed (ex. HIPAA).

J. *Common Format*: Apart from syntactic replication, one can have semantic equivalent data stored in multiple formats. Indeed, an ingestion pipeline, one can define a set of conversions so that a dataset can immediately be converted into different formats as well as multiple resolutions (and abstracts) that can all be searched and cross-referenced together using metadata.

K. *Provenance*: Audit trail is also built into iRODS and can be turned on at various levels to capture a few or all operations performed on datasets. As mentioned before cross-reference metadata can easily capture provenance for derived objects.

II.A. *Fidelity to Consent*: Project level authentication (groups and roles) as well as provision for periodic checks for non-authorized access (audit trails) are helpful for consent provisioning.

B. *Restricted Use Compliant*: One can have policies that can restrict access to few users and then automatically open for larger set of users and finally public. The policies can be encoded as iRODS rules and linked to the “age” of the datasets so that time-bound access can be controlled automatically without any human intervention. Other non-age bound access modifications can also be easily configured by encoding specific rules.

C. *Privacy*: Automatic checks for unauthorized access as well as periodic checks for correct ACL lists are tools that can be used to manage privacy.

D. *Plan for Breach*: iRODS provides a way to store data in an encrypted form with key stored elsewhere. Also, with multiple replicas, one can easily make sure that any malicious changes to datasets can be identified and corrected.

E. *Download Control*: A rich authentication and access framework is part of iRODS. Parallel data transfer benefits large file access.

F. *Clear Use Guidance*: One can associate documents as metadata for each dataset (eg. copyright document, policy document, etc.) as metadata which, and can be made accessible to users.

G. *Retention Guidelines*: Same as above.

H. *Violations*: Since the data management is automated there is a good chance for auto compliance of the policies. With audit trail, one can periodically check for any violations.

Apart from these required and additional qualities that are noted in the Call for Comments, iRODS provide other capabilities that help manage a data repository. We note these broad and useful capabilities can help further in better and efficient data repository implementation.

**Data Virtualization:** Data stored in iRODS is typically accessed through an iRODS client. iRODS clients present files as Data Objects organized into Collections. For the most part, there is little difference between Data Objects and files, and between Collections and subdirectories. However, there are a couple of important distinctions:

- Collections make no reference to the physical storage path. It is possible for two Data Objects in a Collection to be stored in different physical locations
- A Data Object may refer to multiple Replicas. Replicas are exact copies of a file, located in multiple physical locations.

Data Objects and Collections are stored in Storage Resources in an iRODS Zone.

Each Storage Resource has a name (the Resource's logical representation) and a hostname and path (the physical representation of the Resource, where files are kept). The hostname is the network name of the device that serves the data, and the path is the local file system path or object storage bucket that holds the data.

**Data Discovery:** This information about data, called metadata, is extremely useful for Data Discovery, locating relevant data within large data sets. Data Object metadata includes rich, user-defined metadata in addition to traditional system metadata, such as filename, file size, and creation date. This rich metadata allows data to be identified by characteristics such as author names, keywords, case ID, and content type.

Rich metadata can include whatever descriptors you choose to apply to your data. Rich metadata can also be applied to Collections, Users, Resources, and other iRODS Zones. The entire iRODS catalog for a Zone is contained in a relational database. Currently, that database must be hosted in a PostgreSQL, MySQL, or Oracle database management system.

**Workflow Automation:** Each iRODS Server runs a Rule Engine that is an event-triggered background process. The Rule Engine is programmed using iRODS Rules, which specify what actions should be triggered when iRODS initiates a particular system activity.

iRODS event triggers are called Policy Enforcement Points (PEPs). Consider, for example, a rule to transfer ownership of data objects to the project manager when a user is deleted; the trigger — or PEP — is the deletion of the user. Similarly, rules could be written to extract metadata or pre-process data whenever a file is uploaded to an iRODS Resource.

Chaining rules and PEPs allows you to create powerful, customized workflows that save time and prevent human error. Complex multi-step scientific processes can be tightly managed and automated by keeping thorough records of ongoing status and other lab information, and only alerting humans when necessary. Organizational data management policy can be captured in an automated, auditable fashion using iRODS rules.

**Secure Collaboration:** Even in fields where data may not be published, it is usually necessary to share data sets between multiple workgroups. However, as data sets grow beyond several gigabytes, it becomes difficult to impossible to move the data between locations. iRODS provides Secure Collaboration through three technologies:

Tickets, Permissions, and Federation.

- iRODS Tickets provide controlled public access to Data Objects and Collections. The owner of a Data Object or Collection can create a Ticket and share it with non-iRODS users to grant them read or write access. Tickets can be revoked, and they can be set to automatically expire upon a specified date and time or a specified number of reads or writes.

- iRODS Permissions are analogous to UNIX file system permissions. The owner of a Data Object or Collection can assign read or write access for any number of defined iRODS Users and Groups. Group membership is defined by the administrator(s) of a Zone.

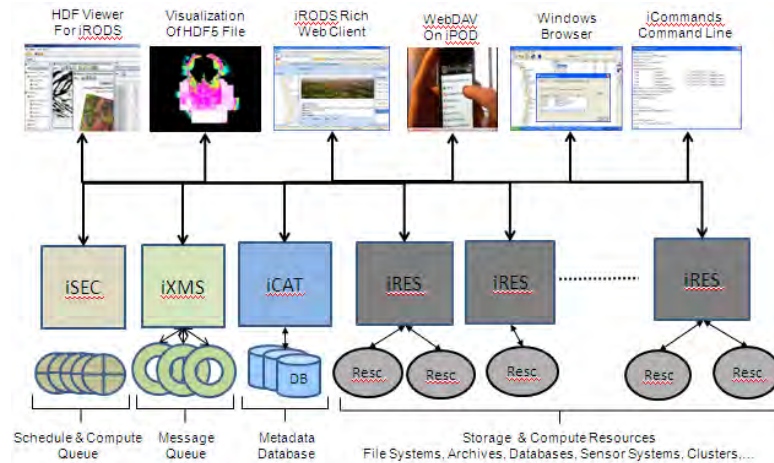
- iRODS Federation extends data sharing and publication beyond a single Zone. In a Federated deployment, once the administrators of two iRODS Zones share a set of keys, the owner of a Data Object or Collection can assign read and write permissions to users from outside Zones. When reading or writing data, the transfer mechanism is analogous to that for a single Zone. Unless the file is very small, iRODS servers broker a connection between the server containing the data and the client requesting it. As a result, Federation enables high performance access to data stored in any other iRODS Zone.



Figure 1 iRODS Capabilities

### Box 1: iRODS Data Grid System

The iRODS Data Grid can be viewed as a network of fully connected nodes of resource servers, called iRES, which provide access to data and computational resources. The servers perform the protocol interchange needed for interfacing with exotic devices, mapping them onto a uniform API used in the client framework. An iRODS system



consists of many servers with the most important being the resource and rule engine servers (iRES) which provide access to storage and compute resources. The iCAT server holds the metadata used by the iRODS system and acts as a persistent store for the system status. The messaging server (iXMS) provides the means for the different servers (and services running in them) to communicate. In this way, services can be distributed, run in parallel, and communicate over time and space. The Scheduling Server (iSEC) allows the system to schedule jobs at a specific time, periodically, or when a resource is

available.

iRODS Features	Description
<b>Logical Collection Hierarchy</b>	Organize distributed data into logical sets
<b>Replication, GUIDs/Object Ids</b>	Unique name/identifier for multiple replicas
<b>Versioning</b>	Version Number support
<b>Rich Authentication &amp; Access Control</b>	Support for multiple authentication schemes including GSI, Shibboleth, etc. Access control data objects, collections, resources for users and user groups.
<b>Discovery Services: Descriptive Metadata support</b>	Associate Attribute-Value-Unit metadata for data or collections. Support for element-based schema such as Dublin Core, FITS, DICOM, Darwin Core
<b>XML metadata Support</b>	Loaded into AVU-Metadata and supports Xpath queries
<b>Policy Execution as Rule Support</b>	System management and domain-specific collection policies can be coded as iRODS rules and executed on demand, on an event, or at periodic intervals
<b>Server-side workflow chains</b>	Rules can be triggered to perform multiple operations such as metadata extraction, format translation, anonymization, apply domain-specific analysis and synthesis of files and collections.
<b>Files, databases , archives &amp; streams</b>	Heterogenous protocols supported
<b>Rich data Transport Protocols</b>	TCP/IP and UDP; parallel stream support
<b>Data management: synchronize, backup, archive, move, copy, ...</b>	Support for distributed data management operations
<b>Integrity &amp; Authenticity Maintenance</b>	Support for checksums, signatures periodic scans to restore damaged replicas
<b>Provenance &amp; Chain of Custody</b>	Support for Audit Trail, lineage analysis & support for execution metadata
<b>Accession, Preservation, Retention, Disposition &amp; Migration</b>	Policy/rule support for long-term preservation
<b>System and User-defined Metadata</b>	Internal catalog (iCAT) in relational database stores object information.

Xx March 2020

Lisa Nichols

Office of Science and Technology Policy

[openscience@ostp.eop.gov](mailto:openscience@ostp.eop.gov)

Re: RFC Response on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

Dear Dr. Nichols:

We appreciate the many ongoing opportunities for continued dialogue with OSTP and the Administration on how to best to promote openness and sharing – consistent with our commitment to promote sustainable Open Science. We especially appreciate OSTP’s recognition that publishers are a valued partner for addressing these questions.

The International Association of Scientific, Technical and Medical Publishers (STM) is the leading global trade association for academic and professional publishers. It has more than 150 members in 21 countries who each year collectively publish more than 66% of all journal articles and tens of thousands of monographs and reference works. STM supports our members in their mission to advance research worldwide. As academic and professional publishers, learned societies, university presses, start-ups and established players, we work together to serve society by developing standards and technology to ensure research is of high quality, trustworthy and easy to access. We promote the contribution that publishers make to innovation, openness and the sharing of knowledge and embrace change to support the growth and sustainability of the research ecosystem. As a common good, we provide data and analysis for all involved in the global activity of research.

The majority of our members are small businesses and not-for-profit organizations, who represent tens of thousands of publishing employees, editors, reviewers, authors and readers, and other professionals across the United States and world who regularly contribute to the advancement of science, learning, culture and innovation throughout the nation. They comprise the bulk of a \$25 billion publishing industry that contributes significantly to the U.S. economy and enhances the U.S. balance of trade.

STM represents publishers across the entire spectrum of science, technology, medicine and the humanities, and is therefore uniquely positioned to discuss the Desirable Characteristics for All Data Repositories (section I). We look forward to continuing our efforts to partner with OSTP, SOS, and individual Federal agencies on these topics.



STM commends OSTP and the SOS for developing these characteristics, which are broadly consistent with those that we are utilizing in our [2020 Research Data Year](#) and also those supported by international initiatives such as the Research Data Alliance (RDA), in which STM is an active participant. We agree with the proposed use and application of the desirable characteristics, in particular that it would be inappropriate to provide “an exhaustive set of design features” or “use these characteristics to assess, evaluate, or certify the acceptability of a specific data repository.” Data sharing is a rapidly-developing field and being too prescriptive at this point could stifle innovation and reduce competition. In addition, specific fields and groups of practitioners may have different needs from those that could be described for all data repositories. Therefore, this flexibility is key.

STM agrees with the SOS that any proposed characteristics of desirable repositories should be consistent with those broadly accepted in research communities. Such criteria would ideally be the result of collaborative efforts by multiple stakeholders in the scholarly ecosystem and are therefore community endorsed. STM’s own efforts to identify and recommend repositories includes the latter requirement as a central characteristic. The identification of ISO 16363 Standard for Trusted Digital Repositories and CoreTrustSeal Data Repositories Requirements as an exemplar. We also greatly appreciate the explicit mention of the FAIR principles in the background section as a motivator for the specific characteristics. STM has been recognized as a member of the FAIRsFAIR project (<https://www.fairsfair.eu/>) in the European Union, and would welcome the opportunity to bring some of these principles and expertise to support OSTP’s efforts in this area.

With respect to the “Desirable Characteristics for All Data Repositories,” we support each of the characteristics that are included. We would like to highlight in particular the importance of “A. Persistent Unique Identifiers” (PUIs), and encourage the use of widely used and interoperable types of DOIs rather than the creation of government- or repository-specific ID types. We encourage the SOS to work with the RDA to ensure alignment of these IDs.

One criterion that the SOS may want to consider softening is “D. Curation & Quality Assurance.” Of course, repositories that offer curation services are to be preferred over repositories that do not. However, these services are not yet developed enough or consistently deployed across the repository ecosystem, even among the higher-quality data repositories. Although expert curation and quality assurance (including peer review) are important themes and are desired in all data repositories, the other items within this list are more fundamental to identifying appropriate data repositories.

It might be useful to add to the characteristic list two organizing ideas that are implicit in the set of desirable characteristics but may not be completely evident to agencies and Federally funded investigators using the list. In particular, although many of the listed features are in line with the FAIR principles, it might be useful to explicitly highlight these principles in the list of criteria as they are accepted as an effective means to communicate the desired characteristics of repositories. In addition, as noted above, it would be constructive for many of the characteristics (e.g. PUIDs, metadata, reuse tracking, security and privacy) to utilize community endorsed standards and approaches. In this context, it might be useful to add a characteristic “Aligned with community endorsed standards” to highlight the importance of non-proprietary approaches to many of the issues shared by data repositories.

Finally, we would like to suggest a few additional characteristics for consideration. These potentially could be included in a supplemental list of “Additional Characteristics for Consideration of Data Repositories,” to which “curation & quality assurance” could also be moved. These selection criteria can be used or seen as “nice-to-haves”:

- “Fit to subject”: Subject specific repositories are usually superior to generic ones. Repositories that are built and designed for specific disciplines are better catered to the specific needs and requirements of academic disciplines, and therefore should be preferred over generic repositories.
- “Size and scalability”: Larger repositories are, in general, to be preferred over smaller ones. The larger a database, the more useful it becomes due to network effects (e.g. it allows its users to find comparable datasets, find connections with related research, and prevents data being distributed over different databases).
- “Mirroring”: To keep data stored safely, repositories should maintain mirror sites, preferably over different geographical locations.

With respect to the feasibility of the proposed list of desired characteristics, we believe this to be a reasonable list that most responsible and appropriate data repositories for agencies and researchers would be able to meet the set of characteristics for. However, the degree to which an individual repository addresses each of the desired characteristics will vary significantly. This remains a key reason to maintain the list as guidance, rather than as requirements. The list is also generally consistent with those used by several certification schemes, as well as supported by the wider scholarly ecosystem.



The global voice of scholarly publishing

A significant challenge going forward will be to support and guide researchers and federal agencies towards the most appropriate repositories to meet their data sharing needs. This RFC, and the ongoing efforts by OSTP and the SOS to support data sharing are an excellent step in the right direction. Such efforts will need to be coordinated across universities, non-federal funders, publishers, scholarly societies, and others who engage in and support the American research enterprise. Publishers stand ready to work with NSTC, OSTP, and Federal agencies on all of these issues going forward, and welcome additional opportunities to engage and collaborate.

Very truly yours,

A handwritten signature in black ink that reads "Ian Moss". The signature is written in a cursive style and is positioned above a horizontal line that extends to the right.

Ian Moss  
CEO

From: [bambacher@verizon.net](mailto:bambacher@verizon.net)

To: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov).

Subject: RFC Response: Desirable Repository Characteristics

This response is submitted by Bruce Ambacher, retired digital preservation systems analyst at the National Archives for more than thirty years and retired Visiting Professor in the iSchool, University of Maryland, College Park. I am currently a research affiliate with the iSchool's Digital Curation Innovation Center

Primary discipline: Social Sciences, Digital Archivist and Data Preservationist.

The Federal Government has been actively involved in digital data preservation since the establishment of a digital preservation unit in the National Archives and Records Administration in the late 1960s. The majority of comments submitted in response to this RFC will stress not only the need for establishing a set of "Desirable Repository Characteristics" but also the significant costs involved in establishing and staffing such repositories. To ensure economy for the ever growing volume of data to be preserved and accessed over time, a uniform set of characteristics must be based on multi discipline criteria and measurable metrics that demonstrate a repository's commitment to long term digital preservation.

This effort, of necessity, must be divided into two somewhat separate different frameworks. One focuses on data created by Federal agencies are subject to Federal law and regulations. The National Archives and Records Administration (NARA) remains the only Federal agency that has the statutory authority to preserve Federal records. Through affiliated archives agreements NARA has authorized the Government Publishing Office (GPO) to preserve reports from Congressional branch agencies. GPO fully embraced its accompanying responsibilities and became the first Federal agency to be certified as meeting ISO 16363's 109 metrics for digital data preservation.

Federal agencies seeking to preserve data should work through their records management programs to have such data appraised for its long term value and to determine the most appropriate data repository once the agency's primary use has ended. NARA has data containing national security classifications to the highest levels, department of Energy restrictions, Title 13 Census information, and a variety of privacy issues relating to individually identifiable information such as health, tax information and survey responses. Restrictions on access are no barrier to transferring Federal information to NARA.

The second focuses on parties using Federal funds to collect digital data collected have wide discretion in selecting a suitable digital repository. In the interests of economy and long term preservation and access, the goal should be to deposit such data in as limited a number of digital repositories as possible. Data preservation is too often an unfunded or underfunded afterthought

leading to makeshift solutions that do not ensure long term preservation and access and/or unnecessary duplication of effort between multiple repositories. It is noteworthy that this FRC seeks to address part of this issue by enunciating “desirable characteristics.” These will ensure data creators adhere to Federal funding requirements, establish a comprehensive set of metrics to which data creators, curators and users must adhere. Were the “desirable characteristics” made mandatory as a condition for obtaining Federal funding, a uniform level of trust in the structure, internal operations and security of the digital information could emerge enhancing data preservation and access into the future. It also would lead to adherence to a broad based set of requirements that can be uniformly measured by professional auditors.

Over the past two decades the Federal Government has been evolving from agency-specific and Federal Government-specific standards and guidelines, to international standards and criteria wherever possible.

ISO issued ISO 14721 Reference Model for an Open Archival Information System in 2002 after seven years of development. OAIS is the seminal document for trustworthy data repositories. One of the “Future Actions” recommended in OAIS was the development of “standard(s) for accreditation of archives.” This was achieved in 2012 when ISO issued ISO 16363, based on a decade of multiple task forces and interagency committees to develop and test the metrics that fulfill the accreditation requirements and specifically test repository compliance with OAIS. In addition to using ISO’s 109 metrics to certify compliant trustworthy repositories, the metrics also can be used as a high level design document for a compliant system, leaving it to the repository to determine which specific hardware and software are best suited to its preservation and access requirements. This will lead to quality data preservation, minimize the costs of operation, ensure data integrity over time, and enhance the reputation of the repository as a “certified trustworthy repository.”

Unlike other contemporary efforts such as FAIR and CoreTrustSeal, ISO 16363 is the only effort that provides measurable metrics to determine the likelihood of digital information being preserved and made available in a usable format over time. It is difficult to imagine how FAIR could become anything more than a set of platitudes like motherhood and apple pie. Who would challenge such lofty goals? But who could actually establish a long term preservation and access repository based on them alone? Equally, a limited number of nonbinding principles such as the fourteen that comprise CoreTrustSeal, which are confirmed by peer to peer review that will vary over time and could be achieved by a spoken or unspoken “you approve me and I will approve you” approach, cannot provide the definitive trust that will emerge from ISO 16363 certification achieved through an extensive review of the repository and its documentation and confirmed by an audit of actual management, operations and security of the data repository.

The list of Desirable Characteristics for All Data Repositories could be separated into those items that pertain to the data and a second set that pertain to the repository. The former would include A, C, and K. The balance of the items relate to aspects of maintaining a long term repository.

- A. Persistent Unique Identifiers. The PUID should also be mandatory when the data it identifies is transferred to another repository.
- B. Long Term Sustainability. These concepts are best developed jointly by the repository administrators and the preservation managers as part of broad repository planning for future access to the data.
- C. Metadata. Metadata should not be institution specific. It should be, at a minimum, discipline wide, and ideally truly universal.
- D. Curation and Quality Assurance. The current statement is, at best, a bare bones enunciation of the myriad issues, approaches, and assurances involved in these complex tasks. These are the most important tasks that must be performed to ensure preservation and long term access to the data. Ideally they will be performed in accordance with international standards such as ISO 16363, with the results fully documented and available to users to enhance understanding of the data and any inconsistencies or gaps that are present.
- G. Reuse. The concept of tracking reuse of data has overtones of control and censorship. As phrased it may limit reuse

#### Additional Considerations for Repositories Storing Human Data (Even if De-Identified)

While these criteria may still be appropriate for research involving individually identifiable information, this set of considerations is not the appropriate place or approach to revise any existing Federal laws or regulations. Restrictions on access to information must be accompanied by legally valid criteria for restricting such data and include the timeframe and conditions for the ending of restrictions, where applicable.

## **Response to Request for Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded or Supported Research.**

Matthew Woollard, Director, [UK Data Service/UK Data Archive](#), University of Essex

The UK Data Archive is a discipline-specific (social sciences) data archive which has been in continuous existence since 1967. The UK Data Service is an ESRC-funded service which is led from the UK Data Archive at the University of Essex, and works in partnership with other UK institutions.

<https://www.federalregister.gov/documents/2020/01/17/2020-00689/request-for-public-comment-on-draft-desirable-characteristics-of-repositories-for-managing-and-sharing-data-resulting-from-federally-funded-or-supported-research>

---

---

### **Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded or Supported Research**

#### **I. Desirable Characteristics for All Data Repositories**

**A. Persistent Unique Identifiers: Assigns datasets a citable, persistent unique identifier (PUIID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The PUIID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.**

We note that an accession number is not necessarily semantically identical to a PUIID. We would suggest that both are necessary. (It is not impossible for a PUIID to include the repository accession number. In the example below, SN (representing Study Number) is the accession number, and that number is clearly identifiable in the full doi.)

Office for National Statistics, University of Manchester, Cathie Marsh Institute for Social Research (CMIST), UK Data Service. (2019). *Quarterly Labour Force Survey, July - September 2018: Teaching Dataset*. [data collection]. Office for National Statistics, [original data producer(s)]. Office for National Statistics. SN: 8499, <http://doi.org/10.5255/UKDA-SN-8499-1>

**B. Long-term sustainability: Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.**

Long-term should include a minimum number of guaranteed years and a mechanism and terms for appraisal over time. Any significant time period where user or technical change may imply a need to change the data or metadata implies a need for active preservation (beyond the bit-level).

Availability does not imply usability. A dataset may be available for reuse in fifty years, but by being stored on punched cards does not allow them to be used by anyone without a punched card reader!

**C. Metadata: Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.**

The digital preservation community tends to use the phrase “independently understandable”. The implication is that a user can discover, access and use the content without additional help from the repository. This places a high overhead on “general repositories” which need to assume a general

user base. Discipline-specific repositories need only make assumptions about the knowledge within their discipline in the long-term.

**D. Curation & Quality Assurance: Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.**

There is a general discussion in the community around quality measures. Our opinion is that the repository should be responsible for the integrity of datasets and metadata, but the original producer needs to be responsible for its quality. Within the social sciences community we often use the “pregnant men” scenario to describe basic checking. If a dataset includes inconsistencies such as pregnant men, we return the data to the data owners and they are expected to recode or correct these inconsistencies. We also provide a check on the level of anonymisation. On occasion a data depositor has included personally identifiable information in a dataset which was expected to be openly accessible. Errors like this are highlighted to the data owner before data is prepared for long-term preservation.

This provision is helpful, but needs clearer pointers to the responsibilities of the repository and the data creator/owner.

**E. Access: Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.**

Methods of accessing data should not drive the access level. The access level (based on the content of the data) should drive the method of accessing data.

**F. Free & Easy to Access and Reuse: Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.**

This statement has some overlap with the previous one. The phrase “documented as being in the public domain” is unlikely to be a consideration for more recent works. Copyright should always be clarified before a repository accepts data, otherwise worldwide copyright laws may be being broken. Note also that copyright is not rescinded on the basis of a CC license.

**G. Reuse: Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).**

This is a knotty problem. The enablement of tracking does not imply that tracking can take place. There is considerable evidence across the globe that data users are not as careful about referencing/citing data as they might be. The responsibility for ensuring that tracking can take place must lie within the hands of the user and not the repository. For example, the UK Data Service can track some use of data which we hold on behalf of others, but this is likely to be a small proportion of the use which actually takes place. In reality a repository will only be able to “Provide the means for the tracking of data reuse through the assignment of adequate metadata and persistent identifiers (PUIDs)” --- which may be more appropriate wording for this characteristic.

**H. Secure: Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>) or the National Institute of Standards and Technology's 800-53 controls (<https://nvd.nist.gov/800-53>).**



International standards like ISO 27001 are a high bar for some repositories. Question: providing documentation to whom? If this is publicly available it might increase the risk of someone attacking the repository. If this is not publicly available it may in effect be of no value.

**I. Privacy: Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.**

We would change the word privacy to confidentiality here. The risks to privacy per se should be built into the data collection process; the risks to confidentiality are bound up with the management of the data.

**J. Common Format: Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.**

We have yet to identify mechanisms such as community-based standards registries to identify appropriate formats. While we would agree that these data (and metadata) should be made available in non-proprietary formats, this should not preclude making them available in proprietary formats if the community desires that.

**K. Provenance: Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.**

Both pre-deposit provenance and post-deposit provenance should potentially be included here. Pre-deposit provenance also provides a mechanism for managing rights information.

## **II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

**A. Fidelity to Consent: Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).**

As an addition to the statement above, it should be clear that it is the responsibility of the data creator to ensure that the consent (of the original data collection) was carried out in compliance with local ethical/institutional review board.

**B. Restricted Use Compliant: Enforces submitters' data use restrictions, such as preventing reidentification or redistribution to unauthorized users.**

This is a perfectly reasonable statement, however there may be times when data submitters need a more "generalist" approach to understanding their restrictions. Often times, data submitters are more risk adverse than they need to be. Data repositories should not make the decisions on behalf of the data submitters, but they may need to provide guidance so that the data submitters' restrictions are appropriate.

**C. Privacy: Implements and provides documentation of security techniques appropriate for human subjects' data to protect from inappropriate access.**

To make this a more overarching criteria, the words "for human subjects" could be replaced with "the". (i.e., : Implements and provides documentation of security techniques appropriate for the data to protect from inappropriate access.") Some data may need restricted access and protection for other reasons than just protection of human subjects --- sensitive commercial information, rights management, the protection of culturally or ecologically sensitive information such as the locations of artefacts or species. So this statement may need to be broader than just human subjects.

**D. Plan for Breach: Has security measures that include a data breach response plan.**

Agreed

**E. Download Control: Controls and audits access to and download of datasets.**

Agreed – and not necessarily just for human subject data.

**F. Clear Use Guidance: Provides accompanying documentation describing restrictions on dataset access and use.**

Again, this should not just be for confidential data. Data which has been anonymised will, for example, have a (very small) risk of disclosure. Making it clear that the user must not attempt to identify individuals is part of our licence regime.

**G. Retention Guidelines: Provides documentation on its guidelines for data retention.**

This may be a language issue but retention may apply to both the length of time that a dataset is expected to be maintained or whether or not there are requests for the withdrawal of personal information in a dataset.

**H. Violations: Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository.**

It might be better to use the word processes as opposed to plans. Having plans might imply that these are only in the development phase and not operational.

**I. Request Review: Has an established data access review or oversight group responsible for reviewing data use requests.**

In the UK all accesses to data which is deemed personal under the Data Protection Act (or any other legal gateway) are approved by Data Access Committees which have different processes for reviewing data use requests.

Conclusion

In general these characteristics are all sensible and valid; in our opinion only some wording changes for clarity or extensions to provide additional meaningful detail are necessary. It is also worth noting that the [CoreTrustSeal](#) is a community-based standard which provides an assessment against trustworthy digital repositories. They already provide detailed guidance on a set of actions/activities which are required for a digital repository to be considered to be trustworthy. So, their requirements overlap with and complement the characteristics here. However, there is little here which is not already within the CoreTrustSeal, and it might be worth considering making the first characteristic to be certified against the CoreTrustSeal.

Two significant omissions from this set of characteristics are noticeable. The first covers **Scope** --- the repositories should have a detailed scope of engagement. This is important because it allows a relationship between the mission (defining scope) and the ability to deal appropriately with the data. Scope also allows for clarity in the skills which are required to manage a repository – and having the correct skills to carry out the activities which are required by these characteristics.

The second covers **user support**. Some user support is generalist, but some is specific. All repositories which are dealing with specialist data, should be in a position to provide some human-level support about the data. (Not just the finding and accessing of data.) Therefore I would also

recommend these two additional characteristics. Repositories may have additional objectives which are not specifically required, but help facilitate the process. The UK Data Service, for example, carries out a fair amount of **training**, which is specific to the data which it facilitates access to. This is not a requirement, of course, but it allows for better (a higher quality) level of support to researchers.

6 March 2020

This document is also available at: doi: [10.5281/zenodo.3698973](https://doi.org/10.5281/zenodo.3698973)

Thank you for the opportunity to respond to the White House Office of Science and Technology Policy (OSTP) request for information (RFI) regarding desirable characteristics for data repositories used to locate, manage, share and use data resulting from Federally funded research.

The Academy for Radiology & Biomedical Imaging Research (Academy) is a non-profit advocacy organization representing stakeholders of the medical imaging (MI) research community, which advocates for federal investment in medical research broadly and medical imaging specifically at the National Institutes of Health (NIH) and across government agencies.

It is important to note that medical images of a patient are one of the most data-rich, complicated, variable and voluminous resources that result from basic and clinical research funded by Federal agencies. With the advancement of artificial intelligence (AI) applications in medicine the need for data repositories that not only demonstrate the accomplishments of past research, but also support future research is critically important. We are honored to provide our perspective.

A crucial first step to support MI research and development is to aggregate large, anonymized medical image datasets, creating a repository at a secure site (or multiple secure sites), neutral to disparate interests, and with a low barrier to access. These so-called “safe havens” are intended to protect the anonymity of patients’ personal health information (PHI) and related regulations (e.g., HIPAA) while creating broad access for technological development.

Some of the key desirable characteristics for data repositories include:

- **Honest Broker:** Establish an intermediary structure that serves as an “honest broker” for users of the data while maintaining confidentiality. Currently, academic and healthcare institutions are reluctant to share patient data with industry or even each other due to concerns about confidentiality, others using their data in ways possibly not intended (e.g., running studies without having relevant information about the cases like the gold standard for the diagnosis), and simply losing their data to outside parties. The repository would remove these concerns and facilitate more collaborations and data sharing in a secure and confidential manner.
- **Provide Reliable and Validated Data Anonymization Tools:** MI and related patient data must be anonymous in order to be stored in a repository for public access. Ideally the repository should provide reliable and validated anonymization tools for users who do not have access to such tools at their institution. The repository should also have a process in place to periodically verify that data are properly anonymized.
- **Protection of IP:** This type of effort requires a neutral, horizontally-structured platform that would encourage stakeholder collaboration and cooperation in an environment where IP and related commercialization concerns are mitigated by the third-party nature of the MI research and development platform ecosystem.
- **User-Friendly Query Interface:** In order to be useful to the broader research community, a repository must be easy to access, navigate and use. It should not require programming or other technical skills that the average non-technical clinical researcher does not possess. The system should allow direct ad-hoc queries (with adequately prepopulated search terms) that would allow for ready cohort discovery, identification

and selection of useful and relevant data elements/cases, and download of (anonymized) data into commonly used database formats (e.g., CSV, DICOM) for data extraction and analysis.

- **IRB interface:** The repository should have a process in place whereby IRB/IACUC and other relevant approvals can be uploaded and verified before users gain access to the repository.
- **Curation:** The repository data must be curated to ensure that data are properly anonymized, acquired under proper IRB/IACUC procedures and comply with other regulatory considerations (HIPAA), and are updated if necessary (e.g., new information becomes available that changes the “gold standard” or other relevant information associated with the data).

The envisioned resource, once created, would be most valuable if it is sustainable into the future as imaging modalities/technologies change with time. It cannot merely be a one-off intermediate endeavor.

From: [jsh416@gmail.com](mailto:jsh416@gmail.com)

To: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov).

Subject: RFC Response: Desirable Repository Characteristics

This response is submitted by J. Steven Hughes, Information Architect for the Planetary Data System. Steve is currently a Principal Computer Scientist at the Jet Propulsion Laboratory.

Primary discipline: Information Architect and Digital Archivist.

The Planetary Data System (PDS) is NASA's official archive for Solar System Exploration science data. It is a federation of science discipline nodes formed in response to the findings of the Committee on Data Management and Computing (CODMAC) [1] that a "wealth of science data would ultimately cease to be useful and probably lost if a process was not developed to ensure that the science data were properly archived."

The PDS started operations in 1990 with the mission statement, "to facilitate achievement of NASA's planetary science goals by efficiently collecting, archiving, and making accessible digital data and documentation produced by or relevant to NASA's planetary missions, research programs, and data analysis programs."

After two decades of successful operations, the PDS transitioned to a more modern system based on foundational principles from ISO 14721, the Open Archival Information System (OAIS) Reference Model, and lessons-learned from two decades of operations. ISO 14721 is the seminal document for trustworthy data repositories. Subsequently an informal "desk" audit was conducted on the PDS using the ISO 16363 standard, a standard designed specifically to test repository compliance with the ISO 14721 standard. The PDS met over 90% of the ISO 16363 requirements, a significant achievement.

As a response to this RFC, requirements from the ISO 16363 standard and principles from ISO 14721 have been mapped to the desirable characteristics presented in the RFC's draft guidelines. The intent is to illustrate how ISO 14721 principles and ISO 16363 requirements can help enable and test that an archive has the desirable characteristics listed in this RFC.

A list of definitions has been provided at the end of this document.

## I. Desirable Characteristics for All Data Repositories

A. Persistent Unique Identifiers: Assigns datasets a citable, persistent unique identifier (PUIID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally

funded research). The PUID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.

4.2.4 The repository shall have and use a convention that generates persistent, unique identifiers for all AIPs. (ISO 16363)

B. Long-term sustainability: Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.

4.2.9 The repository shall provide an independent mechanism for verifying the integrity of the repository collection/content. (ISO 16363)

4.6.2 The repository shall follow policies and procedures that enable the dissemination of digital objects that are traceable to the originals, with evidence supporting their authenticity. (ISO 16363)

5.1.1 The repository shall identify and manage the risks to its preservation operations and goals associated with system infrastructure. (ISO 16363)

3.1.2.1 The repository shall have an appropriate succession plan, contingency plans, and/or escrow arrangements in place in case the repository ceases to operate or the governing or funding institution substantially changes its scope. (ISO 16363)

3.1.2.2 The repository shall monitor its organizational environment to determine when to execute its succession plan, contingency plans, and/or escrow arrangements. (ISO 16363)

C. Metadata: Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.

4.5.1 The repository shall specify minimum information requirements to enable the Designated Community to discover and identify material of interest. (ISO 16363)

4.2.5.2 The repository shall have tools or methods to determine what Representation Information is necessary to make each Data Object understandable to the Designated Community. (ISO 16363)

3.3.1 The repository shall have defined its Designated Community and associated knowledge base(s) and shall have these definitions appropriately accessible. (ISO 16363)

Mandatory Responsibility - Ensure that the information to be preserved is independently understandable to the Designated Community. In particular, the Designated Community

should be able to understand the information without needing special resources such as the assistance of the experts who produced the information. (ISO 14721)

D. Curation & Quality Assurance: Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

3.3.2.1 The repository shall have mechanisms for review, update, and ongoing development of its Preservation Policies as the repository grows and as technology and community practice evolve. (ISO 16363)

E. Access: Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.

4.6.1 The repository shall comply with Access Policies. (ISO 16363)

4.5.1 The repository shall specify minimum information requirements to enable the Designated Community to discover and identify material of interest. (ISO 16363)

F. Free & Easy to Access and Reuse: Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

4.6.1 The repository shall comply with Access Policies. (ISO 16363)

G. Reuse: Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).

4.3.4 The repository shall provide evidence of the effectiveness of its preservation activities. (ISO 16363)

H. Secure: Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001.

5.2.1 The repository shall maintain a systematic analysis of security risk factors associated with data, systems, personnel, and physical plant. (ISO 16363)

5.2.2 The repository shall have implemented controls to adequately address each of the defined security risks. (ISO 16363)

I. Privacy: Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.

4.3.2 The repository shall have mechanisms in place for monitoring its preservation environment. (ISO 16363)



5.2.1 The repository shall maintain a systematic analysis of security risk factors associated with data, systems, personnel, and physical plant. (ISO 16363)

J. Common Format: Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.

4.2.5.1 The repository shall have tools or methods to identify the file type of all submitted Data Objects. (ISO 16363)

4.2.5.2 The repository shall have tools or methods to determine what Representation Information is necessary to make each Data Object understandable to the Designated Community. (ISO 16363)

4.3.2.1 The repository shall have mechanisms in place for monitoring and notification when Representation Information is inadequate for the Designated Community to understand the data holdings. (ISO 16363)

5.1.1.1.5 The repository shall have software technologies appropriate to the services it provides to its designated communities. (ISO 16363)

K. Provenance: Maintains a detailed log file of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.

**Mandatory Responsibility:** Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity. (ISO 14721)

**Provenance Information:** (Listed as information necessary for adequate preservation) Provenance Information documents the history of the Content Information. This tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated, providing an audit trail for the Content Information. (ISO 14721)

**Definitions:**

**AIP: Archival Information Package:** An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAI. (ISO 14721)

**Content Information:** A set of information that is the original target of preservation or that includes part or all of that information. (ISO 14721)

**Designated Community:** An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities. A Designated Community is defined by the Archive and this definition may change over time. (ISO 14721)

**Mandatory Responsibilities:** Mandatory responsibilities are those responsibilities that an organization must discharge in order to operate an OAIS Archive. (ISO 14721)

**Preservation Description Information (PDI):** The information which is necessary for adequate preservation of the Content Information and which can be categorized as Provenance, Reference, Fixity, Context, and Access Rights Information. (ISO 14721)

**Provenance Information:** The information that documents the history of the Content Information. This information tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. The Archive is responsible for creating and preserving Provenance Information from the point of Ingest; however, earlier Provenance Information should be provided by the Producer. Provenance Information adds to the evidence to support Authenticity. (ISO 14721)

**Representation Information:** The information that maps a Data Object into more meaningful concepts. An example of Representation Information for a bit sequence which is a Flexible Image Transport System (FITS) file might consist of the FITS standard which defines the format plus a dictionary which defines the meaning in the file of keywords which are not part of the standard. (ISO 14721)

[1] National Research Council. 1986. Issues and Recommendations Associated with Distributed Computation and Data Management Systems for Space Science, Committee on Data Management and Computing, Space Studies Board, National Academy Press, Washington, DC, pp. 95.

# BRAIN HEALTH ALLIANCE

---

A 501c3 not-for-profit organization  
Email: [admin@BrainHealthAlliance.net](mailto:admin@BrainHealthAlliance.net)  
Website: [www.BrainHealthAlliance.org](http://www.BrainHealthAlliance.org)

8 Gilly Flower Street  
Ladera Ranch, CA 92694  
Tel: 1-949-481-3121

5 March 2020

Kelvin K. Droegemeier, Ph.D.  
Director, Office of Science and Technology Policy  
Executive Office of the President, Eisenhower Executive Office Building  
1650 Pennsylvania Avenue, Washington, DC 20504

Submitted online to [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov) via email with subject  
RFC Response: Desirable Repository Characteristics

Dear Dr. Droegemeier:

With regard to desirable characteristics for data repositories, please refer to our papers on the DREAM principles and the FAIR metrics that we have published in diverse professional organizations and communities including IEEE, AMIA and ASIS&T over the past 13 years. All of our published papers on the PORTAL-DOORS Project have been freely and continuously available since 2007 at [www.portaldoors.org](http://www.portaldoors.org). They can be found at the publicly accessible web page [www.portaldoors.org/PDP/Site/Papers](http://www.portaldoors.org/PDP/Site/Papers) which also provides access to our conference presentations dating back to those at IEEE, AMIA and W3C in the early years 2008-2010 of the PORTAL-DOORS Project.

We support the PDP and NPDS principles from the original PORTAL-DOORS Project that began in 2006. Recently, we have re-named the PDP-NPDS principles as the DREAM principles for the phrase “Discoverable Data with Reproducible Results for Equivalent Entities with Accessible Attributes and Manageable Metadata.” Moreover, we support the FAIR metrics as the truly quantitative numerical metrics that we have defined for FAIR as the logically consistent and self-referential acronym for the phrases “Fair Attribution to Indexed Reports and Fair Acknowledgment of Information Records.”

Those who wish to promote fairness in any ordinary English use of the word *fair* should adhere to the ethical standards promoted by the COPE organization at [publicationethics.org](http://publicationethics.org) as well as many other organizations that promote integrity in science and scholarly research publishing. Thus, being *fair* and promoting *fairness* also should respect the historical record of the published literature with fair citation and discussion of previously published papers with attention to the importance of *equivalent entities*.

Quoting from our recent paper published at IEEE eScience 2019, “we emphasize that science will be neither reproducible nor fair without recognition, acknowledgment, attribution and citation of equivalent entities regardless of whether those equivalent entities are considered to be scientific hypotheses, scientific experiments, scientific data, scientific results or published articles in the scientific literature.”

We recommend that OSTP and government funding agencies adopt a policy that provides better support for data repositories with sufficient attention to and funding for research and development of solutions to the problems of scientific misconduct. In particular, we recommend allocation of funding to support development of software algorithms and software agents for the automated detection and prevention of scientific misconduct, including plagiarism of the data and plagiarism of published papers about the data technologies, as well as other fraudulent misuse of these data repositories.

Sincerely,



Carl Taswell, MD, PhD  
[CTaswell@BrainHealthAlliance.org](mailto:CTaswell@BrainHealthAlliance.org)

Comments from  
THE FUTURE OF PRIVACY FORUM



to EXECUTIVE OFFICE OF THE PRESIDENT  
Office of Science and Technology Policy

Document Number: 2020-00689

*Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from  
Federally Funded Research*

Email Subject: RFC Response: Desirable Repository Characteristics  
Submitted to:  
Lisa Nichols  
Open Science  
% Sean C. Bonyun  
Chief of Staff, Office of Science and Technology Policy  
Email: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

Dr. Sara R. Jordan, Policy Counsel, Artificial Intelligence  
THE FUTURE OF PRIVACY FORUM <sup>1</sup>/<sub>2</sub>  
1400 I St. NW Ste. 450  
Washington, DC 20005  
[www.fpf.org](http://www.fpf.org)

Scientific Disciplines of Submitting Organization: Law, Public Policy, Machine Learning

Dear Mr. Bonyun,

On January 17, 2020, the Executive Office of the President, Office of Science and Technology Policy (hereinafter OSTP) published a Notice for public comments on the characteristics desired for data repositories storing data from federally funded research projects. We thank the OSTP for the opportunity to submit comments to the Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research.

<sup>1</sup> The Future of Privacy Forum (FPF) is a nonprofit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies.

<sup>2</sup> The views herein do not necessarily reflect those of our supporters or our Advisory Board.

FPF are broadly supportive of the draft guidelines. We believe that the requirement for data built through federally funded projects to be made indefinitely available as described in Part I clearly preserves stewardship of public resources and ensures thoughtful data management and data security from acquisition to archiving to de-accession.

We wish to offer suggestions to modify components of Part II, Additional Considerations for Repositories Storing Human Data (even if de-identified) to ensure effective data sharing between organizations, whether public or private. Our comments are intended to encourage the OSTP to adopt a strong, risk-conscious, approach to privacy protections in the context of sharing personal data gathered through Federally funded research projects. Our concern is that stipulations listed in Part II may limit data sharing across organizations due to incompatibilities in privacy law frameworks, due to enthusiastic but misguided efforts to subject all human data to "HIPAA" data requirements, and due to insufficiently articulated enforcement mechanisms that will may limit robust pathways to realization of these desiderate. We outline our recommendations in line with each of the components to Part II on which we comment.

#### Part II.A: Fidelity to Consent

Consent may be an appropriate mechanism for protecting the privacy and data rights of research participants in many cases, but not in all cases. Guidance from the European Data Protection Board (EDPB) reminds that consent may be less appropriate when there is an imbalance of power between data subjects and researchers.<sup>3</sup> FPF encourages OSTP to adopt a nuanced approach to requirements for fidelity to consent that acknowledge the limitations to consent and reinvigorates the use of consent documents to outline which research purposes conform to participants expectations.

Recent discussions by EU states<sup>4</sup> and by the EU Data Protection Supervisor<sup>5</sup> itself suggest that EU member states will permit sharing of de-identified research data under the guide of "broad consent". "Broad consent" permits researchers to use data for almost any form of clinical research when the data was originally given for the purpose of clinical research. Likewise, the 2018 Revisions to the Common Rule, "broad consent for secondary use may be obtained when standard informed consent is obtained for the original or initial primary research when investigators are interacting or intervening with subjects, for example, for a clinical trial".<sup>6</sup> Broad consent requirements give investigators the latitude to request that subjects consider future

<sup>3</sup> Article 29 Working Party (2018). Guidelines on Consent under Regulation 2016/679.

[https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc\\_id=51030](https://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=51030)

<sup>4</sup> Federal Ministry of Justice and Consumer Protection. (2020). Opinion of the Data Ethics Commission. Federal Government of Germany. January 22, 2020.

[https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten\\_DEK\\_EN\\_lang.html;jsessionid=088D6FC6594FF0130AEC723D7A82FEC1.2\\_cid334?nn=11678512](https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.html;jsessionid=088D6FC6594FF0130AEC723D7A82FEC1.2_cid334?nn=11678512)

<sup>5</sup> European Data Protection Supervisor (EDPS). (2020). A Preliminary Opinion on Data Protection and Scientific Research. January 6, 2020. [https://edps.europa.eu/sites/edp/files/publication/20-01-06\\_opinion\\_research\\_en.pdf](https://edps.europa.eu/sites/edp/files/publication/20-01-06_opinion_research_en.pdf)

<sup>6</sup> Office for Human Research Protections. (2018). Revised Common Rule Q&As. July 30, 2018. <https://www.hhs.gov/ohrp/education-and-outreach/revised-common-rule/revised-common-rule-q-and-a/index.html#broad-consent-in-the-revised-common-rule>

unknown uses of their data and give consent to those unknown future uses, within the restrictions that they must set out for the period of time the data may be stored, maintained, or used. Under these terms, investigators do not need to re-approach subjects to notify them if clinically relevant research results emerge from secondary use under broad consent. The requirement that data managed and shared under these guidelines are faithful to the original consent statement is contradictory to present thinking whether in the US or its major research competitors in the EU.

#### Part II.B: Restricted Use Compliant

The restricted use compliance requirement outlines that a data repository will enforce submitters' data use restrictions. Two concerns arise regarding this requirement: 1) requirements for data repositories to reconfirm and "evergreen" data submitters' preferences for data use restrictions and 2) repositories' required responses to change data as the individuals who submitted data change their individual requirements for data use. Particularly as legislation evolves which allows consumers to restrict secondary uses of their data, including removing their information from databases, repositories may become liable for checking to ensure that individuals' data uses restrictions are reflected in the data use restrictions sent by data holders to repositories.

#### Part II.C: Privacy

FPF recommends that the OSTP include a strong statement for the protection of research subjects' data privacy throughout the research data lifecycle. We recommend adoption of a nuanced and targeted approach to privacy protection which recognizes the different risks to participants that arise from storing and sharing research data in the many forms that research data takes. We advise OSTP to consider including stronger language that outlines best practices for de-identification of data for research uses and recommend OSTP to consult our materials developed on this topic.<sup>7</sup> However, HIPAA requirements are both too narrow and too broad to be applied wholesale to research data. A nuanced assessment of the risks based on data types is needed to protect participants privacy and facilitate data sharing.

We are concerned that the language associated with privacy conflates privacy with security in ways that could lead to aggressive management of all forms of repository data through application of the HIPAA privacy and security rule.<sup>8</sup> While cybersecurity and privacy are intertwined, as the NIST Privacy Framework 1.0<sup>9</sup> outlines, security rules for human subjects data as outlined in HIPAA are not appropriate for all forms of individually identifiable data as described in this Notice. Our partners in research institutions report that secondary uses of data are stymied by broad application of HIPAA requirements for safeguarding of data, including

<sup>7</sup> Finch, K. (2016). A Visual Guide to Practical Data De-Identification. <https://fpf.org/2016/04/25/a-visual-guide-to-practical-data-de-identification/>

<sup>8</sup> Department of Health and Human Services, Health Information Privacy. (2013). Summary of the HIPAA Security Rule. July 26, 2013. <https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html>

<sup>9</sup> National Institutes of Standards and Technology. (2020). NIST Privacy Framework, Version 1.0: A Tool for Improving Privacy Through Enterprise Risk Management. January 16, 2020. [https://www.nist.gov/system/files/documents/2020/01/16/NIST%20Privacy%20Framework\\_V1.0.pdf](https://www.nist.gov/system/files/documents/2020/01/16/NIST%20Privacy%20Framework_V1.0.pdf)

HIPAA level security protocols. One of our concerns is that this section could be read to re-interpret the role of research data repositories as “business associates” under the HIPAA security rule would amplify a risk-averse approach to data sharing and collaboration.<sup>10</sup> Although “organization that acts merely as a conduit for protected health information” is not considered to be subject to a Business Associate Contract under the HIPAA Security rule, there is latitude for reinterpretation of this given other obligations listed for data repositories in this notice. Particularly if data sharing repositories are required to ensure continuous updating of data providers’ sharing preferences, there is an argument to be made that these repositories will perform “data aggregation” or “data analysis” functions in order to carry out their normal business activities.

For organizations that encourage data sharing as part of their repository function or through their work with repositories, imposition of HIPAA Security Rule requirements would be onerous, whether *de jure* through specification as such here or *de facto* through adoption of a common risk averse posture. We recommend that the OSTP work with organizations like FPF to carefully craft the language around privacy protections, whether data is de-identified or not, in repositories storing human data.

#### Part II.E: Download Control

We applaud the inclusion of language here to describe control and audit mechanisms for download of datasets that contain data on human subjects. We encourage stronger language to be included that addresses the automated downloading (“scraping”) of datasets from repositories. In particular, we encourage OSTP to include language that encourages software developers, such as the Python Software Foundation, to include dependencies in their scraping and analytics packages that notify users when their scraping violates repository terms of service or that notify repositories that their data is being scraped. We support use of data in development of automated processes and machine learning research, but encourage a more robust set of controls that incorporate software companies as part of the organizations responsible for download control.

In addition, and in conjunction with our remarks for Part II.H. we encourage the OSTP to pursue design of enforcement actions against organizations who create “shadow repositories” for unrestricted uses of research data.

#### Part II.F: Clear Use Guidance

To effectively facilitate use of data in repositories, a clear-language approach, with robust verbal and symbolic descriptions of restrictions and use permissions, should be incorporated into final requirements for use guidance. The Future of Privacy Forum has developed infographics that describe data on a spectrum of fully identified to fully anonymized on which

<sup>10</sup> “A “business associate” is a person or entity that performs certain functions or activities that involve the use or disclosure of protected health information on behalf of, or provides services to, a covered entity. Business associate functions and activities include: claims processing or administration; data analysis, processing or administration; utilization review; quality assurance; billing; benefit management; practice management; and repricing. Business associate services are: legal; actuarial; accounting; consulting; data aggregation; management; administrative; accreditation; and financial. See the definition of “business associate” at 45 CFR 160.103.” (Emphasis added).

we have received excellent user feedback regarding interpretability and explicability.<sup>11</sup> We encourage adoption of our model as one mechanism for description of datasets and terms of their use. Including language that outlines the potential privacy risks for reuse of the data, including results from a well-designed open data risk-benefit assessment, will clarify boundaries to privacy respecting reuse of the data.<sup>12</sup>

#### Part II.H: Violations

With respect to security of the repository itself, we applaud adaptation of the NIST Cybersecurity Framework<sup>13</sup> and NIST Privacy Frameworks for all repositories storing any form of human subject's data acquired through federally funded research projects, whether funding is direct or "flow through". We encourage the OSTP to include strong language and a robust organization architecture for enforcement of violations of the terms of fair use for data repositories. In particular, we encourage the OSTP to collaborate with analytics software companies to develop dependencies in their packages that monitor and report uses of data from repositories.

#### Part II.I: Request for Review

The Future of Privacy Forum welcomes the opportunity to work with the OSTP to develop policies and procedures necessary to implement an oversight group that can be responsible for reviewing data use requests on behalf of repositories storing human subjects data from federally funded research projects. We have received a grant for the express purpose to design an ethical review process for data sharing between corporations and research organizations.<sup>14</sup> We have committed to development of an ethical data sharing review board that broadly meets the mandate described in this Notice for comment. While it is not our intent to develop a data repository, we will provide a framework for review that is compatible with the research ethics and research integrity infrastructure that already governs federally funded research projects<sup>15</sup> and will serve as an independent body to provide review of data sharing arrangements made between for-profit and not-for-profit, non-profit, academic, and other organizations when those data sharing arrangements are made for the specific purpose of research. Our expertise in

<sup>11</sup> Finch, K. (2016). A Visual Guide to Practical Data De-Identification. <https://fpf.org/2016/04/25/a-visual-guide-to-practical-data-de-identification/>

<sup>12</sup> Finch, K. (2018). FPF Publishes Model Open Data Benefit-Risk Analysis. <https://fpf.org/2018/01/30/fpf-publishes-model-open-data-benefit-risk-analysis/>

<sup>13</sup> National Institute for Standards and Technology. (2018). Cybersecurity Framework Version 1.1. <https://www.nist.gov/cyberframework/framework>

<sup>14</sup> Leong, B. (2019). FPF Receives Grant to Design Ethical Review Process for Research Access to Corporate Data. <https://fpf.org/2019/10/15/fpf-receives-grant-to-design-ethical-review-process-for-research-access-to-corporate-data/>

<sup>15</sup> Jordan, S.R. (2019). Designing an AI Research Review Committee. <https://fpf.org/wp-content/uploads/2019/10/DesigningAIResearchReviewCommittee.pdf>



corporate data sharing practices<sup>16,17</sup>, privacy risks for machine learning systems<sup>18</sup> and embedding data protection principles for machine learning<sup>19</sup> puts our organization in an ideal place to serve as a reliable partner for oversight of data use requests.

#### Conclusion

We commend the Office of Science and Technology Policy for their engagement with stakeholders on crafting these draft characteristics for data repositories. We welcome additional engagement with OSTP as these draft desirable characteristics are developed into more robust guidelines.

<sup>16</sup> Harris, L. & Sharma, C. (2017). Understanding Corporate Data Sharing Decisions: Practices, Challenges, and Opportunities for Sharing Corporate Data with Researchers. <https://fpf.org/2017/11/14/understanding-corporate-data-sharing-decisions-practices-challenges-and-opportunities-for-sharing-corporate-data-with-researchers/>

<sup>17</sup> FPF Staff. (2019). Ethical and Privacy Protective Academic Research and Corporate Data. <https://fpf.org/2019/06/07/fpf-companies-academics-developing-best-practices-on-data-sharing/>

<sup>18</sup> Stalla-Bourdillon, S., Leong, B., Hall, P., & Burt, A. (2019). WARNING SIGNS: The future of privacy and security in an age of machine learning. <https://fpf.org/2019/09/20/warning-signs-identifying-privacy-and-security-risks-to-machine-learning-systems/>

<sup>19</sup> Stalla-Bourdillon, S., Rossi, A., & Zanfir-Fortuna, G. (2019). Data Protection by Process: How to Operationalize Data Protection by Design for Machine Learning. <https://fpf.org/2019/12/19/new-white-paper-provides-guidance-on-embedding-data-protection-principles-in-machine-learning/>

## RFC Response: Desirable Repository Characteristics

Organization: National Renewable Energy Laboratory (NREL)

Person(s) filing the comments:

- **Debbie Brodt-Giles: NREL Group Manager Data, Analytics, Tools and Applications (DATA) and participating member of the Federal Data Strategy Working Group**
- **Kris Munch: Acting Director, Computational Sciences Center**
- **Robert White: Sr. Scientist, Research Operations, Materials and Chemical Science and Technology**
- **Courtney Pailing: Scientific Data Systems Project Leader, Data Management, Analysis and Visualization, Computational Science Center**

Primary Scientific Disciplines for all persons commenting: Data science and data management in a scientific research organization

Comments are included in-line colored **red** below:

### **DRAFT Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded or Supported Research**

#### **I. Desirable Characteristics for All Data Repositories**

A. *Persistent Unique Identifiers*: Assigns datasets a citable, persistent unique identifier (PUIID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The PUIID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available. **(This is important, especially for scientific/research data, because the unique persistent identifier is utilized and referenced in publications; therefore, as a publication will persist forever, so should the data that supports the research findings.)**

B. *Long-term sustainability*: Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical

infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events. **(Agreed – this is very important. You may want to request that agencies, offices, and programs consider developing a common repository that would enable an agency-level repository funded similarly as to other key operations resources. Often the hardest part to establish funding into the future. A general site might do well in many cases for supporting a site with extended longevity. However, some data is not as easily stored in simple publication repo style sites (e.g. Time-series repositories, or material science databases). They all can have DOI's as indicated in A., but storing them is a harder process.)**

C. *Metadata*: Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves. **(Extremely important, because the metadata is what turns data into contextual information, particularly as it applies to the reproducibility of experimental data. It is also some of the most elusive and difficult to capture, since much exists only in lab notebooks, if it did not make it into a publication. While it can be easy to require the most basic metadata (e.g. who, what, when, where), other aspects are quite variable depending on the data source generations; different instruments need different metadata to establish context.)**

D. *Curation & Quality Assurance*: Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata. **(NREL has created several curation-based data repositories for the U.S. Department of Energy. These repositories, allow for data input from external sources, enables data to be curated by experts, holds data under moratoriums until the data is acceptable for release, and, once released, the datasets are made accessible and they are federated to other relevant data repositories like Data.gov, OSTI, and others. Additionally, NREL has developed and implemented repeatable processes on public data hubs and repositories to ensure public data undergoes thorough yet streamlined reviews prior to being made public. These processes should be documented, diagramed and available to the public when possible. These applications could be used as examples for others, and/or could be**

leveraged to build new repositories. Examples: Geothermal Data Repository (<https://gdr.openei.org>), Marine Hydrokinetic Data Repository (<https://mhkdr.openei.org>), DuraMat Data Hub (<https://datahub.duramat.org>), HydroGEN Data Hub (<https://datahub.h2awsm.org>), HyMARC Data Hub (<https://datahub.hymarc.org>), ChemcCatBio Data Hub (<https://datahub.chemcatbio.org>) and ElectroCat Data Hub (<https://datahub.electrocat.org>).

E. *Access*: Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.

F. *Free & Easy to Access and Reuse*: Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain. **(The goal should always focus on free and easy access to data, as well as follow the FAIR (Findability, Accessibility, Interoperability, and Reusability) guiding principles for scientific data stewardship, but sometimes data can be complex and very large which can make it more difficult to access. We have experience with providing access to extremely large datasets. For example, we are providing 40-100 TBs of renewable energy resource data to users based on a new model. We are leveraging our partnerships with cloud hosting providers (Amazon Web Services (AWS) and Google) to host our high-value open datasets for free in the cloud. The data is free to all users and they can get the data directly from AWS and Google in a variety of ways. They can choose to access data in the cloud and move it to an environment in the same regional zone free of charge. They can also utilize our Data Lake environment that enables them to mash-up data and do computations on the data free of charge. If a user wants to download the data to their own computer or transfer it to a different regional cloud environment, then the user will incur costs to “transfer” that data elsewhere. This model allows for free access and easy reuse – but puts the data transfer costs on the user (similar to getting a book for free but paying for shipping costs). I wanted to bring up this example, because although the data itself should always be free, sometimes a user may incur a transaction cost for moving the data to various locations. Generally, the availability of cloud services, along with the raw data, support the ease of reuse, although it may cost the user some of their own money.**

G. *Reuse*: Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID). **(Usefulness of this depends on how you plan to use this information. This could be tough to implement, but interesting. A scan of DOIs used in publications after the initial generation of the data would be possible, but if only used as a citing reference then the DOI would generate a false positive on whether the data was re-used or simply providing supporting context.)**

H. *Secure*: Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>) or the National Institute of Standards and Technology's 800-53 controls (<https://nvd.nist.gov/800-53>). **(Good, but we need a common guide within agencies for these same issues and not depend on an aggregate of several other institutions. We also need the guide to be easy enough for researchers to understand what needs to be in place when requesting development of sites to distribute their data either to a private consortium, customers, or a general public release: Bonus application is figuring this out for mixed moderate public data repositories.)**

I. *Privacy*: Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.

J. *Common Format*: Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format. **(Within certain science domains and standardized analysis this might be possible. A better idea would adhere to the best practices of data science where data should be: Non-proprietary, Unencrypted, Un-compressed, common adoption, easily interoperable by machines and humans. Typically, this means simple ASCII or UTF-8, CSV for datasets, simple text files for all other relevant information.)**

K. *Provenance*: Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity. **(Repositories should also try to track new iterations of the dataset. For example, if a user took a dataset, added new data to it, and created a new dataset with the original data as the base, that new dataset should provide provenance that gives proper recognition to the original data owner and informs the public about how the new dataset differs from the original. In many cases this is possible and needed. Provenance in live datasets from a database are harder to control.)**

L. *Licensing*: Documents the proper license terms for each dataset to allow users to properly use, reuse, and attribute data to the data owner (citing formats and license terms). NREL lists DOI as well as the OSTI DOE Data Explorer page (see [Citation Information](#)) alongside public datasets when possible.)

## II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)

A. *Fidelity to Consent*: Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).

B. *Restricted Use Compliant*: Enforces submitters' data use restrictions, such as preventing reidentification or redistribution to unauthorized users.

C. *Privacy*: Implements and provides documentation of security techniques appropriate for human subjects' data to protect from inappropriate access.

D. *Plan for Breach*: Has security measures that include a data breach response plan.

E. *Download Control*: Controls and audits access to and download of datasets.

F. *Clear Use Guidance*: Provides accompanying documentation describing restrictions on dataset access and use.

G. *Retention Guidelines*: Provides documentation on its guidelines for data retention.

H. *Violations*: Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository.

I. *Request Review*: Has an established data access review or oversight group responsible for reviewing data use requests.

# RDAP response to Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

Responder: The Research Data Access & Preservation Association (RDAP)

Response: Discipline Agnostic

Role: Data Practitioner Professional Association

The Research Data Access and Preservation (RDAP) Association offers its comments on the Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research. To put this response in context, RDAP is a community of data practitioners who work in a variety of roles and disciplines. Our goal is to support an engaged community of information professionals committed to creating, maintaining, advancing, and teaching best practices for research data management, access, and preservation. Many of us are actively engaged in assisting researchers with writing and complying with data sharing policies from publishers and funders from a variety of fields and facilitate data submission into institutional repositories. Collectively we possess a wealth of knowledge on how to support data management and sharing as well as the technical expertise to ensure that research data remains usable and accessible.

## The proposed use and application of the desirable characteristics

This document begins with what these characteristics will not be used for: “[f]ederal agencies would not plan to use these characteristics to assess, evaluate, or certify the acceptability of a specific data repository”. This statement needs to be clarified as it can be read in different ways. One possible reading is that federal agencies themselves won’t certify whether a repository is acceptable or not. However, it could also be read as these characteristics should not be used to evaluate repositories used to store federal research, which undercuts the goals of this document. Additionally, this statement can seem contradictory to the remaining proposed purposes. We largely agree with how these criteria *should* be used. Assisting Federally funded investigators with identifying appropriate data repositories is a laudable goal; however, researchers often need substantive help with this process, as they aren’t familiar with the terminology listed below. To mitigate this issue, we suggest the inclusion of resources such as local experts and online educational materials already available to fill these gaps in knowledge.

# The appropriateness of the “Desirable Characteristics for All Data Repositories”

## A. Persistent Unique Identifiers

Persistent Unique Identifiers (PUIDs) are critical for data citation and data access, and consequently, data reuse and reproducibility. Explicitly stating these downstream effects of PUIDs will help researchers understand the importance of this characteristic. Recommendations or rankings for the types of PUIDs would be useful, as there are many competing standards.

## B. Long-term sustainability

This characteristic should be renamed to ‘preservation’ to match with the language commonly used in current Data Management Plans (DMPs). Long-term preservation is not only about keeping the data as-is over the long term, but also to protect against degradation and loss. If the data aren’t also usable long term, the preservation efforts undertaken don’t mean much. This characteristic should refer to the common format criterion and assess whether format migration may be appropriate for the data type stored in a discipline-specific repository.

## C. Metadata

Metadata is critical for understanding and citing data stored in repositories, and thus reuse. The implications for reproducible research and metadata should be emphasized to indicate the importance of this characteristic. Additionally, the word “sufficient” is not adequate guidance for researchers, as metadata standards vary in depth and breadth of use. Once again, pointing to resources that explain these terms and how to evaluate metadata options would improve the utility of this document. We also encourage the Open Science Committee of OSTP consider future guidance / further RFIs about metadata standards for disciplines that currently don’t have them.

## D. Curation & Quality Assurance

This characteristic is straightforward if a repository has data curation staff who ensure that data are curated properly upon submission. However, the phrasing “has a mechanism for others to provide” is unclear. Does it mean that data curation is an allowable grant cost? If so, this seems out of the scope of this document on infrastructure. Please clarify the intent of this clause. Additionally, researchers will not likely have a good idea of what ‘expert curation’ means. This term should be defined.

## E. Access and F. Free & Easy to Access and Reuse:

The distinction between characteristics E and F is subtle, and ultimately not useful. We suggest combining these characteristics or clarifying the intent of E and how it is different than F. We also suggest broaching the concept of licensing to explicitly state conditions for use. This issue



is complicated because data are not copyrightable in all jurisdictions, or equally across formats (e.g. text vs. images).

#### G. Reuse

This characteristic needs to be more specific. Does this mean only that a repository needs sufficient metadata and a PUID? If so, these considerations are already covered in previous characteristics. Does it mean that the repository itself must be able to enumerate where and when data is cited? If so, then this is problematic as access to literature reference and citation metadata is not universally free and open. Furthermore, standards on how to track and count data citations, repository page views, and downloads are still in development. This section should also include mention of what data formats should be used and how to migrate obsolete formats.

#### H. Secure

This characteristic lists a specific ISO and NIST standards, making it clear what technical considerations are in play. However, it is not clear how the average researcher would be able to determine whether a repository complies with these standards, making it less useful.

#### I. Privacy

Privacy is of the utmost concern, especially when dealing with controlled access databases that contain private information. This characteristic contains general cybersecurity concepts that are relevant, but doesn't provide specifics about what is actually necessary for a particular data type. Additionally, the language used in this characteristic would not be understandable by all researchers and is therefore of limited utility to some of your target audiences. Suggesting resources like local IT and data services staff to help evaluate these criteria is critical to mitigate this concern.

#### J. Common Format

This characteristic should be moved up where the metadata characteristic is discussed. Additionally, common formats for data and metadata should be separated into two characteristics, as this concept is subtle and distinctions must be stated explicitly. Adding more details on types of formats that are desirable or where to find standards would help researchers interpret this characteristic.

#### K. Provenance

Logfiles are typically a feature that is hidden from the end user, and thus many researchers are unaware of what they are and why they are important. More detail here would help researchers understand what they are looking for; however, it's unclear how easy it would be to determine if a given repository utilizes logfiles to document changes. Additionally, addition of human readable text for what changes were made and not logfiles will help with the usability of this characteristic.

## II. Additional Considerations for Repositories Storing Human Data

We do not have specific comments on each of the considerations for repositories storing human data, because we are not specialized in this area. We would like to emphasize that most researchers are not fluent in data curation and cybersecurity concepts, and will likely need more guidance than what is listed here. We recommend providing suggestions for where they can get help with evaluating repositories for characteristics that they are not familiar with.

### Additional characteristics that RDAP thinks should be included:

In general, the characteristics listed above are consistent with what is important when thinking about where to deposit data. The existing repositories that we recommend to our faculty and students largely meet the desirable characteristics listed here. However, one of your target audiences, federally-funded investigators, would not find the current definitions helpful, as they are not written in discipline neutral (i.e. non-jargon) language. Assuming that the terms used in this document are widely understood is a mistake.

We appreciate the fact that these desirable characteristics are not intended to change drastically over time, but as technology changes, the specifics must change and evolve with the research landscape, new technologies, and new data security requirements. A criterion regarding how the repository is funded and plans for data preservation in the event that funding is no longer available should be added. Reminding researchers that many institutions have both research data practitioners to answer their questions and institutional repositories to deposit data when a disciplinary repository is not available could assist in reducing confusion and increasing compliance. That said, we are not sure that the stated goal of improving consistency will be met, as the desirable characteristics are similar to those already used for evaluation, and do not add a stricter level of detail that would make them more useful to non-data practitioners.

## RFC Response: Desirable Repository Characteristics

Name: Trevor Stanley

Organization: National Renewable Energy Laboratory

Scientific Discipline: Energy Science and Computer Science

Comments:

- Research funded by taxpayer dollars should be accessible in a timely manner and reasonable (i.e. human readable and or interpretable) format so long as it does not adversely impact national security.
  - This includes data that might be politically sensitive. All data and associated analysis of the data should be open access irrespective of if the findings conflict with political or other interests.
- Automatic metadata analysis should be published with any dataset that is being shared.
- Consider using a public ledger and or blockchain approach for storing, sharing, referencing, and confirmation/validation of all government datasets and ingested datasets from external entities.
  - BurstIQ is an example of a company that does this with Healthcare and Pharmaceutical data
- Include descriptions of how the data was collected and or created and or compiled. This includes the sources, instruments, and method of recording.

Below is a list of comments from offices across the National Oceanic and Atmospheric Administration in response to the Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research.

Submitted by Monica Youngman, Director, Data Stewardship Division in NOAA/NESDIS/NCEI. Primary scientific discipline: Physical Sciences. Role: Federal Manager in Data Management.

#	Commenter Email	Affiliation	Role	Discipline	Section	Item (use N/A if a general comment or other)	Comment
1	tyler.christensen@noaa.gov	NOS	data manager	data manager	Section I	N/A	suggest adding an item on contingency plans to ensure data are not lost if the repository needs to close
2	tyler.christensen@noaa.gov	NOS	data manager	data manager	Section II	F. Clear Use Guidance	should apply to all repositories, not just ones that store human data
3	tyler.christensen@noaa.gov	NOS	data manager	data manager	Section II	G. Retention Guidelines	should apply to all repositories, not just ones that store human data
4	nazila.merati@noaa.gov	NMFS	data manager	data manager	General	N/A	Consider suggesting that repositories and archives have some level of certification (e.g. core trust seal), to insure providers and users that data is "trustworthy"
5	Nazila.merati@noaa.gov	NMFS	data manager	data manager	General	N/A	Provide clear guidance to data providers about what is involved in data submission and a timeline for submission through acceptance
6	nazila.merati@noaa.gov	NMFS	data manager	biological sciences	General	N/A	Many of the characteristics in section 2 apply to environmental and socioeconomic data and should be applied to all data in repositories
7	eugene.burger@noaa.gov	OAR	data manager	data manager	Section I	N/A	Allow for software source code archival, along with compilers.
8	howard.diamond@noaa.gov	OAR	researcher	physical sciences	Section I	Curation and Quality Assurance	Providing, or having the mechanism for others to provide, expert curation and quality assurance is in theory a good thing, but in practice could be problematic if (1) an outside non-Federal person is identified - that could be problematic to get that person access to the Federal archive from an IT security aspected, and second, whether an internal Fed or outside non-Fed is identified, the resources have to be available to support that curation. Such resources are seldom if ever accounted for, they are simply assumed to be in place, and that is not a good assumption.
9	matthew.mahalik@noaa.gov	OAR	data manager	physical sciences	Section I	N/A	Provide guidance for the recommended process and timeframe of data updates. How often should data entries be refreshed with updated information, if at all?
10	chris.krug@noaa.gov	OAR	data manager	data manager	Section I	Reuse	Collaborative institute researchers desire this attribute to evaluate 'value' of data.
11	nancy.ritchey@noaa.gov	NESDIS	data manager	data manager	General	n/a	all repositoryes should provide clear guidance on what data and information should be preserved to ensure independent understanding of the data

12	nancy.ritchey@noaa.gov	NESDIS	data manager	data manager	Section II	G. Retention Guidelines	Retention schedules should be established by all repositories. Transparency on those schedules and the review process needs to be publically available.
13	nancy.ritchey@noaa.gov	NESDIS	data manager	data manager	General	n/a	transparency on repository processes, reviews, appraisals, etc. should be publically available.



March 6<sup>th</sup>, 2020

Lisa Nichols  
Office of Science and Technology Policy  
Executive Office of the President  
Eisenhower Executive Office Building  
1650 Pennsylvania Avenue  
Washington, DC 20504 Washington, DC 20230

Subject: Comments on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research, Document Number 2020-00689.

Dear Ms. Nichols,

The Computing Research Association (CRA) is an association of more than 200 North American academic departments of computer science, computer engineering, and related fields; laboratories and centers in industry, government, and academia engaging in basic computing research; and affiliated professional societies. CRA's mission is to strengthen research and advanced education in the computing fields, expand opportunities for women and minorities, and improve public and policymaker understanding of the importance of computing and computing research in our society. To that end, we write today to submit comments on "Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research" Document Number 2020-00689.

We commend the NSTC Committee on Science's Subcommittee on Open Science (SOS) for developing this set of desirable characteristics of data repositories for data resulting from Federally funded research. Grounding them in the SOS-developed *findable, accessible, interoperable, and reusable* (FAIR) principles goes far in establishing characteristics that will be broadly acceptable and useful.

Data repositories are socio-technical in nature: they provide a service for people, and their utility is tightly intertwined with human behavior in response to the information they provide and the research they enable. This behavior itself changes through the availability of and services provided by data repositories. Focusing on the characteristics of data repositories is vital, but the human infrastructure that needs to be developed around their use is equally vital. Such considerations are outside of the scope for this RFC, and so we encourage the SOS to consider them in future discussions that engage the Research

Librarian Community - such as the Association of College and Research Libraries (ACRL) of the American Library Association (ALA).

Specific to the RFC, we make the following comments:

- To “assist investigators in identifying data repositories”, per this CFP, it is important that repositories document their own collection policies, clearly articulating their self-defined scope and use/reuse policies, including: (a) what does and does not meet the repository’s selection or inclusion criteria (particularly for, but not limited to, human-subjects data); (b) retention guidelines for both human- and non-human-subjects data (related to point II.G); (c) licenses and terms of use that govern both data and metadata where not specified at the dataset-level; etc.
- We would like to see a commitment to supporting requirements for automated access and machine use, including autonomous computational use and reuse of data, by making data and metadata machine-readable and -actionable. There is widespread consensus in the scientific research community (reflected in the FAIR<sup>1</sup> data principles and growing consensus around their implementation across disciplines) that repositories intended to promote reuse must facilitate both human and machine use of data and metadata. (See, for example, “Make scientific data FAIR” by Shelly Stall et al., Nature Comments, June 2019).
  - For example, we recommend that point I.C be amended as: Metadata: Ensures datasets are accompanied by *machine-interpretable* metadata
  - We also recommend that point I.J be amended as: Common Format: Allows datasets and metadata to be accessed, downloaded, or exported from the repository in standards-compliant, *machine-actionable*, and preferably non-proprietary formats
- Supporting the reuse of data in computational workflows will require supporting robust versioning of data that are subject to ongoing change, updates, or growth over the lifetime of research and reuse. Versioning entails more than the adequate identification of individual datasets, and also involves operations such as data cleaning, data reduction, and derivation of secondary data sets from lower level data that may also be archived.

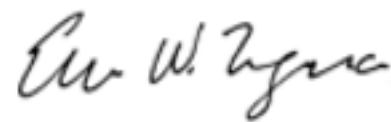
---

<sup>1</sup> <https://www.force11.org/group/fairgroup/fairprinciples>

- In addition, to support computational and human reuse the implicit definition of *provenance* given in these recommendations should be expanded to include not only actions taken during the life of the dataset *after* deposit into the repository, but also lineage or source information for datasets and metadata about actions taken before deposit in the repository.
- Along with the recognition of the importance of restricting access to data in some cases for privacy reasons, a need for recognition of both:
  - The existence of factors that transcend the legal and ethical frameworks that govern *individual privacy*, which may entail restrictions for non-privacy reasons, especially for data that represent human communities or their knowledge
    - E.g., representations of Indigenous populations or their knowledge may be restricted to protect cultural knowledge in accordance with community epistemologies and values
    - The importance of *transparency* as a counterbalance to restriction: Where appropriate, repositories should commit to displaying which data are restricted, under what constraints, and for what reasons.

CRA looks forward to assisting the Department and BIS throughout this proceeding to assess the need for and contours of any changes to this rule. Please contact Peter Harsha of CRA ([harsha@cra.org](mailto:harsha@cra.org)) with any questions concerning these comments, or for assistance on any computing-related technical matter within the scope of this docket. Thank you for your time and attention.

Respectfully submitted,



Ellen W. Zegura  
Chair  
Computing Research Association



Note: These comments were authored by Assistant Professor Katrina Fenlon (University of Maryland College of Information Studies) and members of the CRA [Computing Community Consortium](#) subcommittee.



## Society of Vertebrate Paleontology

7918 Jones Branch Drive, Suite 300

McLean, VA 22102 USA

Phone: (301) 634-7024

Email: [svp@vertpaleo.org](mailto:svp@vertpaleo.org) Web:

[www.vertpaleo.org](http://www.vertpaleo.org)

FEIN: 06-0906643

March 6, 2020

**Subject:** RFC Response: Desirable Repository Characteristics

Dear U.S. Office of Science and Technology Policy,

We represent the Society of Vertebrate Paleontology (SVP: <http://vertpaleo.org/>), a non-profit international scientific organization with over 2,000 researchers, educators, students, and enthusiasts. Our mission is to advance the science of vertebrate palaeontology (a discipline within life sciences) and to support and encourage the discovery, preservation, and protection of vertebrate fossils, fossil sites, and their geological and paleontological contexts. This letter is in response to the White House Office of Science and Technology Policy's (OSTP) for public comment on *Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research* (85 FR 3083; pages 3085–3087; document number 2020-00689). All of our comments concern the middle and right columns on page 3086 85 FR 3083, including “I. Desirable Characteristics for All Data Repositories.” SVP does not have any specific comments on “II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified).”

### **Types of Paleontological Data and Metadata to be Managed by Repositories**

We understand that the “proposed characteristics are intended to be consistent with criteria that are increasingly used by non-Federal entities to certify data repositories, such as ISO16363 Standard for Trusted Digital Repositories and CoreTrustSeal Data Repositories Requirements, so that repositories with such certifications would generally exhibit these characteristics” (page 3086). In addition to the requirement that all digital data from federally funded research should be repositied, SVP suggests that the language of this regulation be expanded to include the physical fossils collected by federally funded research. This is because physical fossils are also a form of data in the field of paleontology besides all associated information and generated data stemming from them, hereafter collectively referred to ‘paleontological metadata.’ Paleontological metadata, include, but not limited to:

- hard copy data (e.g., maps; photographs; field notes, including qualitative and/or quantitative measurements used or taken by researchers; catalog cards; letters containing specimen data; scientific illustrations; publications);
- digital data (e.g., various types of databases, including those that record locality and stratigraphic information, taxonomic and specimen catalogs, measurements, as well as names of land owners, collectors, donors, and/or preparators of fossils; digital photographs; 2-D and 3-D digital scan data; GPS coordinate data; electronic scans of hard copy data; electronic communication containing specimen data; publications);
- replicas (copies of fossils, including molds and digital data to make casts; 3-D prints based on digital data); and

- 'data reserves' for possible future studies, including chemical and microscopic analyses (e.g., rocks and sediment samples; fragmentary fossils; associated fossils collected with primary fossils).

The characteristics of an appropriate repository needed for best practices in paleontology are those that provide long-term preservation and access of not only digital data but also physical fossils and any other forms of paleontological metadata. Because science is an endeavor to make new discoveries, the types of metadata listed above should not be considered comprehensive, where presently unforeseen new types of paleontological metadata may come about in the future that repositories should also accommodate their storage and dissemination. In addition, paleontological metadata to be repositied may even include information in the absence of actual collected fossils. Examples include locality and stratigraphic data of known paleontological sites that have not yet been scientifically explored. Digital data in paleontology include those that represent 'extractions' from physical fossils (e.g., digital scan data as well as field photographs and notes when fossils were surveyed or collected) and therefore are implied pointers to information that is subject to verification. It must be noted also that such information and databases, regardless of whether or not any actual fossil specimens have been collected, often implicitly contain hypotheses or other potential intellectual properties. In addition, restoration and reconstruction of fossils, including physical skeletal mounts, restored fossil elements, digitally reconstructed anatomical elements or skeletons, or even scientifically-based artwork of extinct organisms (including digital images) should also be considered as forms of paleontological metadata where they potentially represent testable hypotheses.

From SVP's perspective, desirable repository characteristics are those that can accommodate management of all types of physical fossils and paleontological metadata. For physical fossil specimen care as well as paleontological metadata storage and dissemination, a wide range of capabilities exists. Efforts should be made by agencies to assist where possible with the ultimate goal of bringing each up to consistent standards. For practical considerations, inadequacies should not exclude granting or maintenance of repository status, but rather additional support should be given to such repository agencies or institutions to help bring them to consistent standards.

We would also like to have a clarification. As noted above, 3-D digital scan data that capture the three-dimensional likeness of objects, such as paleontological (as well as biological and archaeological) specimens, can allow for the reproduction of precise replicas of these objects for scholarly or commercial uses. In cases where these objects are owned by the Federal Government (i.e., original specimens collected from federal lands), reproduction rights are controlled by the permit agreements under which they were collected, and associated federal regulations. How will replica production be restricted, if at all? The rules should allow replica production at least for scholarly and educational purposes.

### **Desirable Characteristics of Paleontological Repositories**

The principle reason for placing scientifically important fossils in a public repository is that vertebrate fossils are rare and often unique. Scientific practice demands that conclusions drawn from the fossils and associated paleontological metadata should be verifiable: i.e., scientists must be able to reexamine, re-measure, and reinterpret them, where such reexamination can happen decades or even centuries after the fact. Furthermore, technological advances, new scientific questions, and opportunities for synthetic research mean that new research often utilizes fossils and associated paleontological metadata that

were originally collected with other purposes in mind. These lines of reasoning mandate that scientifically important fossils be preserved along with their associated paleontological metadata for decades, centuries, and hopefully millennia. Optimal characteristics of suitable repositories include:

- a primary mission that encompasses the preservation of scientifically important fossil specimens and associated paleontological metadata;
- a non-profit organizational structure that is capable of weathering economic changes, political changes, and other changes of fortune
- a demonstrated commitment to preserving specimens and to managing associated metadata such as locality and contextual information (see U.S. Department of Interior's guidelines for federally approved repositories and SVP's *Best Practice Guidelines for Repositing and Disseminating Contextual Data Associated with Vertebrate Fossils* ([http://vertpaleo.org/GlobalPDFS/SVP-Paleo-Best-Practice-Guidelines-\(2nd-Ed\).aspx](http://vertpaleo.org/GlobalPDFS/SVP-Paleo-Best-Practice-Guidelines-(2nd-Ed).aspx));
- a commitment to hiring staff with advanced degrees or equivalent training in paleontological science, curation, and preservation;
- a well-considered policy for keeping fossil specimens and their associated paleontological metadata in the public trust should circumstances change such that the repository no longer able to care for them; and
- a primary mission that includes facilitating active research on the repository's fossil and associated paleontological metadata holdings.

Appropriate repositories therefore include publicly accessible, non-profit museums, universities, colleges, geological surveys, and government agencies whose funding does not hinge on the success of a single company, whose mission statement includes research or education, and whose policies include protocols for keeping material in the public trust if the institution can no longer care for it. Institutions that are set up as non-profit organizations largely independent of the original benefactors would most likely be recognized as credible repositories by peers in the field of vertebrate paleontology.

### **Access and Dissemination of Paleontological Data and Metadata by Repositories**

Reproducibility of paleontological research rests on the premise of permanency and accessibility of examined fossil specimens as well as paleontological metadata, including digital data, deposited in stable repositories under public trust. Because fossils are nonrenewable resources where every fossil specimen is unique, storage of and access to them, along with all associated metadata, must be done with care by repositories. The presumption is that all fossil specimens and paleontological metadata, including digital data, curated by repositories remain permanently stored and accessible to anyone who wishes to access them. However, in some cases, public access to physical fossils and/or paleontological metadata in repositories may need to be controlled, especially if it can result in harm to the fossils, to on-going research, or to the fossil localities. In particular, data pertaining to specific locations of fossil collecting sites must be regarded as 'sensitive' where the following two conditions should be met before placing them in maximally open access data repositories: 1) for fossils collected from U.S. public land, clearance to release the geographic coordinates must be obtained from the relevant secretary as required by the Paleontological Resources Preservation Act (PRPA); and 2) for all paleontological sites, the sensitivity standards outlined in SVP's *Best Practice Guidelines for Repositing and Disseminating*

*Contextual Data Associated with Vertebrate Fossils* ([http://vertpaleo.org/GlobalPDFS/SVP-Paleo-Best-Practice-Guidelines-\(2nd-Ed\).aspx](http://vertpaleo.org/GlobalPDFS/SVP-Paleo-Best-Practice-Guidelines-(2nd-Ed).aspx)) should be followed. In addition to the details about our sensitivity standards, the *Best Practice Guidelines* also provides information concerning the handling of paleontological metadata, including digital data. Much of the following paragraphs come from the document, where the phrase ‘contextual data’ is replaced with ‘paleontological metadata’ for the purpose of this comment letter.

Wherever possible, paleontological metadata stored in repositories, including unpublished forms, should be disseminated freely and widely. However, in some cases, public access to paleontological metadata, especially the precise location of the collecting site, can result in harm to fossils, contextual information (e.g., taphonomic or sedimentologic data), on-going research, or to non-paleontological resources (e.g., endangered species or delicate ecosystems) that remain in the field. In such cases, distribution of information may need to be controlled in compliance with relevant laws and regulations as well as professional ethical standards, although the presumption remains in favor of release. Any restrictions placed on the dissemination of paleontological metadata should be well justified and adhered to rigorously by the repository as well as the collector and all parties with whom the data have been shared.

The sensitivity of all paleontological metadata, especially the location of the collecting site, should be reviewed by the repository as well as by the permitter (i.e., governing body responsible for, or the owner of, the land where the fossils were collected) and the permittee (i.e., collector/researcher) to the best of their ability. In order not to hinder research, curation, and education, the review should be completed as expeditiously as possible. Dissemination of paleontological metadata should be restricted only when there is a genuine risk to the collecting site. Restricting paleontological metadata may affect the precision of research based on aggregated data, such as analysis of fossil occurrences in online public data portals. Therefore, restrictions should be imposed only if absolutely necessary, whereas all paleontological metadata should be made available for research upon request.

Repository managers should consider the needs of users for access to paleontological metadata and other documentation when they evaluate sensitivity and weigh the impacts of disseminating data and restricting their access. For paleontological sites on U.S. Federal lands that fall under the PRPA, this determination is, by law, the responsibility of the agency (permitter) that manages the land. In cases where restrictions are placed on access to paleontological metadata, the original data should be retained intact by the repository, and original data should never be altered, falsified, or discarded. Because research depends on the accuracy of data, repositories should inform the data users about omissions or changes that have been made to metadata in the interest of protecting a site. In cases where redacted data are disseminated, especially cases where the precision of geographic coordinates or stratigraphic placement has been purposefully reduced to protect the location of the collection site, the fact that this has been done should be distributed as part of the metadata for that specimen. In public databases, such as repository catalogs or data aggregators (e.g., online data portals), redacted records should be indicated with appropriate wording, rather than by leaving fields blank or null.

Whenever a repository receives an application for access to restricted data, the assumption of continued sensitivity should be avoided. Rather, the occasion should be used as an opportunity to re-evaluate the determination. Decisions made by government agencies to release previously restricted paleontological metadata must be made in consultation with the repository in order to meet the needs of non-governmental partners, the scientific community, and the general public. Cooperation with relevant governmental bodies is particularly important for repositories or situations where a 'freedom of information access'

law applies in order to discuss potential ramifications of sharing requested sensitive information prior to its formal release.

Repositories acting as data custodians are responsible for receiving, maintaining and preserving all paleontological metadata related to localities, specimens, and collection acquisitions. While these data are maintained in public trust, complete access to data may be restricted at the discretion of the data custodian or as required by law. In the event that data are restricted, the repository manager should disclose this fact to data providers as well as data aggregators and distributors, or should include descriptive language to this effect on their respective online search forms. Should the extent of publicly available paleontological metadata prove insufficient for a given purpose, data users are encouraged to contact individual repositories for more specific inquiries. Repository managers should assess the needs of the user and the fitness for use of the request. Besides their names and institutional affiliations, data users may be asked to provide the following justification to repository managers: 1) a description of the data they seek to obtain; 2) a description of their research, education, resource management, or other public benefit project, and why the requested data are pertinent or essential to their research questions; and 3) a description of how they intend to use and disseminate the data if the request is granted. Repository managers are responsible for relaying institutional policies and specifying any terms and conditions that may be placed on information for release. It should be noted that paleontological metadata are not necessarily always precise, accurate, complete, or reliable. Records may be unverified, vague, contain inherent errors, or reflect incorrect data. Data custodians should impress the importance of not using search results uncritically, as failing to acknowledge these limitations may undermine the legitimacy of certain data interpretations.

We have one question concerning the dissemination of paleontological metadata, including digital data. In the case of data that are exempt from Freedom of Information requests so as to protect in situ scientific (and cultural) resources, such as paleontological (and archaeological) site data, how will these data be protected, and what information would the Persistent Unique Identifier (PUID) or Digital Object Identifier (DOI) point to?

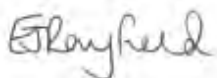
Thank you for the opportunity to comment on this very important issue for scientific advancement. Comments and questions concerning this comment letter and/or our *Best Practice Guidelines* can be addressed to any one of us (our e-mails given below) or Dr. Kenshu Shimada (Chair of SVP's Government Affairs Committee: [kshimada@depaul.edu](mailto:kshimada@depaul.edu)).

Sincerely yours,

Emily J. Rayfield, Ph.D.  
*SVP President*  
[e.rayfield@bristol.ac.uk](mailto:e.rayfield@bristol.ac.uk)

Jessica M. Theodor, Ph.D.  
*SVP Vice President*  
[jtheodor@ucalgary.ca](mailto:jtheodor@ucalgary.ca)

P. David Polly, Ph.D.  
*Past SVP President*  
[pdpolly@indiana.edu](mailto:pdpolly@indiana.edu)



Attn: Lisa Nichols, Assistant Director for Academic Engagement

Office of Science and Technology Policy  
725 17<sup>th</sup> Street, Washington, DC 20501

RE: OSTP RFC: Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

Massachusetts Institute of Technology's comments on Federal Register Document 2020-00689

Chris Bourg, Director, MIT Libraries / cbourg@mit.edu  
Massachusetts Institute of Technology, Cambridge, MA

The Massachusetts Institute of Technology (MIT) Libraries appreciates the opportunity to comment on the merits of the OSTP's RFC on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research. The topic of repositories is of particular interest to the MIT Libraries due to repositories' role in enabling the better world we seek where there is abundant, equitable, meaningful access to knowledge and to the products of the full life cycle of research.

### **Background**

We appreciate the stated goal to improve consistency of recommended guides and practices on the long-term preservation of data from federally funded research, and the references to existing standards. Given the evolving nature of research and data, and particularly computational research, we propose that best practices are challenging to capture and recommend acknowledging this by using the phrase "current good practices," rather than "best practices."

Regarding the use of these characteristics, we recommend in the second paragraph that "could" be replaced with "should" as these characteristics should apply to repositories being used for the storage and preservation of federally funded research data. These are a set of recommended characteristics rather than repository requirements, which is a non-binding phrasing.

Positioning the use of these characteristics as solely a "tool for agencies and Federally funded investigators," circumvents the many partners and local experts that work with Federally funded researchers in the storage, access, and preservation of research data. We encourage the OSTP to think more broadly about the possible users of these characteristics.

Finally, we appreciate that this set of design features is not intended to be an exhaustive list. This supports the previous point that these characteristics are a statement of current good practices, rather than best practices.

### **Section I: Desirable Characteristics for All Data Repositories**

- A. Persistent Unique Identifiers: In addition to mentioning and defining persistent unique identifiers (PUIDs), the description of this characteristic should include that a PUID

needs to be machine actionable and *globally* unique, as per Principle 4 of data citation<sup>1</sup> and needs to align with FAIR principles cited in the Background section. Related, the landing page for the PUID should be both human- and machine-readable and include metadata describing the data and its disposition to enable informed use of the referenced data. This is of particular importance for potentially restricted data.

- B. Long-Term Sustainability: We applaud that this characteristic is not tied to a specific time period of access but rather endorses that the repository should have a plan for sustaining data availability “during and after unforeseen events.” It should be clarified that sustainability is in reference to the repository and not of a singular dataset, and that the repository contents should remain both available and accessible.
- C. Metadata: “Sufficient” is a difficult word to define in the application of metadata. A more appropriate framing may be that the metadata accompanying datasets should be structured according to a standard schema using standardized vocabularies. The chosen metadata schema would ideally be a generalizable and/or community standard (e.g., DataCite, DublinCore, RDF-derived, etc.) as well as machine-readable and -actionable. This would support both the previously cited FAIR standards and the requirement in G of reuse tracking. The section would further benefit from clarification that this should apply to both generalist and discipline-specific repositories.
- D. Curation & Quality Assurance: It is unclear who the “others” are here or the roles explicit to data curation versus metadata curation (e.g., depositors, peer reviewers, repository staff). The former may be outside of the roles of the repository. These need to be broken down and made more clear. Data integrity (e.g. non-corrupt data, check sum, etc. over the data lifetime) is missing from this description yet is crucial in quality assurance and the sustainability of data access.
- E. Access: Further explanation is required for this section to clarify the basic requirements for data to be open and equitably accessible. Repositories should have the ability to assign and communicate specific licenses for datasets that make the conditions of access and reuse clear. This should also be coupled with conditions of access for those datasets that are not able to be open.
- F. Free and Easy to access and reuse: Furthering our comments in Section E, the repositories’ ability to assign and communicate licenses should include that these licenses are machine negotiable.
- G. Reuse: Sections E, F, and G are hard to differentiate as provided (and our comments for these three sections reflect that). These characteristics would be better articulated as a bundled set around access and reuse. This would allow for a related characteristic to be added for metrics associated with data reuse, e.g. downloads, citation tracking, etc.

---

<sup>1</sup> Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. 2015. Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science* 1:e1 <https://doi.org/10.7717/peerj-cs.1>



- H. Secure: It is unclear whether this section is about the release of data or the corruption of data. These aspects should be separated out and clarified. Throughout this document there is a mix of repository-level, content-level, and file-level characteristics that need to be better distinguished to disambiguate the intention of individual characteristics.
- I. Privacy: No comment.
- J. Common-format: To which standards should the repository comply? This needs to be more explicit. Additionally, is this in reference to the data itself or the metadata? Metadata should always be in an open format and non-proprietary. The language of “preferably” here implies exceptions that should not exist.
- K. Provenance: Given the examples in this section, we suggest changing the characteristic name from provenance to file-level integrity to better reflect the intention here. This section speaks more to versioning within the repository versus provenance outside of the repository.

Additional Section I comments: As Section K does not cover provenance, a new provenance characteristic should be added. Core Trust Seal can be a resource for the language to include here.

## **Section II: Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

It is unclear if this section applies to individual-level data or aggregated data or both, or to data that should be restricted for other reasons. This should be clarified. Additionally, there’s an implication in a number of the following characteristics that all human data should be restricted data. This is inaccurate. Restriction must be balanced with risk and the requirements of the consent agreement. This should be made more clear in the introduction of this subset of characteristics. Some reference to the recent NIST Privacy Framework and its framing of privacy as risk management may be useful.

- A. Fidelity to Consent: The repository should have the ability to apply access restrictions when the data requires it. That said, the researcher or depositor, not the repository, has the responsibility to confirm the appropriate access restrictions have been applied per the consent language. The repository responsibility lies in implementing the restrictions which have been applied, which may apply to persons accessing the data as well as uses of the data. This needs to be parsed out here.
- B. Restricted Use Compliant: The use of the word “enforces” here places an undue burden on the repository or implies a requirement that all analysis must happen within the repository under a request/receive model. This seems unlikely and requires clarification. We also suggest that a repository should support the contributor’s rights to remove data from the set.
- C. Privacy: This characteristic is less about privacy and more regarding security. These terms are not interchangeable, and this will cause confusion.

- D. Plan for Breach: This characteristic is not limited to human subject data and should be included in Section I. Additionally, this is a subset of considerations under security and can be bundled as such to diminish confusion.
- E. Download Controls: Download controls are dependent on risk which should be discussed here to provide necessary context for decision making of their application in different scenarios.
- F. Clear Use Guidance: This characteristic is also not limited to this subset of data and should be a general characteristic in Section I.
- G. Retention Guidelines: No comment.
- H. Violations: No comment.
- I. Data request review: This requirement as currently worded assumes a number of conditions around data sensitivity, repository processes, and computational technologies that are not necessarily true now and are less likely to be true in the future. We encourage re-thinking this requirement to focus on what it is attempting to accomplish in the realm of risk management and mitigation.

#### **Other**

With regard to the characteristics discussed above, we would encourage the OSTP to be expansive in their considerations of what future developments may hold for data and the role of data repositories in enabling responsible access and reuse of federally funded data. We encourage exploration of the further implications of FAIR in applications such as the Personal Health Train (<https://www.dtls.nl/fair-data/personal-health-train/>), and how the OSTP's proposed repository characteristics might effectively support these future uses.

Request/Characteristic	Comment
Filers	Nigel Robinson, Megan Force, Patricia Tortosa, Mark Matthews
Organizational Affiliation	Web of Science Group, Clarivate Analytics
Primary Scientific Discipline	Life Sciences (NR), Physical Sciences (MF), Social Sciences (PT), Arts and Humanities (MM)
Role	Director, Content Management (NR); Editorial staff, Data Citation Index (MF, PT and MM)
The proposed use and application of the desirable characteristics	<p>We approach this Request For Comment from the perspective of an indexing service. The Data Citation Index indexes over 10 million datasets and data studies from almost 500 data repositories, linking the data to scientific research publications which create, use and cite them as part of the Web of Science scholarly research platform. Our organization supports the FORCE11 joint declaration on data citation (<a href="https://www.force11.org/datacitationprinciples">https://www.force11.org/datacitationprinciples</a>), as well as the FAIR principles. By linking data and software citations to published works which create or use them, we provide a measure of dataset impact to promote and incentivize data sharing.</p>
<p>I. A. <i>Persistent Unique Identifiers</i>: Assigns datasets a citable, persistent unique identifier (PUID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The PUID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.</p>	<p>We agree with the recommendation. We add that PUIDs are vital to the accurate identification of datasets for citation purposes. The recommended and community PUID gaining traction is DOI, which is most widely supported and provides for citation tracking and linking. DOIs for data are issued for data objects mainly by DataCite and less so by Crossref.</p>
<p>I. B. <i>Long-term sustainability</i>: Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure</p>	<p>We agree with the recommendation. We have observed that when a data repository loses funding or otherwise ceases to be maintained, content is often transferred to a contingency repository that is still operational for long-term preservation. We therefore consider the issuance of PUIDs to be a fundamental component of any plan for long-term</p>

<p>data are available and maintained during and after unforeseen events.</p>	<p>stability, as this guarantees that datasets remain discoverable.</p>
<p>I. C. <i>Metadata</i>: Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.</p>	<p>We consider certain bibliographic metadata elements to be required for accurate, formal data citation, consistent with the Force11 Joint Declaration on Data Citation Principles (<a href="https://www.force11.org/datacitationprinciples">https://www.force11.org/datacitationprinciples</a>) and DataCite recommendations (<a href="https://datacite.org/cite-your-data.html">https://datacite.org/cite-your-data.html</a>). These should be included when determining schema suitability:</p> <ol style="list-style-type: none"> <li>1. Author/Creator - Individuals or organizations that created or contributed to the data set; this metadata element is vital to guarantee attribution and credit for data contributor, and to provide metrics for their nontraditional scholarly output</li> <li>2. Year - The year of “publication” of the data; when it is made publicly available, such as through deposition in a repository</li> <li>3. Title - The title of the data object, which may differ from the title of the parent research paper/project</li> <li>4. Publisher - The data repository that houses the data and/or the governing organization responsible for publishing, (i.e., making available) the data</li> <li>5. Version - Dynamic data sets or those where new editions may be issued (such as with error corrections or new values) must employ proper version control to guarantee accuracy and uniqueness in data citation</li> <li>6. Permanent Identifier - A unique and persistent identifier should be assigned.</li> </ol> <p>Additionally, associated subject keywords and discipline-specific indexing terms improve discovery and reuse. Funder and grant information, as well as author affiliations, enable tracking of funded research output. Many data repositories also archive other content types such as research articles; for this reason, a resource type metadata field aids identification of data objects for indexing and assessment.</p>
<p>I. D. <i>Curation &amp; Quality Assurance</i>: Provides, or has a</p>	<p>We agree with the recommendation. Metadata quality remains a significant hurdle in efforts to</p>

mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.	create complete and accurate citation records for datasets. This is particularly important when crosswalking metadata elements such as authors and publication dates between multiple data sources. Where a data repository accepts data from multiple sources, metadata curation ensures consistency.
I. G. <i>Reuse</i> : Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).	We agree with the recommendation. We add that for some dynamic datasets, further information (such as access date/time) is necessary for reuse. Repositories which archive and distribute such data should anticipate such needs and have recommendations in place. PUIDs and metadata in themselves do not guarantee accurate dataset tracking, which also relies upon publisher and author citation practices. Data citation is increasing but is not pervasive in all disciplines.
I. J. <i>Common Format</i> : Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.	We agree with the recommendation. Metadata formats employed by data repositories should be flexible enough to provide necessary information for citation and reuse. Metadata formats should be curated for consistency when crosswalking content between multiple sources.
I. K. <i>Provenance</i> : Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.	We agree with the recommendation. We also stress the importance of clearly distinguishing between changes to the citable dataset and changes to bibliographic and other metadata.
Additional characteristics that should be included	We recommend that data repositories 1) exhibit a policy or other statement with respect to their mission and scope. 2) Provide a list of editorial board members or repository administrators to help in understanding repository foundations. 3) Provide a point of contact for enquiries

**Further information**

A whitepaper with recommendations for best practice is available from [https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/2019/08/Crv\\_WOS\\_Whitepaper\\_DCI\\_web.pdf](https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/2019/08/Crv_WOS_Whitepaper_DCI_web.pdf)

Further information on the Data Citation Index can be found at <https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index/>



March 6, 2020

Subcommittee on Open Science  
Office of Science and Technology Policy  
1650 Pennsylvania Avenue NW  
Washington, D.C. 20502

RE: Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research [FR Doc. 2020-00689]

Dear Subcommittee on Open Science (SOS) members:

IBM appreciates the opportunity to comment on the *Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research [FR Doc. 2020-00689]*. IBM is pleased with SOS's continued focus on maximizing the quality and utility of data repositories used to locate, manage, share, and use data resulting from federally funded research.

As a global research and development leader, IBM has one of largest corporate research ecosystems in the world. We have more than 3,000 researchers in 19 locations across six continents. Our scientists are charting the future of artificial intelligence, breakthroughs in quantum computing, reshaping how businesses leverage blockchain and much more. IBM offers a unique perspective to help optimize and improve the consistency of agency-provided information for data repositories in order to reduce the burden researchers face in using this data.

Overall, IBM commends the Subcommittee on Open Science for this guidance as it is a welcome and encouraging step towards improving the research data ecosystem and promoting open science. We offer two comments for your consideration.

- **Expand these draft characteristics to include specific examples and resources that can help ensure this guidance is widely adopted and implemented effectively.** For example, the recommendation to use common formats (“J.”) that allow “datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format” could include examples of the problems that arise from the use of proprietary, uncommon formats, as well as highlight examples of open, common formats that repositories could utilize to adhere to this guidance.
- **Recommend the use of open, common licenses for data repositories.** Recommendation “E. Access,” and “F. Free & Easy to Access and Reuse” should include additional detail related to data access and reuse. These recommendations rightly identify the importance of maximizing the accessibility and usability of data. However, while the recommendations imply that data repositories utilize open, common licenses to provide “broad, equitable, and maximally open access to datasets,” they do not explicitly say to use such licenses (for example, the word “license” does not appear in this guidance.) The use of open, common licenses for research datasets is crucial because even the best-supported research teams can find it prohibitive to navigate and evaluate numerous different data licenses, even when these licenses confer similar permissions, limiting researchers’ ability to use and combine datasets for their work. And though recommendation “F.” encourages repositories to make “datasets and their metadata accessible



free of charge...with broadest possible terms of reuse or documented as being in the public domain,” it does not provide specific instruction about how to accomplish this. IBM believes the widespread adoption of common, open data sharing licenses will be a significant boon to researchers, which is why we support use of the Linux Foundation-developed Community Data License Agreement (CDLA).<sup>1</sup> This guidance should specifically recommend data repositories to utilize open, common licenses for datasets and metadata to minimize confusion about how to best adhere to this guidance and maximize the utility of federally funded research data to the broader research community.

Once again, IBM appreciates the opportunity to comment and we look forward to future engagements. For any questions, please contact Mr. Joshua New at [Joshua.New@ibm.com](mailto:Joshua.New@ibm.com)

Sincerely,

A handwritten signature in black ink, appearing to read 'D. Gil', followed by a small square box containing the letters 'D.G.'.

Dr. Dario Gil  
Director of IBM Research

---

<sup>1</sup> <https://cdla.io/>

**Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research**

**Response from Sandia National Laboratories**

**POC: James R. Stewart, Sr. Manager – Computational Sciences and Math**

**[jrstewa@sandia.gov](mailto:jrstewa@sandia.gov); 505-844-8630**

**March 6, 2020**

**Response from Sandia National Laboratories consists of input from management in the following areas: Computational Science, Materials Science, Technology Transfer.**

- **Accessibility:** Data repositories should be accessible (to non-data experts), fully searchable, free.
- **Ease of uploading data:** A repository should not possess too much of a burden to load data into. (1) In the draft characteristics there is a ‘common format’ characteristic. As not all data looks the same, it is important that this characteristic isn’t so rigorously enforced that it prevents some data from being loaded into the repository. (2) While there is a characteristic that the repository should be "Free & Easy to Access and Reuse", there is no corresponding statement about ease of submission. Without this characteristic, the associated costs to the labs will rise with potentially significant compliance issues.
- **Sustainability:** The repository should be designed such that minimal resources are needed to sustain it, thereby ensuring long-term reliability.
- **Reproducibility:** An additional characteristic not listed is "Replication," or "Reproducibility." In the preamble to the document, the definition of Research Data is cited. This definition includes the phrase "as necessary to validate research findings". This means that the repository has to support in some manner links/publications etc. to the analysis codes/methods used to arrive at the research results. The current method with publications is to require the code be available as open source. The repository at data.gov addresses this in a limited manner with some apps for providing access to data but not in a structured discoverable manner.
- **Archival journals:** Published journals being the best repository – they are searchable and maintained by others.
- **Leveraging the Materials Project:** Capitalize on existing infrastructure like the Materials Project (MP; <https://materialsproject.org>). MP has thousands of computationally predicted crystal structures and standard DFT computed properties. But it extends this base of information with “apps” for storing non-standard DFT or other computed



properties that are application specific (i.e., batteries, electrochemical aqueous systems, etc.). In my opinion the best way for Sandia to share/make its data available would be to partner with MP on a hydrogen storage app, hydrogen generation app, etc.

- **Expanding some human data repository characteristics to other use cases:** The characteristics limit some of the additional considerations to only repositories containing human data, but some of these characteristics should be more general. For example, one of the characteristics for human data is "Restricted Use Compliant" which is a requirement that submitter's data use restrictions be enforced. This characteristic would also address many of the issues that are associated with data that arises from tech transfer partnerships such as CRADAs or NFE agreements. Such data may be subject to time-based restrictions such as protected CRADA information which is protected for a period of five years after marking. Furthermore, there may well be information which needs to be protected from general release for a period of time associated with obtaining patent protection.
- **Time/state evolution of access model:** Given the practical matter that the data must be collected and published to a repository during the execution of a project but access may not be permitted until some later time, the repository needs to support this evolution of the access model. The concept of time/state of the data is not adequately addressed by the existing characteristics. There are situations where characteristics and access to data will change over time. The characteristics of the repository need to reflect this.
- **Standardization of data quality characterization:** The issue of quality is not adequately addressed in the current set of characteristics. Data that has been either replicated or validated in some manner has more value in certain contexts. Maybe the plan is to envision that the metadata associated with a dataset would characterize this, but such metadata would need to be standard and the repository itself would have to have a mechanism for indicating and tracking quality.
- **Review and Approval requirements:** There will need to be a significant investment in education and tool development to support the review of data before it is published. You could envision something like the R&A process but with additional requirements on the technical staff to prepare and annotate the data appropriately to meet the data publications requirements. We see a reflection of this in some of the funding opportunity announcements (FOAs) especially from EERE which explicitly call for a data management and access plan to be generated during the project.
- **Copyrightable data:** Unless OSTP gets much more specific, copyrightable 'data' should not be made available in the described manner. Access limits, how to apply restrictions, export issues, etc. (normal things we think of in our copyright licenses) are not addressed.

Response to “Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research” from Space Telescope Science Institute (STScI)

## Authors

Josh Peek, MAST Principal Investigator, STScI (researcher, data manager)

Arfon Smith, Mission Head for Data Science, STScI (researcher, data manager)

## Domain: Physical Sciences (Astronomy & Astrophysics)

### Introduction

Space Telescope Science Institute was founded in 1981 to run the science operations of the Hubble Space Telescope. Since 1990, the Institute (managed by [AURA](#) on behalf of NASA), has been the operational interface for Hubble, serving the global astronomical community who make use of this flagship facility. A key part of our work as the *science operations center* for Hubble and the soon to be launched James Webb Space Telescope (JWST), is to ensure the continued scientific legacy of the missions. Capturing and preserving the data associated with Hubble is the responsibility of the Barbara A. Mikulski Archive for Space Telescopes (MAST) which is the archive for Hubble and more than twenty other mission datasets including Kepler, TESS, IUE, and Galex. MAST currently holds data from 21 missions and surveys and with a data volume of over 2 petabytes is a major infrastructure support effort in and of itself.

### Responses

In the responses below we focus on Repositories for Managing and Sharing Astronomical Data Resulting From Federally Funded Research.

*A. Persistent Unique Identifiers: Assigns datasets a citable, persistent unique identifier (PUIID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The PUIID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.*

We believe this is an appropriate characteristic for a repository of astronomical data. MAST leads NASA astrophysics archives in the use of DOIs, and we have DOIs for many data sets. Further, we allow users to mint their own DOIs using our system to codify a subset of data used in a publication.

*B. Long-term sustainability: Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.*

We believe this is an appropriate characteristic for a repository of astronomical data. NASA funds astronomical archives in perpetuity to maintain these datasets and to have disaster recovery plans. This also holds true for NASA planetary data, which are kept in the NASA Planetary Data System. We note that not all data generated with federal funds should be preserved in perpetuity; some simulation data, for example, can be very voluminous and not terribly valuable over long time periods.

*C. Metadata: Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.*

We believe this is an appropriate characteristic for a repository of astronomical data. MAST uses metadata schema set by the [International Virtual Observatory Alliance](#), the recognized standards body for astronomical data. These metadata standards power key services that enable scientific discovery such as the [MAST discovery portal](#) and our programmatic services (APIs).

*D. Curation & Quality Assurance: Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.*

We believe this is an appropriate characteristic for a repository of astronomical data. In the case of MAST we only provide data that has a publication associated with it in the refereed astronomical literature <http://archive.stsci.edu/hlsp/index.html>.

*E. Access: Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.*

We believe this is an appropriate characteristic for a repository of astronomical data. [It is well documented](#) that providing robust access to archival data enables significantly more scientific value to be derived from missions such as Hubble (see figure below).

In our analysis of publications stemming from Hubble Space Telescope, Spitzer Space Telescope, and Chandra Space Telescope ([Peek+ 2019](#)), we found that more publications came from the community than came from those who proposed the observations in the first place, and that these “archival” publications came from a much broader community demographically. To continue to provide open access in the era of petabyte datasets server-side analytics solutions such as those discussed by the NASA Big Data Task Force will become necessary in astronomy.

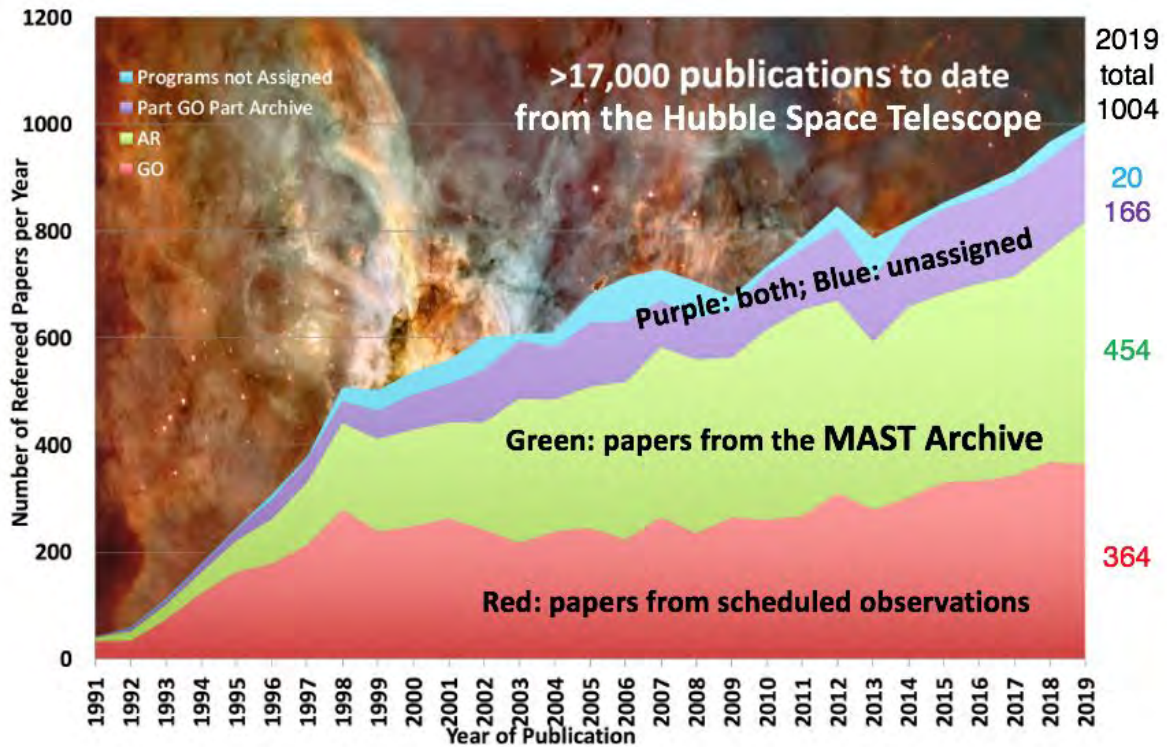


Figure 1: Number of publications (papers) with results derived from Hubble data. 'AR' denotes papers using only archival data, 'GO' denotes papers making use of new observations, and 'Part' denotes papers using a mix of archival and new observation data.

*F. Free & Easy to Access and Reuse: Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.*

We believe this is an appropriate characteristic for a repository of astronomical data. Archives open to all together with clear usage guidelines can only serve to increase the scientific productivity of the datasets we host.

*G. Reuse: Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).*

We believe this is an appropriate characteristic for a repository of astronomical data. STScI actively tracks the usage of data from Hubble which allows us to report metrics such as those in Figure 1. Additionally, MAST offers [Digital Object Identifiers](#) (DOIs) to authors to help them be more explicit about the exact archival data they made use of when publishing a new result.

*H. Secure: Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>) or the National Institute of Standards and Technology's 800-53 controls (<https://nvd.nist.gov/800-53>).*

We believe this is an appropriate characteristic for a repository of astronomical data. MAST handles NASA data that has an exclusive access period, during which only certain users have access to the data. Many astronomical archives only hold public data.

*I. Privacy: Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.*

We follow applicable industry standards in this regard.

*J. Common Format: Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.*

We believe this is an appropriate characteristic for a repository of astronomical data. Nearly all astronomical data are served in non-proprietary formats governed by standards set by the International Virtual Observatory Alliance.

*K. Provenance: Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.*

We believe this is an appropriate characteristic for a repository of astronomical data. Provenance tracking for data hosted at STScI includes maintaining records of (1) the telescopes, instruments, and observing programs that delivered the raw observational data, and (2) the algorithms and software systems associated with creation of higher-level data products based on the raw data.



March 6, 2020

Kelvin K. Droegemeier, Ph.D.  
Director  
Office of Science and Technology Policy  
Executive Office of the President  
Eisenhower Executive Office Building  
1650 Pennsylvania Avenue  
Washington, DC 20504

France A. Córdova, Ph.D.  
Director  
National Science Foundation  
2415 Eisenhower Avenue  
Alexandria, VA 22314

Francis Collins, M.D., Ph.D.  
Director  
National Institutes of Health  
9000 Rockville Pike  
Bethesda, MD 20892

Submitted electronically at: <https://www.federalregister.gov/documents/2020/03/05/2020-04530/request-for-public-comment-on-draft-desirable-characteristics-of-repositories-for-managing-and-sharing-data-resulting-from-federally-funded-research>

**Re: Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research**

Committee on Science Co-Chairs:

AMIA applauds the issuance of this Request for Comment (RFC) seeking information on desirable characteristics of repositories for managing and sharing federally funded research data. We appreciate the opportunity to provide input from the vantage of health informatics professionals.

Health Informatics is the science of how to use data, information, and knowledge to improve human health, the delivery of health care services, and the execution of scientific research.

March 6, 2020

AMIA is the professional home for more than 5,500 informatics professionals, representing frontline clinicians, biomedical researchers, public health experts, librarians and educators who bring meaning to data, manage information, and generate new knowledge across the healthcare system and research enterprise. AMIA members advance health and wellness by implementing and evaluating informatics interventions, innovations, and public policy across settings and patient populations, adding to our collective understanding of health in the 21st century through peer-reviewed journals and scientific meetings.

Developing consensus characteristics for all-data-type repositories, as well as characteristics for repositories that include human data, is foundational to support discoverability, management, and sharing of research data. AMIA appreciates the reference to FAIR (Findable, Accessible, Interoperable, and Reusable) data principles in this document, signifying commitment to these concepts at the highest levels of the Administration. In combination with the Federal Data Strategy,<sup>1</sup> we are encouraged and strongly support the identification of desirable characteristics for data repositories.

AMIA has a long history of policy development related to data sharing for research, and we have developed several policy principles and positions that may be applicable to policymakers at OSTP.<sup>2</sup> The rapid digitization of care and clinical research has ushered in a new era of data-driven research. However, various cultural dynamics, institutional support systems, and policy levers must align to positively impact this new era's ongoing evolution. For example, AMIA has identified several incentives and policies necessary to improve data management and sharing, such as:

- Dedicated funding from research sponsors for data curation and sharing efforts so there are sufficient incentives to share, collaborate, and advance data sharing capabilities.<sup>3</sup>
- Institutional rewards for those who create and/or contribute to public datasets and software that others find useful so that incentives exist for those who create as well as those who analyze data.<sup>4</sup>
- The creation of harmonized regulatory and/or policy frameworks for data sharing, including data use agreements, data sharing plans, human-subjects reviews, and federal, state and local privacy requirements to minimize barriers to sharing data.<sup>5</sup>

While these may be out of scope for this effort, we view the identification of desired characteristics as a foundational first step towards better managing, organizing, and making data resulting from federally funded research more accessible for use and reuse. It will be important

---

<sup>1</sup> <https://strategy.data.gov/>

<sup>2</sup> <https://www.amia.org/sites/default/files/2018-2019-AMIA-Health-Informatics-Policy-Priorities.pdf#page=12>

<sup>3</sup> Borne, P., Lorsch, J., Green, E., "Perspective: Sustaining the big-data ecosystem," *Nature*. November 2015. 527, S16– S17

<sup>4</sup> Piwowar, H., Vision, T., "Data reuse and the open data citation advantage," *Peer J*. 2013. 1:e175

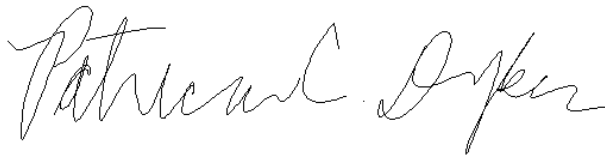
<sup>5</sup> Taichman, D., Backus, J., Baethge, C., et al. "Sharing Clinical Trial Data: A Proposal From the International Committee of Medical Journal Editors," *Annals of Internal Medicine*. 2016. doi:10.7326/M15-2928

March 6, 2020

for OSTP to consider various incentives (positive and negative) to encourage coordination among Agencies and Offices using these characteristics and how best-practices can be identified and promoted across the Executive Branch. Subsequent and significant work will be needed to operationalize use of and adherence to these desired characteristics.

In the enclosure to this transmittal letter, we provide input on the draft characteristics. Thank you for considering our comments. Should you have questions about these comments or require additional information, please contact Jeffery Smith, Vice President of Public Policy at [jsmith@amia.org](mailto:jsmith@amia.org) or (301) 657-1291. We look forward to continued partnership and dialogue.

Sincerely,



Patricia C. Dykes, PhD, RN, FAAN, FACMI  
Chair, AMIA Board of Directors  
Program Director Research  
Center for Patient Safety, Research, and Practice  
Brigham and Women's Hospital

*Enclosed: AMIA response to OSTP Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research*



March 6, 2020

In reviewing this RFC, we note strong support for all the draft characteristics, with slight modifications (see below). Additionally, we recommend OSTP consider two additional characteristics for all data repositories:

1. Governance
  - Experience to-date indicates that several best-practices are emerging from leading repositories around development of standard governance structures and transparency policies. Data deposition policies, dataset descriptions and transparency around funders, advisors, and operations are all important hallmarks of modern repositories. Insofar as repositories have specific kinds of policies in place and are transparent with regards to management and operations, AMIA encourages OSTP to consider a characteristic around Governance. This characteristic need not be prescriptive, and we point to OSTP's own Project Open Data as an example.<sup>6</sup>
2. Feedback
  - Delineating what systems are in place for users of a repository to provide feedback to the owners and the operators would be in keeping with contemporary, private-sector certifications.<sup>7</sup> There is a need to understand which repositories are adaptable, over time, to advances or needs that users identify.

## I. Desirable Characteristics for All Data Repositories

- A. *Persistent Unique Identifiers*: Assigns datasets a citable, persistent unique identifier (PUIID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The PUIID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.

**AMIA Comment:** We support this characteristic. However, we note that datasets may have multiple, associated UIIDs. An example is a dataset from ClinicalTrials.gov that reappears as part of a peer-reviewed journal publication that has its own DOI. As a follow-on data management issue, we encourage OSTP to consider strategies and standard operating procedures to ensure that UIIDs can be appropriately reconciled and traced across literature.

- B. *Long-term sustainability*: Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a

---

<sup>6</sup> <https://project-open-data.cio.gov/governance/>

<sup>7</sup> <https://www.coretrustseal.org/>

March 6, 2020

stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.

**AMIA Comment:** We agree with this characteristic and could be a component of a Governance characteristic or a stand-alone characteristic.

- C. *Metadata:* Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.

**AMIA Comment:** We strongly agree with this characteristic. We recommend OSTP consider adding “such as” examples for how such metadata should be indexed using common vocabularies, including Library of Congress headings (LSH) or National Library of Medicine’s Medical Subject Headings (MeSH).<sup>8</sup> Additionally, standardized metadata schema and so-called “minimal information” models can be used to encourage metadata that is both well-formatted and robust. An example to consider is the HCLS dataset description model:

<https://www.w3.org/2001/sw/hcls/notes/hcls-dataset/>.

We also recommend that versioning and changes tracking be necessary components of metadata.

- D. *Curation & Quality Assurance:* Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

**AMIA Comment:** We agree with this characteristic, and we would add an expectation that there is some consideration of adjudication and oversight of curation and quality assurance. Having confidence that the data has not been altered or censored will be important.

We also note that Data Quality and Data Completeness are overlapping but separate concepts worth considering. There is a need to understand a dataset’s completeness, or the degree of “missingness” in the dataset. We are unaware of a standardized metric or indicator for missingness, but believe a standardized metric or score would be useful to ascertain the completeness of the dataset.

- E. *Access:* Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.
- F. *Free & Easy to Access and Reuse:* Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

---

<sup>8</sup> <https://www.nlm.nih.gov/mesh/meshhome.html>

March 6, 2020

- G. *Reuse*: Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).

**AMIA Comment:** We agree with these characteristics, but we believe there is an opportunity to streamline and make them more applicable. We recommend OSTP combine E, F, and G, as “E. Access, Use, and Reuse:” as they represent a continuum of related concepts, and then add “F. Tracking:”. Characteristic E could be combined with F above as follows:

*E. Access, Use, and Reuse:* Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality, and makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

This characteristic would be followed by a new one (F. *Tracking*:) which could establish the expectation that repositories provide a table of content or index for datasets housed in the repository, as well as mechanisms to track the use of datasets, similar to current language at G.

- H. *Secure*: Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>) or the National Institute of Standards and Technology's 800-53 controls (<https://nvd.nist.gov/800-53>).

**AMIA Comment:** We agree with this characteristic and encourage OSTP to replace “accepted” with “federally approved,” or “published, industry-recognized,” to describe the kind of security criteria that need to be documented.

- I. *Privacy*: Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.

**AMIA Comment:** We agree with this characteristic.

- J. *Common Format*: Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.

**AMIA Comment:** We urge OSTP to consider the implications of replacing “or” with “and” in this characteristic. We do not see a reason why a non-human data repository should provide one, but not the other capabilities. In addition, we recommend OSTP focus on standards-compliant

March 6, 2020

rather than a single form, and we strongly support the non-proprietary aspect of this characteristic.

- K. *Provenance*: Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.

**AMIA Comment:** We strongly agree with this concept, and would suggest that this be part of a subset of characteristics of “Metadata.” We view provenance and versioning as part of robust metadata. As mentioned above, the HCLS dataset description profile covers this nicely. We also note that provenance may include derivation from one or more independently existing datasets.

## II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)

- A. *Fidelity to Consent*: Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).

**AMIA Comment:** We agree with this characteristic.

- B. *Restricted Use Compliant*: Enforces submitters' data use restrictions, such as preventing reidentification or redistribution to unauthorized users.

**AMIA Comment:** We agree with this characteristic.

- C. *Privacy*: Implements and provides documentation of security techniques appropriate for human subjects' data to protect from inappropriate access.

**AMIA Comment:** We recommend OSTP reexamine this characteristic as it describes concepts of security, rather than privacy. Simply renaming to “Security” may be the best option.

- D. *Plan for Breach*: Has security measures that include a data breach response plan.

**AMIA Comment:** We agree with this characteristic.

- E. *Download Control*: Controls and audits access to and download of datasets.

**AMIA Comment:** We agree with this characteristic.

- F. *Clear Use Guidance*: Provides accompanying documentation describing restrictions on dataset access and use.

**AMIA Comment:** We agree with this characteristic.

March 6, 2020

G. *Retention Guidelines*: Provides documentation on its guidelines for data retention.

**AMIA Comment:** We agree with this characteristic.

H. *Violations*: Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository.

**AMIA Comment:** We agree with this characteristic.

I. *Request Review*: Has an established data access review or oversight group responsible for reviewing data use requests.

**AMIA Comment:** We agree with this characteristic and recommend OSTP consider this a global characteristic beyond human data repositories. This may help inform a Governance characteristic as recommended previously.

# University of Massachusetts Amherst Response to Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

Responder: Thea Atwood, University of Massachusetts Amherst

Response: Discipline Neutral

Role: Data Services Librarian, R1 Public research and land-grant university

Dear Chief of Staff Bonyun, Dr. Nichols, and the Subcommittee on Open Science,

The University of Massachusetts Amherst (UMass Amherst) is pleased to offer its comments on the “Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research”. UMass Amherst is a public research and land-grant university with R1: Doctoral Universities -- Very high research activity classification. A generalist data repository, designed to capture the scholarly output of the organization, is managed by the University Libraries.

As described in the RFC, the OSTP seeks comments on key characteristics for data repositories for data resulting from Federally funded research. Specifically, the OSTP is seeking public comment on: I. The proposed use and application of the desirable characteristics; II. The appropriateness of the “Desirable Characteristics for All Data Repositories” (Section I) for data repositories that would store and provide access to data resulting from Federally-supported research; III. Appropriateness of the characteristics listed in the “Additional Considerations for Repositories Storing Human Data (even if de-identified)” (Section II) delineated for repositories maintaining data generated from human samples or specimens; IV. Considerations for any other repository characteristics which should be included to address the managing and sharing of unique data types (e.g., special or rare datasets); V. The ability of existing repositories to meet the desirable characteristics; VI. Consistency of the desirable characteristics with widely used criteria or certification schemes for certifying data repositories; and VII. Any other topic which may be relevant for Federal agencies to consider in developing desirable characteristics for data repositories.

We write on all enumerated topics.

## I. The proposed use and application of the desirable characteristics

The term “preservation:” Digital preservation has particular and rigorous requirements that, based on current infrastructure available for universities and others to adopt, may not be sustainable or achievable. Further, there is confusion around business models of preservation services, worries about the cost of preservation-level storage, unsolved legacy issues, and more.<sup>1</sup> Alternative options to describe the concept of making data available for the long term include providing a precise definition for “long-term preservation,” including minimum number of years, processes necessary to ensure data integrity, stipulations and workflows for format and hardware migration, etc., or using a different phrasing, like “long-term accessibility,” again, with a definition of minimum requirements. Further complicating this issue is that disciplines and other stakeholders have different ideas of what “long-term” means, so some guidance for faculty and repository managers would be welcome.

This document stipulates that “these characteristics are not intended to be an exhaustive set of design features for data repositories” which is admirable -- it helps maintain some flexibility. This paragraph continues: “Federal agencies would not plan to use these characteristics to assess, evaluate, or certify the acceptability of a specific data repository, unless otherwise specified for a particular agency program, initiative, or funding opportunity” which reads as a statement that contradicts itself -- “we won’t use this to assess repositories, except when we will.” And while there is emphasis that the characteristics should guide Federally funded investigators, this might create inequity in the development and use of repositories -- if there is a seal of approval provided by Federal agencies, would researchers be dissuaded from using an otherwise appropriate repository? Further, organizations unable to meet the certification requirements (because of limited resources, administration not understanding the need to pursue a certification, lack of awareness, or other bureaucratic holdups and misinformation), will be negatively impacted.

Finally, because researchers frequently lack the expertise or the training in using digital repositories, we recommend the final report include a section encouraging consultations with local experts and online educational materials. This includes working with local librarians, data curators, security officers, privacy officers, cybersecurity specialists, and other experts. Federally coordinated data repositories need to be encouraged to include guidance on using the data repository. Critically, researchers

---

<sup>1</sup> See Rieger (2018). The state of digital preservation in 2018: A snapshot of challenges and gaps. ITHAKA S+R. <https://doi.org/10.18665/sr.310626>

need guidance on preparing data for deposit, as some of the steps for preparing shareable data need to occur at the beginning of a project -- for example, writing a consent form to appropriately secure consent to share human subjects data digitally.

## **II. The appropriateness of the “Desirable Characteristics for All Data Repositories” (Section I) for data repositories that would store and provide access to data resulting from Federally-supported research**

### **A. Persistent Unique Identifiers**

Persistent, Unique Identifier (PUIs or PIDs) are critical for the success of data citation, access, and reuse. Downstream effects of PIDs should be explicitly stated to help researchers understand the importance of assigning a PID.

We recommend that the final report require support for digital harvesting technologies, like APIs.

The recommendation would be strongest if it includes a recommendation for use of centrally registered DOIs, and exclude accession numbers. DOIs facilitate tracking use of datasets by working off of robust and standardized infrastructure that locally assigned accession numbers lack.

### **B. Long-term sustainability**

Renaming this section “business model and long-term sustainability,” would help clarify its purpose, if indeed this is a section aimed at the business model of the repository.

If so, this section would benefit from providing guidance on what business models look like, or a minimum-viable business model. For example, the University of Massachusetts Amherst uses a hosted repository service, and has one full-time librarian dedicated to managing the repository, as well as some support staff. What *specific* business model characteristics should be included on an “about” page?

Such characteristics might include stipulations on “sunsetting” a repository -- this might be a challenging task to complete, but is an important thought exercise. Again, guidance on this sub-characteristic should be provided.

### **C. Metadata**

Metadata is a critical component in helping others understand a dataset, as well as for findability, and reuse of data. Like with PUIDs, it should be made clear to faculty the



downstream benefit of well-applied metadata.

Furthermore, the word “sufficient” requires additional explanation, as metadata standards vary in complexity and breadth. Researchers will have a very different definition of ‘sufficient’ as compared to librarians or data curators -- in part a reflection of each group’s expertise.

As with other characteristics, the section will benefit from well-defined terminology, with links to resources for further education provided. Using discipline-neutral, jargon-free language in this section is a high priority.

Providing resources on how to evaluate metadata options would greatly improve the usability and reach of this document. Adoption of any metadata standard for a discipline is very poor. Little – if any – guidance on evaluating metadata options is available, further demonstrating a need for providing such guidance.

#### **D. Curation & Quality Assurance**

This section would benefit from definitions of ‘curation’ and ‘quality assurance.’

Further, we are concerned about the ability of repositories to actually meet these requirements. More detail is in section V.

#### **E. Access and F. Free & Easy to Access and Reuse**

The distinction between “Access” and “Free & Easy to Access and Reuse” is subtle. These two categories should be combined.

“Free” may also be a term that requires some expanding -- some repositories charge a fee on data deposit<sup>2</sup> -- would researchers be dissuaded from using repositories that they have to pay to deposit data? Funding structures for data repositories is still very immature.

Consider including guidance on licensing data, which will explicitly state conditions for use and reuse.

#### **F. Free & Easy to Access and Reuse**

Described above.

---

<sup>2</sup> E.g., Dryad charges \$120 as a base fee for data deposit: [https://datadryad.org/stash/publishing\\_charges](https://datadryad.org/stash/publishing_charges)

## **G. Reuse**

This section would benefit from being renamed “use and reuse”, as using a quantifier like download counts doesn’t necessarily demonstrate reuse, but it could demonstrate use (e.g., using a dataset to teach). Further, the infrastructure for tracking and counting data citations, repository page views, and downloads is still immature, and not all repositories will have the ability to buy-in to a program or widget or module with this capability.

The “Use and Reuse” section should be more explicitly defined to include what infrastructure a repository will need to meet this requirement.

## **H. Secure**

It is clear what standards repositories should try to adhere to, but it is unclear how a researcher would assess how this is rolled out in a repository. Will researchers worry that a repository is non-compliant if it uses a different security standard than the two suggested here? Will researchers be deterred from using an otherwise appropriate repository? Will researchers feel anxious about depositing data in repositories that do not seem to require security controls, like those that only publish data that can be made openly available?

If this section refers to *user* data (as in password and login information), this should be made clear.

## **I. Privacy**

This section would benefit from examples of what is meant by “administrative, technical, and physical safeguards,” as well as the “applicable privacy, risk management, and continuous monitoring requirements.”

This section will not be clear to researchers, who won’t have the language to assess a compliant repository. This section would benefit from pointing to information security and cyberinfrastructure officers, or data librarians to help provide more robust guidance.

## **J. Common Format**

The metadata characteristic (c) would benefit from a statement on the capabilities of a repository to export metadata to a common format.

Researchers would benefit from links out resources explaining the different types of

non-proprietary formats, and for what format of data (GIS, images, video, text documents, etc.).

The document can help improve other researchers' ability to reuse data by explaining that some data formats, while non-proprietary, do not facilitate reuse (e.g., PDFs in general, tables stored as images).

#### **K. Provenance**

As stated, this characteristic sounds like back-end functionality for the repository, and not all repository platforms have the capacity to log file differences at the bit level.

This section would benefit from clarity -- Who is responsible for gathering this information -- the researcher, or the repository manager? If it is the researcher, they will need guidance to locate and capture logfiles at the beginning of their research process, or guidance on workflows and programs that capture this information on their behalf. What level of detail is necessary?

Further, "beginning with creation/upload of the dataset," are two entirely different concepts, at two entirely different points in time, and would require two different workflows -- please clarify what is meant here.

### **III. Appropriateness of the characteristics listed in the "Additional Considerations for Repositories Storing Human Data (even if de-identified)" (Section II) delineated for repositories maintaining data generated from human samples or specimens**

This section seems to note that even de-identified data will need to be stored in a repository that includes the controls outlined here. If so, this will be a significant change in current expectations, and may place new obligations on faculty working with human subjects data. For example, faculty may need to secure funding to store their data in a secure repository, like ICPSR. This may also not be in line with what funders are asking -- so some congruence between funders and the OSTP with regard to human subjects data will be necessary.

Furthermore, what will be done with de-identified data that resides in repositories without these considerations? For example, Dryad accepts de-identified human subjects data, and does not have gatekeeping in place.

More generally, we do not submit any specific comments on the considerations for repositories storing human subjects data, but would again reiterate that most researchers do not have a background in cybersecurity or data curation, and will need more guidance than what is provided. As above, we recommend providing suggestions for where researchers can find help for evaluating repositories and characteristics they are not familiar with.

Finally, and noting this is beyond the scope of this document, it may be necessary for consent forms to explicitly state that data will be made available in aggregate into perpetuity, so that subjects can consent to this type of data sharing. Further, if raw subject data needs to be kept into perpetuity, this should be noted in the consent form. The OSTP should consider coordinating with relevant funding agencies, and how consent to share will be executed by the IRB and other relevant offices at organizations that receive grants.

#### **IV. Considerations for any other repository characteristics which should be included to address the managing and sharing of unique data types (e.g., special or rare datasets)**

Some datasets include information that should not be released publicly -- including locations of protected species of plants and animals, or sites that have religious or cultural importance. Of particular concern are images of these items -- images often have GIS coordinates embedded in the metadata. These sharing considerations are not represented anywhere in the guidance, and should either be incorporated in Section II, or in a new Section III relating to special cases.

#### **V. The ability of existing repositories to meet the desirable characteristics**

There is a concern that this will become an unfunded mandate for repositories to meet. In particular, the “Data Curation and Quality Assurance” guideline and the requirement for long-term preservation will be a challenge for anyone but the most robustly staffed to provide. These are laudable goals, but fully realized, will be out-of-reach for many. “Expert curation” requires a great deal of staff and time -- an estimate given at the recent Accelerating Public Access to Research Data Summit held in Washington, DC<sup>3</sup>, stated that for 150 projects, three individuals are needed. This would be compounded if different federal agencies take different approaches to data curation and quality assurance

---

<sup>3</sup> February 19, 20, & 21, 2020 - AAU/APLU Workshop & Summit on Accelerating Public Access to Research Data: <https://www.aplu.org/projects-and-initiatives/research-science-and-technology/public-access/>

## **VI. Consistency of the desirable characteristics with widely used criteria or certification schemes for certifying data repositories**

The guidelines provided seem to be in-step with other used criteria.

## **VII. Any other topic which may be relevant for Federal agencies to consider in developing desirable characteristics for data repositories.**

Consider using the more broadly applicable “open scholarship” over “open science” when referring to open science as a methodology. This can also help expand our definition of scholarly output to include not just data, but curricular materials, teaching outputs, gray literature, primary documents, null data, and more. By using “open scholarship” as the default descriptor, we are inclusionary towards all disciplines and types of data.

Consider creating a template for repositories to use -- something standardized to help researchers quickly assess how their selected repository fits with the recommendations laid out above. This would also be useful for repository managers and institutions to understand what resources they will need to commit to ensure they meet federal guidelines.

Use non-jargon, discipline-neutral language throughout the guide.

Two more complex issues related to data sharing, copyright and intellectual property (IP), are missing from the discussion. Copyright is challenging because it does not uniformly apply to all data types -- e.g., taping an interview of a participant for a language study is considered copyrightable, and ownership can be shared between the researcher and the participant if explicitly stated as such. How IP relates to data is largely unexplored, and is regularly cited as an anxiety for not sharing data. Resources on licensing, copyright, and IP should be included, as these are incredibly confusing topics for researchers, and there is little guidance in existence.

Remind researchers that many institutions have both research data and cyberinfrastructure professionals to answer their questions. Many organizations have an institutional repository to deposit data when a disciplinary repository isn't available. Offering both of these points of guidance will help reduce confusion, spread accurate information, and improve compliance. However, small grantees, e.g., recipients of SBIR funding are unlikely to have such infrastructure or resources; special provisions may be required for this class of federal contractors and grantees.

Finally, Academia is very late in treating data as an asset. For-profit organizations and publishers see this gap in our services. Companies have already figured out how to mine data for profit, and have access to resources and their own proprietary datasets to create shiny solutions for campuses and researchers. Academia is thus subject to for-profit initiatives that will leave us again in the same place we are with publications -- where publishers own the taxpayer-funded scholarly content. If a high degree of data curation and quality assurance is a requirement that receives no additional funding or support from funding agencies, we will be destined to again rely on the deep pockets of for-profit publishers and companies (and publishers, with their extensive profit-margins, will easily be able to buy up third-party options, if they aren't significant backers already) to meet this requirement. This will extinguish any capability we have of truly meeting the promise of scholarship -- to better our lives, the lives of others, and humanity as a whole.

## **Conclusion**

We thank you for the opportunity to comment on this important matter, and hope that our comments prove helpful. Please feel free to contact Thea Atwood ([tpatwood@umass.edu](mailto:tpatwood@umass.edu)) about our comments.

Sincerely,

Thea Atwood, MSLIS  
Data Services Librarian

**Filer:** Keith Webster

**Organizational Affiliation:** Dean of the Carnegie Mellon University Libraries

Response was developed in conjunction with the Research Data Services Team at CMU Libraries, which is comprised of individuals with the following roles: Research Data Management Consultant; Institutional Repository Manager; Liaison to Computational Biology; Liaison to Public Policy and Social Sciences; Data Curation, Visualization, and GIS Specialist; and Digitization Projects Manager.

Our comments are made with the understanding that these characteristics are not intended to be a comprehensive set of specific features for data repositories; rather, we offer feedback on these characteristics as broadly supporting findable, accessible, interoperable, and reusable (FAIR) research data.

## **Section I: Desirable Characteristics for All Data Repositories**

We believe the sections C: *Metadata* and K: *Provenance* could be combined or included in succession, as they both refer to documentation of the research data. It appears that K: *Provenance* is referring to versioning of the research data. We would suggest potentially reframing the title of this section, as many repositories use terms such as “version control” to refer to the characteristics listed in K. For the purpose of identifying specific repositories that a Federal agency might designate for use for particular types of research data resulting from Federally funded research, it would be helpful to mirror the language used in these repositories in relation to provenance.

We believe it would be helpful to clarify the differences between E: *Access*, F: *Free & Easy to Access and Reuse*, and G: *Reuse*, as there appears to be overlap in these sections. G: *Reuse* appears to be referring to metrics on the data, and we recommend amending the name to reflect this and to further distinguish the differences between these sections. Under E: *Access*, we recommend expanding briefly on what “ethical” means. We also recommend clarifying what “access” means - does this mean they can fully download the data, or just view the metadata?

In H: *Secure*, we recommend an additional clarification on closed vs open repositories. Open repositories are not necessarily equipped to set up restrict access to unauthorized users. In closed repositories, this is indeed possible.

In J: *Common Format*, we recommend including language that highlights if the data are in proprietary formats, the repository must require a README which lists all dependencies and contextual information on using the data.

## **Section II. Additional Considerations for Repositories Storing Human Data (even if de-identified)**

In general, we believe this section is comprehensive when considering the sensitive nature of data containing observations on human subjects, and our comments pertain to our own lack of clarity on use cases where these sensitive data need to be stored. It appears this section is referring to *closed repositories* rather than open repositories. With non-government owned data repositories, it is generally not allowable to upload data which has not been deidentified, and the data curation process tied to these repositories generally flags datasets containing this sensitive information. We would love to see a stronger definition of the use cases in which the federal government *would* recommend putting sensitive human data into a repository, when this is not generally the norm with open data repositories. For many open repositories, it would be impossible to implement these types of regulations, especially those with a self-deposit mechanism. We recommend, at the start of this section, defining what specific types of repositories this section's characteristics pertain to.

For section H. *Violations*, if the repository supports self-upload, we would love to see the inclusion of a contingency plan for data which is uploaded that does not meet the requirements of the repository.

For section I. *Request Review*, we suggest providing language on who is qualified to serve as the reviewing group.





# Palantir Technologies' Comments on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

**PREPARED FOR:**

The White House Office of Science and Technology Policy (OSTP)

**PREPARED BY:**

Palantir Technologies Inc.  
100 Hamilton Avenue  
Palo Alto, California 94301

**DATE:**

3/6/2020

**PALANTIR POINT OF CONTACT:**

Katherine Hsiao

[khsiao@palantir.com](mailto:khsiao@palantir.com)

413.388.0056

Palantir Technologies Inc. (“Palantir Technologies”) offers the following comments on the White House Office of Science and Technology Policy’s (OSTP) Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research. Palantir Technologies has extensive experience integrating data across hundreds of federal customers, including agencies specializing in healthcare and scientific research, through our commercial software platform, the Palantir Gotham Platform (“Palantir”, “the Platform”). Palantir integrates critical sources of information to make better-informed decisions and enable secure data sharing across teams with varying skill sets.

The recommendations below reflect the common problems that Palantir Technologies has observed federal agencies encounter when sharing and managing data, including research data. Palantir Technologies would be pleased to answer any further questions the OTSP may have about these recommendations.

### **DESIRABLE CHARACTERISTICS OF REPOSITORIES FOR MANAGING AND SHARING DATA RESULTING FROM FEDERALLY FUNDED OR SUPPORTED RESEARCH**

We recommend that the government focus on provenance and incentivizing research data sharing when considering desired characteristics of repositories for managing and sharing data resulting from federally funded or supported research. By focusing on these two aspects, the government will address robust versioning as a superset of the desired repositories’ characteristics, such as metadata, quality assurance, fast and easy reuse of data, tracking of data reuse, long-term integrity, and more. The Government should emphasize the importance of pursuing a holistic model for both improving existing repositories and creating new repositories. This includes furthering the broader goals of Public Access Plans while also maintaining a high standard for security and privacy to meet the needs of storing human data even if de-identified.

The federal government heavily invests in scientific research each year, yet a large part of the data resulting from this research remains “dark”. In other words, data is described in publications but is unavailable in a raw form for confirmatory analysis or secondary use. This significantly reduces the return on investment in federally funded research and exacerbates the reproducibility crisis facing modern research. As government agencies seek to increase and improve access to data and publications resulting from federally funded R&D through their Public Access Plans, proper data provenance capabilities will be necessary to support discoverability, management, and sharing of data.

## **PROVENANCE AS A FIRST PRINCIPLE**

### **Provenance: Reuse, Free & Easy to Access and Reuse**

Provenance should be clear and accessible for all datasets. All research data should be structured such that data provenance (e.g., user, creation, upload, metadata, logic, transformations, analyses) is stored within repositories. Enabling users to understand the full lineage of their data builds trust in the underlying information and improves data accuracy by simplifying the process of identifying, triaging and fixing errors, leading to more efficient and impactful use and reuse of research data. Data repositories should include the following provenance-related capabilities to promote accessibility and reuse:

- Pair data with compute resources so users can run complex jobs without moving data across systems.
- Datasets and metadata should be downloadable, accessible, or exportable from the repository in a standards-compliant and preferably non-proprietary format, as this makes it easier for users to access and upload research data.
- Allow for the sharing of algorithms and analytical methods with precise versioning to maximize reproducibility and transparency.
- Allow users to open source languages (e.g., sql, python, java) to apply data transformations, provision compute, and more.
- Semantic writeback should exist so that scientists can upload semantic conclusions that they are attempting to assert through their work using a controlled vocabulary, ontology or common data model, improving the accessibility of conclusions.

### **Provenance: Privacy, and PII**

Data provenance is crucial to maintaining granular security and ensuring privacy of sensitive data, such as personally identifiable information (PII) and patient health information (PHI). Data repositories should include the ability to track the full lineage of all data and maintain a detailed logfile of transformations and changes applied to datasets and metadata, beginning with the creation and upload of the dataset. In this way, agencies are able to preserve data integrity and security while also maximizing data sharing.

### **Provenance and Common Format**

Repositories should ensure that data is not compromised despite changes to format, as different researchers need to use the same data in different ways but also without the risk of being unable to trace the data format origin. Reuse requires transforming

data to combine, aggregate or obfuscate the data in some way. Rather than trying to create a common data format or schema, which is notoriously difficult to achieve, data repositories should instead leverage data provenance. Repositories should automatically retain and update metadata within the data itself to preserve integrity and a full inventory of data assets.

## **INCENTIVIZING RESEARCH DATA SHARING**

Incentivizing data sharing is the best way to provide many of the other guarantees articulated in the draft set of desired characteristics, including tracking reuse, privacy, and PII. Researchers currently lack incentives to share their data, especially given the inability to properly acknowledge data owners when data is reused for studies and publications. This is coupled with the technical barrier of ensuring data security in a public access environment and the administrative burden of uploading and maintaining data for reuse.

Data repositories should ideally counter these barriers by including built-in federated and granular privacy capabilities that are approachable to scientists, researchers, and users of all technical skill levels. Granular access controls would enable data owners to have more fine-grained control over the discoverability and accessibility of their data. For instance, data owners should have the option to share data solely with their government funders or with a broader community without having to make an all-or-nothing decision. This will lead to greater visibility into what data is available, whether it can be reused, and how to contact the data owner for access and proper attribution.

Incentivizing research data sharing through desired characteristics of the repositories not only promotes better data sharing and managing, but also assists in fulfilling agencies' Public Access Plans and the February 2013 White House Office of Science and Technology Policy memorandum "Increasing Access to the Results of Federally Funded Scientific Research".

By prioritizing provenance and mechanisms for incentivizing research data sharing, repositories will be better equipped for long-term integrity and storing human data even if de-identified. Through this approach, characteristics of repositories will more quickly reach the goals of Public Access Plans.

**From:** Gulbransen, Tom <[gulbransen@battelle.org](mailto:gulbransen@battelle.org)>  
**Sent:** Friday, March 6, 2020 2:23 PM  
**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>  
**Subject:** [EXTERNAL] Request for Comment on OSTP Repository Characteristics Doc.

Thank you for the opportunity to provide these three comments.  
Tom Gulbransen  
Battelle – NEON

I.C. Metadata records should be provided in machine- and human-readable formats, as per FAIR recommendations.

I.D. Curation & Quality Assurance: Given that the preferred characteristics include the expectation that some agent will “...improve the accuracy and integrity of datasets and metadata”, then the characteristics should also include assignment of ownership for this responsibility. Who owns the dataset of record? In some cases the data generator will be inactive, which may leave dataset ownership with the sponsor agency or maybe even with the repository. In other cases, the repository will simply be an aggregator apart from the organization which is still actively curating the dataset. Ownership of the most current system of record should be documented. Provide a mechanism to document the level of support provided.

I.K. Provenance: The ability to recognize changes to the dataset or metadata is vital. However, there are cases where it would be inadequate to simply mark which data changed. Logfiles of changes should enable descriptions or citations of algorithms if the algorithms influenced the dataset change, including date, user, and, where applicable, changes to code or algorithms used to generate the data update.

## Oak Ridge National Laboratory's Comments on OSTP RFI - Desirable Characteristics of Repositories for Managing and Sharing Data from Federally Funded Research

Katie Knight, Information Science, Data Engineer, Oak Ridge National Laboratory  
Sergei Kalinin, Materials Science, Distinguished Research Staff Member, Center for Nanophase Materials Sciences, Oak Ridge National Laboratory  
Arjun Shankar, Computer Science, Group Leader, Advanced Data and Workflow and CADES Director, Oak Ridge National Laboratory  
Supported/submitted by Jeffrey A. Nichols, Associate Laboratory Director, Computing and Computational Sciences, Oak Ridge National Laboratory

Guidelines and policy statements on data deposition and sharing are only as the extent to which they are put into practice (Schofield et al., 2009). The National Science Foundation requirement that researchers should submit a data management plan has been in place since 2011; in 2013, the Office of Science and Technology Policy distributed a memorandum confirming the need for public access to federally funded research results (Holdren). Furthermore, scientific journals are increasingly requiring that data sharing be integrated with publication submission and distribution (Sturges et al., 2015), but effective searching, storing, and sharing of data is rarely part of a researcher's education. Rather, data management knowledge tends to reside primarily with data managers and librarians (Tenopir et al., 2016). This, coupled with the escalating size and complexity of scientific datasets, underscores the need to shorten the gap between any mandated data lifecycle elements and researcher proficiency. Consequently, any guidelines in place for effective data management should emphasize simplicity in data sharing, reuse, access, searching, and citation.

The value of the data is often difficult to assess without the input of the scientist. Yet, many scientists cannot “curate” for the general scientific audience. How, then, can the data management community help incorporate a scientist's expertise and include their prior knowledge as context? A possible answer emerges when we make the merits of data sharing explicit to the data owners; it should either be obvious as to why a scientist should spend time and possibly extra work sharing data, or the process should be made as simple as possible. We need to recognize that scientific data is Bayesian in nature, and its value requires prior context that is often best understood by the individual scientist collecting it. Therefore, sharing scientific data needs to enable the sharing of this context along with the data, and scientists and organizations should be incentivized to curate and share.

It is clear that there is value in making data and established analysis workflows open, and there is a definite need for tool development that facilitates data interpretation and usability for larger segments of the scientific community. This community is fundamentally heterogeneous and, often, the data will be understandable only to a small number of experts. Therefore, there is merit in broadening the range of scientists who can use data effectively.

The FAIR data principles were developed initially for the life sciences, where data sharing and openness are vital, yet the data are often sensitive and therefore constrained (Wilkinson et al., 2016). Access issues surrounding data ownership, embargo policies, competitiveness, and national security must be addressed in any guidelines provided.

Interoperability of data must be facilitated based on community needs as well as the common techniques, scientific questions, and other forms of association that are common in those communities (Field, 2009). However, “community” is a loose term, as it is often not clear where authority resides. Repository guidelines might focus on recommendations or advice for communities to define or select metadata standards and FAIR data metrics facilitators and/or maturity indicators.

Of the guidelines provided, additional consideration should be made to the “computational narrative” surrounding data creation. Computational processes based on machine learning methods to construct scientific models mean that not just the data, but also the models and processes, must be made available (Brinkman et al. 2019).

## **I. Desirable Characteristics for All Data Repositories**

### *A. Persistent Unique Identifiers:*

Persistent does not mean permanent: a PlosOne study found that of more than one million references extricated from over 3.5 million articles, one in five was a broken link (Klein, 2014). What additional features can be recommended to mitigate "data rot"? Also, there are no current steps in place to prevent multiple identifiers (including DOIs) from being assigned to the same object. It is essential that this is made clear to data management teams, and that this is reflected clearly in the metadata. Additionally, the "persistent landing page" referenced above lists no specific requirements. What information can users expect to find on that landing page if the dataset they are looking for has been deleted, moved, or is otherwise inaccessible? Will the landing page contain redirection information or an explanation as to why the data is no longer available?

### *B. Long-term sustainability:*

How is "long" defined? Are disciplines allowed to determine the "expiration date" of certain datasets? Or can the researchers set an expiration date? The “plan” stated in this guideline does not help with operationalizing this step of data management for a researcher.

### *C. Metadata:*

Not all communities have established standards, and some are more widely adopted than others. What can communities that lack standards do? If an adopted schema is found to be lacking, what steps will be taken to revise and update that schema? How might it evolve, or support interoperability? It should also be clear that the metadata should provide some description of how the data elements in the dataset were collected.

To date, there is no official site that catalogs metadata schema resources for science. Guideline creators may consider creating a set of resources (or invite the community to do so), so that in addition to just guidelines metadata managers and repository creators are aware of what resources are available and/or already in use by their community(ies).

Here is a sample group of metadata standards for science data:

Schema Name	Schema URL
ABCD Schema (Biology)	<a href="https://archive.bgbm.org/tdwg/codata/schema/">https://archive.bgbm.org/tdwg/codata/schema/</a>
AVM Schema (Astronomy)	<a href="https://virtualastronomy.org/avm_metadata.php">https://virtualastronomy.org/avm_metadata.php</a>
Basic Formal Ontology (Gen Science)	<a href="https://basic-formal-ontology.org/">https://basic-formal-ontology.org/</a>
Chemical Methods Ontology	<a href="https://www.ebi.ac.uk/ols/ontologies/chmo">https://www.ebi.ac.uk/ols/ontologies/chmo</a>
CSDGM Geospatial	<a href="https://www.fgdc.gov/metadata/csdgm/">https://www.fgdc.gov/metadata/csdgm/</a>
CSDGM Geographic	<a href="https://www.fgdc.gov/standards/projects/FGDC-standards-projects/csdgm_rs_ex/remote-sensing-metadata/">https://www.fgdc.gov/standards/projects/FGDC-standards-projects/csdgm_rs_ex/remote-sensing-metadata/</a>
CSDGM Biological	<a href="https://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/biometadata/">https://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/biometadata/</a>
CF-NetCDF (Climate)	<a href="https://cf-trac.llnl.gov/trac">https://cf-trac.llnl.gov/trac</a>
Darwin Core (Biology)	<a href="http://rs.tdwg.org/dwc/">http://rs.tdwg.org/dwc/</a>
Data Catalog Vocabulary (Interoperability)	<a href="https://www.w3.org/TR/vocab-dcat-2/">https://www.w3.org/TR/vocab-dcat-2/</a>
DIF 10 (Earth Science interchange)	<a href="https://gcmd.gsfc.nasa.gov/DocumentBuilder/defaultDif10/guide/index.html">https://gcmd.gsfc.nasa.gov/DocumentBuilder/defaultDif10/guide/index.html</a>
MATHML (Mathematics)	<a href="https://www.w3.org/Math/">https://www.w3.org/Math/</a>
NLM MDC (Medicine)	<a href="https://www.nlm.nih.gov/tsd/cataloging/metafilenew.html">https://www.nlm.nih.gov/tsd/cataloging/metafilenew.html</a>
OBI (Biomedical)	<a href="http://obi-ontology.org/">http://obi-ontology.org/</a>
Semantic Sensor Network	<a href="https://www.w3.org/TR/vocab-ssn/">https://www.w3.org/TR/vocab-ssn/</a>
VSO (Astronomy)	<a href="https://docs.virtualsolar.org/wiki/DataModel18">https://docs.virtualsolar.org/wiki/DataModel18</a>

Furthermore, additional guidelines may be useful for staff who need to apply and maintain the metadata. Lacking a local standard (or when looking to apply one), the following considerations will be useful:

1. Who are the potential users and what are their needs?
2. Are there any cataloging/metadata staff, and what is their level of expertise?
3. What is the available time/money for maintaining?
4. How will the resources be accessed (i.e., catalog, command line, website, etc.)?
5. Will any relationships need to be established with other data collections or assets (in other words, will the data be a mixture of disciplines, or will this catalog eventually be linked with other data catalogs)?
6. What is the overall expected scope of the collection?
7. Will the metadata need to be harvested for analysis (data engineering tasks, for example)?
8. What are the interoperability considerations?
9. What are the required levels of maintenance and quality control?

#### *D. Curation & Quality Assurance:*

Data curators gather, prepare, and transfer data and often have a hand in preservation and archiving. Clarifying what is meant by "expert curation" would be useful. Is this section meant to cover all of these aspects, and to what extent? Criteria for what constitutes "quality" should



also be addressed. Does this include built-in capabilities for creating data management plans, metadata application profiles, and a means to define "quality" in the context of the repository?

*E. Access:*

Ease of access needs to be clarified here: as stated above, scientific data is often context-specific and best understood by the collector. "Accessible under well-defined conditions" is probably a better interpretation for the scientific community, primarily for reasons surrounding national security, competitiveness, and privacy (Mons et al., 2017).

*G. Reuse:*

Data citation ought to be included here. It should be a requirement (or strongly suggested guideline) that any data repository provides a means to easily cite data. Provision of just a PUID may not be enough, especially if the PUID is only relevant to local data users (i.e., does not adhere to a nationally or internationally recognized standard). Mechanisms for ease of retrieval, translation, and extraction must be provided.

*J. Common Format:*

This may be a problem for the sciences, where each discipline can have its own special data format. To what standard are these guidelines referring? Is the standard decided by the scientific community or the data managers? What if the data standard changes? Will older data need to be converted to the new accepted format?

*K. Provenance:*

Provenance is a tricky issue with data and is linked with the equally tricky issue of data citation. If a document cites a particular dataset D, what happens if the owner of D updates that dataset? While these guidelines are not meant to be design features there should be some better guidance on how to handle provenance, as a logfile may not be sufficient (this assumes that data citation includes a date of use; currently there are no official data citation standards).

## **II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

To what end does this section also apply to data that may be sensitive but not human data?

*B. Restricted Use Compliant:*

Restriction of use can become extremely tricky if multiple parties are involved in dataset management. For instance, if a researcher creates a new dataset via analytics, who owns that new dataset? The guidelines should encompass such scenarios.

#### F. Clear Use Guidance:

See the restricted use comments (item II.B). Not all data management plans are created by parties who interact with the data. Again, while these guidelines are not design requirements, it would be advisable to provide or recommend resources for researchers and data managers on methods for mapping data use.

#### References

Brinkman, A., et al. (2019). Computing environments for reproducibility: Capturing the “Whole Tale.” *Future Generation Computer Systems*, 94, 854-867.

Holdren, J.P. (2013). Memorandum for the heads of executive departments and agencies: Expanding public access to the results of federally funded research. Executive Office of the President, Office of Science and Technology Policy; Available: [http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf)

Klein M., Van de Sompel H., Sanderson R., Shankar H., Balakireva L., et al. (2014) Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLOS ONE* 9(12): e115253. <https://doi.org/10.1371/journal.pone.0115253>

Mons, B., et al. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, 37, 49–56.

National Science Foundation. (2011). Significant Changes to the GPG. Available: [http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg\\_sigchanges.jsp](http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_sigchanges.jsp)

Schofield, P.N., Bubela, T., Weaver, T., Portilla, L., Brown, S.D., Hancock, J.M., et al. (2009). Post-publication sharing of data and tools. *Nature*, 461, 171–173. pmid:19741686

Sturges, P., Bamkin, M., Anders, J.H., Hubbard, B., Hussain, A. and Heeley, M. (2015), Research Data Sharing: Developing a stakeholder-driven model for journal policies. *Journal of the Association for Information Science and Technology*, 66, 2445-2455. doi:10.1002/asi.23336

Tenopir, C., et al. (2016). Data Management Education from the Perspective of Science Educators. *International Journal of Digital Curation*, 11(1), 232–251. doi:10.2218/ijdc.v11i1.389

Wilkinson, M. D., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3. Available: <https://www.nature.com/articles/sdata201618>

Field, D., et al., (2009). ‘Omics Data Sharing. *Science* 09 Oct 2009: Vol. 326, Issue 5950, pp. 234-236. DOI: 10.1126/science.1180598

# **NIEHS Response to the Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research**

## **Name/role/affiliation**

Richard Woychik, Ph.D., Acting Director of the National Institute of Environmental Health Sciences (NIEHS) and the National Toxicology Program (NTP), National Institutes of Health (NIH), U.S. Department of Health & Human Services (HHS).

## **Primary scientific discipline(s)**

Life sciences, environmental health sciences.

## **Introduction**

The mission of the National Institute of Environmental Health Sciences (NIEHS) is to discover how the environment affects people in order to promote healthier lives. NIEHS works to accomplish its mission by conducting and funding research on human health effects of environmental exposures, developing the next generation of environmental health scientists, and providing critical research, knowledge, and information to citizens and policymakers, to help in their efforts to prevent hazardous exposures and reduce the risk of preventable disease and disorders connected to the environment. Success in the NIEHS mission requires strong stewardship of resources, including scientific research data and data infrastructure.

This document is NIEHS' response to the Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research. These comments represent input from the NIEHS as a whole, with specific suggestions and observations from NIEHS' three major scientific divisions:

- Division of Extramural Research and Training (DERT),
- Division of Intramural Research (DIR), and
- Division of the National Toxicology Program (DNTP).

The NIEHS DERT administers the institute's grant program, which funds research and research training in environmental health. DERT grantees conduct research in a variety of Environmental Health Sciences (EHS) fields, including but not limited to toxicology, epidemiology, exposure research, and social determinants of health. These efforts address complex environmental health problems that are enhanced by and dependent on diverse data management and sharing mandates and practices.

Additionally, the NIEHS DIR and DNTP conduct both broad programmatic and investigator-led research across a wide range of disciplines including, toxicology, pathology, epidemiology, genomics and epigenomics, structural biology, computational biology, reproductive and developmental biology, and clinical research.

Research conducted by NIEHS and supported by NIEHS grants involves an extremely broad range of data types of variable size and structure as well as diverse security, privacy, and compliance mandates. Data from NIEHS-supported research are preserved in wide range of repositories, including repositories operated by NIEHS, NIH, and/or other government entities, repositories supported through NIEHS and/or NIH grant awards, as well as repositories operated by non-governmental entities.

Comments on each topic are listed below.

## **The proposed use and application of the desirable characteristics (as described in the Background section)**

NIEHS is strongly supportive of the efforts of the Subcommittee on Open Science (SOS) to advance open science and improve the consistency of guidelines and best practices that agencies provide about the long-term preservation of data from Federally funded research. The proposed “Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research” is a worthwhile first step in ensuring publicly funded data are readily findable, accessible, and secured for long-term preservation. In general, NIEHS agrees with the proposed use and application of the set of desirable characteristics of data repositories for data resulting from Federally funded research.

NIEHS believes that a formal definition of “repository” is needed, as the scope of coverage for applying the desirable characteristics is not sufficiently defined. Clarification is needed on whether these characteristics are intended to apply to only those publicly-funded sites which exist to accept/manage/disseminate data from deposition by others, or whether these characteristics also apply to databases found in the primary data producer’s laboratory (or somewhere in between).

**The appropriateness of the “Desirable Characteristics for All Data Repositories” (Section I) for data repositories that would store and provide access to data resulting from Federally-supported research, considering: Characteristics that are included, and additional characteristics that should be included.**

- A. *Persistent Unique Identifiers (PUIID)*: Assigning datasets a PUIID is an essential characteristic. However, additional consideration, guidance, and best practices are needed around procedures for assigning PUIIDs for objects at different levels, including raw and processed data and objects within a dataset. For example, using standard identifiers (e.g., for gene IDs, test article names) permits interoperability of data within and between datasets. With interoperable identifiers, related data may be more easily discovered, and the ease of data-mining across datasets is enhanced. Additionally, careful consideration is needed on how each repository assigns PUIIDs to fit into the overall data ecosystem in a way that allows seamless references across repositories and avoids duplication of PUIIDs.
- B. *Long-term sustainability*: Plans for long-term sustainability are important, however there are challenges and inconsistencies with the current NIH funding structure, with most awards being made for 5-year periods. If NIH is the sole funder of a repository, then the long-term sustainability plan is a major challenge, and it is difficult to gauge the likelihood of success of such sustainability plans. Cooperation is needed across stakeholders (including funders, publishers, and repositories) to better understand the issues, expectations, and potential solutions for addressing long-term sustainability.
- C. *Metadata*: Ensuring that datasets are accompanied by sufficient metadata is an essential characteristic, however this area is currently a major challenge. Investigators need a way to identify the appropriate schema for their data, and resources are needed for communities to define standards. In defining these standards, it is desirable for metadata terms to be interoperable with external lexicons and across repositories.
- D. *Curation & Quality Assurance*: The current wording for this characteristic is unclear. Data quality standards are needed that align with community-accepted metadata schema. Common quality standards need to be consistently applied to data, and repository staff are needed to adjudicate data quality.
- E. *Access*: Ensuring broad, equitable, and maximally open access, as appropriate, is important. However, access will need to be appropriately regulated when dealing with Human Subjects or other controlled data, and in these situations, repositories need to consider who is accessing data and how they are going to use the data.
- F. *Free & Easy to Access and Reuse*: Care should be taken around the use and definition of ‘free & easy’. Going back to the long-term sustainability issue, repositories need to be able to recoup costs for doing business, including costs for supporting curation and access. In certain situations, service at completely no cost may not be reasonable if there are not adequate federal mechanisms for sustaining

repositories. Repositories should strive to minimize costs and restrictions for access and reuse, as appropriate, but in certain situations may need to reflect the true costs.

- G. *Reuse*: Enabling tracking of data reuse is very important. However, a definition of “reuse” is needed. It would be beneficial to create a distinction between utilization metrics (e.g., downloads, webpage hits) and outcome metrics (e.g., data citation, publications). It is important to understand how data are being reused, for example as part of a meta-analysis, for secondary data analysis, for reproducibility, or for training. There is opportunity for repositories to play a more active role in promoting reuse. Partnership with publishers to encourage/enforce citation of data is also critical for tracking reuse.
- H. *Secure*: The definition of ‘security’ is broad and unclear. Repositories realistically must balance between desired characteristics of being “free & easy to access and reuse” and high cost features like “security”.
- I. *Privacy*: Compliance with applicable privacy requirements is important but presents many challenges and costs.
- J. *Common format*: To support data science best practices, it is recommended that the repository support not only human query, but also machine accessibility through APIs. On a broader point, the entire list of characteristics should be reorganized such that related terms are grouped together. This point goes hand-in-hand with the above point on metadata.
- K. *Provenance*: Tracking provenance is important, but this may not be an essential characteristic for all repositories.

Additional characteristics recommended:

1. *Mission/Purpose Statement*: The desirable characteristics should be expanded to include a mission/purpose statement per the CoreTrustSeal and ISO Standard.
2. *Statement of Compliance*: NIEHS recommends that repositories communicate their adherence with the Desirable Characteristics of Repositories to the public using a standard format to permit comparison of different repositories. Eventually communication of this adherence should be a prerequisite for publication of publicly-funded data using a given repository. This can be assessed by the funding agency, especially if a standard reporting form is used.

**Appropriateness of the characteristics listed in the “Additional Considerations for Repositories Storing Human Data (even if deidentified)” (Section II) delineated for repositories maintaining data generated from human samples or specimens, considering: Characteristics that are included, and additional characteristics that should be included.**

- A. *Fidelity to Consent*: For human data, restriction of dataset access to appropriate uses consistent with consent is essential. This must be documented appropriately.
- B. *Restricted Use Compliant*: This point should emphasize restricting unauthorized access and use purposes rather than unauthorized users.
- C. *Privacy*: This point seems duplicative of other points in the Human Data list as well as the security and privacy characteristics in the general list.
- D. *Plan for Breach*: no comment.
- E. *Download Control*: In certain situations, repositories may consider the use of common compute spaces to access and analyze data in the cloud (without downloading or transferring data), which could help with security and privacy controls.
- F. *Clear Use Guidance*: The point around guidance for data access and use, together with the point on “request review”, needs to be emphasized higher in the list and further expanded, as this is an important issue with many complexities.
- G. *Retention Guidelines*: no comment.
- H. *Violations*: An overarching area of challenge is addressing violations of terms-of-use and data mismanagement. The repository’s plans should include how to address situations where data are

accessed or used in ways inconsistent with consent. For situations where the repository is mismanaging data, then plans should include an independent, third party assessment of the situation.

- I. *Request review*: Transparency in the review for data access/use requests is very important; a repository needs clear criteria that are as objective as possible. Repositories should strive to prioritize the purpose for which the data are being requested rather than limiting access to a small academic/research community.

### **Considerations for any other repository characteristics which should be included to address the management and sharing of unique data types (e.g., special or rare datasets)**

- Many repositories house preexisting or legacy data that do not currently meet these desired characteristics. Issues with legacy or preexisting data are not sufficiently addressed.
- When considering appropriate data use guidance, repositories should be aware of the potential impacts of vested economic interests with respect to data linked to a particular commercial product.
- Nothing in these characteristics addresses access by foreign entities (foreign universities, foreign corporations, etc.), and this needs to be considered, particularly in relation to security, privacy, and data breach.

### **The ability of existing repositories to meet the desirable characteristics**

- In general, the draft characteristics are brief and broadly written, which means that repositories can be in 'compliance' with varying degrees. For example, no specific guidance is provided on what constitutes 'privacy', 'quality assurance', or 'timely manner' for data release. Each repository could interpret and implement much of the language in very different ways resulting in lack of consistency across repositories.
- Not all existing repositories will meet all these desirable characteristics. While some are close, there are no repositories that are perfect. The National Institute of Mental Health Data Archive (NDA) National Database for Autism Research (NDAR) is a great model.
- The metadata standard, metadata structure/schema, and quality assessment characteristics are going to be the hardest to address in many cases, because community standards do not currently exist. It would be beneficial for Federal agencies to provide sustained support for community data standard development efforts. These efforts to recruit experts in a particular field to set up community standards/guidelines can use fields with established success in data standards implementation (e.g., genomics) as guiding examples.
- Given the wide range of repositories and their level of sophistication, it is recommended that implementation be gradual, moving toward an implementation target date. A recommended timeline for adoption including milestones would be useful. By establishing a timeline and request for a 'compliance report,' repositories will be prompted to act in a timely way and make it easier for funders and users to determine how adherent repositories are with the guidelines. As a result, those repositories in compliance will rise to the top of usage.
- Currently, these characteristics are 'desirable.' As written, the guidance provides no incentive for repositories to adopt these characteristics. It is currently unclear whether repositories will move toward meeting these characteristics unless these are set as mandatory or minimal requirements.

### **Consistency of the desirable characteristics with widely used criteria or certification schemes for certifying data repositories**

- Upon comparison to the CoreTrustSeal criteria, there is a high degree of overlap, but there are some specific areas from CoreTrustSeal that are missing in the characteristics listed, including documentation of an explicit mission statement (that would be worth adding to the desirable list), defined workflows,

preservation plans, among others. Efforts to align with these widely-used criteria or certification schemes should be encouraged.

**Any other topic which may be relevant for Federal agencies to consider in developing desirable characteristics for data repositories.**

- To promote transparency for repositories that provide/preserve government data, Federal agencies could consider requesting creation of a publicly available dashboard/report that states each repository's level of 'compliance' with each of these characteristics.
- Moving forward, it may be beneficial to define a subset of minimal required characteristics that repositories are expected to meet.
- An additional desired characteristic that is not specified is the ability of repositories to ensure that their data/metadata can be machine accessible.

**Concluding Remarks**

NIEHS appreciates the opportunity to provide feedback and thanks the Office of Science and Technology Policy for considering the points of clarification, additional guidance, and other considerations raised above. These cross-agency efforts represent a major step forward in advancing open science and improving access to data from Federally funded research.

**Response from the National Solar Observatory to “Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research”**

**Author:** Robert Tawa, National Solar Observatory (Data Center Project Manager, Daniel K. Inouye Solar Telescope)

**Domain:** Physical Sciences (Astronomy & Astrophysics)

## Introduction

***The National Solar Observatory (NSO) is the United States’ national center for ground-based solar physics (studies of the Sun, its environs, and its impact throughout the Solar System). NSO is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation (NSF) Division of Astronomical Sciences. NSF’s Daniel K. Inouye Solar Telescope (Inouye Solar Telescope, or Inouye), is the largest solar telescope in the world, with a focus on understanding the Sun’s dynamic behavior. Inouye’s 4-meter mirror provides views of the solar atmosphere as never seen previously. NSO’s data repositories make images and spectra of the Sun available to scientists and the public doing research that advances America’s leadership in the study of the Sun. These images have long-term scientific legacy value for studying solar dynamics over time.***

In mid to late 2020, NSF’s Inouye Solar Telescope will begin to operate and will deliver an unprecedented petabyte-scale solar data database which will be accessible in real time to all US astronomers and international participants. Inouye promises to transform the field of heliophysics (study of the Sun), and at the same time strives to increase participation in forefront solar research with its associated data archives and data services. Inouye Solar Telescope observations use detectors that convert solar light into digital imaging and spectroscopy. Inouye will, when fully operational, generate 6 Petabytes of data and 150 Million images yearly. Hence, Inouye operates computerized data archives and data services as an integral part of its science-enabling mission. The Inouye data repository serves a broad range of needs:

- Transfer of raw data from the mountaintop at Haleakala Observatory on the island of Maui where observations are made to a centralized location for storage and processing.
- Implementation of redundant, geographically-distributed backup and disaster-recovery systems.
- Integration with computing capabilities that transform raw data and calibrations into science-ready data products.
- Enabling data search, discovery, and open access for research investigators to enable the primary science for which the observations were conducted.
- Long-term curation of Inouye data, along with associated metadata, software, documentation, and expert knowledge.



## Responses regarding the ability of the Daniel K. Inouye Solar Telescope to meet the desirable characteristics

- A. *Persistent Unique Identifiers:*** NSF's Daniel K. Inouye Solar Telescope Data Center addresses this characteristic through the use of a unique "Proposal ID" for the telescope observing proposal with which individual data sets are associated. As well, each individual frame and dataset have a unique Frame ID and Dataset ID respectively, all which can be used for attribution, usage tracking, and bibliometric analysis.
- B. *Long-term sustainability:*** Inouye Solar Telescope has a long-term mission (~ 44 years) and purpose that includes stewardship of Inouye-produced data for the length of the project. While long-term sustainability is a desirable characteristic of a data repository, the paragraph detailing how that may be achieved may be insufficient: specifically, requiring data repositories to simply have a long-term plan for (among other things) "availability of datasets" may not address the core issue of actually implementing such a plan. Historically, while the National Science Foundation has requested such plans, the needed funding may not always be available for curation of the data beyond the planned lifetime of the observatory that created the data. In the past, once an observatory was no longer funded, the primary repository of the data was decommissioned, along with the decommissioning of any backup or mirror site. The National Science Foundation now recognizes that long-term data storage is important, and is working with groups such as ours to ensure effective storage and curation of data from federally-funded facilities.
- C. *Metadata:*** Data repositories at our Data Center extract standard metadata from raw data file headers; validate and standardize it as appropriate for each instrument; and ingest it into online metadata databases for use in search and discovery. Metadata queries are supported through interactive web interfaces as well as application programming interfaces (APIs) that can be used by Virtual Observatory protocols and application written by external researchers.
- D. *Curation and Quality Assurance:*** Our data repositories are developed, operated, and maintained by integrated teams of scientists and software engineers with deep expertise in the data sets that they serve. Quality Assurance of the data begins at ingest of the data by verifying that headers contain all required data, and is continuous through processing as quality metrics are calculated for inclusion into dataset quality reports and for trend analysis.
- E. *Access:*** Our data repositories are fully accessible for free by the global solar community. The data are supplemented by on-line capabilities for data discovery and exploration, as well as downloadable user tools for visualization and analysis.
- F. *Free & Easy Access to Reuse:*** Data obtained by the Inouye Solar Telescope are discoverable as soon as they are ingested into the repository, and are automatically made available for

download as soon as the proprietary period - if one exists - of the original investigators expires.

- H. Secure:** Data access controls are automatically enforced by user authentication and authorization services. These controls are supplemented by traffic monitoring and load balancing systems that can throttle or stop unauthorized or excessive (bot driven) requests.
- I. Privacy:** Our data repositories are operated within a broader comprehensive organizational cybersecurity framework at the National Solar Observatory. In addition, and by design, the Daniel K. Inouye Solar Telescope's data repositories contain no Personally Identifiable Information (PII) for anyone.
- J. Common Format:** Our data repositories serve data in Flexible Image Transport System (FITS) format, which is the most open and widely accepted data-file standard within astronomy worldwide. We also serve data sets in the Advanced Scientific Data Format (ASDF), which is a modern replacement to the FITS format that permits easier search and discovery of data on laptops and workstations.
- K. Provenance:** Provenance for data hosted in our repositories is based on maintaining and including in with the data, records of the telescope, instruments, and observing programs that delivered the raw observational data, and the software programs, versions, and systems associated with creation of higher-level data products based on the raw data.

### **Another topic which may be relevant for Federal agencies to consider in developing desirable characteristics for data repositories**

***Proximity of data to analysis capabilities:*** In the age of terabyte to petabyte data sets, scientists can no longer just download datasets and process them on their laptop or workstation. The memory, storage capacity, and processing power of a laptop or workstation is no longer sufficient for the data rates generated by today's facilities. For these very large data sets, moving them around has become expensive, inefficient, and increasingly restricted to users with access to large facilities with the ability to consume and process large data sets. Young and aspiring scientists in high school and college, as well as amateur ("citizen") scientists working at home have very limited access to such data.

To make these large data sets available to all, data and processing capability must be (1) co-located and (2) free to users. Co-location of the data with the processing infrastructure eliminates the requirement to transfer large amounts of data to where it will be processed. Making the processing free to users would increase access of the data repository to those who do not have the technical facilities at hand to store and process large volumes of data.

**RFC Response: Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded or Supported Research**

**Authors:**

Louise M. Prockter, Lunar and Planetary Institute (PDS Chief Scientist)

Matthew S. Tiscareno, SETI Institute (PDS Ring-Moon Systems Node)

Charles H. Acton, NASA Jet Propulsion Laboratory (PDS Navigation and Ancillary Information Facility)

Raymond E. Arvidson, Washington University in St. Louis (PDS Geosciences Node)

James Bauer, University of Maryland (PDS Small Bodies Node)

Nancy Chanover, New Mexico State University (PDS Atmospheres Node)

Daniel J. Crichton, NASA Jet Propulsion Laboratory (PDS Engineering Node)

Lisa R. Gaddis, U.S. Geological Survey (PDS Cartography and Imaging Systems Node)

Edward A. Guinness, Washington University in St. Louis (PDS Geosciences Node)

Trent M. Hare, U.S. Geological Survey (PDS Cartography and Imaging Systems Node)

John S. Hughes, NASA Jet Propulsion Laboratory (PDS Engineering Node)

Jordan Padams, NASA Jet Propulsion Laboratory (PDS Engineering Node)

Mark R. Showalter, SETI Institute (PDS Ring-Moon Systems Node)

Susan Slavney, Washington University in St. Louis (PDS Geosciences Node)

Thomas Stein, Washington University in St. Louis (PDS Geosciences Node)

Raymond Walker, University of California at Los Angeles (PDS Planetary Plasma Interactions Node)

Email: [prockter@lpi.usra.edu](mailto:prockter@lpi.usra.edu)

**All authors are planetary scientists and/or data scientists who are affiliated with the NASA Planetary Data System (PDS).**

## Introduction:

This RFC Response is presented on behalf of the Planetary Data System (PDS), a distributed data archive that hosts and serves data collected by Solar System robotic missions, and ground-based support data relevant to those missions (Appendix A). The PDS is managed by the NASA Science Mission Directorate's Planetary Sciences Division as an active archive that makes available well documented, peer-reviewed data to the research community. The main objective of the PDS is to maintain a planetary data archive that will withstand the test of time such that future generations of scientists can access, understand and use historical planetary data. The PDS ensures compatibility across the archive by adhering to strict standards of data archiving formats and required documentation. The PDS4 archiving standard has been required for data archives from NASA-funded missions since 2011, and provides simple, standardized formats for long-term stability and interdisciplinary use.

The PDS is divided by science discipline into six teams (called "nodes"), each of which curates data holdings relevant to its discipline's community of researchers and actively interfaces with its discipline's research community to understand and meet its needs. Each node is led by an active planetary science researcher. Technical support is provided by the Engineering Node and the Navigation and Ancillary Information Facility, both at NASA Jet Propulsion Laboratory. PDS project management is provided by the Solar System Exploration Data Services Office at NASA Goddard Space Flight Center.

The PDS is a founding member of the International Planetary Data Alliance (IPDA), a group supported by the international Committee on Space Research (COSPAR), which actively works for common data standards and open planetary archives. The IPDA membership includes representatives from the space agencies of most spacefaring nations. The IPDA has adopted PDS4 as the international archiving standard for planetary mission data, and the archives of international missions are increasingly interoperable with PDS.

The PDS already exhibits many of the desirable characteristics of (non-human) data repositories as described in the RFC's draft guidelines, or is in the process of implementing those characteristics. Our services to NASA and the planetary community are evolving as the community's needs evolve (McNutt et al., 2017).

In this RFC Response, we offer comments and recommendations regarding each characteristic in the draft guidelines and will also comment on how those guidelines apply to PDS and related archives, and how PDS is already addressing those guidelines. If desired, the authors would be glad to continue a conversation with OSTP regarding archiving best practices.

---

*A. Persistent Unique Identifiers:* Assigns datasets a citable, persistent unique identifier (PUIID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The PUIID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.

**Suggested changes:** None

**PDS Application:** PDS assigns DOIs to datasets, fulfilling this requirement. This is currently optional for data providers, but we expect it to become required for all datasets as early as 2021. We are currently finalizing a streamlined DOI procedure that will enable this change. Furthermore, each individual data product within PDS holdings is assigned a Logical Identifier (LID) that uniquely points to it via multiple hierarchical fields. The LID can be appended by a Version Identifier to become a LIDVID, which uniquely identifies a particular version of the data product. Although LIDVIDs do not point to a persistent landing page (this would be impractical at the data-product level, due to volume), they provide all other benefits of PUIDs with a level of granularity that exceeds the capabilities of DOIs.

---

*B. Long-term sustainability:* Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.

**Suggested changes:** None

**PDS Application:** This is a core value of the PDS. Our information model and archiving standards are engineered to ensure that future generations of scientists will be fully able to understand and use the data. PDS4 formats are designed to be both robust over the long term and difficult to misunderstand, even decades from now when computing standards will have dramatically changed. We observe information security measures to guard against tampering, and we keep multiple, distributed copies of our holdings that are readily accessible for restoration if necessary.

---

*C. Metadata:* Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.

**Suggested change:** Change “metadata” to “metadata and documentation,” both in the title and in the body of this guideline.

**Justification:** Repositories should include documentation sufficient to understand instrument function, data collection methodology (operational and contextual), and calibration/processing applied to the data.

**PDS Application:** Discipline scientists within PDS work with data providers to prepare both documentation and metadata, which are crucial for enabling other researchers to discover and use the data. PDS also produces additional metadata of its own, which in many cases is essential for enabling cross-platform search tools to guide potential data users to datasets that serve their needs.

---

*D. Curation & Quality Assurance:* Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

**Suggested change #1:** Change “improve” to “ensure”

**Justification:** Without expert curation, datasets and metadata may not have desired attributes at all. The job of expert curation is not to “improve” desired attributes, as if they can be assumed to exist, but to ensure that they exist.

**Suggested change #2:** Change “accuracy and integrity” to “accuracy, integrity, discoverability, and usability”

**Justification:** Data might be difficult to find or to use, even if it is accurate and intact. The job of expert curation is to ensure all these attributes for the benefit of the research community.

**Suggested change #3:** Add a concluding sentence: “Data should be peer reviewed by discipline experts for integrity and usability.”

**Justification:** Scientific review of data upon archiving is important for ensuring usability of both the metadata and data itself. Peer review ensures that the metadata can effectively be used to annotate and understand the data. For the data itself, it helps to ensure scientific usability. All data repositories should include mechanisms for data validation and peer review.

**PDS Application:** Expert curation by discipline scientists is at the core of PDS’ value to the community. PDS actively curates its holdings to ensure that documentation, file formats, citeability, and discoverability remain current, and to provide individual support to users. As prominent experts in their fields, PDS discipline scientists create a bridge between the scientific community and the public archives, ensuring that the archive is scientifically useful and that the local policies are consistent with the scientific needs of the community. Furthermore, PDS ensures that all data are peer reviewed by scientists for integrity and usability. Reviewers are drawn from the community and have expertise relevant to the science discipline of the data.

---

*E. Access:* Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.

*F. Free & Easy to Access and Reuse:* Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

**Suggested change:** These two guidelines should be combined.

**Justification:** These two guidelines substantially overlap

**PDS Application:** We agree that access to data should be broad, equitable, maximally open, and as easy as possible. To facilitate this, we host search tools that allow users to discover data using a variety of parameters across data sets. We develop the metadata that enables such cross-platform search, and we are expanding its coverage to more datasets. Furthermore, we are implementing an API to support access and interoperability.

All PDS holdings are in the public domain, and are made accessible free of charge. We release data in a timely manner, according to schedules announced beforehand. These schedules are designed to balance the need of data providers to publish their own work and the need of the larger community to have timely access to data.

*G. Reuse:* Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).

**Suggested change #1:** Change “reuse” to “use” (also in guidelines C and F)

**Justification:** The guidelines appear to mean use of archived data by researchers other than the data provider. However, this usage is not clear. Furthermore, it should not be assumed that the data providers used the data for research before archiving, nor that their further use of archived data constitutes “reuse.”

**Suggested change #2:** Change title from “Reuse” to “Usage Tracking”

**Justification:** This guideline is not about enabling data use, but about tracking it with metrics.

**PDS Application:** PDS recognizes the need for metrics to track the use of archived data in published research. Citation of data in public archives should be a standard, scientific practice, and the production of a high-quality citable dataset should be considered equivalent to the production of a scientific publication. However, agencies should be aware as they evaluate archives that both community standards and archive capabilities are still evolving in this direction, and that metrics may substantially undercount the actual usage of archived data.

---

*H. Secure:* Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization’s ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>) or the National Institute of Standards and Technology’s 800-53 controls (<https://nvd.nist.gov/800-53>).

*I. Privacy:* Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.

**Suggested change #1:** These two guidelines should be combined.

**Justification:** These two guidelines substantially overlap

**Suggested change #2:** If “privacy” refers to the privacy of archive users, or privacy of data, this should be clarified.

**Justification:** It is not clear what type of privacy is meant.

**PDS Application:** Because all PDS holdings are in the public domain, there is no need to worry about “unauthorized access or release” of current holdings. On the other hand, PDS does practice responsible information security to protect its software, holdings being prepared for release, etc. In the near future, PDS may begin to collect user data in order to improve the experience of data users. If and when this is done, the collected data will be safeguarded using current best practices for handling Personally Identifiable Information (PII).

---

*J. Common Format:* Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.

**Suggested change:** Change the title from “Common Format” to “File Formats”

**Justification:** The current name of this section implies that it is desirable to use a format in common use. Such formats are often proprietary or otherwise inappropriate as an archive format. A format suitable for a long-term archive is often not consistent with today’s popular formats.

**PDS Application:** PDS works to ensure that all data are as useable as possible. Our datasets come from a large number of instruments and spacecraft and are not in one common format, although we do insist on using non-proprietary formats. When the data are not in a format easily accessible by user tools, PDS provides transformation tools to assist users with data access and/or provides additional versions of data in formats that support browsing.

---

*K. Provenance:* Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.

**Suggested change #1:** Change “of” to “describing the heritage or source of the data,”

**Suggested change #2:** After “user,” add “substance and reason for all changes,”

**Justification:** In order to fully state the provenance of a dataset, logfiles should specify the starting point and should summarize the changes.

**PDS Application:** Change logs and versioning are core components of PDS standards.

**References:**

McNutt RL *et al.* 2017. Planetary Data System Roadmap Study for 2017–2026. NASA. 110pp.  
[https://pds.nasa.gov/home/about/PlanetaryDataSystemRMS17-26\\_20jun17.pdf](https://pds.nasa.gov/home/about/PlanetaryDataSystemRMS17-26_20jun17.pdf)



## **Comments of the Medical Library Association and Association of Academic Health Sciences Libraries**

### **Re: OSTP Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research**

**Document 85 FR 3085**

**Page 3085-3087**

**Document Number: 2020-00689**

**Submitted March 6, 2020**

These comments were prepared by health sciences librarians who are members of the Medical Library Association (MLA) and Association of Academic Health Sciences Libraries (AAHSL). As health information professionals, their primary scientific discipline is in health and bio-sciences, and they engage in the practice of librarianship, data management, bioinformatics and other areas of information research, and library administration.

#### **Desirable Characteristics for All Data Repositories**

- Persistent Unique Identifier (PUID).

The PUID must

- Be resolvable
- Point to the data object or a persistent landing page to indicate and, if possible, link to alternative access when the dataset is de-accessioned or no longer available
- Manage the DOI of related publications in order to manage the linking from the datasets to related articles - at the time of submission - to provide context for the datasets

It would be ideal to track data reuse. Supporting periodic maintenance that links the data to future publications (e.g., studies that cite the article where the dataset is used) would be useful to researchers.

- Long Term Sustainability.

The preservation policy should be transparent and easy to find. The funding agency should provide clear/explicit guidance on all aspects of the preservation, and not leave these to the repository to decide. There should be minimum or required standards for compliance.

- Metadata.

Ideally, the use of common coding standards and associated resources should be established to promote the creation of more effective and interoperable biomedical information repositories. Two examples are,

- The Unified Medical Language System Metathesaurus which contains over one million biomedical concepts from over 100 source vocabularies that brings together many health and biomedical vocabularies and standards enabling interoperability between computer systems  
[https://www.nlm.nih.gov/research/umls/knowledge\\_sources/metathesaurus/index.html](https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html); and
- NISO's Recommended Practices for Online Supplemental Journal Article Materials that provides the minimum metadata for digital objects  
[https://groups.niso.org/apps/group\\_public/download.php/10055/RP-15-2013\\_Supplemental\\_Materials.pdf](https://groups.niso.org/apps/group_public/download.php/10055/RP-15-2013_Supplemental_Materials.pdf)

In addition, the repository should provide a mechanism that supports continuous improvement or updating to keep the data fit for purpose (e.g. updating nomenclatures, gene symbols)

- Curation and Quality Assurance.

Given the large variety in types of data and types of research, it is hard to predict the effort, expertise, and infrastructure to ensure that all datasets are properly curated (correct metadata, documentation, format, etc). Repositories should provide mechanisms or point to mechanisms/services so that researchers with different levels of data curation can take their datasets through the process of curation and QA before uploading.

- Access.

The Medical Library Association and Association of Academic Health Sciences Libraries support providing broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality. Our organizations maintain that open access facilitates scientific collaboration, strengthens biomedical research, accelerates innovation, and supports better patient care.

- Free and Easy to Access and Reuse.

We recommend that

- Establishing an embargo period (no longer than 12 months) that gives researchers the time needed to go through what is sometimes a long peer review process. During this period, manuscript reviewers should be able to access the data.

- Adding guidelines that clarify the intention of: “With the broadest possible terms of reuse,” in order to ensure privacy and security of research subjects; and
  - Requesting (or having) repositories to implement an interface that allows users to interact/preview level 3 data (e.g., data that has been de-identified and normalized). This will enhance discoverability without the need to download the data. Examples of this are cBioPortal, GEO2R, Allen Brain Atlas.
- Common Formats.

MLA and AAHSL recommend that data repositories conform to the FAIR (Findable, Accessible, Interoperable, Reusable) principles, thus admitting datasets in proprietary formats would impinge on its interoperability. Open, documented formats would ensure that interoperability is maintained whether the data is accessed directly by humans or programmatically by some future system. Allowance for proprietary formats, even if the necessary computing environment is co-submitted, is not desirable.

<https://www.force11.org/group/fairgroup/fairprinciples>

- Provenance.

We recommend clarifying that the term provenance includes the origin of the dataset, not only the changes to the dataset.

### **Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

- Fidelity to Consent.

We recommend that repositories storing human data provide access to de-identified data for Level 3 which typically represents aggregated, normalized, and/or segmented data. This will enhance discoverability by allowing researchers to test and generate new hypotheses as well as to validate their results without the need to request access for controlled access data.

- Request Review.

The funding agency must provide clear guidance/standards as required by regulatory agencies in order for the repositories to comply with the fidelity to consent.



**Vivli response to the White House Office for Science and Technology Policy request for public comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Results from Federally Funded Research.**

**Sector scientific focus:** This response is primarily focused on desirable characteristics of repositories as they relate to the sharing clinical trial data at the participant-level.

Vivli ([Vivli.org](https://vivli.org)) is a non-profit organization founded in 2018 that manages a clinical trial data sharing platform. We provide a single point of search and request to participant-level data from more than 4,900 trials representing 2.7 million participants from 111 countries. Our comments are restricted to clinical trial data sharing, which we believe has the broadest and most immediate impact on advancing human health by accelerating new findings through data sharing and re-use. Clinical trial data sharing respects trial participants' assumption of personal risk to contribute to science by maximizing the value of their contributions. We share data from more than 20 data contributors including members from the pharmaceutical industry, academic organizations and non-profit foundations. The Vivli data sharing platform operates on the FAIR principles for data sharing and strongly supports the adoption of FAIR as a guiding set of desirable underlying principles adopted by federal research repositories.

We welcome the leadership of the Subcommittee on Open Science (SOS) of the National Science and Technology Council's Committee on Science has shown by developing a draft guidance for characteristics of repositories for managing and sharing data from federally funded research. The US federal government, particularly the NIH, is the world's largest funder of clinical research and we hope this will influence the NIH's data sharing policy as well as influence other federal agencies who fund clinical research.

The federal government can take the lead to significantly impact data sharing by updating and aligning its data sharing policy with contemporary best practices. The overall draft guidance is a step in the right direction, although it places much of the responsibility and burden for safe and responsible data sharing on repositories, instead of on the researchers who have contributed the data. This approach is not reflective of the current realities of who has the most control, or funding, to meet this guidance as much of the levers of control are in the hands of funders, such as the federal government, not repositories, to require researchers to share data in the most responsible fashion. Repositories play the role of enabling best practice, not necessarily mandating how data must be shared.

We would encourage the committee to also reference [The National Academy of Medicine Report](#) on sharing of clinical trial data. This consensus document highlights best practices in sharing of clinical research data and the responsibility of all parties involved.

**Vivli limits our comments on “Desirable Characteristics for All Data Repositories” (Section I) to our area of scientific and technical interest – clinical research data**

We support the desirable characteristics denoted as universal for all data repositories

Persistent unique identifiers, long term sustainability, requirements for metadata, curation & metadata, access, ease of access, re-use, security, privacy, common format and provenance. Specific comments below.

***Curation & Quality Assurance: Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.***

The obligations to curate and provide quality assurance of data and metadata are not always optimally met by a repository. Repositories can provide guidance and best practices for how metadata and data curation should be done, but are not always positioned to mandate this approach, which may also be performed done by the either funders or researchers.

***Sustainability: Long term sustainability is important for any repository. Oftentimes government platforms may be funded for a finite period of time through a grant-mechanism. The renewal mechanism is unknown or non-existent at the time of award. Public-private partnerships or leveraging other models could enable long term sustainability of repositories.***

***Free & Easy to Access and Reuse: Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.***

The current guidance leaves open the timeframe for when data would be made available to users at the discretion of researchers, other than it should be timely. We recommend that federally funded clinical trials require reporting of completed clinical trial datasets to an approved repository within a reasonable time period. [The National Academy of Medicine Report](#) has suggested a practical timeframe of 18 months post-trial completion.

***Common Format: Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.***

A repository does not have control over how data is collected or how it is provided to them. While repositories may recommend minimum standards for how data should be shared, funders are in a stronger position to designate the standards that should be employed in data collection. For clinical trials data, these are collected years prior to deposit and therefore the format is often pre-determined 5 or more years before a repository or platform has jurisdiction over the data. Clearly interoperability is facilitated by standards. Acceptable standards continue to evolve, require considerable cooperation among multiple stakeholder groups and has proven to be extremely challenging. Therefore, Vivli has recommended but not mandated a common format for data contributed to our platform.

Vivli comments on “Additional Considerations for Repositories Storing Human Data (even if de-identified)” (Section II)

The responses below assume this section refers to clinical trial datasets. We comment on specific items as per highlighted below.

***Fidelity to Consent: Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).***

We at Vivli have taken the approach that those who contribute data to a platform or repository (and therefore are responsible for the acquisition of the data) ensures that the data use is consistent with the original consent. Data repositories are typically one step removed from data acquisition process and therefore it may be difficult for repositories to have access to the contents of the consent form to restrict access to appropriate uses consistent with the original consent.

***Download Control: Controls and audits access to and download of datasets.***

Download controls are laudable; however, there are many other legal and technical restrictions that are especially useful to restrict and control the access of sensitive human clinical data. Increasingly, cloud technology has been utilized to share data securely through managing access and deploying a common set of analytical tools to multiple researchers (for example, recently the National Academies hosted a workshop to explore opportunities to further research using cloud platform technology<sup>1</sup>). This utilization of a research environment or “secure sandbox” through which data is retrieved allows the appropriate balancing of privacy and a managed access approach.

***Request Review: Has an established data access review or oversight group responsible for reviewing data use requests.***

Often repositories can offer this as a benefit for those who choose to deposit their data, but at times repositories may partner with those that deposit the data or funders who mandate how access is granted.

In conclusion, while the guidance provides an overview for federal agencies, more should be done to outline the responsibilities of the data repository, the data contributor and the funder of the research to work together. In many cases, the data repository is playing an enabling role and has a limited ability to mandate that a data contributor must meet its requirements. This is a role better played by the funder, which in this case is the federal government. We urge the OSFT to refine its language around the responsibilities of a repository, data contributor and a funder in its next draft.

---

<sup>1</sup> [http://www.nationalacademies.org/hmd/Reports/2020/neuroscience-data-in-the-cloud-pw.aspx?utm\\_source=HMD+Email+List&utm\\_campaign=bfab7d0320-ncpf-pw-Dec1\\_COPY\\_01&utm\\_medium=email&utm\\_term=0\\_211686812e-bfab7d0320-&mc\\_cid=bfab7d0320&mc\\_eid=%5bUNIQID%5d](http://www.nationalacademies.org/hmd/Reports/2020/neuroscience-data-in-the-cloud-pw.aspx?utm_source=HMD+Email+List&utm_campaign=bfab7d0320-ncpf-pw-Dec1_COPY_01&utm_medium=email&utm_term=0_211686812e-bfab7d0320-&mc_cid=bfab7d0320&mc_eid=%5bUNIQID%5d)

From: David Giarretta [david@giarretta.org](mailto:david@giarretta.org)

To: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov).

Subject: RFC Response: Desirable Repository Characteristics

This response is submitted by David Giarretta, Director of the Primary Trustworthy Digital Repository Authorisation Body Ltd (PTAB, [www.iso16363.org](http://www.iso16363.org)) and Chair of the CCSDS Data Archive and Interoperability Working Group which wrote OAIS (ISO 14721) and ISO16363.

Primary discipline: Digital Preservation and previously Physical Sciences.

This response is informed by a number of caveats about repositories and digital preservation, especially when applied to data. These are caveats which are usually not stated or are simply ignored. However remembering that what is at stake is information which has been collected at the cost of many millions, or even billions, of dollars (euros/pounds etc.), many research careers and which potentially could benefit mankind's future wellbeing, it is worth stating the caveats clearly here.

- 1) It is easy to make rather vague claims about how good a repository is in terms of preservation.
- 2) Repositories almost certainly will have finite lifetimes, even if embedded within a longer-lived organization.
- 3) Data is fundamentally different from the digital equivalent of printed documents. As long as one can print the paper version or display the latter, then it can reasonably be assumed that it can be understood by anyone who can read the language. The same cannot be said for data. Even a simple spreadsheet can be unusable if one does not know the meaning and units of the columns. A single element or even a single bit can mean anything even if a standard format is used, if the researcher feels inhibited by the "normal" use of that format. This has implications for the next point.
- 4) Data is different from Gold. Gold is valuable because it is rare and does not rust i.e. does not easily combine with other elements. Data on the other hand is valuable because it is increasingly plentiful and becomes massively more valuable when combined with other data.

To respond to caveat (1) the claims of the repository must be testable, and must be tested. This is a fundamental concept of OAIS and ISO 16363. The update of OAIS makes this even clearer. A related point is that the use of the word "metadata" often, perhaps one should say usually, causes confusion and misunderstandings because it is so ill-defined. The word is useful as a collective term if used sparingly but if that is the only term available then one cannot ask whether one has enough of the different types (whatever they are) of "metadata". OAIS introduces a much finer taxonomy of terminology covering the information needed for preservation.

Caveat (2) implies that the repository should collect together all the information that is needed for preservation and able to be handed over to a successor repository (or repository system

within the same organisation). This collection of information is what OAIS terms the Archival Information Package (AIP). A point which can be overlooked is that repository systems are likely, in my experience, to have undocumented knowledge held by staff or embedded in software, and so is impossible to hand over unless it is made explicit.

Caveat (3) requires that the process of creating or collecting the data, which may involve not just a single individual but many, separate teams of people and many stages of work over many years, is accompanied by the creation of enough of the different types of “metadata” which should be collected about the data in order to ensure that it can be re-used and preserved. This process of collecting information should be part of the data management plan. The terms (re-)use and preservation are closely connected terms in OAIS and ISO 16363. In particular enough Representation Information, a specific type of “metadata” required to be able to understand the data, must be collected.

This brings us to caveat (4) which requires that the Representation Information a repository has can be supplemented by other sources and can be linked to different disciplines, for humans and ideally or software, in order to facilitate interoperability.

While, broadly speaking, the repository characteristics listed are consistent with ISO 16363, some specific comments about repository characteristics are:

[C] one should beware of using the term “metadata” other than as a collective term. It may be clearer to use the phrase “metadata, as described in OAIS”.

[J] need to take caveat (3) into account if possible, otherwise appropriate Representation Information should be provided with the data.

[K] There is much more to Provenance that is covered in the current text. For example, to create a dataset one may combine, using complex algorithms, with data from many sources. These are important items of Provenance if one is to understand from where a data set comes. Some of this may be included in the publications about the data.

My overall message is that “the devil is in the detail” so one needs the finer grained taxonomy provided by OAIS rather than simply using the word “metadata”. One also needs the level of detailed inspection provided by ISO 16363 audit certification, which is backed up by the ISO processes and procedures on which we all depend in many aspects of our lives.

As to the ability of existing repositories to meet the desired characteristics, one can state that not all the repositories will have the resources and skills needed to preserve data properly. Even those that do have the skills and resources may not wish to plan for their own demise, and are, in our experience, not perfect. However, ISO 16363 audit and certification does not demand perfection but instead ensures that the repository does not have preservation threatening flaws, supplemented by a process of continuous improvement.



# Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing

*Administrative Data Accelerator, Pennsylvania State University*

The Administrative Data Accelerator of the Pennsylvania State University collaborates with researchers, data analysts, policymakers, practitioners, and private industry to acquire and securely store and utilize administrative data. Our team consists of researchers, data analysts, and project managers in the interdisciplinary field of social science working to inform policy through evidence-based research. Thus, the comments on the draft of desirable characteristics of repositories from the Data Accelerator reflect both the supply and demand side of data repositories.

To enhance data accessibility:

- Metadata catalog with unified documentation and format would make the data searching process more efficient and effective.
- Guidance on data sharing including a unified format of documentation, codebook, and data format across data should be provided to researchers. The guidance should include the standard format of supporting documents (including the length and scope of description), codebook, and data files. Often there are variations across data on the length and depth of documentation. Each field may have a different standard in file format (e.g. economics field encourages to submit data in dta format), but if there is not, guidance should suggest acceptable file format that can maximize the accessibility and reusability of the data.
- List of data anomalies / artifacts of the data, especially administrative data.
- List of research using each data would inform users about the validation and usefulness of data. List of publications or ongoing research using the data to help user's understanding of the utilization of data can be provided along with documentation and codebook. Publication repository linked to data repository would be preferred.
- Webinars for users and/or uploaders would be useful to help researchers both in demand and supply side.
- List of identifiers restricted but available upon request: Some identifiers to merge with other datasets may available upon request/application. Providing a list of identifiers restricted but available upon request/application would enhance data reusability.

To enhance data sustainability:

- Data management and sharing plan (DMP): some research funders (mostly in the UK) requires DMP along with format and checklist for researchers. Unified format of a management plan and checklist provided to researchers sharing their data to the repository would make the data storing process more efficient.
- Point of contact for each data (or upon request) for future questions and communication on the data usage

To enhance efficiency in repository management:

- Tracking application process of restricted data usage: Often the process of application to restricted data usage is a black box and not able to track the process. Making the process of applications trackable by adding application status (such as the application received – under review – revision requested – accepted), and estimated time for each stage would enhance both understanding from users and efficiency in data management.

## Comments to OSTP Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

Point F says: F. *Free & Easy to Access and Reuse*: Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

This point is the only one that carries the connotation of timeliness, which is very important because:

- Data quickly becomes “stale” when it is not publicly released in a timely fashion. The reason why they quickly become stale is that

1. A lot of cutting-edge research is performed by scientists in non-permanent positions, including collecting and curating the data from federally funded research.
2. If data is released too long after it has been used, these scientists will likely have moved on to other jobs and possibly have lost interest in the original product. As a consequence it might become hard to correct mistakes or refine the data as suggested in point D.
3. New (typically still proprietary) data is being analyzed and the older data loses its appeal. This leaves scientific potential untapped and prohibits time sensitive analysis common to observational fields.

- When data becomes stale it ceases to be scientifically useful which means that the effort of making it accessible does not come to fruition.

So timeliness in the release should be an important guiding principle and suggest to stress this by :

- moving point F to a higher position in the list (suggest D)
- rephrasing point F as: *Timely, Free & Easy to Access and Reuse*: Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

A good example of a federally funded project not releasing data in a timely manner is the NSF-supported LIGO. Their next data release of 6 months of data is planned for 1.5 years after that data was produced and 1 year after the dissemination of scientific results from that data.

## **RFC Response: Desirable Repository Characteristics**

Janis Geary, Arizona State University, Social Sciences, Researcher

Mary Majumder, Baylor College of Medicine, Ethics and Health Policy, Researcher

Christi Guerrini, Baylor College of Medicine, Ethics and Health Policy, Researcher

Jill Oliver Robinson, Baylor College of Medicine, Ethics and Health Policy, Researcher

Adrian Thorogood, Centre of Genomics and Policy, McGill University, Researcher

Robert Cook-Deegan, Consortium for Science, Policy & Outcomes, Arizona State University, Public Policy Studies, Researcher

1. *Comments on characteristics included in Section I (Desirable Characteristics for All Data Repositories):*

Characteristic E (Access) indicates that access must be “consistent with legal and ethical limits required to maintain privacy and confidentiality”. We suggest that this definition should include language that acknowledges and respects sovereignty of Tribal Nations over their own data. This limit extends beyond privacy and confidentiality and should be made explicitly. Additionally, there should not be a presumption that the only data that qualify for heightened restrictions on access are human subjects data. Some datasets might be sensitive for reasons other than their implications for individual privacy. For example, datasets that include information about the location of endangered species are sensitive because they might be used by poachers to harm those species. Information about Indigenous or vulnerable groups might be used to make inferences about disease, environmental conditions, socio-economic status, or stigmatizing conditions. Finally, we urge repositories to consider allowing access of data by investigators who are not affiliated with traditional scientific institutions and other citizen scientists, especially when the underlying research describes itself as citizen science, when doing so is unlikely to risk privacy or other harms to individuals or communities.

Example: Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality, **Tribal data sovereignty, and protection of other sensitive data.**

Characteristic G (Reuse) should include tracking who has been granted access to controlled data, and how it has been used. This could help verify that data reuse will abide by relevant restrictions. Tracking should not be simply enabled through assignment of adequate metadata, but users of data should be required to submit their publications and similar research outputs back to the repository to be linked to the original dataset record.

Example: **Requires** tracking of **data access** and reuse (e.g., through assignment of adequate metadata and PUID).

Characteristic F (Free & Easy to Access and Reuse) should include making the documentation regarding use guidelines easily and freely accessible in a timely manner along with the data.

Example: Makes datasets and their metadata **and clear use guidelines** accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

*2. Characteristics that should be included in Section I:*

Violations: The potential for data misuse is not limited to human data. All repositories should have publicly accessible plans in place to describe what constitutes misuse and include sanctions.

*3. Comments on characteristics included in Section II (Additional Considerations for Repositories Storing Human Data (Even if De-Identified)):*

Characteristic C (Privacy):

Example: Implements and provides **public** documentation of security techniques appropriate for human subjects' data to protect from inappropriate access, **and provisions for filing notice of privacy concerns to an independent oversight body.**

Characteristic D (Plan for Breach):

Example: Has security measures that include a **publicly available** data breach response plan, **which includes an external independent monitoring of compliance that is not controlled by the breached party.**

Characteristic H (Violations):

Example: Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository **that include possible sanctions imposed by a credible authority that monitors compliance.**

Characteristic I (Request Review): Note: Needs an independent “lifeguard” that is not controlled by the data-hosting institution.

Has an established data access review or oversight group responsible for reviewing data use requests **with publicly accessible terms of reference and membership that includes appropriate representation from human data contributors.**

*4. Characteristics that should be included in Section II:*

Transparency: While we have suggested several ways to improve transparency within other characteristics, transparency is a core principle that should be considered within all aspects of human data governance (All of Us Research Program 2015; Knoppers 2014; Cook-Deegan, Majumder, and McGuire 2019).

*5. Other characteristics which should be included to address the management and sharing of unique data types.*

It is unclear if Tribal Nations have been targeted for consultation. Ideally, Tribal Nations should be supported in developing their own repositories. Until this is feasible, they should be supported in developing their own list of characteristics for repositories, or consulted to ensure that current guidelines for repositories do not unwittingly hinder Tribal data sovereignty. For guidance, a group of Indigenous scholars has developed Indigenous data governance principles intended to work in parallel with the FAIR principles, called the CARE principles (Research Data Alliance International Indigenous Data Sovereignty Interest Group 2019).

Many of the characteristics in Section II are relevant to all repository types, as concerns around protecting sensitive data and transparent governance are not limited to human subjects data. Restricted Use Compliant, Plan for Breach, Download Control, Clear Use Guidance, Retention Guidelines, Violations, and Request Review all could be included for non-human subject data that is sensitive for reasons beyond individual privacy and consent.

*6. Consistency of the desirable characteristics with widely used criteria or certification schemes for certifying data repositories*

The World Data System guidelines have organizational requirements to ensure there is adequate funding and staff to enable the organization to carry out its mission. However, organizational requirements are missing from the suggested characteristics.

## References

- All of Us Research Program. 2015. "Precision Medicine Initiative: Privacy and Trust Principles." 2015. <https://allofus.nih.gov/protecting-data-and-privacy/precision-medicine-initiative-privacy-and-trust-principles#precision-medicine-initiative-privacy-and-trust-principles-2>.
- Cook-Deegan, Robert, Mary A. Majumder, and Amy L. McGuire. 2019. "Introduction: Sharing Data in a Medical Information Commons." *The Journal of Law, Medicine & Ethics* 47 (1): 7–11. <https://doi.org/10.1177/1073110519840479>.
- Knoppers, Bartha Maria. 2014. "Framework for Responsible Sharing of Genomic and Health-Related Data." *The HUGO Journal* 8 (1): 3. <https://doi.org/10.1186/s11568-014-0003-1>.
- Research Data Alliance International Indigenous Data Sovereignty Interest Group. 2019. "CARE Principles for Indigenous Data Governance. The Global Indigenous Data Alliance." [GIDA-global.org](http://GIDA-global.org).

**Response from NSF's Optical Astronomy Lab to "Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research"**

**Author:** Adam S. Bolton, Ph.D. (Director, Community Science and Data Center, NSF's Optical Astronomy Lab)

**Domain:** Physical Sciences (Astronomy & Astrophysics)

***NSF's Optical Astronomy Lab<sup>1</sup> runs an international network of state-of-the-art astronomical telescopes that take high-definition digital images of billions of stars and galaxies throughout the Universe. The Lab's data repositories make these images available to all qualified scientists doing forefront research that advances America's leadership in the study of the cosmos. These and all other astronomical images are irreplaceable. They have long-term scientific legacy value for studying the changing nature of diverse astronomical objects over days, months, years, and even centuries.***

NSF's Optical Astronomy Lab is the US national center for ground-based observational nighttime astronomy. The Lab operates multiple optical and infrared research telescopes with apertures ranging from 1 to 8 meters in diameter in Arizona, Hawaii, and Chile. These telescopes provide merit-based open access to all researchers in the astronomical community without regard to institutional or collaborative affiliation.

All experimental research with the Lab's telescopes uses digital detectors that convert optical and infrared light from astronomical objects into digitized imaging, spectroscopy, and catalogs. Hence, the Lab operates computerized data archives and data services as an integral part of its scientific mission, as do many other modern astronomy research centers. These data repositories serve a broad range of needs:

- Facilitating transfer of data from remote mountaintop sites at which observations are conducted to centralized locations for storage and processing
- Implementation of redundant, geographically-distributed backup and disaster-recovery systems
- Integration with computing capabilities that transform raw data and calibrations into science-ready data products
- Provision of data access for research investigators to enable the primary science for which the observations were conducted
- Hosting of derived data products generated by research teams
- Enabling data discovery and open access for other investigators after the expiration of original data proprietary periods, to support reproducibility of published results as well as new scientific applications of archival data (*see figure at end of this document*)

---

<sup>1</sup> NSF's National Optical-Infrared Astronomy Research Laboratory (full name) is operated by the Association of Universities for Research in Astronomy under a cooperative agreement with the US National Science Foundation.



- Deploying high-level tools for query, exploration, visualization, and analysis to maximize the scientific return from open-access data holdings
- Supporting data-intensive scientific analyses requiring the combination of multiple data sets in a single archive or across multiple archives
- Long-term curation of astronomical data, along with associated metadata, software, documentation, and expert knowledge

NSF's Optical Astronomy Lab was launched on 01 October 2019 through the combination of three centers: the National Optical Astronomy Observatory (NOAO), the Gemini Observatory, and the Vera Rubin Observatory (currently under construction). This restructuring will lead to increased integration and coordination between the Lab's data repositories over the next several years.

In the early 2020's, the Lab will begin to operate the Vera Rubin Observatory's 10-year Legacy Survey of Space and Time (LSST) using the Simonyi Survey Telescope on Cerro Pachón in Chile. The LSST will deliver an unprecedented petabyte-scale astronomical database covering the entire sky visible from its southern-hemisphere site, which will be accessible in real time to all US astronomers and designated international participants. The Rubin Observatory's LSST promises to transform astrophysics, and at the same time to radically democratize participation in forefront astronomy research. *All science with the Rubin Observatory will be done via data repositories and real-time data streaming services.*

### **The Proposed use and application of the desirable characteristics**

In the context of ground-based astronomy, the desirable characteristics are especially appropriate for the application of "Developing Federal agency repositories to store data resulting from Federally funded research". Capable modern data repositories are essential to enabling NSF's mission to promote the progress of science in the era of data-intensive astronomy.

### **The appropriateness of the "Desirable Characteristics for All Data Repositories"**

All of the draft characteristics are well aligned with best practices in astronomy, and with the goals of NSF's Optical Astronomy Lab for the data repositories that it operates.

The chief concern, particularly in the era of petabyte-scale astronomical data sets, is in identifying the resources necessary to fully realize all these characteristics for all data sets of scientific value. In a funding-constrained environment, organizations such as the Lab must set priorities for data sets and data-repository characteristics based on the principles of maximizing scientific return and broad-based research participation per dollar. Likewise, scientific investigators must be incentivized and resourced to expend the additional effort necessary to make their data products understandable and usable by other teams.

## **Considerations for any other characteristics which should be included to address the management and sharing of unique data types**

*Support for active experimental science:* As outlined above, astronomical repositories are not just passive storage locations for data resulting from completed research, they are also critical for providing active support to ongoing experimental research through data transfer, staging, processing, and access. Data repositories in astronomy and other domains must accommodate the experimental support and integration requirements of each discipline.

*Open-source technology:* An additional desirable characteristic for all data repositories is implementation using open-source technologies, with supporting documentation. If the software and other intellectual property upon which a data repository is built is not fully open-source, then hosted data that are public in principle can be rendered proprietary in practice through exclusive control of the associated storage and interface. Repositories based on open-source technology are furthermore preferred for their replicability and adaptability.

*Bringing the analysis to the data:* Major astronomy data sets are rapidly becoming too large for individual astronomers to download and analyze with their own local resources. In the era of petabyte- and exabyte-scale data, it is crucial for data repositories to be co-located with computing resources that can provide the processing and analysis power needed to obtain scientific results. To achieve this goal, astronomy data repositories should be empowered to leverage major Federal investment in fundamental cyberinfrastructure for networking, storage, and computing in both academic and commercial environments.

## **The ability of existing repositories to meet the desirable characteristics**

Many current data repositories in astronomy meet many of the draft desirable characteristics. Here we address the repositories of NSF's Optical Astronomy Lab, which include the Science Data Archive, the Astro Data Lab, the Gemini Observatory Archive, and the (under development) LSST Science Platform.

*Persistent Unique Identifiers; Reuse:* Data repositories currently operated by the Lab currently address this characteristic through the use of a unique "Proposal ID" for the telescope observing proposal with which individual data sets are associated. Published online instructions specify that researchers using archival data should acknowledge data reuse via these Proposal IDs, which enables tracking and bibliometric analysis. Future planned developments include the issuance of DOIs through the Lab to allow for more fine-grained and customized tagging of data sets.

*Long-term sustainability:* As the Federally Funded Research and Development Center (FFRDC) for ground-based nighttime astronomy, the Lab has a long-term mission and purpose that includes stewardship of astronomy research data. Predecessor organizations of the Lab have been in continuous operation since 1958, and have been engaged at the forefront of archiving of digital astronomy data for over 25 years (e.g., Seaman et al, 1994, ASPC, 61, 119).

*Metadata:* Data repositories at the Lab extract standard observational-astronomy metadata from raw data file headers, validate and standardize it as appropriate for each telescope and instrument, and ingest it into online metadata databases (an “extract-transform-load” pattern). External metadata queries are supported through interactive web interfaces as well as application programming interfaces (APIs), including standardized “Virtual Observatory” protocols such as Table Access (TAP) and Simple Image Access (SIA). Future development is planned to support other standardized astronomical data-access and data-discovery protocols such as Simple Spectrum Access (SSP) and Common Archive Observation Model (CAOM).

*Curation and Quality Assurance:* The Lab’s data repositories are developed, operated, and maintained by integrated teams of scientists and software engineers with deep expertise in the data sets that they serve. This expertise is rooted in substantive scientific and technical collaboration with multiple community-based research teams using Lab telescopes to obtain new data and generate new data products.

*Access; Free & Easy to Access and Reuse:* The Lab’s data repositories are fully open without cost to the global astronomical community. Data obtained at Lab telescopes are world-discoverable as soon as they are ingested into the repository, and are automatically made available for download as soon as the proprietary period (typically 12-18 months) of the original investigators expires.<sup>2</sup> Data repositories at the Lab are furthermore supplemented by rich capabilities for high-level data discovery, exploration, visualization, and analysis.

*Secure:* Data access controls are automatically enforced by user authentication and authorization services associated with data repositories at the Lab. These controls are supplemented by generally accepted ethical principles in astronomy that hold unauthorized data access to be a form of research misconduct.

*Privacy:* Data repositories at the Lab are operated within a broader comprehensive organizational cybersecurity framework.

*Common Format:* The Lab’s data repositories serve data in Flexible Image Transport System (FITS) format, the most open and widely accepted data-file standard within astronomy worldwide.

*Provenance:* Provenance tracking for data hosted by the Lab’s repositories is based on maintaining records of (1) the telescopes, instruments, and observing programs that delivered the raw observational data, and (2) the algorithms and software systems associated with creation of higher-level data products based on the raw data. Multiple successive changes to data sets are generally not supported.

---

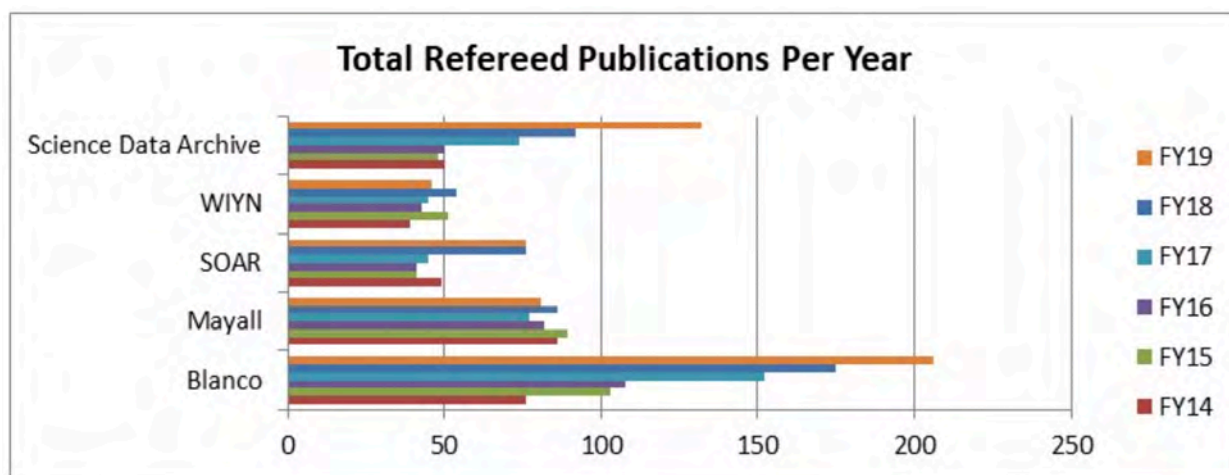
<sup>2</sup> Rubin Observatory is still planning for how to make its immense holdings of publicly sharable data available to the worldwide community after a two year proprietary period during which the data will be available only to the US community and designated international partners.

### Consistency of the desirable characteristics with widely used criteria or certification schemes

In areas where they address similar scope, the desirable characteristics are consistent with the recommendations of the International Virtual Observatory Alliance (<http://www.ivoa.net>), the main international organization for astronomical data standards.

### Any other topic which may be relevant for Federal agencies to consider in developing desirable characteristics for data repositories

Software and computing technologies relevant to scientific data repositories are flourishing and evolving rapidly. To ensure that these technologies benefit the largest possible community of researchers, Federal agencies should consider the importance of training and workforce development for both the operators and the users of scientific data repositories.



*Total peer-reviewed scientific publications per year from each of NSF's Optical Astronomy Lab 4-meter class telescopes (WIYN, SOAR, Mayall, and Blanco), along with total peer-reviewed scientific publications based entirely on analysis of archival data. This figure illustrates how data repositories can magnify the scientific productivity and impact of astronomical research telescopes.*

**OFFICE OF SCIENCE AND TECHNOLOGY POLICY (OSTP) REQUEST  
FOR PUBLIC COMMENT**

**DRAFT DESIRABLE CHARACTERISTICS OF REPOSITORIES FOR  
MANAGING AND SHARING DATA RESULTING FROM FEDERALLY  
FUNDED RESEARCH**

**Name:** Rajni Samavedam, MPH, Principal/Director

**Organizational Affiliation:** Booz Allen Hamilton, Inc.

**Role:** Institutional Official

**Primary Scientific Discipline:** Multiple domains of life sciences

## **BOOZ ALLEN’S RESPONSE TO THE DRAFT DESIRABLE CHARACTERISTICS OF REPOSITORIES FOR MANAGING AND SHARING DATA RESULTING FROM FEDERALLY FUNDED RESEARCH**

### **I. The proposed use and application of the desirable characteristics (as described in the “Background” section)**

Booz Allen applauds the efforts of OSTP’s Subcommittee on Open Science (SOS) to improve access to data generated from federally funded research and development (R&D) by seeking to establish desirable characteristics for data repositories. Establishing – perhaps even prescribing – specific characteristics and criteria for both generalist and specialized data repositories that are developed and maintained using federal funds is not just timely but long overdue. With the ever tightening budget for biomedical research, having to do more with less, and the sheer volume of research data generated on a daily basis from federally funded studies, it is imperative that these data be preserved and shared through data repositories for broader use – this is a critical first step in maximizing the value of already collected research data.

Towards the goal of advancing open science through repositories and maximizing the value from collected data, we recommend the following to the intended use of the characteristics:

1. SOS has indicated that the intended use of these proposed (and additional) characteristics of data repositories is for use by federal agencies to primarily *inform* their respective stakeholders, including federally funded investigators, repository developers, and data users. However, we would advocate that OSTP consider *proffering these characteristics more as a requirement than merely as ‘for your information’ – if federal funds are used to develop the data repository*. This would align and actuate in tangible ways OSTP’s memorandum of “Increasing Access to the Results of Federally Funded Scientific Research” that has been in place since 2013 and calls for improved access to data to advance open science. We understand that this is a cultural shift and it requires bold action from OSTP, but federally funded repositories are essentially tax payer funded repositories and members of the public are key stakeholders of such repositories.
2. Making the repository characteristics as required elements for federally funded repositories would also require that the SOS consider questions such as:
  - How to enforce these characteristics across federal agencies?
  - How to measure compliance to the required characteristics?
  - What are some key performance indicators by which repositories can evaluate that they are meeting the required elements? We propose that such metrics be provided in advance of the development of repositories.

3. We propose that these characteristics be mapped to the FAIR (Findable, Accessible, Interoperable, and Reusable) principles so that all agencies work off of the same fundamental principles for sharing data. This will ensure that federally funded repositories are not siloed and can move towards an ecosystem where cross-repository data integration and analysis can be done. For example, one can envision the use and propagation of biomedical research data and results through a pipeline that flows across various agencies: biomedical research data (NIH) → trial data (FDA) → health data/EHR (VA/CMS/etc.) → public use/improved public health. Of course, such interoperability of data and systems will require the use of common data elements and controlled vocabularies, and we propose that the federal government mandate the use of these as much as possible.

## II. The appropriateness of the “Desirable Characteristics for All Data Repositories” for data repositories (Section I) that would store and provide access to data resulting from Federally-supported research, considering:

### 1. Characteristics that are included:

- a. **Metadata:** Data are only as good as the accompanying metadata. Having proper and accurate metadata associated with the datasets is essential for making data FAIR and for meaningful use. We advocate that federal agencies require that data deposited to repositories be packaged with appropriate essential metadata that are based on standards so that data can be harmonized and integrated for analysis and used for more advanced data science-based approaches such as machine learning (ML), predictive modeling, and other Artificial Intelligence (AI) applications. While both research funders and investigators understand the value of metadata, development and use of metadata standards have been severely lagging. This is one area where federal agencies can hold a critical role in establishing federal-wide data standards so that data are collected uniformly and can be pooled, integrated, and analyzed effectively.

Part of having good metadata is also requiring that data submitters provide associated study documents such as data dictionaries and study protocols so that secondary users can understand and use the data meaningfully. How studies are conducted varies widely and cannot be controlled within or across agencies; however, certain essential elements such as having standard templates for protocols and tools for developing data dictionaries, especially for clinical trials, can go a long way for meaningful reuse of the data. With accompanying documentation, proprietary instruments need to be protected and handled appropriately.

- b. **Secure:** We recommend that security of the data is as important as FAIR data principles, and should not be an afterthought, especially as precision medicine

moves into an era where certain data will be clinically actionable and thus cannot be deidentified. Therefore, security need to be considered even before data collection and sharing. To mitigate this risk, it may soon be feasible using techniques like homomorphic encryption to analyze data while it remains encrypted. Government should consider providing guidance for security of such non-deidentifiable data.

## 2. Additional characteristics that should be included:

Booz Allen proposes the following additional characteristics for all repositories:

- a. **Policy and governance:** Establishing appropriate repository policies and governance relating to the data that is submitted (who, when, and how) and requested (who and how) will ensure fairness and transparency. For data access, the repository will have to provide means to public access versus restricted access based on policy, level of de-identification, and other considerations. Each repository should also have a governing body that can efficiently oversee repository operations, data submissions and requests, and will ensure accountability of the repository.
- b. **Digital data:** Data deposited into repositories must be digitized or collected in a machine-readable manner so that it is consumable by and computable with analytic tools; PDF scans or unsearchable images must be avoided. Stored data should be at the individual-level; although summary/aggregate datasets are acceptable.
- c. **Training and guidelines for data deposition:** To ensure the deposition of high quality and reusable data, some form of Government-offered training on data management, best practices, and even publicly available computational resources must be developed and propagated – this would make data sharing via repositories more palatable, especially for low-funded researchers and those at smaller institutions. In addition, common training including notions of what constitutes good metadata and documentation would improve data structures and the use of metadata, which are critical for data harmonization.
- d. **Acknowledgement of original data contributor(s):** Requiring data users to acknowledge the original Principal Investigator/s that conducted the study and collected the data will incentivize and promote data sharing from federally funded research.
- e. **Application Programming Interfaces (APIs):** Government should strongly encourage or require well-documented APIs from the repository whenever possible to promote interoperability of data – a key element of the FAIR principles. These APIs must be documented so users can extract and analyze metadata from the repository.
- f. **Scalability:** To accommodate the growing volume of data, a repository must consider scalability as one of their design principles and consider using a cloud environment to host the data. To accommodate the increase in storage costs over



time, which becomes an issue once funding for data collection, analysis, and sharing ceases, archival tiers of storage provided by cloud service providers can be used that would substantially lower these costs. The archival tiers will depend on the estimated or demonstrated value to the community, expressed demand for continued access to the data sets, the amount of continuing costs, and projections for how much longer the data are likely to remain useful.

### III. Consistency of the desirable characteristics with widely used criteria or certification schemes for certifying data repositories

Booz Allen recognizes the growing field of repository certifications and the value of a certified repository to data submitters and data users. We propose additional measures to ensure consistency of the SOS proposed desirable characteristics with certification schemes:

- a. To be more consistent with CoreTrustSeal requirements, add documented workflows for archiving to the list of characteristics. This would ultimately help to ensure that the FAIR-ness of archived data is consistent across the repository and could prove to be cost effective for maintaining a repository in the long-term by promoting efficiency and avoiding ad hoc actions. Data contributors should be made aware of workflows, especially the safeguards and procedures in place for archiving human data.
- b. Consider criteria established by other entities, beyond certifications such as CoreTrustSeal and ISO16363, for desirable characteristics for all data repositories. For example, criteria based on policies established by Journals and Publishers, such as PLOS ONE and Scientific Data (Springer Nature), to ensure data associated with publications are shared via Recommended Repositories. Linkages between the archived data and associated publications should be provided to enable data users to examine hypotheses not tested by the original investigators.

## CONCLUSION

Booz Allen is pleased to have the opportunity to respond to the request for comments from OSTP on the desirable characteristics for managing and sharing data. We recognize that the greatest value of data is realized when it is shared – and shared in a manner that is reusable. Data reusability is fundamental to crowd sourced scientific discovery and clinical outcomes. OSTP’s goal to develop these (required) characteristics will establish robust and effective data repositories which can serve as the underpinnings to an open science platform.

March 6, 2020

Dr. Kelvin K. Droegemeier  
Director  
White House Office of Science and Technology Policy  
Executive Office of the President

Submitted electronically to: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

**RE: Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research**

Dear Dr. Droegemeier:

On behalf of The University of Texas MD Anderson Cancer Center, thank you for the opportunity to comment on the Office of Science and Technology Policy's (OSTP) Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research.

MD Anderson is one of the world's most respected centers focused on cancer patient care, research, education and prevention. Since 1944, more than 1.2 million patients have sought out MD Anderson's expertise. The institution pioneered a multidisciplinary approach to research-driven care and has more than 10,000 patients enrolled in 1,250-plus clinical trials exploring innovative treatments. The institution invested almost \$863 million in research in Fiscal Year 2018.

As the recipient of the most cancer research grants from the NCI, MD Anderson is invested and deeply interested in further partnering to facilitate approaches in support of new discoveries and accelerated advancements in cancer care while also guaranteeing the integrity of data generated through our care and research efforts. In response to the Request for Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research, we have provided below comprehensive responses to each of the posed questions.

In addition to these point-by-point responses, we have also suggested re-casting some of the proposed characteristics of the data repository. These are captured in the comments below. Furthermore, we would like to emphasize that we have identified two important characteristics that are not addressed explicitly in the report:

CARING      INTEGRITY      DISCOVERY

- (i) **Identity management:** This is a critical enabler of any platform for robust security and digital rights management. The proposed digital repository needs to support integration with emerging standards in identity management such as the W3C's Sovrin initiative. To be clear, the identities being managed should be sufficiently general to allow security and digital rights management. This would include individual investigators, patients, and even commercial entities.
- (ii) **Ontological support** for the data stored within or indexed through the repository is a critical requirement to maximize the value of the insights extracted. The adoption of the Ontology Web Language (OWL) or other standards is key characteristic that is lacking.

In closing, we would like to thank the OSTP for identifying the need to develop next generation approaches to facilitate science through the implementation of repositories that rigorously protect data governance and provenance. As an organization committed to accelerating our research and care mission through leadership in data governance, we look forward to further dialog and collaborations with the OSTP on this important topic. Please contact me at [DAJaffray@mdanderson.org](mailto:DAJaffray@mdanderson.org) if you have any questions.

Sincerely,



David Jaffray, PhD  
Sr. VP & Chief Technology and Digital Officer  
Professor, Radiation Physics and Imaging Physics  
University of Texas M.D. Anderson Cancer Center

## **I. Desirable Characteristics for All Data Repositories**

### *A. Persistent Unique Identifiers*

*Comment: The development of a robust DOI for identification and localization of data is a desirable characteristic. The use of a 'landing page' to support this effort is a limiting illustration of the nature of the DOI and PUID paradigm. A 'pointer' and 'services' paradigm should be adopted to generalize the concept further to make it more futureproof.*

### *B. Long-term sustainability*

*Comment: While the proposed elements are completely reasonable, substantial thought should be put into the curation process to assure the context of the data is maintained. It would be appropriate and potentially beneficial to architecturally separate the digital objects from the curation system. This would allow other systems to interact directly with both sub-systems independently: the context/curation and the data objects.*

### *C. Metadata*

*Comment: This is critically important. Metadata is becoming as critical as the data itself. The maintenance of rich metadata serves to increase the value of the data by allowing it to be placed in context. A standard schema is ideal but must be based on graph-based data recording systems.*

### *D. Curation & Quality Assurance*

*Comment: The curation and quality assurance capabilities are also critical. That said, they should not be divergent from the methods used to manage metadata. A parallel architecture to store and estimate the quality of the data either manually, or preferably, automatically (possibly via arguments of provenance) is critical. A means to store quality scores and revise these scores without over-writing previous estimates should be supported. This will allow the critically important capability to perform retrospective evaluation of data source 'value' based on future derived benefit.*

### *E. Access*

*Comment: While this is an attractive characteristic, it is far too vague and theoretical to be translated to robust implementation. This characteristic should be re-cast to be focused on the currently missing but foundational characteristic of having integrated 'data governance technology' to enable record-level data access and use rights machinery that assure accessibility is reflecting the rights of the data owners. Open access without rigorous tracking of consent will ultimately lead to unintentional breaches of privacy.*

### *F. Free & Easy to Access and Reuse*

*Comment: This characteristic should also be re-cast to "A data governance architecture that enables the rigorous management of data rights to enable the generation of publicly available datasets." As written, the characteristic would be similar to that found in countries with very weak privacy laws.*

### *G. Reuse*

*Comment: This characteristics should also be re-cast as “A data governance architecture that enables the provenance of derivative works to be traced to source data and its associated metadata. The methods of derivation should also be captured in UID/PUID forms to understand the degree of contribution from source data.”*

#### H. Secure

*Comment: The security frameworks proposed are reasonable. However, the security should operate as distinct from the underlying data governance framework that controls the rights of use of the data. Every effort should be made to separate ‘privacy’ into ‘security and consent’.*

#### I. Privacy

*Comment: As noted above, privacy should be decomposed into parallel but overlapping layers for security and consent (or ‘digital rights’). These two layers should also be managed from a common identity management framework (i.e. knowing who or what entity can have access from a security perspective and also knowing who has rights controls to govern use of the data for what purpose).*

#### J. Common Format

*Comment: Where possible data should be formatted in open standards. Ideally, data would not be ‘exported’ but rather accessed with appropriate tracking of rights. Possibly, the derivatives of these efforts would also be required to be returned to the same digital storage architecture.*

#### K. Provenance

*Comment: This definition is very limited. A more complete perspective on provenance and its linkage to the underlying data governance framework (see comments above) needs to be developed.*

## **II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

#### A. Fidelity to Consent

*Comment: In general, creating a separate set of characteristics for human data is not advised. It is better to build a single data governance framework that has broader capabilities for all data forms and apply the controls as appropriate.*

#### B. Restricted Use Compliant

*Comment: This is important and highlights the need for underlying architecture for data governance and provenance.*

#### C. Privacy

*Comment: See privacy comments above.*

#### D. Plan for Breach

*Comments: Breach mitigation subsystems such as integrated transparency and data access show-back approaches should be designed into the architecture from the beginning. These approaches are a superset of the requirements associated with audit trails etc.*

#### *E. Download Control*

*Comments: In general, data downloads should be minimized. Direct access with right controls possibly using tokens is a more rigorous approach and supports provenance maintenance for derivative works.*

#### *F. Clear Use Guidance*

*Comment: This is record keeping for the data governance architecture that include linkage to records of the consent agreements for those that have participated/contributed data (e.g. patient consent).*

*Encumbrances held against the data stored in the repository need to be respected (e.g. contractual rules associated with the funding body or joint funding or GDPR for European residents).*

#### *G. Retention Guidelines*

*Comments: These are supported by the combination of data governance and a data life-cycle management framework.*

#### *H. Violations*

*Comments: An overarching data governance framework needs to be established for support decision-making and exceptions. This should follow best practices including the establishment of the role of Chief Data Officer.*

#### *I. Request Review*

*Comment: This process should sit within the data governance framework. A major void in the characteristics that needs to be addressed is the creation of a system to manage the ontology of the data managed within the repository.*

## Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research

---

<b>Name of person filing comment</b>	Michael Hofmockel
<b>Institution</b>	Pacific Northwest National Laboratory
<b>Primary scientific discipline(s) in which they work</b>	Multi-program, U.S. Department of Energy, Office of Science national laboratory
<b>Role</b>	Research Computing, Data Capability Lead

---

PNNL is the nation's premier laboratory for scientific discovery in chemistry, earth sciences, and data analytics, and for technology solutions to the nation's toughest challenges in energy resiliency and national security. Based in Richland, Washington, PNNL is one of ten United States (U.S.) Department of Energy (DOE) Office of Science (SC) national laboratories.

PNNL greatly appreciates the authoring committee's efforts in creating this recommendation. The published version shows the committee invested significant thought and effort pulling together a foundation for building and assessing repositories of the future. The desired characteristics of data repositories are critical to the advancement of quality science and to the value the U.S. Government and science institutions achieve for research investments. Quality repositories will enable production of higher quality science and greater innovation in research.

### I. Desirable Characteristics for All Data Repositories

#### I.B. *Long-term sustainability*

1. Deletion is a pragmatic decision to assist financial stability that must be allowed when datasets no longer have value and funding is limited. A transparent decision-making process for assessing when data should be deleted must be available. However, provenance should never be deleted.
2. Institutional commitment is required for sustainability, but institutions need support in order to faithfully make these commitments.

#### I.C. *Metadata*

1. Repositories might need to allow incomplete uploads of datasets that do not yet have the minimum set of required annotations to be considered "complete." When this is allowed those incomplete datasets need to be clearly identified as a quality metric.
2. Multi-disciplinary institutions with a wide variety of data types and emerging domains may not be able to find a single schema that is standard to the many interested communities. In addition, many communities and/or disciplines are still in formative stages and have not yet established community standards.

#### I.D. *Curation & Quality Assurance*

1. Curation is not defined, and its common definition is extremely broad. Curation could be considered part of the analysis for many domains. Curation and Quality Assurance are distinct enough to warrant separate entries.
2. Mechanisms for “others” to provide input about existing datasets makes sense for some domains or data types but not all. Allowing “others” to provide input could cause confusion and a distrust of the data, putting the integrity of the data at risk. Transparent policies should be defined on the curation process. Consumers of the data should be able to view it without secondary input beyond the original data author.
3. Data quality assessment metrics need to be available when possible to enable consistent curation of data.

#### I.E. *Access*

1. Data should be described using web-compliant technologies like RDFa<sup>1</sup>, and the repository should provide SPARQL<sup>2</sup> endpoints to facilitate discovery. These approaches greatly expand visibility and interoperability, making broad access achievable.

#### I.G. *Reuse*

1. Tracking downloads or other metrics of use is vital, but this would better fit under a characteristic on “Use Metrics” avoiding confusion with the way “Reuse” is defined in the FAIR principals.
2. The repository should support information sharing about use and data characteristics. Researchers not understanding the data is often a limitation in reuse; communication helps build understanding and trust.
3. Clear instructions on how to best cite data should be available to incentivize data authors to continue to upload their data.

#### I.J. *Common Format*

1. A common format is ideal if one is available, but they are not always available. This is common in the applied sciences.

#### I.K. *Provenance*

1. While datasets may be deleted for pragmatic reasons, provenance should never be deleted as an enduring record of what was.
2. Provenance of data transformation defining parent and child data sets is a preferred approach for addressing curation and data lineage.

---

<sup>1</sup> <https://www.w3.org/TR/rdfa-primer/>

<sup>2</sup> <https://www.w3.org/TR/rdf-sparql-query/>



## II. Additional Considerations for Repositories Storing Human Data...

1. This section should be expanded to ‘Moderate Impact Data’ as defined by FIPS Publication 199<sup>3</sup>. Through graded approaches these same characteristics address Human Subjects Data but also extend to include protections for data from areas where there is Personally Identifiable Information (PII) or other sensitive data.

### II.H. *Violations*

1. A mechanism for detecting potential violations is needed. Suggest requiring a defined audit and assessment process to assure users are behaving appropriately.
2. Data mismanagement violations by the repository should be a separate sentence from user violations because they are very different. While repositories can self-audit for bad user behavior, a repository violation requires a third-party auditor to assess without bias.

### Suggested new characteristics

#### I *Data Acquisition*

1. All of the characteristics are focused on managing or downloading data from a repository as a human. Uploads, editing, and appending data and its associated metadata should be easy and accomplished at a web interface or systematically through data APIs.

#### II *Sensitivity Transition*

1. Repositories that hold ‘Moderate Impact Data’ should have a clear process to transform and/or transition data from sensitive to open under certain conditions.

---

<sup>3</sup> <https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.199.pdf>



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

March 6, 2020

**Subject:** RFC Response: Draft Desirable Repository Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research (FR Doc. 2020-00689)

**Respondent:** American Psychological Association

APA is the leading scientific and professional organization representing psychology in the United States, with more than 121,000 researchers, educators, clinicians, consultants and students as its members.

**Scientific Discipline:** Social Sciences/psychology

**Role:** Member organization, society publisher

Sean C. Bonyun  
Chief of Staff, Office of Science and Technology Policy  
*Transmitted via internet*

Dear Mr. Bonyun:

The American Psychological Association (APA) commends the Office of Science and Technology Policy (OSTP) for its efforts to help Federal agencies provide more consistent information on desirable characteristics of data repositories.

Psychologists work with a wide range of data, including data from surveys, laboratory experiments, government statistics, administrative records, imaging, genomics, and social media (Alter & Gonzalez, 2018). As a member organization with a publishing program in the psychological sciences, APA is committed to the promotion, education, and communication of open science and transparent practices.

Since the passage of the Open Government Data Act in 2016, the federal government's commitment to open data has been institutionalized. And as the Act has been implemented, the federal government has taken additional steps internally and with non-governmental partners and stakeholders to improve the use of data assets for decision-making and accountability for the federal government. This RFI will allow OSTP to further those aims.

APA offers its support to your efforts to refine and finalize the draft set of desirable characteristics of repositories. Please find our feedback organized according to four general categories: 1) items for inclusion in the list of desirable characteristics, 2) procedures for

750 First Street, NE  
Washington, DC 20002-4242  
(202) 336-5500  
(202) 336-6123 TDD



Please Recycle

[www.apa.org](http://www.apa.org)

handling the misuse of shared data, 3) the implementation and vetting of repositories, and 4) policy setting and review.

## 1. Items for Inclusion in the List of Desirable Characteristics

**Definition of Data and Code:** Given the breadth of research funded by the Federal Government, there should be a clear definition of data to be shared, with examples. This definition should include distinguishing between raw data and primary data. Guidance about the data should also be provided by an accompanied codebook that serves as a key for the file (Shönbrodt, Gollwitzer & Abele-Brehm, 2016).

**Timeline for Deposit and Embargo:** For clarity, grant recipients will need to know when they are required to deposit the data from the point of project completion. Further, specification about whether or not embargo periods will be allowed before secondary use is needed should be supplied (Shönbrodt, Gollwitzer & Abele-Brehm, 2016).

**Machine Readability and Interoperability:** We stress that metadata should be specifically machine readable and interoperable consistent with the principle of interoperability defined by the FAIR Guiding Principles working group (Wilkinson et al., 2016).

**Collaborative Data Sets:** Clarification is needed on what aspects of a dataset are to be shared for grant recipients who manage datasets that come out of domestic or international collaborations.

**Guidance for Non-Proprietary Data:** If research is funded that requires reuse of a non-publicly accessible dataset, guidance for compliance is needed.

**Consent Agreement:** We ask for clarification including how updates to datasets will be made and tracked when an individual participant changes their consent after publication.

**Sharing Rights:** More specification on which rights grant recipients need to assign for reuse are needed. For example, whether data owners conveying simple rights of use (such as the right to archive) to the repository, while retaining the exclusive rights of use to third parties (Shönbrodt, Gollwitzer & Abele-Brehm, 2016).

## 2. Procedures for Handling the Misuse of Shared Data

Researchers who work with human participants have seen the effects of inappropriately shared data, for example, data mined from dating sites that led to personal identifiability (Resnick, 2016). Similarly, video data from animal research could be misused for political reasons such as ending non-human animal research. Guidance on the checks that repositories should consider to ensure appropriate data reuse is needed. Plans should be included for violations in terms of use as well.

The rights and responsibilities of data users should be outlined in the desirable characteristics. Using a repository that enables tracking of data reuse is not sufficient.

### 3. Implementation and Vetting of Repositories

We recommend the OSTP consider how it will ensure that repositories are able to scale up for the amount of data that will be publicly shared or shared with protections. There are repositories to use as a model, including the Australian Data Archive and the Medical Research Council in the UK.

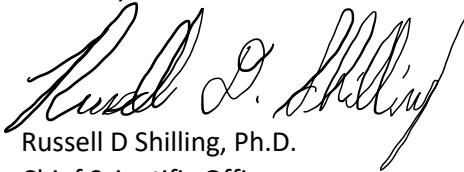
Given the variability in repositories, their funding models, and long-term viability and ability to house large datasets, grant recipients should have a list of vetted repositories to consider for their datasets. Guidance on how to pay for deposits is also needed; some datasets (such as neuroimaging) are quite large and cannot be included in free repositories. Similarly, curation is resource intensive and requires trained staff; if a dataset will have any protections, the researcher will likely need to pay a one-time or annual fee for depositing data.

### 4. Policy Setting and Review

Given the pace of scientific advancement, we request that a formal procedure for regular review of these characteristics be considered and shared along with the final desirable characteristics of data management and sharing. For example, future considerations might address recommendations for preregistering with the same repository one plans to use for data sharing. A review of the implementation of the desirable characteristics and whether there are adequate options among data repositories will need routine monitoring.

APA thanks OSTP for this opportunity to share our comments on the draft set of desirable characteristics of repositories for managing and sharing data resulting from federally funded or supported research. If you have any questions, or if we can provide any further information, please feel free to contact me at [rshilling@apa.org](mailto:rshilling@apa.org)

Sincerely,



Russell D Shilling, Ph.D.  
Chief Scientific Officer

### References

- Alter, G., & Gonzalez, R. (2018). Responsible practices for data sharing. *American Psychologist*, 73(2), 146-156. <http://dx.doi.org/10.1037/amp0000258>
- Resnick, B. (2016). Researchers just released profile data on 70,000 OKCupid users without permission. *Vox*. [vox.com/2016/5/12/11666116/70000-okcupid-users-data-release](http://vox.com/2016/5/12/11666116/70000-okcupid-users-data-release)
- Shönbrodt, F., Gollwitzer, M., and Abele-Brehm, A. (2016). Data management in psychological

science: Specification of the German Science Foundation (DFG) Guidelines.  
Wilkinson, M. D., Dumontier, M., IJssbrand J. A., Appleton, G. Axton, M., Baak, A., Blomberg, N.,  
Boiten, J.-W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J.,  
Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R. . . . Mons,  
B. (2016). The FAIR Guiding Principles for scientific data management and stewardship.  
*Scientific Data*, 3, Article 160018. <https://doi.org/10.1038/sdata.2016.18>

March 6, 2020

## RFC Response: Desirable Repository Characteristics

Comments provided by: John Allison, William F. Hosford Collegiate Professor of Materials Science and Engineering, University of Michigan, Ann Arbor, MI. Email: [johnea@umich.edu](mailto:johnea@umich.edu)  
I am providing these comments individually and as Director of the Center for PRedictive Integrated Structural Materials Science (PRISMS Center) which developed and since 2014 has maintained the Materials Commons, an open-access information repository and collaboration platform for the materials profession funded by DOE-BES. These are my personal opinions, informed by approximately 15+ years of experience and observations on the topic of "Open Science" (data sharing, repositories, open-source software). If additional information/clarification is required, please feel free to contact me at the above email address.

The desirable characteristics contained in the OSTP list are all reasonable and certainly desirable, with the caveat that several are currently beyond the scope for repositories in the materials science field, and I suspect for much/all of the physical sciences. This limitation is primarily due to funding constraints. Currently in the materials field, the only viable, sustainable funding mechanism for repositories is federal agency funding. Despite the existence of the Materials Genome Initiative which supports and anticipates the development of such repositories, federal agencies have provided, at best, limited financial support for repositories in the materials field. While the financial support that has been provided has been essential for establishing the repositories that are currently available, it has not been sufficient to address all of the desirable characteristics listed in the OSTP list.

In the materials field the minimum viable repositories have (or should be expected to have) the following characteristics (with some caveats)

1. PUID
2. Data security and back-up (an aspect of the long-term sustainability item on the OSTP list)
3. Metadata
4. Access
5. Free & Easy Access and Reuse
6. Privacy
7. Reasonable security (but perhaps not to the extent anticipated by the Standards listed in the OSTP list, I am not familiar with these Standards and they appear to be very detailed)
8. Common formats
9. Provenance

Specific areas that are not currently able to be addressed in the materials field are:

1. Sustainability. In the materials field, it is my opinion that repositories are adequately protected for unforeseen circumstances, etc, however, they are currently only feasibly operated with the availability of federal funding. In the event of a decline in federal

funding, it is not clear that these repositories would be able to continue. While contingencies for cold data storage (meaning the data are protected but inaccessible to the general public) in the event of loss of federal funding are desirable and reasonably to be expected, federal government should establish means of ensuring long term sustainability of this infrastructure either via federally provided facilities or modest maintenance funding.

2. Curation. Curation, while generally desirable, is currently beyond the resources provided by the federal government in the materials field. It may also be an impediment to data sharing, which is currently a significant limitation in the materials field (see Incentives below).

Items that are not included in the OSTP list but should be considered are:

1. Interoperability. This is one of the FAIR principles. In the materials field where there are a number of public repositories, interoperability would allow more complete access to all materials data available within repositories. However, this is currently beyond current resources in the materials field, where repositories have developed in a fragmented manner without standards of interoperability.
2. Incentives/mandates for data sharing via repositories: Although this is not specific to repository characteristics, data sharing has not “caught on” within the materials field. Incentives and/or federal mandates for providing data via these data repositories may be required to improve this.

# RFC: Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

Anubhav Jain  
Chemist Staff Scientist  
Lawrence Berkeley National Laboratory  
Berkeley, CA 94720

3/6/2020

## 1 Background

I am a computational materials scientist at Lawrence Berkeley National Laboratory and help lead two data efforts at LBNL. I serve as a thrust lead for the Materials Project ([www.materialsproject.org](http://www.materialsproject.org)), a DOE-BES funded program to calculate the properties of all inorganic materials and make the data available online. I also serve as a thrust lead for DuraMat, a DOE-EERE Energy Materials Network and for which I help advise the development of the DuraMat Data Hub (<https://datahub.duramat.org>) for storing various data sets related to solar photovoltaics.



Figure 1 Landing pages for the Materials Project ([www.materialsproject.org](http://www.materialsproject.org)) and DuraMat data hub (<https://datahub.duramat.org>) - two data repositories related to the materials science domain.

The Materials Project is largely composed of homogeneous data (density functional theory calculations) and the database is developed and maintained by the same team that generates the data (the research team also uses the database for their own research). First released in 2011, the Materials Project now has a community of over 100,000 registered users and has stimulated many downstream studies (>2000 citations). The Materials Project has also recently released MPContribs (<https://mpcontribs.org>), a platform for users to submit their own (small) data sets. A major motivation for the MPContribs framework is that data can be linked to existing entries in the Materials Project database - for example, a user can add a report of an experimental band gap which can then be shown alongside the computational band gap from the Materials Project. Thus, the central data set forming the core of Materials Project can be linked to user contributions.

The DuraMat Data Hub, released in 2018, supports user data contributions for a wide variety of heterogeneous data types (images, time series data sets, spreadsheet analyses, etc.) that relate in some way to solar photovoltaic module degradation. There is a dedicated team that develops and maintains the database; these team members are largely separate from the data contributors and the users of the data. The data hub is based on the CKAN platform (<https://ckan.org>) and many of the various DOE Energy Materials Networks (e.g., HydroGen, ElectroCat) use the same platform to consolidate development effort under a unified platform. A primary goal of these data repositories is to preserve the scientific output of the research conducted by the various Energy Materials Network projects.



## 2 Recommendations

### 2.1 Data upload

- **Batch uploads:** many data sets require the upload of both metadata (e.g., instrumentation settings, calculation parameters) as well as the measurement data itself. In some cases, the metadata might be common to multiple, even hundreds of measurements. A "batch upload" feature that allows one to reuse the same metadata information for multiple data contributions can be a time-saving feature. Such a feature was recently introduced into the DuraMat Data Hub.

### 2.2 Data access

- **Easy download in common formats:** e.g., a "Download All Data" button that provides one-click download of entire data sets in a common format that can be parsed out-of-the-box by many programming languages (e.g., CSV, JSON, XML).
- **API-based data access:** An Application Programming Interface (API) that allows users to write computer programs that query and access subsets of the data. For example, if a data set is large, a computer program can be written to perform queries and download only the subset needed rather than download the entire data set to a user's local hard drive. Or, if a data set is rapidly changing, an API allows for the computer program to automatically download the latest version prior to performing an analysis. An API can be specific to a programming language or it can be HTTP-based; the latter is accessible to essentially all programming languages and thus is becoming the more popular option. A HTTP-based API is used by the Materials Project and DuraMat Data Hub, which both expose a type of API called REpresentational State Transfer (REST). Another modern example of an HTTP API is GraphQL.

### 2.3 Data download statistics

- **Data download statistics:** Researchers typically need to report the impact of their data generation effort to funding agencies. Thus, the ability to track the number of unique visitors to a data set, the number of unique downloads, etc. is useful in establishing the impact of a work.
- **Site registration considerations:** Users wishing to access data sets typically prefer to avoid a registration (i.e., sign-up and login) process and instead download directly from a web page. However, data providers often use registration as a tool to count users, which serves as an invaluable metric to demonstrate impact of the repository to funding agencies and is typically seen as more reliable than pageviews (e.g., pageviews may be triggered by web crawlers). In cases where registration is needed, being able to use a common third-party authentication service (e.g., OAuth) is suggested versus creating a site-specific account.

### 2.4 Integrated data analytics and visualization

- **Data visualization:** The ability to perform basic data analysis (e.g., visualizations of data distributions) within the scope of the online repository can be useful - e.g., to spot outliers in the data or quickly verify if certain common assumptions in statistical analysis (e.g., data is normally distributed) hold true or not.
- **Data analysis via interactive web applications:** In the case of focused, homogenous databases like the Materials Project, it is possible to integrate very specific "apps" that provide scientifically relevant analyses of the data. For example, one of the most popular apps in the Materials Project generates a phase diagram for a chemical system of interest using the current data set (users often copy this phase diagram into their papers, with attribution). Another app uses the energies computed in the database to calculate reaction energies between chosen sets of compounds.
- **Data analysis functions via APIs:** Experience with the Materials Project indicates that exposing data analysis functions via a programmatic API is also useful to users. For example, users can write a computer program that instructs the Materials Project to generate a phase diagram via a single endpoint in the REST protocol; the Materials Project returns an object representing the phase diagram to the user. Thus, users are

not restricted to generating phase diagrams with the web app, they can also perform the same analyses by writing computer programs that loop over the chemical systems of interest.

## 2.5 Data versions and snapshots

- **Data versioning:** In some cases, data sets are appended to over time as more measurements are taken. Or, data that is found to be erroneous may be modified or removed to prevent further problems. In such instances, a method to version data and view data from past versions can be useful, especially if the community publishes research results with a certain snapshot of the data. The simplest versioning system is to upload a separate data set for each version, but this duplicates data that is common between versions and may require large amounts of disk space. Nevertheless, some form of data versioning is often needed to ensure the reproducibility of published research results (in the same way that software used to perform the analysis is versioned).

## 2.6 Privacy aspects

- **Access control:** In some use cases, data may fall into one of multiple classes: public, private (visible and accessible by only a set of individuals, e.g., those working on a project), and embargoed (initially private, with an agreement to make the data public after a set period of time).
  - The DuraMat data hub supports all the use cases above via the CKAN framework
  - The Materials Project has the notion of "sandboxes", which distinguishes between public and private data. When performing an analysis like generating a phase diagram from the data or performing a search query over the data set, the Materials Project will use data from the public sandbox as well as all private sandboxes granted to the user.
- **User credentials:** Ideally, user credentials (usernames, passwords) should be handled by already developed libraries rather than managed and handled by researchers developing a system. Systems like OAuth can be used to avoid the problems associated with improper handling of user credentials.
- **Sensitive data:** Neither the Materials Project nor the DuraMat Data Hub contain sensitive data on users. However, should a database include such information, then the mechanism of differential privacy may be one way to protect user confidentiality for downstream analyses. The differential privacy technology is already used by many technology firms such as Apple and Google, and may become easier to deploy in practice over time. Such techniques might also allow researchers to publish research involving private data (e.g., data provided by companies in a "private" section of the repository) with a strong guarantee that no company's involvement in the study is exposed.

## 3 Other examples of data repositories and data storage

The journal *Scientific Data*, introduced in 2014 by Nature Publishing Group, has had considerable success in attracting scientists to publish data sets. The impact factor of *Scientific Data* is 5.9, which is fairly high given that the journal does not publish new scientific findings, but mostly neutral descriptions of data sets along with a link to the original data. This encourages scientists to publish in the journal as a data contribution can be cited as a paper, counting towards metrics like *h*-index and yearly publications that serve as scientific performance indicators.

Note that *Scientific Data* does not host the data itself, but instead provides a list of recommended repositories for that accept user data in many domains (<https://www.nature.com/sdata/policies/repositories>). The recommended repositories include both "generalist" repositories that do not contain much customization or integration with other community data (e.g., Figshare, Dryad, Zenodo) as well as "community" repositories that more specifically target a given data type and community.

Another resource for data repository examples is the Registry of Research Data Repositories (<http://re3data.org>), which provides a comprehensive list of scientific data repositories.

**From:** Giulia Galli <[gagalli@uchicago.edu](mailto:gagalli@uchicago.edu)>  
**Sent:** Friday, March 6, 2020 7:48 PM  
**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>  
**Cc:** Graf, Matthias <[Matthias.Graf@science.doe.gov](mailto:Matthias.Graf@science.doe.gov)>; Biven, Laura <[Laura.Biven@science.doe.gov](mailto:Laura.Biven@science.doe.gov)>  
**Subject:** [EXTERNAL] RFC Response: Desirable Repository Characteristics

Dear All,

**Published papers, in particular those supported by federal funding, should be much more than a pdf!**

**They should be living, searchable stories with all data and appropriate metadata available to the community.**

We propose a strategy and created a simple tool to facilitate scientific data reproducibility **by making available, in a distributed manner, all data and procedures presented in scientific papers, together with metadata to render them searchable and discoverable.** In particular, we have created a graphical user interface (GUI), Qresp ( [>http://www.qresp.org/<](http://www.qresp.org/)) to curate papers (i.e. generate metadata) and to explore curated papers and automatically access the data presented in scientific publications.

I include a pdf explaining the idea behind Qresp.

Please see [>https://paperstack.uchicago.edu/<](https://paperstack.uchicago.edu/) for examples of curated papers (select Explorer and then Search)

I'd be happy to discuss the project and strategy in more detail.

Best Regards,

Giulia Galli

--

Giulia Galli  
Liew Family Professor of Molecular Engineering  
Professor of Chemistry  
The University of Chicago  
+1 773.702.0515  
[gagalli@uchicago.edu](mailto:gagalli@uchicago.edu)  
[>http://galligroup.uchicago.edu/<](http://galligroup.uchicago.edu/)

Senior Scientist  
Argonne National Laboratory

Director  
Midwest Integrated Center for Computational Materials (MICCoM)  
[>http://miccom-center.org/<](http://miccom-center.org/)

Executive Assistant  
Lisa Vonesh  
+1 773-702-0714  
[lvonesh@uchicago.edu](mailto:lvonesh@uchicago.edu)

From: [garrett@his.com](mailto:garrett@his.com)

To: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov).

Subject: RFC Response: Desirable Repository Characteristics

John Garrett

Co-chair Data Archive Interoperability Working Group

Consultative Committee for Space Data Systems (CCSDS) and ISO TC20/SC13

I've reviewed the proposed set of characteristics and I thank your group for identifying and circulating them. They do set forth a number of items that are important considerations when making use of data and I do agree with their thrust. I would, however, also suggest that you consider including a recommendation for assessment of repository via a widely recognized set of criteria. The option I suggest for this is ISO 16363.

My understanding of the purpose for these characteristics is that they are intended to be a set of repository-oriented characteristics that repositories would exhibit if they were consistent with and supported the data-oriented FAIR principals. In that vein, I think the proposed characteristics do a fairly good job. One reservation I have is that this is approached as if a repository's holdings were significantly the same. Many excellent repositories hold a variety of datasets with different requirements and policies. This should perhaps be noted in the introduction to these characteristics and made clear that they are considerations for the current datasets. Another aspect, that is perhaps outside the scope, is recognition that the datasets may over time move from repository to repository, for example from a project repository to an active domain repository and finally to some long-term repository. Or the same dataset may live in more than one repository at the same time. In those circumstances, an individual repository should consider the other repository's handling some of these characteristics, e.g. A. PUIDs, C. Metadata, and G. Reuse. For example, if the PUIDs used by each repository are different, then everything is basically reset. This problem could be alleviated through coordination between the repositories or at least by expanding understanding of characteristic K. Provenance to incorporate information inherited from the other repository.

In addition, you do indicate that you are attempting to make these characteristics "consistent with criteria that are increasing being used by non-Federal entities to certify repositories, such as ISO 16363 Standard for Trusted Digital Repositories and CoreTrustSeal Data Repositories Requirements." I believe that the proposed characteristics are compatible with ISO 16363 (and by extension Core Trust Seal, which is essentially a subset of ISO 16363) with the understanding that ISO 16363 does allow for different datasets having different policies. However, many aspects covered by ISO 16363 and Core Trust Seal are not covered by these proposed characteristics. While either ISO 16363 or Core Trust Seal can be used for self-auditing or peer review, note that a major difference is that Core Trust Seal relies on peer to peer review while

ISO 16363 is aimed towards professional, impartial third-party review. It seems an easy solution would be to add a characteristic that encouraged some type of certification. I would suggest ISO 16363 criteria as a possible useful set of considerations.

The proposed characteristics do set forth a number of items that are important considerations when making use of data and I do agree with their thrust. I would, however, also suggest that you consider including a recommendation for assessment of repositories via a widely recognized set of criteria. The option I suggest for this is ISO 16363, which allows for a third-party examination of a repository resulting in certification that is consistently applied and recognized world-wide. Of course, outside, independent assessment would incur costs, but especially for repositories that are large or that hold data of significant value it would be useful to have such review. For smaller data programs where the cost of outside evaluation is prohibitive, I would still suggest that the ISO 16363 criteria still be used for self or even peer-reviewed evaluations of the repository (although I understand there is starting to be a charge even for the peer-review of Core Trust Seal now). The Core Trust Seal (same high-level organization and effectively a subset of ISO 16363 criteria) also provides for peer-review. While useful for smaller projects, peer-review of self-prepared materials does allow for blind-spots to be overlooked and unevenness of application. So overall, I still recommend application of the wider ISO 16363 set of criteria with outside certification.

Overall, I feel that ISO 16363 is the most comprehensive set of metrics for establishing the value-added services of the repository. Even if certification will not be pursued, the ISO 16363 metrics can constitute a set of design criteria for a digital access and preservation system. Use of ISO 16363 will lead to the preservation of data, cost reduction, data integrity over time, and enhanced reputation of the repository.

As noted, your 11 proposed characteristics are related to and overlap with other standards such as OAIS (6 mandatory responsibilities for archives) and ISO 16363 (109 criteria at a few hierarchical levels). The Open Archival Information System (OAIS) Reference Model Standard is one of the most widely recognized standards for repositories with long-term preservation of information as their mission. It also provides some underlying concepts for the criteria defined in ISO 16363. In the limited space here, I will note some limited questions that arise from mapping between the OAIS mandatory responsibilities and the proposed characteristics. (Similar but more extensive comments could be made regarding ISO 16363 criteria in relation to the proposed characteristics.)

The 6 Mandatory Responsibilities from the OAIS standard are:

*The [repository] shall:*

- *Negotiate for and accept appropriate information from information Producers.*

None of the proposed characteristics address this. Although this is a fundamental aspect of a repository, you may have felt this was outside the scope of the characteristics identified. Perhaps that is correct, but perhaps knowing a repository is reaching out to enlarge its collections in your domain may be a discriminator in your decision of where to deposit data.

– *Obtain sufficient control of the information provided to the level needed to ensure Long Term Preservation.*

None of the proposed characteristics directly address a repository's ability to make changes to the data or its supporting metadata over time. It is possible that characteristic D. Curation & Quality Control is approaching some of the issues. Is there an assumption that any Federal-funded data deposited to repositories always will be granted with sufficient control for long-term preservation?

– *Determine, either by itself or in conjunction with other parties, which entities should become the Designated Community, that is, the communities that should be able to understand the information provided. Definition of the Designated Community includes a determination of their Knowledge Base.*

Again, this is not directly addressed by the proposed desirable characteristics and it should be. Characteristic C. Metadata could help in addressing this, but in Characteristic C. Metadata determination of the target communities is implicit and has already taken place before the metadata schemes are developed. It is not true that the target communities are always well known and understood. Another misconception possibly exhibited in characteristic C. Metadata is that repositories have a single "community that the repository serves." Many repositories support datasets that are aimed at different communities. And increasingly there is multi-disciplinary work. The community that should "understand" each dataset needs to be identified at the onset and re-evaluated over time. A particular repository may contain datasets that are appropriately targeted by different communities. A distinction should also be made between "users" of the dataset and the "designated community" for which it is being preserved.

– *Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.*

In your characteristics, a distinction needs to be made between "users" of the dataset and the "designated community" for which the dataset is being preserved.

– *Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.*

Most of the proposed characteristics could be seen as addressing this responsibility. However, the issue of deletion of data is not addressed. Characteristic A. PUIDs only addresses the PUID for deleted (or deaccessioned) data and the underlying and more important issues concerning deletion of Federal information by repositories are not addressed.

– *Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Content Information with evidence supporting its Authenticity.*

Several proposed characteristics -E. Access, F. Free and Easy to Access and Reuse and J. Common Formats, address some availability aspects and some aspects of other characteristics, A. PUIDs, B. Long-term sustainability, C. Metadata, D. Curation & QA, and K. Provenance, addresses, address the Authenticity issues.

As can be seen in this quick comparison, the proposed characteristics are broadly compatible with this widely-used standard, but there are also significant gaps that could also be addressed.

Finally, I just have some question on a couple of the characteristics.

I am unsure of the intent of Characteristic G. Reuse. Is this intended to simply be enabling of reuse or to actually track it? How much tracking? While I worked at NSSDC, we did have options to add metadata to reference papers that were based on individual datasets. However, we did not track (other than by count) and maintain information on every download of the dataset. I don't think any large Federal repository can afford to do that. And even trying to do that level of tracking may run into legal questions.

Characteristic F. Free & Easy to Access and Reuse. Many Federal Archives do have some cost recovery requirements imposed on them.

Characteristic K. Provenance covers items related to a single dataset within the current repository. You may want to expand the concept to also carry information provided when the dataset was deposited You may want to expand the concept to cover datasets that are derived (change of format, sub-setting, super-setting, mashing together of datasets, etc.) from the current dataset.

I would note that the concept of different types or levels of support for different datasets seem to be envisioned by your inclusion of the “Additional Considerations of Repositories Storing Human Data (Even if De-Identified)”. I again point out that a single repository may contain more than a single level or type of data. Distinction should be made whether these characteristics are applicable to the repository as a whole or to the individual datasets. Making



this distinction also makes it easy to have additional considerations (and possibly union of other considerations) not only for human-data but also for other data categories, e.g. proprietary, various levels of classification, financial, etc.

Thank you for the chance to comment on this. I wish you best of luck with them and hope that they can serve as a springboard for continuing improvement in long-term preservation and use of Federally-funded data. Hopefully, we can collaborate in the future in developing useful guidance for producers, curators and users of authentic data.

#### Author and Organizational Background:

My professional career spanned more than 30 years originally in Federal agencies and then as a contractor to Federal agencies, primarily NASA. The earliest portions of my career were generating, analyzing and using Federal data and then moved onto archiving digital information primarily at NASA's National Space Science Data Center, one of the earliest and at the time, largest repositories of digital information. While there, I participated and helped lead CCSDS information preservation standards projects. I am currently mostly retired, but continue to lead CCSDS efforts aimed at advancing and ensuring long-term information preservation.

The Consultative Committee for Space Data Systems (CCSDS) is an international standards development organization addressing space communications and on-ground data and information archiving standards. CCSDS also acts as ISO Technical Committee 20 / Subcommittee 13 – Space Data and Information Transfer Systems. CCSDS' Data Archiving Interoperability Working Group has developed a number of archival ISO standards that are widely recognized and respected in the information preservation community. Those standards include the Reference Model for an Open Archival Information Systems (OAIS) which is familiar to most professional archives world-wide. Most seek to be and claim to be compliant with OAIS. Another CCSDS standard and a more stringent hurdle is the ISO 16363 Standard for Trustworthy Digital Repositories which sets out 109 specific metrics used for third-party certification, but which can also be used internally by a repository for quality improvement.

From: [conradsireland@gmail.com](mailto:conradsireland@gmail.com)

To: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

Re: RFC Response: Desirable Repository Characteristics

My name is Larry “Mark” Conrad I retired from the National Archives and Records Administration (NARA) on October 31, 2019 with 28+ years of service working as an Archivist/Archives Specialist. I spent my entire career at NARA working with electronic records. From 2012 to my retirement I served as NARA’s representative to the NITRD Subcommittee. I was also an active participant in the NITRD Big Data Interagency Working Group and the HCI&IM Task Force. I am currently an instructor for the Digital Curation for Information Professionals (DCIP) Certificate Program at the University of Maryland, College Park, iSchool. Since 2009 I have been a member of the Data Archive Interoperability Working Group of the Consultative Committee for Space Data Systems (CCSDS) and ISO TC20/SC13. This working group developed and maintains ISO 14721, ISO 16363, ISO 16919, and related standards.

Primary disciplines: social science and information science. My roles have included appraisal and accessioning archivist for electronic records, program officer for funded research, researcher, peer reviewer of proposals and educator.

I appreciate the opportunity to comment on the Desirable Repository Characteristics. I have organized my comments below by the sections of the draft document.

## **Background**

It is unclear what is meant by “consistent with” FAIR, ISO 16363, and CoreTrustSeal. Does this mean that these are incorporated by reference or does it simply mean does not contradict the text of these documents? It would be important to clarify what is intended.

The Desirable Repository Characteristics, without the addition of requirements such as those found in [ISO 16363 – Audit and Certification of Trustworthy Digital Repositories](#), are inadequate to serve the purposes the SOS proposes for the Desirable Repository Characteristics. ISO 16363 is specifically designed to meet many of the SOS objectives. “This document is meant primarily for those responsible for auditing digital repositories and also for those who work in or are responsible for digital repositories seeking objective measurement of the trustworthiness of their repository. Some institutions may also choose to use these metrics during a design or redesign process for their digital repository.” (ISO 16363 Section 1.2)

The draft document does not reference ISO 14721 [Reference Model for an Open Archival Information System](#) (OAIS). This standard is the seminal document for trustworthy digital repositories. Before it was even published as an ISO standard, repositories began claiming “OAIS compliance.” I would recommend including this standard in the Desirable Repository Characteristics document.

ISO 16363 is designed specifically to test repository compliance with the OAIS Reference Model. It was written and is maintained by the same working group that wrote and maintains ISO 14721. This same working group wrote and maintains ISO 16919 - [Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories](#). This latter standard, in combination with ISO/IEC 17021 Requirements for bodies providing audit and certification of management systems, are used to accredit audit and certification bodies that carry out audits of trustworthy digital repositories. In other words, there is an internationally recognized ISO framework for accrediting auditors to carry out the audits (ISO 16919 and ISO/IEC 17021), an ISO standard to assess the trustworthiness of a digital repository during the audits (ISO 16363), and an internationally recognized “gold standard” for developing repositories that can provide long term preservation of digital information (ISO 14721).

No other suite of internationally recognized standards exists that cover everything from ensuring the auditors are competent to carry out the audit, to metrics to be used for assessing the repository, to a standard to guide the development of a repository fit for the long-term preservation of digital information. These standards do not require a particular implementation for the repository and are flexible enough to be used no matter what the discipline is of the data producer. Given the international scope of many of today’s research projects, it would make sense to use international standards for data repositories. An ISO-certified trustworthy digital repository would be the best place to deposit federally funded research data.

CoreTrustSeal is mentioned in the same sentence with ISO 16363. There is no real comparison between the two. CoreTrustSeal has 16 high level, general requirements -including some that are not directly related to the OAIS Reference Model. ISO 16363 has over 100 metrics directly related to OAIS. CoreTrustSeal certification consists of a self-assessment followed by peer review of the self-assessment results. ISO 16363 certification requires the end-to-end international standards-based process described in the previous paragraphs. I would recommend removing the reference to CoreTrustSeal or making it clear that there are substantial differences between it and the suite of ISO standards.

FAIR is mentioned in the same paragraph with ISO 16363. FAIR is concerned with requirements for the data that will be stored in the repository rather than requirements for the repository that will store that data. It would be a good idea to make this distinction clear.

If the SOS wishes to consider best practices for preparing the data, it might also want to consider a few other initiatives. “[The Turing Way](#)” is another set of best practices for creating reproducible data science in a manner that it can be used over the long term. The [Data Documentation Initiative](#) (DDI) provides a standard, best practices, and tools to “document and manage different stages in the research data lifecycle, such as conceptualization, collection, processing, distribution, discovery, and archiving.” DDI is widely used in the social, economic, behavioral and health sciences. It has been in use for decades and has an active international community of users.

The Background section of the document lists a number of Federal authorities, laws, regulations and other requirements. Some of the research data made or received by Federal Agencies in the course of business may, in fact, be Federal Records and subject to the [Federal Records Act](#) and [related regulations](#). It would be a good idea to consult with the [Chief Records Officer of the United States](#) concerning the implications of the law and regulations for this document.

## I. Desirable Characteristics for All Data Repositories

### B. Long-term sustainability:

NOTE: These comments are primarily issues for the funders of federal research rather than the repositories. Many of these issues were highlighted during the NITRD Big Data IWG Workshop, [Measuring the Impact of Digital Repositories](#). See especially, the [publications](#) resulting from the workshop.

Digital preservation is not a “one and done” operation. It requires continuous actions to keep the data viable in a rapidly changing technological environment. It also requires vigilance to ensure the information remains understandable to the end users as their knowledge base changes over time. **A repository needs a continuing stream of resources** – funding, personnel, expertise – to carry out these responsibilities. **The Federal Government** spends billions of dollars on producing research data. At least some of that money should be allocated to ensuring the data remains usable and understandable for as long as that data may be needed for reuse.

**Funding agencies** should carefully consider the length of time that the research data will need to remain available and understandable and ensure that this is taken into account in data management plans. Many data management plans associated with recent proposals only commit to making the data available for a few years after the project terminates. Much of the data would be useful – and may have to be reproduced at additional expense – well beyond that time frame.

## II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)

It would be good to acknowledge that data containing PII or other restricted content may be subject to laws and regulations that would supersede the Desirable Repository Characteristics. For example, data held in a repository run by a Federal Agency, might be subject to review and release under FOIA or might need to be registered as a Privacy Act system.

Thank you for allowing me to comment on the Desirable Repository Characteristics document.

Mark Conrad

Name: Hunter Moseley

Affiliation: University of Kentucky

Title: Associate Professor

Scientific Disciplines: bioinformatics, computational biology, systems biology

Specialty: omics data analyses, metabolomics data analysis, ontology analysis and utilization, structural bioinformatics

Role: researcher

Degrees: PhD in Biochemistry, BA in Computer Science, Mathematics, and Chemistry

The current “Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research” is the best description of desirable characteristics of scientific repositories that I have read so far. It covers not only FAIR principles, but also issues dealing with scientific rigor and reproducibility. With that said, I see two minor issues in section I.C. Metadata. I would suggest including the use of “controlled vocabulary” that is standard to the community. A data schema without established controlled vocabulary has a much lower reusability. Also, I would suggest explicitly mentioning “reproducibility” as a desired target that the metadata should support. There needs to be explicit balance of support for “reuse” and “reproducibility” in these desired characteristics of scientific repositories. I would rewrite this section as follows:

“C. Metadata: Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, reproducibility, and citation of datasets, using a schema and controlled vocabulary that is standard to the community the repository serves.”

## **Environmental Data & Governance Initiative (EDGI) Response to the Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research**

Submitted by:

Gretchen Gehrke, EDGI, physical sciences, researcher

Grace Poudrier, EDGI, social sciences, researcher

Steven Gentry, EDGI, information sciences, researcher

Rob Brackett, EDGI, computer engineering, developer and database manager

Kelsey Breseman, EDGI, information sciences, researcher

The [Environmental Data & Governance Initiative](#) (EDGI) is a North American network with members from more than thirty different academic institutions and ten nonprofit or grassroots organizations, as well as caring and committed volunteers who come from a broad spectrum of work and life backgrounds. EDGI promotes open and accessible government data and information along with evidence-based policy making. EDGI supports this OSTP effort to make data from federally funded research more available and accessible. Our comments focus mostly on the importance of version-control, data accessibility, and facilitation of data utilization. We address specific RFC elements as enumerated in the published Request.

*I.A. Persistent Unique Identifiers:* EDGI strongly supports the adoption of persistent unique identifiers (PUIDs) and additionally suggests version control for datasets. Version control would include a PUID for each version of a dataset and allow for automatic identification of changes made between versions, including checksums to identify changes or errors made in the processing of the dataset. This would assist researchers conducting secondary analyses to ensure they use the most accurate data, and would support scientific research integrity efforts emerging across the country. To facilitate collaboration and efficient field-wide research progress, EDGI also supports the creation of preliminary data repositories with PUIDs, which would allow researchers to share work in a timely manner throughout the research process, and spur related work without the risk of being scooped. Below (I.C.) we also recommend that research aims and methods be included alongside produced data, both to contextualize the data and to further support researchers sharing their progress without risking their work being improperly appropriated.

*I.B. Long-term sustainability:* Long-term sustainability is critical. Particularly crucial is the establishment of contingency plans so that researchers maintain access to data in the context of unforeseen events (such as a government shutdown), and that research outcomes can be verified and further utilized for decades into the future. Plans for long-term maintenance of repositories can be checked against a [data risk matrix](#), as all data without plans for risk

management must be assumed to be at risk. Safeguarding repositories begins with simple steps such as plans for automated backup of all data. As outlined in FAIR, metadata should be maintained and available even if the data itself is no longer retained.

*I.C. Metadata:* In addition to being structured and using a community-standard schema, repositories should anticipate any metadata crosswalks data users might need to do to use a repository's data, and provide metadata in simplified expressions accordingly. Metadata should include a brief, plain language description of the data; description of research study purpose and design that created these data, highlighting topic, specific research aims, keywords, and any significant constraints (e.g. specific geographic location, specific geologic age); a full description of methods for data collection, with linked SOPs where possible; exact instruments used for qualitative methods; uncertainties for data points where appropriate (such as standard deviations where multiple measurements were taken to produce one recorded data point); data dictionaries and other tools necessary to fully understand and contextualize data. Repositories should support the archiving of code, and the code used to process or create data should be available and citable.

*I.E., I. F., I.G. Free and Easy to Access and Reuse:* EDGI strongly supports free and easy access to datasets. All data, metadata, and supporting information (e.g. methods) should be freely available via web-based search and download. Data should, at a minimum, be searchable by keyword, topic, location, dataset size, funding agency, and year of completion. Publications that cite a given dataset should be findable from the dataset site. Those citations should be available immediately, and full access to publications should be available to citing publications after one year post-publication. Data repositories should be navigable from agency websites and from publications that cite them. Where data can't be published in the public domain, using a menu of standard licenses should be preferable to custom licensing terms. As described above (I.C.), extensive metadata is also crucial for accessibility and reuse, especially descriptions of research design such that other researchers and the public can gauge the appropriateness of a dataset for their aims and be aware of any data quality issues that may hinder their reuse of a given published dataset. As outlined in the FAIR principles, any data not open and accessible to the public should have openly accessible metadata which includes clear protocols and contact information for gaining access to the data.

*I.J. Common Format:* EDGI strongly supports the use of common data formats, and particularly non-proprietary formats. The Library of Congress [has listed several formats](https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml) and their descriptions that could serve as a basis for format requirements.<sup>1</sup> For rare or specialized data

---

<sup>1</sup> Library of Congress, "Sustainability of Digital Formats - Planning for Library of Congress Collections: Format Descriptions," <https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml>. Accessed on March 6, 2020.

formats, the rendering software should be indicated and (where possible) included in the repository.

I.K. *Provenance*: Provenance is critical. A repository with good provenance maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity. Please see the comment in I.A. suggesting unique PUIDs per version with changes between versions identified and tabulated, including changes due to preservation actions. Unique versions of datasets should be linked from each other version. It also would be helpful to see links to archived copies of the data management plans for projects that contributed data to a repository; even after a study concludes, this makes it easier to verify methods used.

II.G. *Retention Guidelines*: Retention guidelines are important not just for repositories storing human data, but for all data repositories. These should be developed and applied universally. OSTP or lead funding agencies should institute periodic system-wide monitoring processes to ensure data and any requisite software remain functional and available.

In sum, EDGI supports OSTP's formation of desirable characteristics for management and sharing of data from federally funded research. The FAIR guidelines are an excellent starting point. EDGI draws on a rich experience across environmental science research, website monitoring, and use of federal data sites to suggest further desirable characteristics of repositories that would aid in research, reuse, archiving, and community access to data.



# Population Association of America Association of Population Centers

## Office of Government and Public Affairs

1436 Duke Street • Alexandria, VA 22314

www.populationassociation.org • www.popcenters.org • 301-565-6710 x 1006



### Population Association of America President

**Dr. Eileen Crimmins**

University of Southern California

### Vice President

**Dr. Sara Curran**

University of Washington

### President-Elect

**Dr. Robert Hummer**

U. of North Carolina-Chapel Hill

### Vice President-Elect

**Dr. Marcia Carlson**

University of Wisconsin-Madison

### Secretary-Treasurer

**Dr. Bridget Gorman**

Rice University

### Past President

**Dr. John Casterline**

Ohio State University

### Board of Directors

**Dr. David Bloom**

Harvard University

**Dr. Deborah Carr**

Boston University

**Dr. Jennifer Dowd**

King's College, London, UK

**Dr. Pamela Herd**

Georgetown University

**Dr. Emily Hannum**

University of Pennsylvania

**Dr. Jennifer Johnson-Hanks**

University of California, Berkeley

**Dr. Hedwig Lee**

Washington University in St. Louis

**Dr. M. Giovanna Merli**

Duke University

**Dr. Mary Beth Ofstedal**

University of Michigan

**Dr. James Raymo**

University of Wisconsin, Madison

**Dr. Jenny Trinitapoli**

University of Chicago

**Dr. Kathryn M. Yount**

Emory University

### Association of Population Centers President

**Dr. Kathleen Cagney**

University of Chicago

### Vice President

**Dr. M. Giovanna Merli**

Duke University

### Treasurer

**Dr. Marcia Carlson**

University of Wisconsin-Madison

### Secretary

**Dr. Jeffrey Morenoff**

University of Michigan

March 9, 2020

Sent via email to: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

To whom it may concern:

The comments below are submitted on behalf of the over 3,000 members of the Population Association of America (PAA) ([www.populationassociation.org](http://www.populationassociation.org)) and the over 40 federally supported population research centers at U.S. based research institutions comprising the Association of Population Centers (APC) in response the notice in the Federal Registrar, [“Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research.”](#)

We are gratified to see that many of the stated desired characteristics of data repositories for storing data resulting from federally funded research align with the priorities of our organization and its members. We agree particularly with the following characteristics:

(A) *Persistent Unique Identifiers*. Unique identifiers will support data discovery, reporting, and research assessment. Repositories that support such identifiers would provide a standardized way of citing data products, aligning the incentives of academic rewards for principal investigators with the scientific community’s data sharing needs.

(B) *Long-term sustainability*. Experience in the population sciences has highlighted that technologies change quickly, and the value of historical data is sometimes limited because researchers cannot quickly and easily use such data with modern computing technologies.

(C) *Metadata*. We also agree that the distribution of metadata is crucial. Many fields have developed standards for the distribution of metadata, and such standards are crucial to enabling data discovery and reuse.

(E) *Access*. The call for data to be broad, equitable, and maximally open.

(F) *Free and Easy*. Making data free and easy to access are in keeping with traditions in the population sciences, as exemplified by our leading archive the Inter-university Consortium for Social and Political Research (ICPSR).

(H-K) *Secure. Private. Common Formats. Provenance*.

We also agree that storing data in repositories that are *secure* (H), *private* (I), have *common formats* (J) and clarifies the *provenance* (K) are critical objectives that can maximize the utility of data to researchers while also ensuring research participants that their data will be handled in ways that can produce societal good without personal harm.

We would also like to highlight certain tensions that exist within the proposed framework. Achieving the objectives of building repositories that are "Free & Easy to Access and Reuse" (F) and provide "Long-term sustainability" (B) have real costs, especially for large complex data sets that are often not adequately supported on Federal grants that fund data collection. Another tension is that designing repositories that make data more accessible and easier to reuse often means relying on existing technologies (e.g., for formatting and storing data), while ensuring the long-term sustainability of data in repositories often favors less efficient but more durable technologies. Achieving both of these objectives often involves considerable resources (e.g., because data must be stored in multiple formats). We believe the Federal government should make a modest investment to institutions to assure the long-term preservation and viability of research data.

Although we see the value of developing repositories to store data resulting from federally funded research, we are not in favor of placing these repositories within Federal agencies themselves. In the population sciences, so much of our data depend on the voluntary participation of citizens. Trust is an essential feature in gaining this voluntary participation, and the advent of massive computation and databases has left the public fearful of data sharing. We worry that the creation of massive Federal data repositories could be viewed suspiciously by the public, especially in the face of historical examples of Federal agencies sharing data to the detriment of the participating subjects. The Census Bureau is a model, having established data sharing firewalls across Federal agencies to address the public's concerns about how their data could be accessed and used and to ensure the agency's ability to collect complete and accurate data. The value of central data storage and distribution is enticing, but we recommend a cautious approach that balances a need to maintain the public confidence with the data sharing needs of the research community. We note that strengthening and maintaining well-known and stable repositories that have been created by non-governmental organizations, such as ICPSR might be a viable alternative to Federal data repositories as these repositories may not raise public fear of inappropriate data sharing within the Federal government.

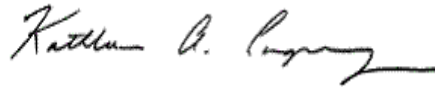
Finally, we would like to advocate for another characteristic of data repositories that was not mentioned in the Federal Registry notice, which is that there should be more mechanisms to protect sensitive, confidential data stored in repositories from being obtained through Freedom of Information Act requests and/or court orders. The Federal government currently provides two such mechanisms that protect confidential and sensitive human subjects data: the National Institutes of Health can issue Certificates of Confidentiality for funded and unfunded health-related survey efforts, while the Department of Justice can provide Privacy Certificates for DOJ-funded research projects. There are many examples of Federally-funded research projects that fall outside these narrow definitions and thus cannot be protected against the risk of disclosure, which compromises the ability of researchers to guarantee confidentiality to human subjects and also makes it difficult for researchers to obtain data from some government agencies. We advocate for the creation of a unified that would provide legal protections for all federally funded human subjects research.

Thank you for providing the opportunity for PAA and APC to comment on this important topic. We are eager to be a resource to the Office of Science and Technology Policy as it proceeds with any plans to encourage enhanced Federal data sharing and management policies.

Sincerely,



Eileen Crimmins, Ph.D.  
President  
Population Association of America



Kathleen A. Cagney, Ph.D.  
President  
Association of Population Centers

RFC Response: Desirable Repository Characteristics  
Lawrence Livermore National Laboratory

The Lawrence Livermore National Laboratory is pleased to respond to the White House Office of Science and Technology Policy's (OSTP's) [Request for Public Comment](#) on a draft set of desirable characteristics of data repositories used to locate, manage, share and use data resulting from Federally funded research. Additional guidance and policies cited in the Request for Comment that were consulted as part of our response were the U.S. Department of Energy's Public Access Plan (July 24, 2014), Federal Data Strategy Practices (<https://strategy.data.gov/practices/>) and FAIR metadata principles (<https://www.go-fair.org/fair-principles>).

Overall, we greatly appreciate that OSTP is taking thoughtful, consultative approach. Our overall sense is that the research community, computer and information science communities will need more time to continue to digest and respond as plans continue to develop. Our Laboratory and our research communities acknowledge that research data management (RDM) is an emerging field in information science, with strong connections in computer and library sciences. We encourage OSTP to engage experts in these fields from laboratories that would either contribute to or consume stored data, as well as those that might serve as repository hosts to not only formulate operational models, but also durable collection development policies and governance structures.

Our Laboratory acknowledges that some data has value well beyond its original purpose, much like published information, and that exposure of certain forms of data to the broadest audience (when possible and appropriate) can accelerate scientific discovery and innovation in unforeseen ways. Research Data is like a new form of literature, to be managed and curated over time; indeed, there are important explorations underway among libraries to determine what roles they might play in this space. Certain early assessments already suggest an emerging division of labor among well-funded institutions to potentially serve as repositories while others develop new consultative capacities to guide researchers on where to deposit.

<https://escholarship.umassmed.edu/jeslib/vol4/iss2/4/>

Our response reflects certain assumptions about the 3–10 year future of research data repositories, including:

1. A vision of repository management that assumes **an ecology of shared repositories** (multiple repositories, not a single repository; repositories supported “above the institution”).
2. A finite number of **repositories can serve distinct functions** and purposes in that ecology (e.g. collection and organization repositories, preservation repositories, discovery and access repositories); investments in repositories should focus on developing the best-in-class in that function.
3. **Trust in shared repositories**, and consequently buy-in to use and sustain them, **will require sustained community engagement and governance** to define their scope, operations, policies and ongoing investments.

## General Recommendations

In that context, our overarching recommendations are:

1. To consider preparing **initial minimal and desirable characteristics** for repositories that serve different purposes (**archetypal functional repositories**) and to refine them over time.
2. To create an initial **“list” of repositories that comply** with those characteristics already and a DOE guidance toolkit that not only outlines the components of a Data Management Plan but also includes guidance on data repository selection. Consider grading repositories against the criteria (rather than a “meets” or “does not meet” response) and publicizing the grade. The purpose of this is to not only inform depositors as they make choices about where to deposit, but also to encourage gradual development of repository services through achievement of a “better grade.” Consider enhanced or DOE specific versions of: <https://rdmtoolkit.jisc.ac.uk/share-and-publish/where-should-i-deposit-my-data/>.
3. To develop **assessment criteria** for shared repositories that not only address repository **operations** but also **business continuity, governance and clarity of collection development scope**.
4. To create a **DOE steering committee for research data management (DOE RDM Steering)** comprised of representatives from Data Archive Holders, Consumers and Depositors to formulate plans for a repository network. The steering committee may simply recommend use of existing repositories and/or may include recommendations for new repositories where there are gaps. If new repositories are recommended, consider the very successful parallel governance models that exists for distributed shared print or shared preservation repositories among research libraries (e.g. [WEST](#), [LOCKSS](#), [HathiTrust Digital Library](#) and HathiTrust Research Center).
5. To establish an **advisory committee** comprised of representatives from leading repositories and digital preservation services (e.g. DataOne/DataCite, Dryad, NIST Data, DOE Data, HEP Data, ArXiv, Portico, CLOCKSS).

Initial activities that a steering committee might undertake include:

1. **Surveying STEM scientists** and research groups in DOE Laboratories to gain a high-level scope of **data inventories, current storage and growth projections and begin characterizing types of datasets**.
2. Developing an **education program to educate scientists** at all levels on expectations and practices for data management and to communicate the value of RDM. Consider library outreach programs to connect scientists with publishing and data repositories.
3. **Conducting an environmental scan and formulating initial recommendations**. Assess the current landscape of shared data repositories in STEM fields, as well as a survey of Research Data Service skills and capacities at DOE Laboratories (in computer science

RFC Response: Desirable Repository Characteristics  
Lawrence Livermore National Laboratory

departments and libraries.) Identify gaps. Make recommendations about participating in existing repositories, developing new ones to fill certain gaps, and about skill sets to cultivate locally.

4. Engaging time limited task force(s) to develop **1) governance and business models, and 2) operational models for repository types**. Bring together a **broader range of experts** in library and information science, computer science, data science, cybersecurity, information architecture, and research data management.

### Barriers to Effective Research Data Management

Small groups within our Laboratory have explored RDM issues in the literature and have begun to experiment with pilot efforts (e.g. among HPCCs, DOI minting, Dryad participation, etc.). From these efforts, several key barriers to effective RDM have been identified, which are to some extent explored in the RFC.

#### Technical Barriers

- **Definitions:** how do we define data? Scope definitions? Size definitions? Relational definitions? Raw, analyzed, synthesized data, lab notebooks. Data associated with publications. Data requiring software to interpret.
- **Selection criteria and collection development:** what should be kept? Which datasets will be valuable to the future? Who will make those decisions? Should preliminary work be included? failed experiments? Irreproducible data? Is there a timeline requirement for deposit?
- **Software associated with data:** Under what circumstances can/should software and datasets be retained together and both curated over time? Implications for repository design?
- **Inventory and Growth:** what is the current scope of datasets to be managed? How can they be characterized? What is the anticipated growth?
- **Size and technical capacity:** what is the “right size” of a repository? And what organization(s) have the capacity in terms of storage and expertise to maintain them?
- **Standards and Interoperability.** Infrastructure Standards, Metadata Standards, Authority Controls to support interoperability: What are the minimal infrastructure standards, metadata standards and authority controls a repository should meet given its declared function (e.g. preservation, discovery, access)?
- **Ease of deposit:** priority should be placed on simplifying deposit for the user.
- **Curation:** ensuring requirements are met at deposit and curating/migrating datasets over time beyond their original purpose. When should datasets be kept indefinitely?
- **Quality Control:** What framework will be created to support the integrity of data maintained within the repository? Will OCR and AI/machine learning tools be utilized in the primary design of the data repository?
- **Dashboard:** Will the reporting requirements be standardized for institutional data reporting? Ensure ease of customization for obtaining KPIs and other metrics through user defined/created workflows.

## Governance

- **Guidance and timing.** What is the best form of guidance to achieve buy-in: Executive Order, Principles, Policy? when is it needed?
- **Organization and coordination.** What organizational model and central organization is needed to oversee a data repository system or network?
- **Retaining institutions.** Which institutions will maintain the repositories, even if—or especially if—shared repositories? What incentives are there to sustain repositories that serve a wider audience?
- **Costs to sustain a repository and repository system.** What does it cost to sustain a shared data repository over time? And a system of repositories? What are the resource requirements? What costs of RDM shall be borne by the retaining institutions and what costs borne by a coordinating organization?

## Federal Research Data Management and Repository Strategy Development

In terms of strategy, we encourage OSTP to develop repository characteristics in the context of an ecosystem of a finite number of shared repositories that are consortially supported for multiple years.

We suggest the functional role of a repository as a primary defining feature above others:

- Collection repository (disciplinary, general, institutional, publisher).
- Preservation repository (dark/dim archive).
- Discovery repository (public search, expert search, machine-level search).
- Access repository (public access, limited access, expert access).

## Repository oversight and policies

We offer suggestions for roles that OSTP might play as an actor and as a catalyst for development of shared repositories. With a steering committee, OSTP can be well positioned to:

- **Develop initial repository characteristics** for a small number of **repository types** (e.g. minimal characteristics of preservation repositories, minimal characteristics of discovery repositories, etc.) to be adopted by initial pilot repositories or affirmed with existing repositories.
- **Develop collection development guidance** for repository types and across repositories, monitor and report gaps over time across the landscape of repositories.
- Invest not only in defining technical characteristics, but also in **creating effective consortial governance structures** to oversee and sustain shared repositories over time.
- Develop **community review committees** to assess repositories for business continuity and operational excellence.
- Emphasize **interoperability and standards** as core characteristics. Recommend initial standards and establish review group to refine them over time.
- As distributed consortial and national repositories gain momentum and begin to fill appropriate niches, OSTP might transition its leadership efforts to establishing

RFC Response: Desirable Repository Characteristics  
Lawrence Livermore National Laboratory

connections between repositories (at governance and operational levels) or even facilitating consolidation of some repositories.

### Additional technical feedback

Specific technical feedback on elements in the RFC include:

- The **globally unique identifiers** need to be global across all domain spaces and not just unique within a given repository or domain space. Make this more explicit.
- Repositories should have a plan for **how data links are managed over time**, including testing and planning for handling dead links.
- An **attribution policy** needs to be specified and supported so that data providers get appropriate credit for their work.
- **Metadata standards** such as FAIR, FRBR, RDA, Premis, and domain-specific metadata should be adopted to enable data queries within and across repositories. They can be lightweight, usable by all participating DOE Laboratories.
- Raise awareness of established **national and global ontologies**, provide training and guidance for communities and repositories that choose to **supplement with custom** metadata fields, terminology, glossaries and ontologies.
- **Tagging and linking** should be supported by repository interfaces. This allows users (including but not just the data provider) to augment datasets with ad hoc, emerging, supplemental, or missing information, or post-processing and analysis.
- **Fine-grained access controls** are required for many projects to participate in a broader repository. Transparent mechanisms for how security risks would be identified over time as new data is added, and constraints that could prevent those risks.

### Supporting the Data Management Workforce

Finally, as an emerging interdisciplinary profession, we recommend that OSTP or its steering committee:

- Provide guidance on skills sets and expertise needed to manage data repositories.
- Define typical roles/responsibilities for RDM scientists and support staff.
- Identify research and development areas in RDM and funding support for RDM research.
- Develop RDM residencies to allow staff at existing Laboratories to conduct research or participate in steering or working groups for a period of time.
- Develop communication assets to socialize research data management among scientists (assets for use locally by DOE Laboratory computer scientists and librarians to get the word out).

Lawrence Livermore National Laboratory thanks OSTP for the opportunity to submit comments and to collaborate on matters of national importance.





**Association of  
American Medical Colleges**  
655 K Street, N.W., Suite 100, Washington, D.C. 20001-2399  
T 202 828 0400 F 202 828 1125  
www.aamc.org

March 10, 2020

Office of Science and Technology Policy  
Executive Office of the President  
Eisenhower Executive Office Building  
1650 Pennsylvania Avenue  
Washington, DC 20504

**Re: Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research (85 FR 3085)**

Submitted electronically to: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

The Association of American Medical Colleges (AAMC) appreciates the opportunity to comment on the White House Office of Science and Technology Policy (OSTP) request for information on desirable characteristics of data repositories, as proposed by the National Science and Technology Council's Subcommittee on Open Science. The AAMC is a not-for-profit association representing all 155 accredited U.S. medical schools, nearly 400 major teaching hospitals and health systems, and more than 80 academic and scientific societies. Through these institutions and organizations, the AAMC represents nearly 173,000 faculty members, 89,000 medical students, 129,000 resident physicians, and more than 60,000 graduate students and postdoctoral researchers in the biomedical sciences.

The AAMC strongly supports improved access to data resulting from federally funded research. The development of consistent guidelines and clearly defined characteristics for repositories to preserve and provide access to research data are critical in enabling academic institutions to achieve this goal. AAMC encourages harmonizing these guidelines for investigators and institutions across agencies as much as possible, while still allowing for flexibility to accommodate different fields of research and agency objectives. We also agree that in some

instances, it is most effective for the agency to designate a specific repository for particular research initiatives or data types.

Additionally, as AAMC noted in recent comments<sup>1</sup> in response to the National Institutes of Health’s draft data management and sharing policy, many institutions are planning on building and/or expanding their own repositories as agencies institute new requirements for researchers, and “without guidance from the agency on standards for data storage and discoverability... holding data in such disparate platforms and systems will place a significant technical burden on anyone who wants to access the data, thwarting the agency’s laudable goals to increase and improve data reuse.”

The AAMC is generally supportive of the proposed desirable characteristics of data repositories, many of which we note align with community-driven criteria proposed last year.<sup>2</sup> Given the rapidly developing importance of data in scientific research, these guidelines should be flexible enough to keep pace with technological advances, as well as the increasing volume and diversity of scientific data.

We strongly agree with the recommendation (C) that repositories assign datasets a “citable, persistent unique identifier (PUIID), such as a digital object identifier (DOI) or accession number.” Attaching a PUIID to a dataset would not only support data discovery and research progress reporting, as noted by OSTP, but is a critical step in tracking data reuse, crediting investigators for their work, and ultimately developing a more comprehensive understanding of research outputs. We also note that the use of PUIIDs has previously been suggested by several federal research funding agencies, including the National Science Foundation.<sup>3</sup>

A recently published initiative from AAMC and other research stakeholders to promote effective data sharing describes a path to connect researchers to their datasets, based on the use of PUIIDs.<sup>4</sup> While the use of PUIIDs for datasets is key, we recommend that the subcommittee consider specifying that repositories provide the option to attach additional unique identifiers to the dataset, including ORCID ID for investigators, and in the future, grant and/or organizational IDs.

---

<sup>1</sup>AAMC Response to NIH NOT-OD-20-013: “Request for Public Comments on a DRAFT NIH Policy for Data Management and Sharing and Supplemental Draft Guidance” (2020). <https://www.aamc.org/system/files/2020-01/ocomm-ogr-AAMC%20Response%20to%20NIH%20draft%20data%20sharing%20policy.pdf>

<sup>2</sup> Sansone, et al. Data Repository Selection: Criteria That Matter (2019). <https://osf.io/m2bce/>

<sup>3</sup> NSF 19-069: Dear Colleague Letter- Effective Practices for Data (2019). <https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp>

<sup>4</sup> Pierce, et al. Credit Data Generators for Data Reuse. *Nature* 570, 30-32 (2019). <https://www.nature.com/articles/d41586-019-01715-4>

Such connection between identifiers, beginning with the repositories, is necessary if the goal to fully and effectively track data reuse is to be realized.

AAMC agrees with the recommendation (E) that repositories should provide “maximally open access to datasets, as appropriate, consistent with legal and ethical limits to maintain privacy and confidentiality.” We suggest that this recommendation also include providing access to metadata, in agreement with FAIR data principles.<sup>5</sup> We also recommend that the language in (E) replace the recommendation (F) that repositories should make datasets “accessible free of charge in a timely manner after submission,” which does not seem to allow for restricted use cases.

We appreciate the subcommittee’s recognition that repositories which store data from individuals require additional considerations in order to ensure adequate privacy and security, as well as controls on use and access, even when those data are considered de-identified. However, some of the proposed characteristics, including (A) restricting dataset access to appropriate uses consistent with original consent and (B) the need for a repository to enforce submitters’ data use restrictions, while imperative considerations for human subjects data, may be outside of the traditional purview of a repository. We urge the subcommittee to consider specifically which of these recommendations are suitable for a list of recommended repository characteristics, and which would be better addressed under a separate agency policy or guidance and be the responsibility of the investigator depositing the data. Regardless of the mechanism, we agree that specific promises made to human subjects through consent documents about the use or sharing of research data should be honored and that repositories should facilitate, not create barriers to, the ability for investigators to ensure those promises are kept.

We strongly encourage, in addition to these guidelines on repository characteristics, the creation of a clearinghouse for federal research data policies and related resources, such as tools for metadata creation. Investigators may also find helpful a comprehensive list of agency-supported repositories, as is currently maintained by the National Library of Medicine,<sup>6</sup> as well as links to other commonly used repositories to store the results of federally funded research. In order for data to be successfully reused, it must not only be deposited in an appropriate repository, but also meet several other criteria, including adequate metadata, curation, and the use of common standards. Providing additional guidance on these topics is essential to meeting the end goal of effectively sharing the results of federally funded research.

---

<sup>5</sup> Wilkinson, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://www.nature.com/articles/sdata201618>

<sup>6</sup> Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)- Data Sharing Resources (Accessed March 2, 2020). [https://www.nlm.nih.gov/NIHbmic/nih\\_data\\_sharing\\_repositories.html](https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html)

The AAMC appreciates OSTP's efforts to seek input from stakeholders and looks forward to continued engagement as the federal government develops guidance relevant to data management and sharing. Please feel free to contact me or my colleagues Anurupa Dev, PhD, Lead Specialist for Science Policy ([adev@aamc.org](mailto:adev@aamc.org)) and Heather Pierce, JD, MPH, Senior Director for Science Policy and Regulatory Counsel ([hpierce@aamc.org](mailto:hpierce@aamc.org)) with any questions about these comments.

Sincerely,

A handwritten signature in blue ink that reads "Ross E. McKinney, Jr., MD". The signature is stylized and cursive.

Ross E. McKinney, Jr., MD  
Chief Scientific Officer



# American Economic Association

*Office of the Data Editor, Email: [dataeditor@aeapubs.org](mailto:dataeditor@aeapubs.org)*

The Data Editor of the American Economic Association (AEA) is pleased to respond to OSTP's "Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research", as invited in the Federal Register of January 17, 2020 (85 FR 3085).

*Thank you for your consideration.*

*Questions on this document can be directed to the Data Editor of the AEA, Lars Vilhuber at [dataeditor@aeapubs.org](mailto:dataeditor@aeapubs.org).*

## Primary discipline and roles

The American Economic Association (AEA), was founded as a professional society in 1885. Current membership is comprised of over 20,000 economists in academia, business, and government service. The AEA publishes eight journals, including the most prestigious academic journals in economics, as well as an electronic bibliography that serves as a comprehensive index to peer-reviewed journal articles, books, book reviews, collective volume articles, working papers, and dissertations.

In January 2018, I was appointed as the first Data Editor of the American Economic Association, with the mission to "design and oversee the AEA journals' strategy for archiving and curating research data and promoting reproducible research."

## Comment

The importance of sharing data (and computational instructions, "code") for the purpose of transparency and reproducibility of science is paramount to AEA and for science in general. Repositories used by scientists to deposit the inputs, tools, code, and outputs of research, whether funded through federal funds or other, play a key role.

We in the AEA emphasize that the scope of these considerations should include research created by scientists in the direct employ of the federal government, data created for public and research use with federal funds as part of the business of the 13 [federal principal statistical agencies](#), as well as any data created for research and evaluation under [H.R.4174 - Foundations for Evidence-Based Policymaking Act of 2018](#). All of the above are federally funded, and are frequently used to validate research findings. It is as important to include the preservation of such data in the considerations of the SOS, and to ensure consistency of application of any guidelines issued across all these different domains.

We support the reference embodied in the cited standards (ISO16363 Standard for Trusted Digital Repositories and CoreTrustSeal Data Repositories Requirements). In what follows, I comment on specific aspects of the characteristics as outlined in the RFC.



# American Economic Association

Office of the Data Editor, Email: [dataeditor@aeapubs.org](mailto:dataeditor@aeapubs.org)

## I. Desirable Characteristics of Data Repositories

### A. Persistent Unique Identifiers

We agree that persistent identifiers are an important attribute of data in repositories. However, we also suggest that the federal government set aside funds specifically to support the registration of persistent unique identifiers in central registries. While the individual price seems low (as of February 2020, [CrossRef](#) charges \$0.06 to assign digital object identifiers (DOI) for datasets or components, and the lowest tier at [DataCite](#) another registrar, is 500€), the associated cost of implementing robust integrated systems to perform the initial registration and maintain the associated landing pages is probably non-trivial. Assignment of DOI to specific (reproducible) queries or data extracts in interactive systems can quickly escalate. Costs for maintaining such systems typically extends beyond initial funding periods, but must in principle be supported “permanently”.

*Recommendation 1: Allow for funding in grants and research contracts for the maintenance of persistent identifiers.*

### B. Long-term sustainability

Maintaining data assets for a sufficient long time is *critical* to ensure reproducibility. Two aspects are worthy of consideration here. First, most federal funding does not provide clear guidance that would allow for the expenditure of funds beyond the funding period. For instance, most research grants allow for expenses for the 2-5 years of the grant period, but are unclear about the use of funds to pay for storage or maintenance costs beyond the end of the grant period. In Europe, recent [funding guidance](#) clearly identifies data management costs as eligible costs, and [explicitly allows](#) for the costs of deposit of research data in an open access data repository (run by an external organization).

*Recommendation 2: Explicitly allow for deposit costs as line items in federal funding vehicles, clarify usage of such funds when benefits accrue beyond the funding period.*

Second, we also note that not all data needs to be preserved into perpetuity. The question of how to identify when data can be de-accessioned or even destroyed is one where very little guidance exists in practice. Proper tracking of re-use (I.G) can provide some guidance, but is inherently a backward looking metric, whereas de-accessioning requires forward-looking analysis. We would encourage providing research funding to better understand how and when de-accessioning of data should be considered.

*Recommendation 3: Fund research into the measurement of the long-term value of data.*



# American Economic Association

Office of the Data Editor, Email: [dataeditor@aeapubs.org](mailto:dataeditor@aeapubs.org)

Finally, we recommend that whatever the preservation or retention policy may be, repositories should clearly state both a general policy as well as an object specific policy. Such policies can be recorded with central registries (e.g., Registry of Research Repositories, [re3data](#)) and within object-specific metadata, for instance the DOI record (DataCite Metadata Working Group 2017). Having this information easily available allows researchers to immediately assess the utility and robustness of a particular data item for their research, contributing to its reproducibility.

*Recommendation 4: Require that information about dataset persistence be easily available in human and machine-readable form.*

## C. Metadata

We strongly endorse the requirement of sufficient metadata. Much of economic research uses datasets which for a variety of reasons (ethical, commercial interests, security concerns) cannot be made available as public use data, and yet may be accessible through a variety of tiered access mechanisms ([Federal Statistical Research Data Centers](#), [licensing agreements](#), non-disclosure agreements, etc.). In order to make such access mechanisms more efficient, and to allow for re-use (I.G.), metadata is critical. Metadata allows researchers to prepare analysis code prior to accessing the restricted-access data (examples from [Norway](#) and Germany (Müller and Möller 2019) illustrate such procedures), making such procedures much less costly to researchers, and supporting ease of access (I.F).

However, we would also suggest that there are various degrees of metadata. We would strongly suggest that a minimum (and cheap) requirement for such repositories is to provide **data citations**. Data citations enable more consistent tracking of usage (by data providers) and of provenance (for scientific reproducibility), see (Martone 2014). Persistent identifiers (I.A) like DOI are not a requirement for proper data citation and attribution. Much more helpful is for repositories (in the broad sense) to provide suggested citations, and strongly encourage researchers to use them. An excellent example are the data citation practices of [IPUMS](#). Even before the (relatively recent) implementation of DOI, IPUMS had an excellent track record of getting researchers to cite the (federally funded) data that they have prepared. Thus, the much simpler implementation of “suggested data citations” (prior to implementation of DOI) is a critical element to support

*Recommendation 5: Require provision of a suggested data citation as the required minimum for metadata.*

## D. Curation and Quality Assurance

We believe that there are various levels of appropriateness for curation and quality assurance. While heavily re-used data should be professionally curated, it should be possible to improve curation over time. To the best of our knowledge, there is currently no robust mechanism to



## American Economic Association

*Office of the Data Editor, Email: [dataeditor@aeapubs.org](mailto:dataeditor@aeapubs.org)*

allow for continuous improvement in curation over time, in particular of metadata. In part this is technological (most existing repositories do not support such activities) as well as legal (unclear responsibilities and permissions of data owners). For instance, many entities -- [IPUMS](#), [FRED](#), [NBER](#)) have, over time, improved the metadata and curation of federally created data (data from Bureau of Labor Statistics and U.S. Census Bureau), but rely on that data being in the public domain. It is much harder to find examples where such data is freely available under open licenses, and yet being improved by entities other than the original data owners.

### E. Access, and I. Privacy

Proper access description is key to broad re-use of data. While reasonable safeguards are necessary, they can take many different forms. The “Five Safes” framework (Desai, Ritchie, and Welpton 2016) highlights that many factors contribute to making data access safe, and can be balanced. Combining legal constraints (entering into enforceable confidentiality agreements), statistical data protections (anonymizing data) with physical constraints (accessing data only from safe rooms) allows data repositories to optimize the access protocol for the broadest possible access. It may be desirable for repositories to allow for multiple access protocols. For instance, allowing remote access to data for individuals with high trust, while allowing safe-room access to individuals who are building their trust, can increase the acceptability of stringent safety requirements.

Similar to our earlier point regarding the visibility of sustainability policies, whatever the access protocols for a particular dataset may be, they should be clearly and visible recorded. Access restrictions should be clearly outlined (for instance on dataset landing pages), and any conditions clearly described (e.g. citizenship or physical presence requirements). These should also be recorded as part of the metadata on the repository (aforementioned re3data) and the object (DOI).

### F. Free and Easy to Access and Reuse

While there is little doubt that metadata should be free - a key tenet of the [FAIR data principles](#) - it is less clear that access itself needs to be free at the point of service. While free access for downloadable data seems to be a standard, it intersects with the (costly) long-term preservation (I.B.). More onerous but necessary access restrictions to enforce ethical or privacy concerns (I.E.) are generally much more costly. Sustainability in the absence of user fees is thus a concern that needs to be balanced with those aspects. A model that is seemingly practiced in the bio-medical community is for repositories to be developed, with federal funding, by third-parties, implementing access mechanisms, protocols, and policies. Once such repositories are stable, federal institutes (NIH) take over the continued maintenance of the repository, internalizing the maintenance cost. However, neither federal institutes nor funding for external activities are immune from the vagaries of the federal budget cycle, and are at risk of short-term funding cuts.





# American Economic Association

Office of the Data Editor, Email: [dataeditor@aeapubs.org](mailto:dataeditor@aeapubs.org)

Alternative models see cost-recovery or user fees at the point of service, with such user fees being allowable on federal grants or other funding sources. An example of such a [pricing scheme](#) can be found for the French administrative data center (CASD). Such pricing schemes must balance the inequities that could be generated across the research landscape.

## G. Reuse

We believe tracking of data reuse is a key metric to incorporate into any repository. And yet, the current, mostly manually curated bibliographies and other metrics are an inefficient mechanism for doing so. Leveraging persistent identifiers (I.A.), encouraging simple metadata (I.C. and our recommendation 5), and using existing registry infrastructure should automate such processes. However, all such mechanisms are ineffective if researchers do not actually cite the data used. We thus suggest that federally funded researchers be required to cite data, and that this requirement be enforced and rewarded.

*Recommendation 6: Require data citations.*

Positive reinforcement can come from making data citations a measurable metric in federal funding. For instance, when grant outcomes are reported, automatic mechanisms, fed by data citations in researchers' publications, can populate reports automatically. Use of data citations in grant evaluations and "prior outcomes" would incentivize researchers to adopt and use data citations.

*Recommendation 7: Measure data citations in reporting mechanisms*

## References

- DataCite Metadata Working Group. 2017. "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.1." Edited by Jan Ashton, Amy Barton, Noris Birt, Stefanie Dietiker, Jannean Elliot, Martin Fenner, Wim Hugo, et al. <https://doi.org/10.5438/0014>.
- Desai, Tanvi, Felix Ritchie, and Richard Welpton. 2016. "Five Safes: Designing Data Access for Research." University of the West of England. <http://eprints.uwe.ac.uk/28124>.
- Martone, Maryann. 2014. "Joint Declaration of Data Citation Principles." Force11. <https://doi.org/10.25490/a97f-egyk>.
- Müller, Dana, and Joachim Möller. 2019. "Giving the International Scientific Community Access to German Labor Market Data: A Success Story." In *Data-Driven Policy Impact Evaluation: How Access to Microdata Is Transforming Policy Design*, edited by Nuno Crato and Paolo Paruolo, 101–17. Cham: Springer International Publishing.

To: Sean C. Bonyun,  
Chief of Staff, Office of Science and Technology Policy

Mr. Bonyun,

Thank you very much for the opportunity to respond to this Request for Comments on Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded or Supported Research. This is an important topic. Repositories will frame how scientific data are shared and reused (or not) and thus will determine how science is conducted in the 21st century.

To frame our response to this Request for Comments, we at Sage Bionetworks believe it is helpful to make the purpose of data repositories explicit. We want data repositories not merely because we want data to be available for reuse; we want data to actually be reused. The characteristics of data repositories should be determined by the practices of reuse we envision, not by the practices of sharing, *per se*.

Scientific data isn't like other kinds of data. Its reuse - and non-reuse - has been studied. Pasquetto, Randles, and Borgman (2017) make the case that the purpose of data sharing is data reuse, and that different scientific endeavors have different priorities and practices for data reuse. For example, "data policies that favor reproducibility may undermine data integration, and vice versa. Similarly, data policies that favor standardization may undermine exploratory research or force premature standardization."

As such, while Section I intends to identify "Desirable Characteristics for All Data Repositories," we do not recommend any one-size-fits-all characteristics for data repositories. While some characteristics may serve all needs, many others will be particular to different kinds of applications. Standards that do not recognize different needs for data reuse will not meet any of the needs particularly well. Below, we detail a few important reuse distinctions, and recommendations for each.

### **Foreground vs Background**

Rather than defining characteristics for open data repositories that are the same for all users, and all ways of reusing data, there should be different characteristics for different types of reuse. **We recommend that federal policy recognize two distinct kinds of research done with a scientific data repository - *background research* and *foreground research* - and support the growth of repositories in each class..**

Wallis, Rolando, and Borgman (2013) distinguish these two data uses thusly: "Foreground data are those that are the focus of research questions for a given study, whereas background data are those that provide context or calibration. The same data can be foreground to one

researcher and background to another, even on collaborating teams; the distinction is in the use of the data.”

Most data reuse is background rather than foreground. An investigator will typically search a database to see what is there, what has been done, to find the lay of the land. This kind of use is broad in scope and does not often result in new scientific findings. Rather, it is the substrate for new work. Foreground reuse - where reuse is focused and deep, to generate new scientific findings - is rarely done by downloading a data set and doing the work. Instead, the dataset is a social connector that leads to collaboration between the reuser and the data depositor.

Repositories should meet the needs of foreground research, to create new knowledge in emerging areas, and to be integrated into larger scientific paradigms. This means focusing on collaboration tools, local curation, and being flexible on standards to allow communities to experiment with reuse practices and develop norms over time.

Repositories should meet the needs of background research, to contextualize studies within larger paradigms, connecting the local knowledge of investigators into wider circles of the research community. This means focusing on query tools that enable search and discovery, and rigorously implementing tight standards for quality control.

Regardless of whether a repository’s principal goal is background or foreground, each repository should support enough standards (in data, science, and governance) to form connections to other repositories. Scientists may search in one repository in background mode, then move to another repository more tuned to foreground research.

Confusing the repository needs of background use with those of foreground use, or vice versa, would be a mistake. Moreover, making general standards in an attempt to satisfy both needs would hamper both the reproduction/replication of current research findings, as well as collaborative efforts to create and validate new research findings.

### **Centralized vs Decentralized Science**

Attending to the needs of foreground and background use, necessarily means attending to the needs of science conducted at different scales. **We recommend defining characteristics separately for repositories intended for *centralized science* and *decentralized science*.**

“Centralized science” is characterized by a small number of large investments in infrastructure, usually for observations at scale beyond a typical, hypothesis-driven research grant.

“Decentralized science” (a.k.a. “long tail science”) is characterized by a large number of independent, distributed research efforts, each producing their own unique material at low levels of investment.

We can understand centralized and decentralized science in terms of their foreground and background practices. The foreground needs for centralized science are often met by “observatories,” such as the the AllofUs Research Program for precision medicine, the Human Genome Project for genetics, NEON for ecology, or the Sloan Digital Sky Survey for astronomy. Observatories are funded to cover the costs not only of data collection, but of data preparation, documentation, and interpersonal negotiation.

These investments can be justified because there is high confidence that a large, defined research community (e.g. geneticists, ecologists, and astronomers) will use the types of data collected. And the types of data collected are those that are most useful to the most members of those communities. Because the data they produce are generally applicable, observatories and other repositories can also meet the background research needs of centralized science. These centralized repositories have more institutional support over time than many other repos, and as such, should support rigorous standards to allow them to work also as “search engines” where researchers can find datasets and connect to their creators. Centralized repositories therefore need to support both background and foreground research.

In decentralized science, such as the R01 grant in medical science, or similar research grants in qualitative social science and geography-specific environmental science, investigators cannot rely on observatories for foreground research because their data needs are various and sundry. While they often use data from observatories or repositories for their background research to interpret their findings, their data needs are too particular and unpredictable to warrant such a large investment. Decentralized science Investigators collect limited batches of data themselves that, in aggregate, actually create the bulk of all research data. Moreover, decentralized science investigators tend not to share their data broadly because, as individuals or small teams, they cannot bear the costs of creating and maintaining repositories for datasets with a low probability of reuse.

There is, therefore, a great opportunity in creating repositories specifically for “decentralized science” to use for foreground research. This is where the tragedy of the commons in open data is greatest. Such repositories should be cheap and easy to use for a diversity of investigators, and that enable the ingest, discovery (as a tie to background use repositories), integration, and analysis of diverse datasets. These decentralized repositories could live on a small set of platforms, and serve primarily as a place where context can be higher amongst data users in communities. Researchers would thus navigate between centralized and decentralized repositories as their data use takes them from background to foreground research.

### **Local vs Global Disciplines**

In addition to accounting for foreground and background uses, and science at different scales, data repository characteristics will need to account for the cultures of different disciplines, as well. **We recommend defining new characteristics to enable the interaction among repositories, particularly between the disciplinarily *local* and *global*.**

The scientific community cannot and should not be treated as a singular thing. There are many scientific communities across many domains. Each has different practices of community review and benchmarking to validate the research findings of their peers, and has different cultures surrounding data curation, the use of metadata, and establishing provenance, as well as of validation. As such, the characteristics of repositories will - and should - depend on the scientific community being served.

Although local characteristics are essential to let scientific disciplines function in their own unique ways, it's essential that global search, referencing, and partnering is supported. As such, some repository characteristics must also acknowledge and support how disciplines are related - some more closely and others more distantly. Each discipline has its own context, which can have unexpected connections to the context of other disciplines (image recognition in computer science now supports diabetic diagnosis in medicine) - federal policy should recognize and support this kind of deep, cross-disciplinary linking.

### **Hub-and-spoke policy models**

While we should aim for repositories that satisfy unique contextual needs, we also do not want repositories that create fragmentation among disciplines, or that entrench existing divisions in the greater landscape of scientific knowledge. Instead, we want repositories that let local norms of data sharing and reuse coexist with global norms, and let emerging communities experiment with sharing practices while legacy communities maintain traditional practices. Thus, we need repositories that function simultaneously at several levels. Indeed, what we need is not just a batch of random repositories, but a network of repositories, each of which draws on the various characteristics of centralized/decentralized, local/global, and more. This interconnection of repositories will best support ongoing federal investments in scientific data.

The modern air traffic network can serve here as a metaphor. "Hub-and-spoke" can describe the network of data and references that support validity claims across disciplines within a scientific paradigm. The major "hubs" in our network will support the most commonly used data from observatories in big science (such as the Cancer Genome Atlas). But there must also be "spokes" to small science at the local level (such as teams at two laboratories collaborating with data from the Cancer Genome Atlas). In a system where big science repositories at hubs have different characteristics from small science repositories at spokes, both must have additional characteristics that enable them to be used together. Hubs and spokes must be able to handoff to each other, so that investigators can navigate the data space, just as an airline passenger can navigate a multi-flight trip.

### **Universal vs Variable Standards on Each of the Characteristics**

Given the discussion above, we offer the following assessment of each of the proposed characteristics with respect to foreground and background use. Some can be safely understood as essential to all needs. Others, per the discussion above, are best understood as project-specific, per the needs of different communities and different situations.

	Background use	Foreground use
<b>I. Desirable Characteristics for All Data Repositories</b>		
A. Persistent Unique Identifiers	Essential	As necessary to connect to other data
B. Long-term sustainability	Essential	Project-dependent
C. Metadata	Essential - standardized	Essential - blend of external standards and internal project metadata
D. Curation & Quality Assurance	Essential - standardized	Essential - very project and data-type specific
E. Access	As open as possible, while protecting privacy	Project-specific
F. Free & Easy to Access and Reuse	Essential	Project-specific
G. Reuse	Essential	Project-specific
H. Secure	Essential	Essential
I. Privacy	Essential	Project-specific
J. Common Format	Essential	Project-specific
K. Provenance	Essential	Project-specific
<b>II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)</b>		
A. Fidelity to Consent	Essential	Essential
B. Restricted Use Compliant	Depends on governance type - download v sandbox	Essential
C. Privacy	Essential	Project-specific
D. Plan for Breach	Essential	Essential
E. Download Control	Depends on governance type - download v sandbox	Project-specific
F. Clear Use Guidance	Essential	Essential
G. Retention Guidelines	Essential	Essential
H. Violations	Essential	Essential
I. Request Review	Depends on data type	Essential



# CoreTrustSeal Board Response to the OSTP Draft Desirable Characteristics of Repositories for Managing and Sharing Data

**Responding organization:** CoreTrustSeal Standards and Certification Board (<https://www.coretrustseal.org/>)

**Domain:** Digital preservation

**Role:** Certification Body

The CoreTrustSeal Standards and Certification Board appreciates the opportunity to provide comments to the Draft Desirable Characteristics of Repositories for Managing and Sharing Data. The CoreTrustSeal began as a World Data System–Data Seal of Approval effort under the Research Data Alliance to develop a core set of trustworthy data repository (TDR) requirements in response to stakeholder demand for a single, common, community standard and process with a low barrier to entry. The CoreTrustSeal Board and community welcomes all work to further define the desirable characteristics of repositories and seeks to engage with such processes whenever possible. The Board would also welcome feedback on the CoreTrustSeal TDR Requirements<sup>1</sup>, which remain community driven and subject to ongoing review and approval.

We are pleased that the proposed characteristics align strongly with the content of the CoreTrustSeal TDR Requirements. In our response, we seek to identify those alignments, comment on the characteristics as provided, and clarify any cases where the specifics are out of the current scope of the CoreTrustSeal.

## Overall comments

In addition to the direct responses below, there are one or two overall comments and a suggestion for an additional characteristic.

L. *Evidence of alignment.* Provides public evidence (documentation) sufficient to demonstrate alignment with the desirable characteristics.

As a part of the certification process and to avoid resource intensive site visits, the CoreTrustSeal requires that applicant self-assessed responses to the Requirements are supported by documented links to evidence on the web. This transparency of evidence helps ensure that the claims made by repositories are visible to their peers, as well as to the reviewers. The benefit here is that publicly visible evidence not only helps ensure honesty, but also makes available to the community a valuable resource of repository practice information. Some clarity on the level of expected evidence to support the desirable characteristics would be valuable to repositories and their depositors and users. The

<sup>1</sup> CoreTrustSeal Standards and Certification Board. (2019, November 20). CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 (Version v02.00-2020-2022). Zenodo. <http://doi.org/10.5281/zenodo.3638211>.



evidence sought for the CoreTrustSeal TDR Requirements is not onerous and seeks to align closely with what a repository would need both to offer a consistent, high-quality service and to ensure service sustainability.

We note also that while B. *Long-term sustainability* implies the sustainable availability of the dataset, there is no explicit statement about the ongoing curation of the data (including ongoing access through format migration or emulation) or metadata (including updates to meet evolving metadata standards and the changing needs of data users). In addition to minimizing the risk of unintended change through integrity measures, sometimes referred to as bit-level preservation, many communities seek more active preservation of resources. These might include deposit, storage, and access functions provided through a generalist repository that retains rights to address changes to formats or general metadata (for discovery, identification, reuse, etc.). A further tier of care is provided by disciplinary or domain repositories with the expert knowledge necessary to offer specialist deposits, curation, access, and reuse services. Research data funders, creators, and users may benefit from a range of curation and preservation levels, but the level of care a dataset will receive should be transparent to all.

In addition to the above, we would like to encourage that repositories provide information about their governance and expert input (see, for example, CoreTrustSeal Requirements R1. Mission, R5. Organizational Infrastructure, R6. Expert Guidance). This can help demonstrate how repositories ensure the maintenance of documentation and evidence relevant to their mission, and how they support managed change over time. The information is also important to support cooperation and interoperability.

## I. Desirable Characteristics for All Data Repositories

A. *Persistent Unique Identifiers*: Assigns datasets a citable, persistent unique identifier (PUID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The PUID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.

**CoreTrustSeal Board comment:** Aspects pertaining to this characteristic are addressed in CoreTrustSeal Requirement R13. Data discovery and identification.

“Accession number”: If there is a set of agreed criteria for an accession number/system that would align with the criteria for a PID, this could be mentioned here.

“Persistent landing page”: As written it is suggested that a “landing page” is mandated. Landing pages are a common choice of implementation, whether because there is some additional barrier to access the data (e.g., authentication/authorization) or to present the user with metadata about other versions of a dataset. So, in the absence of a standard landing page design, they can present an extra step or even a barrier to access (e.g., not supporting machine accessibility). This has been noted in work to develop automated tests of FAIRness. A more neutral statement might be to say that the PUID consistently resolves to an informative target even in the absence of the original dataset. This leaves room for data





stewards to link directly to a dataset when this is the desired functionality.

It may be worth pointing out explicitly that the data repository must have the capabilities for maintaining the persistent unique identifier to ensure that the PUID continues to point to the correct location of the dataset, even if this changes. This seems particularly relevant in cases where only an internal identifier such as an accession number is used.

*B. Long-term sustainability.* Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.

**CoreTrustSeal Board comment:** Aspects pertaining to this characteristic are addressed in CoreTrustSeal Requirements R3. Continuity of Access, R7. Data integrity and authenticity, R10. Preservation plan, and R16. Security.

This criterion appears to address different dimensions of sustainability: technology (including capabilities for both routinely maintaining bitstreams and disaster recovery) and financial/organizational measures to ensure that data remain available in the long term. In the experience of the CoreTrustSeal Board, these dimensions sometimes tend to be confused or to be regarded as interchangeable rather than complementary. Therefore drawing attention to the fact that, for example, sound technology without a long-term financial and organizational commitment is not sufficient for long-term sustainability may be helpful. It may also be useful to consider moving technological aspects to H. *Security*.

“Long-term”: We suggest specifying what is meant by “long-term”; for example, by pointing to an agreed definition such as the one from the OAIS Reference Model.

For repositories that are looking to obtain the CoreTrustSeal, we are seeking evidence that they have the ongoing ability to take preservation actions in response to changes to technology or to the needs of the user community. A sustainable repository can then provide assurance that it can repeatedly curate the data beyond the next round of change. This is an aspect not clearly addressed here. To clarify the scope of the suggested characteristic, we therefore suggest specifying what is meant by “availability of datasets”: Is it the availability of the “original” bitstream in unchanged form, or does it entail measures such as format migrations to ensure that the data remain usable despite technological change. You may want to refer to Requirement 10 “Preservation Plan” of the CoreTrustSeal TDR Requirements.

*C. Metadata:* Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.

**CoreTrustSeal Board comment:** Aspects pertaining to this characteristic are addressed in CoreTrustSeal Requirements R11. Data quality, R13. Data discovery and identification, and R14. Data reuse.



The effective use of metadata is implied throughout the CoreTrustSeal, with references from relevant Requirements including appraisal, reuse and discovery (including citation). The adoption of metadata schemas that are supported by the user community are encouraged, but as yet there is neither a clear mechanism to define an acceptable community standard, nor a clear reference (e.g., registry) to record them as such. This may negatively affect the usefulness of this criterion.

D. *Curation & Quality Assurance*: Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

**CoreTrustSeal Board comment:** Aspects pertaining to this characteristic are addressed in CoreTrustSeal Requirements R8. Appraisal, R10. Preservation plan, and R14. Data reuse. In addition, see R0. Background information for a definition of curation levels.

We suggest taking into consideration that not all curation levels may include curating for the accuracy of dataset content. It may be helpful to differentiate between technical quality (standardization and compliance with norms) versus research quality assessment and action, which ideally should be carried out following the guidance of the community served.

The CoreTrustSeal Requirements (R11. Quality, in particular) also demand that a repository employs means for users to access information about data quality and other documents describing aspects of the data to support users.

Given the cost of curation and preservation, and the limited resources available to repositories, we strongly encourage that repositories should perform an appraisal function. Appraisal, following a set of selection criteria also conveyed to the depositors, is needed to ensure that the data is within scope of a particular repository and that the data has value to the community the repository serves. See CoreTrustSeal TDR Requirement R8. Appraisal for more on this aspect.

E. *Access*: Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.

**CoreTrustSeal Board comment:** Aspects pertaining to this characteristic are addressed in CoreTrustSeal Requirements R2. Licenses, and R4. Confidentiality/Ethics.

The aspect of Open Access is largely beyond the remit of the CoreTrustSeal and therefore not covered in the CoreTrustSeal TDR Requirements. However, both E. and F. can be considered subsets of "Rights Management", including not only considerations of privacy and confidentiality, but also of Intellectual Property Rights (c.f., R2. Licenses and R4. Confidentiality/Ethics).

Echoing CoreTrustSeal Requirement R1. Mission/Scope, we also recommend that the repository should have a publicly accessible policy stating it provides access to open data that are only restricted for legal and ethical reasons. This ensures that users and depositors are made aware of the repository's commitment to Open Science.

While outside the scope of CoreTrustSeal, evaluation E. and F. touch on issues covered in



the “WDS Data Sharing Principles” (<https://www.icsu-wds.org/services/data-sharing-principles>). These, for example, also explicitly encourage that repositories enable international reuse of data.

**F. *Free & Easy to Access and Reuse*:** Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

**CoreTrustSeal Board comment:** See response to E. above. In addition, it may be helpful to specify what is meant by “easy”.

**G. *Reuse*:** Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).

**CoreTrustSeal Board comment:** Aspects pertaining to this characteristic are addressed in CoreTrustSeal Requirement R13. Data discovery and identification.

We suggest that this is renamed “Monitoring Reuse” or similar, rather than only “Reuse”. It does not address the capability of the dataset to be processed by software and understood by researchers, but instead the repository functions for monitoring impact, such as download statistics, altmetrics, or citation tracking.

To facilitate monitoring, we recommend for repositories to provide a suggested citation for the dataset that includes a persistent identifier. Users can then easily record the use of the dataset in published reports.

**H. *Secure*:** Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>) or the National Institute of Standards and Technology's 800-53 controls (<https://nvd.nist.gov/800-53>).

**CoreTrustSeal Board comment:** Aspects pertaining to this characteristic are addressed in CoreTrustSeal Requirement R16. Security.

From the experience of the CoreTrustSeal Board, ISO standards can present quite a high bar for many repositories, which can be difficult to meet and which also may not always be necessary depending on the types of data curated. We therefore suggest replacing “such as” with “for example” to express stronger optionality.

We additionally recommend that the repository should provide documentation on how it protects its infrastructure to ensure continuity of service.

**I. *Privacy*:** Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.



**CoreTrustSeal Board comment:** Aspects pertaining to this characteristic are addressed in CoreTrustSeal Requirements R4. Confidentiality/Ethics, and R16. Security with a dependency on R5. Organizational infrastructure.

J. *Common Format:* Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.

**CoreTrustSeal Board comment:** Aspects pertaining to this characteristic are addressed in CoreTrustSeal Requirement R14. Data reuse.

We suggest taking into account here that a format commonly used in a community may not always be “standards-compliant” and/or non-proprietary. We recommend expanding the statement to require that data, metadata, and documentation should be provided in formats that are consistent with the needs and practices of the community served.

K. *Provenance:* Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.

**CoreTrustSeal Board comment:** Aspects pertaining to this characteristic are addressed in CoreTrustSeal Requirement R7. Data integrity and authenticity.

Important provenance information should also be made available to repository stakeholders. For this purpose, repositories also should employ version control of data, with naming conventions that communicate when a new version of the dataset has been released. We recommend in addition that pre-deposit provenance should be sought and available with the dataset, including information about the collection and processing of the dataset, prior to ingest.

“Logfile”: As this suggests a specific implementation of collecting provenance, it might be stated in a more abstract way (e.g., “methodology to record” or “audit trails”).



## II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)

**CoreTrustSeal Board comment:** As “de-identification” and “anonymization” are not always considered synonymous<sup>2</sup>, we suggest referring to both terms here.

A. *Fidelity to Consent:* Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition)

**CoreTrustSeal Board comment:** This aligns with part of R4. Confidentiality/Ethics, which includes the guidance that “the repository should ensure that data collection or creation was carried out in accordance with legal and ethical criteria prevailing in the data producer’s geographical location or discipline”. The clarification of these rights falls under R2. Licenses.

B. *Restricted Use Compliant:* Enforces submitters’ data use restrictions, such as preventing reidentification or redistribution to unauthorized users.

**CoreTrustSeal Board comment:** For the CoreTrustSeal, the rights issues would be clarified through R2. Licenses. Evidence for other CoreTrustSeal Requirements is expected to align with the permissions, obligations, and prohibitions defined by those rights. Other protective measures would be applied through R16. Security. The repository’s responsibility to ensure that users understand and follow requirements for the use of restricted data also aligns with R4. Confidentiality/Ethics. The repository should also communicate to users of restricted data how best to ensure they are protected.

C. *Privacy:* Implements and provides documentation of security techniques appropriate for human subjects’ data to protect from inappropriate access.

**CoreTrustSeal Board comment:** The title here of “Privacy” (duplicative of I. *Privacy* above), has a number of overlaps with, and distinctions from “Confidentiality” in some geographic/legal areas. But, the content is primarily focused on information security.

CoreTrustSeal applicants holding human data would identify this under R4. Confidentiality/Ethics, document it through R2. Licenses, and demonstrate information security measures through R16. Security. Threats to confidentiality are not static, so as part of sustainability and preservation, a repository also should document how it maintains capabilities to ensure privacy as technology evolves.

<sup>2</sup> See, for example: Fullerton SM, Anderson NR, Guzauskas G, Freeman D, Fryer-Edwards K. Meeting the governance challenges of next-generation biorepository research. *Sci Transl Med.* 2010 Jan 20;2(15):15cm3. doi: 10.1126/scitranslmed.3000361. PMID: 20371468; PMCID: PMC3038212.



D. *Plan for Breach*: Has security measures that include a data breach response plan.

**CoreTrustSeal Board comment:** The need to plan and ideally conduct periodic tests for breach scenarios is valuable for all data collections, not only with those containing human subject data. This also applies to items E to H below, which are valuable measures to have in place for all digital assets. To restrict this recommendation to 'human' data might devalue the investment repositories make in these matters for all of their data. It is also important to note that non-human data, such as ecologically or culturally sensitive geographic data, is at risk without these measures.

E. *Download Control*: Controls and audits access to and download of datasets.

**CoreTrustSeal Board comment:** Provision of Access as a function is stated in CoreTrustSeal TDR Requirement R1. Mission. Control is implied through the application of R2. Licenses, and technical measures for Authentication and Authorization are mentioned in R16. Security. CoreTrustSeal would expect more rigorous access control for personal data, but appropriate technical and organizational measures to ensure that data can only be accessed and used in accordance with the stated license would have to be in place for all data.

F. *Clear Use Guidance*: Provides accompanying documentation describing restrictions on dataset access and use.

**CoreTrustSeal Board comment:** Conditions would be included under CoreTrustSeal Requirement R2. Licenses and communicated to users at the point of reuse. The license and any accompanying documentation should describe the conditions of use for any data, regardless of whether the data contain personally identifiable information (PII). But, it is especially critical for data that contain PII.

G. *Retention Guidelines*: Provides documentation on its guidelines for data retention.

**CoreTrustSeal Board comment:** It is not clear whether this applies to a general retention period (e.g., minimum number of years) or rules for re-appraisal of collections over time to make disposition/retention decisions. This would apply to all types of data in a collection, though for personal data there may be additional criteria to define the circumstances under which a dataset, or data related to one or more individuals, might be withdrawn; for example, in response to a request from a data subject.

H. *Violations*: Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository.

**CoreTrustSeal Board comment:** The guidance for CoreTrustSeal Requirement R2. Licenses includes a request for "Documentation on measures in the case of noncompliance



with conditions of access and use”, while R4. Confidentiality/Ethics asks whether there are “measures in place if conditions are not complied with?”

I. *Request Review*: Has an established data access review or oversight group responsible for reviewing data use requests.

**CoreTrustSeal Board comment:** This is not explicitly requested by the CoreTrustSeal Requirements, but would usually form part of the license and rights negotiation at the point of deposit. The Requirements could usefully add a specific question about this process for applicants curating personal (or otherwise sensitive) data.



# Office of Science and Technology Policy (OSTP)

Request for Information: Public Access to Peer-Reviewed Scholarly  
Publications, Data and Code Resulting from Federally Funded  
Research

**Submitted To:**

Lisa Nichols, Assistant Director for Academic Engagement

**Tyler Point of Contact:**

Contact Name: Michael Donofrio

Phone: 703-403-3373

Email: [michael.donofrio@tylertech.com](mailto:michael.donofrio@tylertech.com)



Socrata is the national leader in software-as-a-service (SaaS) for self-service data management, analytics, and information sharing for governments, with over 400 customers Nationwide. Socrata is the flagship solution in the Data & Insights division of Tyler Technologies, the largest software company in the U.S. exclusively focused on software for the public sector.

We power some of the largest data sharing and analytics platforms across Federal, State, County, and City governments including:

- Federal - Department of Transportation, Department of Commerce, Department of Veterans Affairs, US Agency for International Development, Centers for Medicare and Medicaid Services (CMS), Centers for Disease Control and Prevention (CDC), NASA, etc.
- 31 States - California, Texas, Washington, New York, Pennsylvania, Maryland, Michigan, etc.
- Counties - Los Angeles, San Diego, King (WA), Fulton (GA), Montgomery (MD), etc.
- Cities - New York, Chicago, Los Angeles, San Francisco, Seattle, Dallas, Miami, Austin, etc.

We look forward to collaborating and adding value to this important asset; data. Below is our response to the questions:

What current limitations exist to the effective communication of research outputs (publications, data, and code) and how might communications evolve to accelerate public access while advancing the quality of scientific research? What are the barriers to and opportunities for change?

In our opinion, the biggest barrier to change is rationalizing how to effectively consolidate a massive cobweb of distributed mechanisms for finding and accessing research outputs, that so many depend upon every day, and cost the government billions, yet impedes our collective ability to advance research.

Government has created countless mechanisms for finding and accessing research outputs both in public and secure environments. Some mechanisms are good, some bad, and some non-existent. Providing a common catalog and metadata to index all the existing mechanisms, and the research outputs within them, would greatly improve discovery.

Once a user discovers the research outputs they are looking for, there remains opportunities to improve their ability to access the publications, data, and code. Data is commonly locked behind query tools, presented as text in a website, or embedded in a PDF table or chart. This makes access to and reuse of data inconsistent, time consuming, and often times impossible. This approach is also costly to government to maintain all the search and query tools that are preventing users from accessing the raw data.

Research outputs should be discoverable in a machine-readable way and leverage application programming interfaces (API) to facilitate search across all the distributed mechanisms. Government should leverage a common catalog that can securely govern access to research outputs, or appropriately redacted versions thereof, for diverse stakeholders including programs, internal teams, other government organizations, grantees, research partners, private sector, the public, and others. Users should be able to access all data, not just the filtered results of query, in an interoperable API format.

Stakeholders should be able to leverage the research outputs, and in particular the data, to continue the effort to improve the quality of scientific research. Future stakeholders should be

enabled with capabilities to connect their analytical tools of choice to API-enabled data, and reduce the time and cost of accessing, replicating, normalizing, transforming, joining, and storing data. Stakeholders should also be able to contribute new data that will be used in their work and leverage platform API's to make the data interoperable to power their analytical tools of choice. This would reduce the costs stakeholders incur for using and managing data today.

Additionally, stakeholders should be enabled to build and submit their research outputs in a secure, collaborative, yet controlled manner. Providing an intuitive interface to create interactive and machine-readable reports can replace the current proliferation of PDFs that lock data and insights away.

We can overcome these barriers with Socrata by leveraging the existing dissemination mechanisms, making the research outputs discoverable in a central catalog, and do so quickly with technology that's already proven across all levels of government. Over time we can supplement or replace those dissemination mechanisms that don't work and expand the utility for offering stakeholders a collaboration space to build and submit their research outputs.

- What more can Federal agencies do to make tax-payer funded research results, including peer-reviewed author manuscripts, data, and code funded by the Federal Government, freely and publicly accessible in a way that minimizes delay, maximizes access, and enhances usability? How can the Federal Government engage with other sectors to achieve these goals?

In the short-term, leveraging a common catalog and metadata to index and search all the existing mechanisms, and the research outputs within them, would greatly improve discovery and access to existing resource outputs. Usability would not change and continue to be an impediment.

Over the longer-term, the opportunities to create incremental efficiencies in the end-to-end process will minimize delay, maximize access, and enhance usability; systemically. Leveraging Socrata from end-to-end will create efficiencies throughout the process enabling stakeholders to create and contribute their research, government stakeholders curate, redact and govern, then disseminate research outputs back to stakeholders in a controlled manner.

- How would American science leadership and American competitiveness benefit from immediate access to these resources? What are potential challenges and effective approaches for overcoming them? Analyses that weigh the trade-offs of different approaches and models, especially those that provide data, will be particularly helpful.

Leveraging the outputs of past research for subsequent efforts will propel innovation forward and improve American competitiveness. We suspect many of the challenges to be cultural and contractual related to the ownership, governance, and reuse of data.

The initial challenge, that can be solved quickly, is deploying a consolidated catalog to improve the discovery of existing resource outputs. This would immediately make it easier to find resource outputs in a consistent manner and leverage existing mechanisms for access and usability.

The mid-term challenge is assessing and retroactively improving accessibility and usability of existing research outputs and mechanisms. The scope of this is monumental, so it makes sense to undertake this phase incrementally. Converting Excel files to Socrata datasets will provide for APIs quickly. Extracting data from PDF documents will take much more time. It makes sense to prioritize where existing resource outputs should be made more accessible and usable.

In parallel, we would look to deploy capabilities for stakeholders to create and contribute their research outputs, and associated data, to the platform. This would provide an optimal scenario for new research outputs to comply with any new requirements and provide for clean data from end-to-end.

- Any additional information that might be considered for Federal policies related to public access to peer-reviewed author manuscripts, data, and code resulting from federally supported research.

We thought it might be helpful to see an example of some of our relevant work with the US Agency for International Development. Here is an excerpt from their launch in November 2018:

*The [Development Data Library](#) (DDL) is USAID's publicly available repository for Agency-funded data-on-demand. As a best practice digital archive, the new platform strives to preserve and accelerate the re-use of valuable data to advance international development and improve program development and performance.*

*Actively managed by a staff of data curators, USAID's new DDL is a true data repository, suited for internal Agency analytics as well as sharing with the general public. New features in the DDL can be used to visualize data, download in its raw form, track changes over time, or create dynamic connections via an Application Programming Interface (API) to filter, query, and aggregate data.*

*There is an immense richness in the data collected by USAID partners around the world, and this data holds the potential to improve the lives of some of the world's most vulnerable people. When a development project ends, the data can yield new insights for years or decades into the future. Rather than risk losing access to this data, USAID partners and staff upload their data to the DDL, ensuring its preservation and making it easier to discover, share, and reuse this data over time.*

We look forward to collaborating with your team to find opportunities to expand research and provide an American competitive advantage.

Kind regards,

Michael Donofrio  
Sr. Advisor Federal Solutions  
Phone: (703) 403-3373  
Email: michael.donofrio@tylertech.com

March 13, 2020

Lisa Nichols, PhD  
Assistant Director for Academic Engagement  
National Science and Technology Council  
Subcommittee on Open Science

RE: Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research [FR Doc. 2020-00689]

Transmitted electronically via e-mail: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

Dear Dr. Nichols,

The Federation of American Societies for Experimental Biology (FASEB) appreciates the opportunity to provide feedback on [Request for Comments](#) (RFC) seeking input on a draft set of desirable characteristics of data repositories used to locate, manage, share, and use data resulting from federally funded research released on January 17, 2020. As a coalition of 28 biological and biomedical scientific societies collectively representing over 130,000 individual scientists and engineers, FASEB recognizes the critical role of data preservation and accessibility in facilitating scientific rigor and reproducibility.

Our comments in response to the draft characteristics posted by the Office of Science and Technology Policy (OSTP) in this RFC reiterate FASEB positions on core issues of interest to our members: coordination and harmonization of data policies across federal agencies, long-term sustainability of data management, and data accessibility while upholding the standards of scientific peer review.

## **I. Desirable Characteristics for All Data Repositories**

### *A. Persistent Unique Identifiers:*

To ensure that large volumes of data are of the greatest potential utility to researchers, clinicians, and the public, FASEB supports the use of unique identifiers. Consistent with the [FAIR principles](#) (Wilkinson et al., 2016), identifiers such as digital object identifier (DOI), accession numbers, or ORCID ID will aid in researchers' ability to identify and access data even if the metadata URL has changed since its publication. Potential efforts OSTP may want to consider

include: (1) developing tools to improve search functions and the aggregation of data, and (2) creating formatted citations associated with each dataset, preferably including a DOI. These improvements can also incentivize researchers to share quality data. Greater reuse and citation of datasets will encourage investigators to optimize the formatting and organization of their data and metadata for reuse by others, rather than merely fulfilling minimal reporting requirements.

Successful implementation of interoperable data management practices will require training for all research team members. Institutions should also foster an atmosphere where quality data management and appropriate data sharing are standard practice. To establish and maintain such an environment, institutions should encourage investigators to collaborate on improving data practices within their discipline and ensure data management resources can be easily identified and utilized.

### *B. Long-term sustainability*

Responsible data stewardship requires a long-term plan. Data management plans (DMPs) are an important tool for promoting quality data management and appropriate data access.

Consideration of potential opportunities for data reuse at project initiation also ensures retention of all appropriate data. Inclusion of DMPs as a component of grant applications clarifies expectations between investigators and research sponsors. Flexibility and adaptability can be achieved by having individual investigators develop a DMP specific to their research area, data types used, and resources available. Research sponsors may also enlist DMPs for secondary uses of benefit to the research community, such as identifying common resource needs and other barriers.

To attain the benefits of DMPs without creating unnecessary burden, DMPs should be short summary documents that address the most essential aspects of data management and access. In most cases, a brief (one-to-two pages) summary should be sufficient, although additional information could be requested just-in-time for select circumstances. FASEB recommends the following DMP content requirements across federal agencies:

- Description of the data and metadata to be collected
- Overview of data management practices
- Summary of any data sharing restrictions (confidentiality, intellectual property, etc.)
- For shared data, information about when it will be made available, where it will be stored, how it will be maintained, and how others will be able to find, access, and reuse it
- For data that will not be shared, justification for not making it accessible (which many include considerations of feasibility, data utility, etc. as well as sharing restrictions)

### *C. Metadata*

Research reproducibility depends upon rigorous experimental design and appropriate analysis of resulting data. Metadata provide essential information for determining appropriate use.

Unfortunately, robust, consensus-based metadata standards do not exist for many fields or many data types. Furthermore, minimal metadata standards have not been established or deployed

across all scientific agency databases. Therefore, FASEB encourages OSTP to support the development of community-based metadata standards. Scientific societies can support these efforts by identifying and convening subject matter experts and disseminating consensus standards. We also urge OSTP to foster trans-agency development of automated tools for assigning metadata to files and datasets. Development of these tools can begin before or in parallel with the establishment of consensus standards. Automation would streamline efforts associated with tracking and updating metadata to meet current standards, accelerating adoption of new standards and changes to existing standards reducing investigator burden.

Repository tools are also indispensable for promoting data citation and attribution to investigators responsible for generating datasets. Data citation enhances the findability and accessibility of datasets and incentivizes data sharing. Currently, tools supporting citation of journal articles are more robust and readily available than tools for data citation. If researchers must look up a new citation format and manually assemble citation information, they will cite the associated journal article because it is simpler and more expedient. Tools that export dataset information, similar to what is provided for articles indexed in PubMed, lower the “activation energy” for data citation and provide a visible reminder to do so. To further promote such recognition, OSTP may want to consider collaborating with scientific journals to develop manuscript submission tools that prompt, facilitate, and standardize reporting of repository use.

#### *F. Free & Easy Access and Reuse*

FASEB understands and supports the development of an IT ecosystem that facilitates access to large, high-value datasets, as this will ensure these datasets are consistent with FAIR principles.

To effect positive change, research sponsors must carefully balance the costs and benefits of data access when developing and amending policies. Making datasets accessible – including the skilled human labor necessary to prepare and maintain data and metadata, technological infrastructure, and continued development of effective search platforms – is costly. Some datasets have little value for reuse or a short “shelf-life”; requirements to share and preserve such data could create inefficiencies in research funding and resource distribution. Therefore, FASEB recommends that sponsors ensure data access policies prioritize data with the highest potential for reuse

#### *G. Reuse*

The diversity of data types, research areas, and resources available make it challenging to identify data accessibility strategies that are practical and relevant for all fields of research, challenges that are further amplified within the biological sciences. Regular assessment of data utilization will allow investigators and federal agencies to evaluate usage and outcomes in the context of past performance and project future needs. Such utilization assessments would be further enhanced by the creation of time series data, when feasible. Analysis of user communities may also reveal patterns in how usage expands to new disciplines, thus informing scientific programs at federal agencies.

### *J. Common Format:*

Data standards are necessary to ensure adherence to the FAIR principles; without standards, large volumes of data cannot be reused or even reassessed. Several issues that may hinder users from submitting data include limited data formats, heavy reliance on manual entry, and insufficient tools available to export and import data and metadata.

To encourage deployment of user-friendly platforms FASEB recommends coordinating with funding agencies such as NIH and NSF to develop metrics that evaluate and offer guidance about such barriers. Additionally, FASEB encourages OSTP and colleagues to measure the extent to which automation is incorporated in the submission process. Automated features such as auto-fillable fields and saved templates can enhance the submission experience and circumvent several sources of data corruption and loss.

### *K. Provenance:*

Understanding the context by which data is obtained, processed, and analyzed is essential to its appropriate interpretation and application. Because datasets are often reformatted to pursue new research inquiries, data provenance allows researchers to trace newly designed or repurposed data back to their original settings.

Implementation of strong data provenance ensures data creators are held accountable for their work and enables systematic data tracking for a wide range of scenarios that utilize and apply research data. For example, researchers frequently share and adapt data for their individual purposes when collaborating with fellow investigators on research projects. With clear data provenance guidelines, end-users will be able to visualize how a specific dataset was derived and thus more appropriately employ the information that is suitable for their research.

FASEB supports responsible data management and encourages OSTP to engage with the stakeholder community to incorporate data provenance best practices across federal agencies.

### *L. Other relevant topics*

The emergence of “big data” is allowing investigators to pursue more lines of inquiry that could ultimately lead to transformative discoveries. However, as larger quantities and more types of data can be combined in new ways, we must also be cautious of spurious correlations and “over-mining” of datasets. The Federation is concerned that analytical methods and tools do not always keep pace with research opportunities. Rigorous research practices will depend on coordinated efforts among federal agencies, and research stakeholders, ranging from single investigators to large institutions, to generate and support “big data” analytical methods and best practices.

FASEB encourages OSTP to take the lead in coordinating these efforts to ensure parity across agencies and scientific disciplines.

## **II. Additional Considerations for Repositories Storing Human Data (Even if De-identified)**

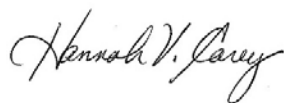
### *C. Privacy*

Current U.S. policy frameworks and privacy proposals are insufficient to ensure the privacy of human research subjects in perpetuity. In comments on the [NIH Genomic Data Sharing Policy](#), FASEB stated that “de-identification cannot be guaranteed for certain types of data, including whole genomic sequences.” FASEB, therefore, recommended the consideration of alternative models to protect human research subjects, such as shifting from a privacy-protection paradigm to “one that provides research subjects with substantive legal protections against the misuse of or inappropriate access to their data.”

OSTP should also consider the risk of harm from inaccurate re-identification or speculation of the identities of participants and their outcomes. There are many other types of data misuse, and OSTP must proactively work with federal agencies and the research community to mitigate these risks.

FASEB appreciates the opportunity to provide input on this important topic. In addition to the comments provided in response to the specific elements of this RFC, links to recent organizational statements on this issue are provided below the signature line. We look forward to working with OSTP, federal research agencies, and other stakeholders on development of a feasible strategy to foster data sharing and reuse across scientific disciplines.

Sincerely,



Hannah V. Carey, PhD  
FASEB President

### **Related FASEB Statements of Interest**

1. [FASEB Comments in response to NIH Request for Information \(RFI\) on Draft Data Management and Sharing Policy and Guidance Documents](#) (Issued December 10, 2019)
2. [FASEB Comments on Draft NIH Strategic Plan for Data Science](#) (Issued April 4, 2018)
3. [FASEB Response to NIH RFI, “Registration and Results Reporting Standards for Prospective Basic Science Studies Involving Human Participants”](#) (Issued November 8, 2018)
4. [FASEB Response to NIH RFI, “Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research](#) (Issued December 10, 2018)
5. [FASEB Comments on Next-Generation Data Science Challenges in Health and Biomedicine](#) (Issued November 8, 2017)
6. [FASEB Statement on Data Management and Access](#) (Issued March 1, 2016)
7. [FASEB Response to NIH RFI: Metrics to Assess Value of Biomedical Digital Repositories](#) (Issued September 7, 2016)
8. [Comments on NIH RFI: Strategies for NIH Data Management, Sharing, and Citation](#) (Issued December 7, 2016)



**TO:** Office of Science and Technology Policy (OSTP)  
**DATE:** March 12, 2020  
**RE:** RFC Response: Desirable Repository Characteristics

**Primary scientific discipline**

North Carolina State University's research enterprise is broad and interdisciplinary, encompassing, among other areas, a wide range of genomics, health, and life sciences disciplines such as bioinformatics, environmental health science, genetics and genomics, molecular biology, translational regenerative medicine, and all aspects of veterinary medicine. Scholars and researchers from diverse backgrounds collaborate with each other and with public and private sector partners to address a wide range of critical research questions. As the largest academic institution in North Carolina, the university enrolls over 36,000 students, offering bachelor's and master's degrees in more than 120 fields of study and doctoral degrees in 67 disciplines.

Librarians at NC State collaborate intensively with university researchers in all disciplines and on emerging tools and technologies for research and scholarly communication in a changing environment. We offer consultation and guidance during all phases of the research data lifecycle, from developing data management plans for grant proposals, to consulting on best practices and appropriate infrastructure for data storage and preservation, to optimizing the sharing and discovery of data. We also advise on copyright and intellectual property issues.

**Name:** Greg Raschke  
**Role:** Senior Vice Provost and Director of Libraries  
**Institution:** NC State University

We thank the OSTP for the opportunity to respond to "Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research." We are pleased that OSTP is taking steps to address key characteristics of data repositories. As the number of available repositories continues to grow, we anticipate that consensus and consistency at OSTP and its agencies will help direct the support we provide our researchers when faced with identifying the most appropriate and sustainable option for their data. We agree with OSTP that these criteria should be reviewed and updated periodically and recommend outlining a process describing who should oversee this review and how often it will occur.

The Background section states that "the set of characteristics is intended to be used as a tool for agencies and Federally funded investigators" and specifically not used to "assess, evaluate, or certify the acceptability of a specific data repository." We support this statement and its intent and

strongly suggest this remains in any final documentation output(s). Additionally, we suggest clarification about the intended audience. If there are multiple audience groups, as suggested, we recommend multiple versions of the documentation that provide unique language, context, examples, and guidance targeted at those audience groups and their specific needs. This documentation presents an opportunity to educate researchers/data producers that the ultimate goal and purpose of these “desirable characteristics” is to make data as reusable and reproducible as possible.

The Background section states that the proposed “desirable characteristics” are meant to be consistent with repository certification criteria (specifically citing ISO16363 Standard for Trusted Digital Repositories and CoreTrustSeal Data Repositories Requirements). These certification processes can be very resource-intensive and are therefore inaccessible for many data repositories. We strongly suggest that any final documentation output(s) clearly explain that repository certification is intended as a guide and not a requirement.

We would like to highlight that there remains a lack of consistency, guidelines, and language around other fundamental aspects of data sharing, including how much data should be made accessible, for how long it should be retained, and how it should be preserved. There is also a noticeable gap for how to handle “big” data access. We applaud the Federal agencies for not becoming too prescriptive in these areas, but we also believe that basic guidance would be helpful for funded investigators and the institutional staff that support their research. Multiple Federal agencies have issued closely related Requests for Information and Requests for Comment<sup>123</sup> in the last several months. We strongly recommend that information be shared among these groups to assist in the development of consistent guidelines and recommendations across agencies when possible.

Lastly, we endorse the responses to this document that we have seen to-date from well-recognized entities within our community, including the Confederation of Open Access Repositories and the Research Data Access & Preservation Association. We would also like to call attention to prior work that should be considered, as it directly relates and is relevant to this content: the Enabling Fair Data Project<sup>4</sup>, Make Data Count<sup>5</sup>, and “FAIRsharing Collaboration with DataCite and Publishers: Data Repository Selection, Criteria That Matter”<sup>6</sup> (which was open for comment until January 31, 2020). The work of several of the Research Data Alliance’s (RDA) interest groups should also be considered, particularly the RDA/WDS Certification of Digital Repositories Interest Group, the Domain Repositories Interest Group, and the Repository Platforms for Research Data Interest Group. Additionally, we recommend following the progress

---

<sup>1</sup> <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>

<sup>2</sup> <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-013.html>

<sup>3</sup> <https://www.federalregister.gov/documents/2020/02/19/2020-03189/request-for-information-public-access-to-peer-reviewed-scholarly-publications-data-and-code>

<sup>4</sup> American Geophysical Union. (2017). American Geophysical Union Coalition Receives Grant to Advance Open and FAIR Data Standards in the Earth and Space Sciences. Retrieved from <https://news.agu.org/press-release/agu-coalition-receives-grant-to-advance-open-and-fair-data-standards/>

<sup>5</sup> <https://makedatacount.org/>

<sup>6</sup> McQuilton, P., Sansone, S.-A., Cousijn, H., Cannon, M., Chan, W. M., Carnevale, I., ... Threlfall, J. (2020, January 27). FAIRsharing Collaboration with DataCite and Publishers: Data Repository Selection, Criteria That Matter. <https://doi.org/10.17605/OSF.IO/N9QJ7>

of the GO FAIR USA Support and Coordination office, which is hosted by the US National Data Services at the San Diego Supercomputer Center, and “focuses on all knowledge domains and the general goal of increasing FAIR data stewardship”<sup>7</sup>. The American Geophysical Union<sup>8</sup> is a leader in promoting data management programs and enabling FAIR data, and we recommend consideration of their continued work in this area.

## **The appropriateness of the “Desirable Characteristics for All Data Repositories” (Section I) for data repositories that would store and provide access to data resulting from Federally-supported research**

General comments:

- The term “desirable” is subjective and therefore problematic. If the OSTP has already identified repositories that meet all or some of these “desired characteristics”, it would be instructive to include these repositories as examples. If the agencies that support R&D have lists of their own that meet these criteria, it would be helpful to share them.
- The OSTP should determine if there is a repository index or catalog that is deemed appropriate for researchers to use. Two well-known examples in the field include re3data.org<sup>9</sup> and fairsharing.org<sup>10</sup>.
- There should be a stand-alone section or language within an existing section stating that versioning at the dataset level is a desirable characteristic that is important for data integrity.

### 1A: PUIDs

- We recommend including the requirement that PUIDs be global (GUIDs). We suggest listing minimum required metadata elements that should be associated with the ID.

### 1B: Long-term sustainability

- This section describes both preservation of data and sustainability of the repository. We recommend breaking these into two distinct sections.
- Regarding preservation:
  - Long-term integrity is important but goes beyond just the infrastructure. Repositories should be transparent about the preservation policies.
  - Provide an explanation for what is meant by “authenticity”.
- Regarding planning for sustainability of a repository:
  - We recommend the document should characterize the minimum expectations for contingency plans.
  - It is important that a repository have plans in place should the repository need to cease operation.

### 2C: Metadata

- We agree that quality, *machine-readable* metadata is fundamental in aiding discovery and reuse, and of critical importance in web-discoverability. While some disciplines have mature and robust standards, others do not. When no standard exists, we suggest that OST should recommend a general standard (e.g. DataCite Metadata Schema). When they

---

<sup>7</sup> <https://www.go-fair.org/go-fair-initiative/go-fair-offices/go-fair-usa-office/>

<sup>8</sup> <https://www.agu.org/Learn-About-AGU/About-AGU/Data-Leadership>

<sup>9</sup> <https://www.re3data.org/about>

<sup>10</sup> <https://fairsharing.org/>

exist, we also suggest that the use of persistent unique identifiers should be encouraged, such as ORCiDs and RORs (Research Organization Registry).

### 3D: Curation & Quality Assurance (QA)

- Data curation adds value to a dataset in a number of ways, and ultimately aids reusability. Data curation can be resource-intensive. We suggest language about what level of data curation is either required (if any) or desirable. Since many reputable repositories do not currently have resources to provide curation, we recommend data curation is listed as a desired and not a required characteristic.
- Curation and QA require access to scripts and code used to generate the data. These outputs should be required if curation and QA is desirable.
- The document states the repository “provides, or has a mechanism for others to provide, expert curation and quality assurance”. It is unclear what type of mechanism is being referenced and what is meant by “others”. More context would be helpful. For example, does “other” refer to repository staff, an external curator, the researcher to perform curation within the repository platform, other?

### 1E (Access) and 1F (Free & Easy to Access and Reuse)

- These two sections would make more logical sense combined into one section about access and reuse.
- In Section 1F, we recommend defining what is meant by “timely manner”. Does this mean that the repository should support embargo (for peer review or another process)? Should the repository support or require publication of data within a certain time period after publication of a paper?
- In Section 1F, we recommend language and resources about dataset licensing. We recommend that the repository support machine readable licenses (e.g. Creative Commons), which enable reuse. We also recommend including license information with the dataset, such as in a readme file, so that the licensing information moves with the dataset as it is downloaded or exported.

### 1G: Reuse: Enables tracking of data reuse

- There is ongoing development within the field to facilitate tracking of data reuse. We recommend utilizing the work of Make Data Count (footnoted on page 3).

### 1H: Secure

- Researchers are often unfamiliar with this language, so providing additional context or referencing additional resources would be useful.

### 1I: Privacy

- We recommend language that states policy and procedures should be in place for handling or removing sensitive or private information.

### 1K: Provenance

- This section states that the repository must maintain a “detailed logfile of changes to datasets and metadata, including date and user...” Further clarification about who the “user” is in this statement is recommended. Is it the person who ingests the data, the data creator, other?
- This information is often not made available to an end user. More context is needed to determine if this means that a repository should provide this information to an end user.

**Appropriateness of the characteristics listed in the “Additional Considerations for Repositories Storing Human Data (even if de-identified)” (Section II) delineated for repositories maintaining data generated from human samples or specimens**

General Comments:

- We suggest including considerations for other sensitive data beyond human subjects, such as geographic locations of endangered species.

IIC: Privacy

- This privacy requirement should be mandated for all data repositories, not just human subjects' data repositories. We suggest moving it to Section I.

IIG: Retention Guidelines

- Retention guidelines should be a requirement for all data repositories, not just human subjects' data repositories. We suggest moving it to Section I.

March 13, 2020

Lisa Nichols  
Assistant Director, Academic Engagements  
Office of Science & Technology Policy

RFC Response: Desirable Repository Characteristics

We appreciate the opportunity to comment on this important issue. We are aligned with the Subcommittee's goals to improve guidelines and promulgate best practices on the long-term preservation of data from Federally funded research. Effective repositories and guidance/standards for their use is important to the success of validation, transparency in research finding, and supporting rigor and reproducibility. We hope that this effort will result in standards and best practices that do not contribute to researcher burden.

We agree with many of the proposed characteristics for repositories, and have suggestions and questions we hope will provide clarity and improve guidance. We hope that the policy itself will have a core set of requirements, recommendations for best practices, align with policy requirements in other countries, and have room for updates as technology and data management standards change.

Our comments and questions are in line, below, under Sections I and II.

***Section I: Desirable Characteristics for All Data Repositories***

- A. Persistent Unique Identifiers (PUIs). We agree with this requirement, and feel that the policy should be agnostic in terms of type of PUID. We would like to see examples of downstream use of PUIDs besides data citation and data access, and how common scenarios would be handled, like in a case where one dataset has multiple DOIs that point to the same information. This case is not an issue today, and this should not change under any new policy.
- B. Long-Term Sustainability. Concepts in this section should be divided into separate topics. Preservation of data, sustainability of the repository, and emergency planning, while related, should be handled separately. Will the policy address data degradation, loss, and migration of data from a proprietary to an open format for access? If a repository claims that access is free and easy, their sustainability/business model should be reviewed.
- C. Metadata. Metadata requirements for discovery, citation, reuse, and preservation are different. General repositories will not be able to support a range of standards for these uses. The policy should address minimum requirements for each, and how different community/domain standards would supplement these. Reference to general purpose metadata standards would be welcome (e.g., DataCite Metadata Schema or Dublin Core). Relevant documentations, such as codebooks and readme files should be included.
- D. Curation and Quality Assurance. The policy should identify appropriate quality assurance vs. expert creation. It must be clear that the data creator/researcher has a responsibility for curation beyond basic levels. A division of responsibilities between the researcher and repository should be clear, including responsibilities for compliance. The policy should further define "other curators."

- E. Access. AND
- F. Free and Easy to Access and Reuse.
- G. Reuse. E, F, and G have overlapping goals. Please include guidance/requirements around fee and open access, continuous availability, and use of APIs. If researchers opt to deposit data for sharing amongst the research team, it should be clear when/if that data is expected to be available for access, and free, to external users. We would also appreciate additional information on intent around the combination of “free”, “easy” and “indefinite”. These characteristics will not be possible without a discussion of cost recovery. Please also clarify reuse in sections F and G. Is there a difference in intent based on the audience (users vs. publishers)?
- H. Secure. The repository should provide documentation of its procedures and best practices, including terms of use and access, prevention of unauthorized access, manipulation of data, and provenance/versioning of data.
- I. Privacy. Please add language to distinguish repositories that only collect data that will be made openly available vs. those that will include sensitive data. The researcher is responsible for compliance.
- J. Common Format. Researchers will be responsible for the collection and format of data. This should be included in a data management plan.
- K. Provenance. We suggest that terminology be changed from “logfile” to “record” of changes.

## ***Section II: Additional Considerations for Repositories Storing Human Data (Even if De-Identified)***

Compliance requirements around storing human data, de-identifying it, and ensuring the purpose of data use is reflective of consent obtained for the study is the responsibility of the researcher, not the repository.

### ***Additional Comments/Concerns***

- The policy should include a glossary, documentation, and guidance for who to contact for help, such as local research data staff and online educational resources.
- Data repository policies should be made available online, including terms of service and terms of use so that it's easy for researchers to evaluate different data repositories.
- We would appreciate clarification of intent once a repository policy is active: can we expect that the researcher/agency would review for compliance?
- Will there be an examination of possible conflicts with institutional/organizational disposition (retention) requirements and policies?
- Institutions of higher education are capped on recovery of administrative costs, and there are many questions about the allowability of direct-charging repository costs to federal grants, particularly after the performance/project period. Uniform Guidance (2 CFR 200) should be updated to reflect charges for repositories and a timeline for allowed costs after the project end date.

The University of Arizona enjoys being part of a broader higher education community discussion about responsible stewardship of data resulting from federally funded research. Thank you for the opportunity.

Sincerely,

Gerald J. Perry  
Associate Dean & Librarian, University Libraries  
University of Arizona

Lori Ann M. Schultz  
Sr. Director, Research, Innovation & Impact  
University of Arizona

## RFC Response: Desirable Repository Characteristics

We are representatives of, and contributors to, the Open Bio Ontologies (OBO) Foundry project [1], a major effort in the life sciences to provide a suite of interoperable ontologies for data and metadata annotation, including major resources such as the Gene Ontology [2]. We are part of this effort because many of us also work with large databases, knowledge bases or repositories, and recognize the need for ontology efforts to make data FAIR [3].

Firstly, we want to state our strong support for the goal of this RFC. We completely agree that there should be clear guidelines on what characteristics a data repository that is used to store data from federally-funded research should have. The draft document captures several important characteristics that we completely agree with, many of which are reflected in the FAIR principles paper [3].

Our comment is directed to one specific part of the draft, Section I- Subsection C, which lists the desirable characteristics for all data repositories and states:

C. Metadata: Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, **using a schema that is standard to the community the repository serves.**

We suggest that this recommendation should be modified, as it does not take into account that a metadata standard should not only be chosen to serve the community that deposits data in a repository. Rather, metadata should ensure that data in the repository is findable, accessible, interoperable, and reusable (=FAIR) for all investigators. This is accomplished by recommending a metadata standard that has been developed with cross-community applications in mind.

Significant efforts to develop cross-community metadata standards have been undertaken in several fields. A pioneer in this area was the effort to develop a shared metadata standard to annotate genes across different organisms, which gave rise to the Gene Ontology [2]. The success of this effort led to the development of the OBO Foundry [1], which established a set of principles that member ontologies should meet in order to be usable as interoperable metadata standards for cross-community data representation and analysis. Many of the OBO principles for ontologies are mirrored in FAIR principles for data. For example, the OBO principles include that ontologies should have an open license, should have globally unique and persistent identifiers for the terms they contain, should be registered in a centrally-indexed resource, should be versioned, and so on.

We thus suggest modifying the recommendation above to explicitly state:



C. Metadata: Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using **a metadata standard that itself is designed to be findable, accessible, interoperable, and reusable for experts across scientific domains, such as ontologies following the OBO Foundry principles [3].**

We believe that this addition to the recommendation will make it clearer to repository developers and funders how they should choose a metadata standard, if multiple are available. We also believe that the recommended adoption of interoperable metadata standards across repositories will greatly enhance the value of data in each repository, as it allows one to interlink data and thus enable integrated analysis.

This comment was compiled and approved by the people listed in the table below, all of whom contribute to and/or use data from repositories that have been supported by federal funding.

Sincerely,



Bjoern Peters  
Professor  
Division of Vaccine Discovery  
La Jolla Institute for Immunology

On behalf of the following contributors (signing on behalf of themselves, which not necessarily implies endorsement of this comment by their affiliated institutions):

Person	Affiliation	Primary Discipline	Role
Bjoern Peters	1) Division of Vaccine Discovery, La Jolla Institute for Immunology, La Jolla, CA 2) Department of Medicine, University of California San Diego, La Jolla, CA	Immunology	Professor
Chris Mungall	Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA	Bioinformatics, Genomics	Research Scientist

Yongqun Oliver He	Unit for Lab Animal Medicine, Department of Microbiology and Immunology, and Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI	Bioinformatics	Associate Professor
Bill Duncan	Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA	Bioinformatics	Software Developer
Deepak Unni	Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA	Bioinformatics, Genomics	Software Developer
Lindsay G. Cowell	Department of Population and Data Sciences, UT Southwestern Medical Center, Dallas, TX	Data Science, Cancer Immunology	Associate Professor
Nicole A. Vasilevsky	Oregon Clinical & Translational Research Institute, Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR	Bioinformatics, Biocuration	Research Assistant Professor
Randi Vita	Division of Vaccine Discovery, La Jolla Institute for Immunology, La Jolla, CA	Immunology, Biocuration	Lead Ontology and Quality Manager
Lynn M. Schriml	University of Maryland School of Medicine, Institute for Genome Sciences, Baltimore, MD	Epidemiology, Microbiome, Genomics, Bioinformatics	Associate Professor
Nicolas Matentzoglu	European Bioinformatics Institute (EMBL-EBI), Hinxton, UK	Bioinformatics	Senior Semantic Web Developer
Leigh Carmody	Jackson Laboratory for Genomic Medicine, Farmington, CT, USA	Bioinformatics, Biocuration	Scientific curator
Darren Natale	Protein Information Resource Georgetown University Medical Center Washington, DC	Bioinformatics, Biocuration	Research Assistant Professor
Alexander	Department of Biomedical Informatics,	Biomedical	Associate

Diehl	Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY	Ontology Development	Professor
Lawrence Hunter	Department of Pharmacology University of Colorado School of Medicine Aurora, CO	Pharmacology	Professor
Melissa Haendel	Dept. of Environmental and Molecular Toxicology, Oregon State University, Corvallis, OR	Developmental Biology, Clinical Informatics, toxicology	Director of Translational Data Science
Jonathan Bisson	Institute for Tuberculosis Research, Department of Pharmaceutical Sciences, University of Illinois at Chicago, Chicago, IL	Bioinformatics, Chemistry	Research Assistant Professor
Ruth Duerr	Ronin Institute for Independent Scholarship, Westminster, CO	Geoinformatics, Earth and Space Science Informatics	Research Scholar
Valerie Wood	University of Cambridge, Cambridge, UK	Biocuration	Project Manager

## References

- [1] Smith, B., Ashburner, M., Rosse, C. et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25, 1251–1255 (2007). PMID: PMC2814061. <https://doi.org/10.1038/nbt1346>
- [2] Ashburner, M., Ball, C., Blake, J. et al. Gene Ontology: tool for the unification of biology. *Nat Genet* 25, 25–29 (2000). PMID: PMC3037419. <https://doi.org/10.1038/75556>
- [3] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). PMID: PMC4792175. <https://doi.org/10.1038/sdata.2016.18>

RFC Response: Desirable Repository Characteristics  
Response to OSTP Request for Public Comment on Draft Desirable Characteristics for  
Managing and Sharing Data Resulting From Federally Funded Research

SOFTWARE SIMPLICITY

Edward S. Lowry, [eslowry@alum.mit.edu](mailto:eslowry@alum.mit.edu)

Primary scientific discipline: Computer Science (now retired)

There is another topic which is highly relevant for Federal agencies to consider in developing desirable characteristics for data repositories:

technology for software simplicity.

Details of how software can be simplified using improved language are discussed in "[Software Simplicity, Banished for 45 Years](#)".

Summary

Progress in simplifying software by using improved language has been obstructed for 45 years - - causing many disasters.

Progress in language for expressing software more simply comes to an end when it becomes impossible to further simplify large applications without changing their functionality. The way the process ends is likely to provide final answers to fundamental design questions.

Sound computer language for rich applications needs both:

- a common implicit iteration mechanism and
- flexible data structures.

Failure to provide both can be regarded as a **severe flaw** in ALL current substantial computer languages. Fixing the flaw enables broad use of compact plural expressions combined with flexible data. The fix appears to clear a path toward a single core language semantics for rich applications where simplicity of expression approaches an enduring practical optimum and a large improvement over current languages.

Fixing the flaw can be done by using a small number of connective information building block structures. Those structures can be merged into a single connective structure. It is hypothesized that further refinement of that structure can move toward a practical permanent optimum structure. The simplicity, stability, design convergence, and subject matter generality point toward major improvements.

Data processed or produced by a computer language will most easily be represented in the data model of that language. If the language is designed to express rich applications in the simplest possible way, then that data model will maximize many of the identified Desirable Characteristics for All Data Depositories:

*B. Long-term sustainability.*

If simplicity is already maximized, the data model can remain durable indefinitely.

D. *Quality Assurance.*

Simplification of software can improve quality of the software and its output in many dimensions. See the above reference, page 1.

E. *Access.*

Access to many kinds of data can be written with simplicity in a common language which is likely to become widely known.

F. *Easy to Access and Reuse.*

Queries to access the data will be simple and easy to express. Reuse will be simple and easy.

H. *Secure.*

Simplicity of expression will reduce bugs that create security holes.

I. *Privacy.*

The above security improvement will also improve privacy.

J. *Common Format.*

Maximizing simplicity of expression will lead to natural format standards.



March 16, 2020

Lisa Nichols  
Subcommittee on Open Science  
Office of Science and Technology Policy

Dear Dr Nichols,

**COMMENT ON DRAFT DESIRABLE CHARACTERISTICS OF REPOSITORIES FOR MANAGING AND SHARING DATA RESULTING FROM FEDERALLY FUNDED RESEARCH**

The two data organizations of the International Science Council ([ISC](#)), the Committee on Data ([CODATA](#)) and the World Data System ([WDS](#)), together with the Research Data Alliance ([RDA](#)), all international data organizations with official presences and missions to support US and international stewardship of scientific data, are pleased to provide comments in response to your *Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research*.

First, some general comments:

- We applaud and support the publication of these important guidelines as a statement. The intent and concepts in this document are an excellent frame of reference for management of government-funded research outputs.
- A first level statement of commitment to support the need for desirable characteristics is most important. We strongly support early publication of your statement, but in parallel, we recommend development and publication of detailed implementation guidelines. It is likely that the US research data community will request such detailed guidance following the publication of your statement.
- Given that this is a first cut at identifying desirable characteristics, there is a need for better definitions of many terms and ways to evaluate or measure what “compliance” or “achievement” means since they are voluntary guidelines and not mandatory. This would advance the overall objective of the initiative.
- Many of the specifics of providing sustainable, trusted, well-managed, and FAIR data to the community are addressed in the CoreTrustSeal Trustworthy Data Repositories

Requirements. WDS and the Data Seal of Approval, have developed a [set of requirements](#) under the auspices— and as an official output—of RDA, and these have now been operationalized in the [CoreTrustSeal](#). These requirements play a vital role in the quality management of repositories. CoreTrustSeal will be providing a detailed technical response to your request, which we endorse.

- In addition, there are ample community-developed guidelines, best practice, and specifications that support the intent of your statement, developed by RDA, [GOFAIR](#), Group on Earth Observations ([GEO](#)), Earth Science Information Partners ([ESIP](#)), the Inter-university Consortium for Political and Social Research ([ICPSR](#)), and many others. These should ideally form some major sources of your detailed guidance to federal research grant recipients.
- One area that is not covered in the characteristics deals with copyright or more broadly rights management. Please consider adding something on the appropriate licensing regime and notation. This discourages plagiarism and encourages data publishing and use.
- We foresee that follow-up work will be required to address non-data research outputs that are government-funded. Code, algorithms, methodologies and protocols, and similar products of research are currently largely outside the scope of this guidance, and this aspect will attract additional focus in years to come.
- Finally, as a statement of commitment to support the need for desirable characteristics is fundamentally important. We therefore strongly support making the most salient changes and moving ahead with issuing guidelines to put a stake in the ground on the importance of FAIR data management and preservation.

In the Annex to this letter we share some high-level comments from our three organizations. In addition, we cross reference and endorse the much more detailed response provided by CoreTrustSeal to the draft text of the Desirable Characteristics of Repositories, which includes input by volunteers and staff of RDA and WDS.

Finally, CODATA, RDA, GOFAIR, and WDS have recently formed an initiative called “Data Together”. The intent of the initiative is to promote our joint missions, and these, in turn, overlap significantly with the scope of your statement. As such, in addition to the comments provided, we offer our support to the Subcommittee on Open Science. We stand ready to provide further input and assistance to your initiative and to the promotion of well-managed, trusted, open, and FAIR data.

Please do not hesitate to contact us if you have questions or would like additional information.

Respectfully,



Alex de Sherbinin  
Scientific Committee Chair  
ISC WDS  
[adesherbinin@ciesin.columbia.edu](mailto:adesherbinin@ciesin.columbia.edu)



Leslie D. McIntosh  
Executive Director - US  
RDA  
[leslie.mcintosh@rda-foundation.org](mailto:leslie.mcintosh@rda-foundation.org)

[Bonnie C. Carroll](#)  
Secretary General  
ISC CODATA  
[bcarrolltn@gmail.com](mailto:bcarrolltn@gmail.com)



## Annex: Some Specific Comments

### I. Desirable Characteristics for All Data Repositories

B. *Long-term sustainability*: Has a long-term plan for managing data, including guaranteeing ...

Suggest you add: “With evidence that the plan can be delivered”. (Such evidence forms part of the [CoreTrustSeal](#) criteria)

C. *Metadata*: Ensures datasets are accompanied by metadata sufficient to enable discovery ...

FAIR has definitions and measures that can be used to assess this. We suggest that this be referenced <https://doi.org/10.1038/sdata.2016.18>

F. *Free & Easy to Access and Reuse*: Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

What does this mean in practice and how is it measured?

J. *Common Format*: Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.

The requirement for non-proprietary formats should be voiced more strongly; for example, by stating that it should be the default.

K. *Provenance*: Maintains a detailed log file of changes ...

The statement potentially limits implementation options. We suggest making the statement technology agnostic/more generic; for example, “Maintains a detailed *record* of changes ...”. As a case in point, metadata changes are increasingly managed by way of versioning and a provenance chain stored in the metadata itself.

### II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)

A. *Fidelity to Consent*: Fidelity to Consent: ADD: “has consent documentation and” restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).

# Response to Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research

Responder: Sherry Lake on behalf of the University of Virginia Library  
Role: Scholarly Repository Librarian  
Domain: Discipline Agnostic

On behalf of the University of Virginia Library, I thank you for the opportunity to provide feedback to the OSTP draft characteristics of repositories for managing and sharing data. Good data management is critical for ensuring validation, transparency of research findings, as well as to support data reuse. As manager of the University of Virginia's institutional data repository, a local instance of Dataverse, I can attest from professional experience as to what criteria generally meet the desirable characteristics listed in this RFC.

As we seek to expand our capacity to support research data management, we need to develop repositories that are using best practices, while at the same time, ensure that any repository requirements are not overly onerous and result in excluding many (potentially general purpose and institutional) repositories.

I applaud the Subcommittee on Open Science's (SOS) goal "to improve the consistency of guidelines and best practices that agencies provide about the long-term preservation of data from Federally funded research." This is an ambitious undertaking given the many different stakeholders. But creating generic and easy-to-comply-with guidelines will go a long way of ensuring that data will be shared and reused.

## General Comments:

- The current repository landscape includes domain and general purpose repositories. An implicit assumption in the current OSTP draft is that all data repositories are domain repositories (and have community standards). However, general repositories (most often managed by university libraries) play a critical role in the landscape by providing sustainable services for researchers that do not have access to an appropriate domain repository.
- It is unclear exactly how the set of characteristics would be used. The background states that the characteristics would assist PIs in identifying data repositories, but that the agencies themselves would not use them to evaluate the use of a repository. This statement seems in conflict with another section of the RFC that states the characteristics would be used to evaluate a data management plan and its proposed repository.

- Assisting Federally funded investigators with identifying appropriate data repositories is a worthy goal; however, researchers often need substantive help with this process, as they aren't familiar with the terminology listed in these characteristics. To help with this issue, our response suggests the inclusion of resources such as local experts and online educational materials already available to fill these gaps in knowledge.
- In some cases, the characteristics proposed in the draft would fall under the responsibility of the data creators/providers (access and reuse rights, data format) or their institutions, making it difficult, if not impossible, for repositories to enforce these in the context of the repository.
- OSTP may hope that these desirable characteristics “be enduring”, but because this is a rapidly evolving landscape and technology and standards for data management will change over time, it will be important for OSTP to review and update these characteristics regularly.

## I. Desirable Characteristics for All Data Repositories

### A. *Persistent Unique Identifiers:*

Assigning PUIDs to data is a MUST; preferably DOIs for citation, linking and discovery. However, this characteristic also implies IDs for identifying the outputs of Federally funded research. A simple DOI (data identifier) cannot do this alone. Repositories need to make use of the Federal Government's **agency identifier**; this type of PUID should be specified independently of the PUID for the object.

### B. Long-term sustainability:

This section is currently a mix of requirements, (preservation practices, sustainability of operations, emergency planning), Divide these into two characteristics (1) preservation (data), and (2) sustainability of the repository. “Preservation” is the language used in current Data Management Plans (DMPs) for a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets. “Sustainability of the Repository” is where the organization (or repository) has a long term plan for managing and funding the data repository.

### C. Metadata:

Metadata is required to support a number of objectives (discovery, citation, reuse, and preservation). Clarification of what is meant by “sufficient” is needed, since metadata needs to be of “good” quality and as well as be comprehensive. Note that most communities do not have “community standards” either in the type of fields described, nor a minimum set of fields.

Metadata requirements are different for each purpose stated (discovery, citation, reuse, and preservation) and it would be valuable to outline the distinct requirements for each

objective. In addition, while some domains already have well developed standards for metadata, others do not. Therefore, I suggest a reference that general purpose metadata standards is also acceptable (e.g. DataCite Metadata Schema or Dublin Core). This characteristic also assumes a “discipline” repository, but many institutions have local data repositories that would very much fit this characteristic, but are generic without a discipline community, thus the metadata will be general.

#### **D. Curation & Quality Assurance:**

A basic level of curation for both metadata and data should be a requirement. This characteristic is straightforward if a repository has data curation staff who ensure that data are curated properly upon submission and/or if it is undertaken by the data creators. The phrasing “has a mechanism for others to provide” is unclear what that “mechanism” is. Additionally, researchers will not likely have a good idea of what ‘expert curation’ means. I suggest a requirement of basic curation at the repository, and a recommendation for the repository to support data curation by the creators and/or curators.

#### **E. Access:**

The distinction between characteristics E and F is not useful and can be dropped. I also suggest mentioning the concept of licensing to explicitly state conditions for use. This issue is complicated because data are not copyrightable in all jurisdictions, or equally across formats (e.g. text vs. images). Finally, a repository that supports (provides) an “open” license still requires that the depositor opts to select and use such a license.

#### **F. Free & Easy to Access and Reuse:**

This characteristic and the previous one are very similar – the prior one is more along the lines of choosing an “open license”. If you keep this characteristic, include specific requirements related to the availability and how to access such as including open free access, continuous availability and open APIs. “Timely manner” can be eliminated and just stated “accessible free of charge without an embargo”, i.e., no time – immediate access.

#### **G. Reuse:**

This characteristic needs clarification. There are three requirements needed to support reuse: citation metadata, permanent unique identifiers, and the use of machine readable, standardized licenses. Include all of these as requirements to support data reuse and re-label this as “tracking reuse”.

#### **H. Secure:**

This characteristic lists specific ISO and NIST standards, making it clear what technical considerations are in play. However, it is not clear how the average researcher would be able to determine whether a repository complies with these standards, making it less useful. I suggest including: “Repository provides documentation of its practices that prevent unauthorized access/manipulation of data.”

### **I. Privacy:**

If this characteristic is for repositories for data that will be made openly available, this requirement should be clarified. It seems to be about general cybersecurity concepts. The language used in this characteristic would not be understandable by all researchers and is therefore of limited utility to some of your target audiences. Suggesting resources like local IT and data services staff to help evaluate these criteria is critical to mitigate this concern. As above, it is not clear how the average researcher would be able to determine whether a repository complies with these standards.

### **J. Common Format:**

Although repositories can recommend formats, it is the data creators that determine the format of the data they collect. I suggest that this is a responsibility of the researchers and data creators and should be a requirement included in a data management plan and NOT a characteristic of a data repository.

### **K. Provenance:**

Provenance of data is important for data integrity and assurance. Logfiles are typically a feature that is hidden from the end user, and thus many researchers are unaware of what they are and why they are important. Many repositories “record” changes to datasets and metadata by the way of “Version Control.” I propose that this characteristic be labeled as “Provides Version Control”.

## **II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

In terms of storing human data (or other sensitive data), it is the responsibility of the researcher to ensure that access conditions reflect consent and ensure that human data is appropriately de-identified. The role of the repository may be to support a variety of access levels (including restricting access to authorized users) and adopt practices that ensure secure management of data. It should be noted that not all repositories collect sensitive data.

As there seems to be a lack of repositories that collect sensitive data, perhaps the following characteristics could be used by a Federal Agency to develop a Federal or national repository to store data resulting from Human Data research.

### **Missing Characteristics from Draft**

- What is not covered in the above characteristics, but should be included, is **Repository Contact** and **Repository Documentation**. Every repository must have a contact point or helpdesk to assist data depositors and data users. In addition, repositories should provide documentation about the scope of data accepted by the repository such that a researcher could make a decision on appropriateness for their data.

- In the increasing requirement of interoperability, another missing repository characteristic is the **existence of an API** that allows for automatic uploading, sharing and using.
- A criterion regarding how the repository is funded and plans for data preservation in the event that funding is no longer available should be added.

## In summary

I want to highlight that any list of desirable characteristics for selecting repositories:

- **Should not overlook the local the local support systems** - I suggest encouraging researchers to seek out local experts and online educational materials already available to assist in the selection of a repository for managing and sharing data. Many institutions have both research data practitioners to answer their questions and institutional repositories to deposit data when a disciplinary repository is not available could assist in reducing confusion and increasing compliance.
- **Should not ignore general repositories** – These repositories, most often managed by university libraries, play a critical role in the landscape by providing sustainable services for researchers that do not have access to an appropriate domain repository.

Since these proposed characteristics would apply to all federal funded research, terminology in such a document should match the terminology in current Data Management Plan requirements (or DMP requirements would need to be consistent with the terminology here).

If the OSTP and Subcommittee on Open Science feels that it is important to share data in repositories by stating desirable characteristics, then the Federal Government should consider verifying that researchers are putting their data in repositories according to what was stated in their Data Management Plans.

I want to underscore that many of the University of Virginia Library's comments align with the COAR and SPARC's joint response [https://sparcopen.org/wp-content/uploads/2020/03/COAR\\_SPARC-Joint-Response.pdf](https://sparcopen.org/wp-content/uploads/2020/03/COAR_SPARC-Joint-Response.pdf)

I thank you for the opportunity to provide feedback to the OSTP on the draft characteristics of repositories for managing and sharing data.



# AMERICAN ASTRONOMICAL SOCIETY

American Astronomical Society public comment on “*DRAFT Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research*” (Document 2020–00689; Posted 2020-01-17)

Physical Sciences

16 March 2020

Authors:

August Muench, AAS Data Editor

Greg Schwarz, AAS Data Editor

Julie Steffen, AAS Director of Publishing

The American Astronomical Society (AAS) is the major organization of professional astronomers in North America. Its membership of over 8,000 individuals also includes physicists, mathematicians, geologists, engineers, and others within the broad spectrum of subjects comprising contemporary astronomy, planetary science, and heliophysics. The mission of the AAS is to enhance and share humanity’s scientific understanding of the universe.

As a 501(c)(3) nonprofit corporation, the AAS owns, operates, and publishes the most widely read and cited journals in the field, *The Astronomical Journal*, *The Astrophysical Journal*, *The ApJ Letters*, *The ApJ Supplements Series*, and *The Planetary Science Journal*. Since creating electronic editions starting in 1995, the AAS has encouraged researchers to submit data critical to their research result *along with their manuscript*. Machine-readable tables (MRT) and data-behind-figures (DbF) are examples of the research data integrated into and hence preserved for posterity in many thousands of research articles published in AAS journals. The AAS has employed trained astrophysicists as data editors and adopted publishing workflows that help researchers share their data for the past twenty years, which has led to the inclusion of a significant amount research data in the literature. Additionally, the AAS has spearheaded efforts to link to important, related data sets in federally funded data repositories and will continue to develop and deepen these connections.

## **The proposed use and application of the desirable characteristics**

Astronomers share many of the problems experienced by other researchers in the physical sciences, although some of our common problems have been solved. Original data products resulting from NASA space missions are well-curated and stored in repositories that follow many (but not all) of the proposed characteristics. This culture of saving and sharing original data has made astronomy a leader among physical science disciplines in sharing research data. However, astronomy has a shortage of domain-specific repositories for most derived data products, for simulation and modeling results, and for data not resulting in a research publication (e.g., null result data). There are few existing repositories in astronomy that accept researcher data and satisfy all proposed characteristics. Requiring them for new repositories may narrow rather than expand the already limited options for federally funded investigators.

## **The appropriateness of the “Desirable Characteristics for All Data Repositories” (Section I)**

The AAS finds the proposed characteristics noteworthy and valuable and endorses them uniformly. Further commentary, informed by long experience working with researchers, is intended to highlight specific issues and to reflect on the current repository landscape for astronomy researchers.

In addition to endorsing the entire set of proposed characteristics, the AAS strongly endorses the need for curation and quality assurance mechanisms in data repositories (**Section I; Characteristic D**). Data submitted without curation to generalist repositories are of limited value and may be missing critical details, e.g., units, that are necessary for either human or machine reuse. Standard data review conducted at AAS uncovers errors in the tabulation of results that would otherwise not be detected, especially if those data are archived ex post facto in a generalist repository. Useful metrics should correlate successful compliance of an open data mandate with enhancement of the scientific record. Unreviewed data may distort these metrics, rendering them useless.

Adding open curation platforms to repositories would improve the quality and success of data sharing by researchers. Experience indicates that supporting and assisting researcher data submissions increases the likelihood of data sharing and improves the overall result. Generalist repositories tend to lack workflows for external review and improvement of submitted data and even domain-specific archives struggle with managing data review efficiently and expeditiously. Enabling curation by external teams, such as data scientists, or other stakeholders, such as data librarians, would make the sharing process more efficient and accurate.

The AAS also strongly endorses the need for domain-specific metadata (**Section I; Characteristic C**). It may be valuable to enable search and discovery across repositories using abstracted or “common” metadata; however, community-specific metadata schema, e.g, the standards of the International Virtual Observatory Alliance (IVOA), are even more vital for successful reuse and interoperability.



## **The ability of existing repositories to meet the desirable characteristics**

Astronomy is well positioned as a result of data archiving and release mandates put in place for NASA space missions and their data repositories (examples include the Infrared Science Archive [IRSA] and the Mikulski Archive for Space Telescopes [MAST]). The Astrophysics Data System (ADS) is a federally funded repository of bibliographic data that is fundamental for astronomy and astrophysics, providing links between the literature and data archives. As previously mentioned, however, more repositories are needed for derived data products not covered by the current scope of these NASA repositories, such as ground-based observations, model and simulation data, and laboratory astrophysics data.

The AAS is in active collaboration with NASA data repositories that are engaged in improving their functionality to support many of the proposed characteristics, including generating persistent identifiers for data. None of these NASA data repositories, however, are currently CoreTrustSeal certified. Some repositories do not accept data in advance of journal publication. Negative or “null” result data may go unarchived in the current repository landscape.

## **Summary Conclusions**

- The AAS has actively supported research data sharing for over twenty years by encouraging researchers to submit the data critical to their research result along with their manuscripts to our flagship research journals.
- The AAS believes that the most successful research data sharing involves curation and researcher support and has employed professional astronomers as data editors for this purpose.
- The AAS supports domain-specific repositories and metadata over more general solutions.
- The AAS actively collaborates with existing federally funded data repositories and would welcome further adoption of the OSTP proposed repository characteristics at these repositories.

## **END OF COMMENTS**

If you would like to follow up on any of the above comments, you may contact the AAS at [public.policy@aaas.org](mailto:public.policy@aaas.org).



Trust Farm, LLC  
Allen L. Phelps, CEO  
501 Church Street NE | Suite 210  
Vienna, VA 22180 USA  
(703) 989-3894  
[www.trustfarming.com](http://www.trustfarming.com)  
[allen.phelps@trustfarming.com](mailto:allen.phelps@trustfarming.com)

- **Primary Scientist Discipline:** Advisor to Biopharmaceutical, Biotechnology, Healthcare, and Defense industries, including universities and corporations.
- **Role:** Research Security Manager

Thank you for the opportunity to provide input into the debate for workflows resulting from the 2013 White House Office of Science and Technology Policy (OSTP) memorandum entitled “Increasing Access to the Results of Federally Funded Scientific Research” that called for improved access to data and publications resulting from Federally-funded R&D. I appreciate the opportunity to comment, as I have working with research organizations in the academic and corporate sectors to create security management system to safeguard innovations.

The National Institutes of Health (NIH) is the largest source of public funding for medical research in the world. NIH’s mission is to seek fundamental knowledge about the nature and behavior of living systems and apply that knowledge to enhance health, lengthen life, and reduce illness and disability. Although NIH is among the largest of the grant-making entities in the Federal Government, it is not alone in facing the threat to program integrity from foreign influence and manipulation of the grant-awarding enterprise from the misuse, loss, theft or misappropriation of research assets by external threat actors. The compromise of research program integrity, and the compliance-driven controls that are designed to safeguard America’s interests, is a National Security problem.

Trusted insiders are ignoring long-standing compliance and ethics rules and are converting their access to Federally-funded research programs for lucrative, foreign government-backed opportunities. Our trusted researchers and other research program professionals are largely compromising our innovation for personal enrichment (i.e., greed). The actions taken are by individuals, often those fully-immersed in the research community. Because of the willingness of individuals, foreign state-sponsored efforts have found fertile ground to exploit. But, the grant making community is not without responsibility for this environment. Although policy and agreements are clear about reporting requirements, there is enough latitude in the language to allow for broad interpretation, plausible deniability and provides individuals who would wittingly or unwittingly compromise program integrity, the ability to cognitively rationalize their activity.

The grant-making and grant-receiving community lack imagination and a willing strategic vision that their community could be exploited in this way. They have been overwhelmingly naïve and believed that researchers would jeopardize research integrity in exchange for compensation. Adding to the lack of awareness, there exists an organizational culture attitude that oversight and internal control offices often are considered a necessary inconvenience for compliance purposes, rather than the *guardians of stewardship* they should be.

Although this problem is not unique to NIH, NIH was among the first Federal agencies to identify and respond to the threat. My comments and opinions outlined here were framed by my experience with NIH-related research integrity compromises, but my recommendations are applicable across the Federal grant enterprise.

Upwards of 70 to 80 percent of the estimated \$39 billion NIH receives from Congress is awarded in the form of extramural grants to more than more than 2,500 universities supporting more than 300,000 principle investigators at health science centers, medical schools, and other research institutions. Research principle investigators may use their grant funds to support a variety of needs, including staffing laboratories, purchasing supplies and equipment, and attending national and international conferences to discuss research findings. To further address their research needs, some investigators who apply for (or receive) NIH grants may also seek research support from other organizations, including foreign entities.

Recipients of these funds are responsible for soliciting and reviewing participants and investigators associated with the grant application financial interests from all sources of support, financial interests, and affiliations and then certifying whether those “significant financial interests” constitute financial conflicts of interest (FCOIs). According to NIH policy, research support includes all financial resources—whether Federal, non-Federal, commercial or institutional—available in direct support of an individual’s research endeavors, including, but not limited to, research grants, cooperative agreements, contracts and institutional awards.

Concern about foreign threats to the United States biomedical research, grant process and associated intellectual property have always been present; however, vastly under appreciated by the grant-making and grant-recipient community. Beginning in late 2016 and rolling into early-2017, indicators became more visible in part due to insider threat activities at a major cancer center in Texas. There were indicators that research integrity had been compromised by insider behaviors, leading to investigative inquiries within the NIH grant audit program and extending into concerns regarding the integrity of the NIH Peer Review.

In an August 2018 letter to institutes receiving National Institutes of Health (NIH) funding, the Director of NIH acknowledged that “threats to the integrity of the U.S. biomedical research exist” and that NIH was concerned about three areas:

- (1) diversion of intellectual property;
- (2) sharing of confidential information from grant applications; and

- (3) the failure by some NIH-funded researchers to report substantial financial support from other organizations, including foreign governments.

In October 2018, Congress sent a letter to the Director of NIH expressing concern about foreign threats to the integrity of United States biomedical research. Specifically, the letter highlighted concern regarding “cases in which researchers supported by Federal grants may have failed to disclose financial contributions from foreign governments.”

Given the scope and scale of the research program integrity program, there needs to be an organized system to determine the readiness of grant-receiving research institutions to safeguard innovations and protect research assets from misuse, loss, theft, and misappropriation from compromised insiders and nefarious external threat actors.

**Proposal: Research Program Integrity Stewardship Scoring System:** I propose that there are ways to meet compatible goals of making publicly-funded research available sooner and increasing overall stewardship in all stages of biomedical research prior to public release. The ideas are presented as general concepts and would require development and an integrated grantee and grantor implementation plan. Grantor and grantee offices of oversight and compliance would be essential to the success of this effort.

Grant making requires several steps from application to decision. The decision to award and ability of awardees to retain, public funding should require some level of stewardship consideration. Throughout the award process there are opportunities for grant applicant compromise. **Stewardship Scores** can consist of many components that would allow the Federal Government to determine if the level in which research organizations have the necessary compliance controls and research integrity standards to protect research assets. The Steward Score would be generated by a weighted assessment to identify readiness in terms of grant policy compliance, responsiveness to audits and inquiries, how well the recipient organization conducts vetting of researchers, staff and projects. Other input areas could also include internal processes to review and certify the accuracy of grant applications, due diligence on external partners and suppliers, insider threat incident management, and the capability to conduct effective compliance investigations into research integrity and compliance policy violations and related allegations.

Obviously, additional work is needed to fully-bake the concept of the Steward Score and the implementation plan. Subject Matter Experts (SMEs) from the grant-making entities, grant-receiving entities, and other internal and external stakeholders could convene to frame the concept. The score could either be validated by the Federal grant making agency or an impartial third party like an accreditation process.

The Stewardship Score should be a consideration for award. After Peer Review and prior to final award decision, the Stewardship Score should be applied to the decision candidates. If an awarding Agency wishes to award to an applicant who has a score below the acceptable threshold, they would have to mandate some additional stewardship measures to ensure

accountable and appropriate use of those funds. This would not change the award decision based on the science deemed most viable, but would result in greater rigor and confidence in the stewardship of the funds.

The process to develop the Stewardship Score would require funding to implement and manage. However, an increase to facilities and administrative (F&A) allowable costs in research grant financial planning that are specifically applied to maintaining stewardship would provide the necessary financial resources to promote research integrity. Adding Stewardship to F&A budget plans effectively creates **FA&S budget lines** on grant expenditure allowances. Federal granting authorities need to emphasize the role of oversight and compliance offices in protecting the integrity of research programs in the United States. Stewardship should be a line item allocation from grant awards. Grantees will see an increase in available funding from F&A for those expenses now attributed to oversight and internal controls. The Stewardship line item would be dedicated to these costs and should be accounted for annually and verified by the grantee and grantor offices of oversight and compliance.

I appreciate the opportunity to share these thoughts and would be happy to assist further as needed.

Best Regards,

A handwritten signature in black ink, appearing to read 'APhels', with a long horizontal flourish extending to the right.

Allen Phelps  
CEO – Trust Farm, LLC

**Statement of the American Economic Association's (AEA)**

**Committee on Economic Statistics** <https://www.aeaweb.org/about-aea/committees/economic-statistics>  
and

**Committee on Government Relations** <https://www.aeaweb.org/about-aea/committees/government-relations>

**on Desirable Characteristics of Repositories for Managing & Sharing Data from Federally Funded Research,  
as invited in the Federal Register of January 17, 2020 (85 FR 3085)**

We commend the OSTP for having proposed a set of necessary characteristics of repositories for managing and sharing data from Federally-funded research, which would apply equally as well to data repositories for research funded by or originating from any source.

We also endorse the comments from the AEA Data Editor, Lars Vilhuber, submitted to you by separate package: <https://www.aeaweb.org/content/file?id=11689>, and wish to reiterate his opening comments on:

- The importance of sharing data (and computational instructions, “code”) for the purpose of transparency and reproducibility of science and the key role of data repositories in this endeavor; and
- The need for the scope of OSTP considerations to include research created by scientists in the direct employ of the federal government, data created for public and research use with federal funds as part of the business of the 13 federal principal statistical agencies, as well as any data created for research and evaluation under the Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act).

Beyond this, we would appreciate some more explicit cross-walking to assure consistency between what OSTP is proposing and the work on data repositories and data access that are part of the Federal Data Strategy Action Plan, the 2019 guidelines issued by the Office of Management and Budget (OMB) to update the Information Quality Act <https://www.whitehouse.gov/wp-content/uploads/2019/04/M-19-15.pdf>, and the Evidence Act.

One feature of the Evidence Act on which OMB is working is the implementation of tiered data access to protected federal data. Tiered access can simultaneously widen access to data while limited the risk of nondisclosure by setting up a hierarchy of users. Low level users can access only a limited set of information, whereas the highest-level users can access the most sensitive data on the system. Incorporating this into the definition of OSTP’s *access* characteristic would demonstrate compatibility across Executive Branch efforts.

Another OMB and federal agency effort that may benefit from incorporation into OSTP’s proposals is the common application for data access being piloted by the Inter-university Consortium for Political and Social Research (ICPSR) under a contract with the U.S. Bureau of the Census. The pilot sets up a single portal and standard application process for requesting access to restricted data, across multiple data repositories or centers. This might be considered as an additional feature for OSTP’s *access* characteristic.

A final item that would benefit by clarification or coordination with OSTP’s proposal is the Evidence Act’s establishment of a Data for Evidence Advisory Committee charged with setting up a National Secure Data Service facilitating access to all government data. What might be the similarities and differences between a set of data repositories as prescribed by OSTP and a National Secure Data Service?

Clearing up confusion among similar national efforts within the Executive Branch would be beneficial to organizations that, like the AEA, wish to support and facilitate federal efforts that improve data storage, curation, and access for research purposes.

Thank you for the opportunity to comment.

*John C. Haltiwanger*  
Chair, Committee on Economic Statistics

*Kenneth Troske*  
Chair, Committee on Government Relations

To: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

Name: Danielle Kinkade

Affiliation: Biological and Chemical Oceanography Data Management Office

Primary Scientific Discipline: Earth Science: Oceanography

Role: Repository Director / Data Manager

Below please find comments related to the OSTP document: *Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research* (2020-00689.pdf) provided from the perspective of a grant-supported geoscience digital data repository.

**General comments:**

Many of these criteria are already identified in repository certification processes (e.g., CoreTrustSeal), so it is unclear why separate overlapping information is now being aggregated to achieve what is essentially the same result: to identify trustworthy repositories. Therefore this process seems redundant, and could be simplified to support repository certification and its associated certification badge display to identify trustworthy repositories. However, it should be noted that building the capacity to meet these criteria may require funding support on behalf of many repositories; not only to fulfill the requirements, but also to apply for the certification process.

These criteria are also being presented in a manner that alludes to their use in decision making activities. Within the geosciences, a rather mature digital data curation landscape already exists consisting of domain and institutional repositories that have evolved to meet the needs of their specific research communities. The creation of the FAIR Principles (Wilkinson et. al, 2016) and subsequent desire by many data stakeholders to see their implementation among repositories is a recent and rapidly evolving phenomena. Many repositories currently exist within grant-supported funding cycles that often preclude nimble adaptation to such rapidly evolving strategies and technologies. What has not been articulated, is how development and implementation of desired repository criteria will impact the relationships between existing data repositories and their served communities. Could circumstances arise where these criteria are, in fact, used in an assessment capacity? Likewise, will strategies be developed to bring repositories, who may not currently meet these criteria, but play a critical role in supporting their specific research communities, up to a capacity commensurate to this criteria list?

**Background Section:**

**Para3:** “Federal agencies would not plan to use these characteristics to assess, evaluate... a specific data repository, unless otherwise specified for a particular agency program, initiative, or funding opportunity.”

Again, this statement appears to indicate that the proposed draft characteristics could, in fact, be used to assess and/or, evaluate repositories under certain circumstances including those related to funding, and should be clarified as to what types of programs, initiatives or funding opportunities might specify and/or warrant use of the criteria for assessment or evaluation, and any implications thereof.

**Bullets 1 and 2:** These statements indicate the criteria may be used by both Federally-funded investigators and Federal agencies to identify or designate a particular repository for use when looking to host data.

Again, this process seems redundant, and could be simplified to support repository certification and associated certification badge display to identify trustworthy repositories who meet many of the criteria listed through certification.

**Bullet 5:** The use of the criteria for evaluating data management plans (DMPs) proposing to deposit research data in a repository not supported by a Federal agency would imply that such criteria are in fact being used to assess repositories on their ability to fulfill the data curation role. This seems contradictory to the statement in Introduction Para3. In addition, repository use as described in a DMP is only one small piece of a broader DMP assessment that should be fully developed before any one particular piece of a DMP is assessed. Educating proposal/DMP reviewers to the terms and nuances of these criteria (in addition to other DMP assessment criteria) would need to happen as well.

**Para4, sentence1:** "...in a user-friendly manner", implies that these criteria, if used directly by investigators and non data-savvy individuals would be easily accessible and understandable. However, much of the subject matter and content are not familiar to most researchers, nor are they topics that are easily understood or determined by looking at the public information provided by repositories.

**Para4, sentence3:** If the proposed criteria are in-line with those existing for trustworthy digital repository certification, it would be easier for researchers to simply look for a repository within their domain that exhibits a certification badge or otherwise displays proof of such certification. Why are redundant pieces of information being presented in a repackaged form?

**Request for Comments Section:**

**Para1:** What are the additional requirements that must be met by repositories operating under Federal agencies wrt security, privacy and accessibility that are NOT included here, and why were they excluded if these criteria are to be applied to such repositories?



**Bullet 2:** Additional Characteristics for data repositories that would store and provide access to Federally-funded research results should include versioning along with Provenance. Preparing data for integration may result in valuable data products. Establishing links from original datasets to data products is important for reuse and new research.

**Bullet 4:** Considerations for repository characteristics pertaining to facilities that manage physical samples may have value in being included (possibly special PIDs for sample metadata, e.g., IGSN).

**I. Desirable Characteristics for All Data Repositories:**

**C. Metadata:** Although there is mention of sufficient metadata to enable discovery, reuse and citation, it is not explicitly stated that such metadata must also be machine readable (in addition to human readable) to facilitate discovery and increased interoperability.

**D. Curation and Quality Assurance:** What constitutes an “expert” in this criteria? Curation and quality assurance of data is much different than that for metadata. Many data curation and quality assurance activities require expertise beyond that available in generalist, some institutional, and even some domain repositories. This term is currently vague and should be clarified.

**G. Reuse:** Although attractive and adds value to the researchers (and potentially other data management stakeholders), the tracking of data reuse statistics in no way impacts the ability for a data repository curate and steward digital objects. This should be considered within the drivers behind such broader desirable criteria and clarified (i.e., are these criteria being used to ensure repositories are capable of stewarding Federally-funded data, or providing value-added services beyond robust curation?).

**H. Secure:** It should be noted that the examples provided for the security of digital data by ISO and NIST are extremely complex and require considerable effort to undertake and fulfill. The ISO standard is only viewable through purchase of the document and the NIST full criteria for “Low Impact” (i.e., minimum compliance) consists of 115 controls. At this time, there are very few, if any, existing geoscience repositories who have applied for and achieved compliance with these standards, even though they may implicitly comply with many of the individual components through their existing operations. The cost and effort associated with successfully demonstrating these is typically prohibitive for grant-sponsored repositories. The CoreTrustSeal certification has been leveraged as a cost effective proxy for satisfying many of these criteria and should instead be used to determine this specific criteria.

**Response to OSTP's Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research**

*Response drafted by Varsha Khodiyar (Data Curation Manager) on behalf of Springer Nature.*

As a research publisher, Springer Nature is home to trusted brands including Springer, Nature Research, BMC, Palgrave Macmillan and Scientific American. As the largest publisher of open access primary research, we have a firm commitment to the drive to openness and open research, in all its manifestations, seeing it as one of the major forces reshaping the way that researchers communicate and collaborate to advance the pace and quality of discovery. Sharing data and research is a central part of open research and open science, facilitating debate and collaboration and advancing progress, and we have a vision of an open research future where every element of the research process is instantly available, discoverable, usable, re-usable and widely shareable.

To make widespread FAIR data a reality, collaboration is essential across fields and between funders, institutions, librarians, researchers and publishers. Therefore, we welcome the opportunity to contribute to the Office of Science and Technology's Request for Information to ensure that we, collectively, continue to drive forwards sustainable goals around data sharing and open research.

Varsha Khodiyar, Ph.D. is an Executive Advisor of the repository indexing service FAIRsharing.org and a member of the CODATA Data Policy committee. As part of the Springer Nature Research Data team, Varsha maintains and curates a list of recommended repositories for use by Springer Nature editors and researchers, and developed the standardized questionnaire currently used to gather information about each repository wishing to be recommended via this list.

**General comments on the proposed Desirable Characteristics for All Data Repositories:**

At Springer Nature we primarily work with researchers in the process of publishing their work, and so the Springer Nature Research Data team work to facilitate sharing of those data which are associated with the article under consideration. Guiding researchers to the most appropriate repository for their data is not always straightforward, and the Springer Nature recommended repository list (<https://www.springernature.com/gp/authors/research-data-policy/repositories/12327124>) was created to meet this need. Thus the development of a common set of desirable characteristics for research data repositories is a welcome advance which will assist those of us working with researchers to encourage the sharing and publishing of research data.

**Specific comments:***A. Persistent Unique Identifiers*

The assigning of Persistent Unique Identifiers (PUIDs) is essential for good data management. We would be supportive of further guidance on what constitutes an acceptable PUID.

With regard to Digital Object Identifiers (DOIs), we understand that both CrossREF and DataCite recommend repositories to use DataCite DOIs, since the metadata format developed by DataCite is specific for datasets. We suggest recommending the use of DOIs from agencies which use a data-specific metadata schema. Further, we suggest DOI-minting agencies be encouraged to allow access to the metadata for each DOI to facilitate data citation.

It is unclear to us whether non-DOI using repositories are under the same obligation as DOI-minting repositories to retain a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available. We suggest clarifying in the PUID paragraph to state whether PUIDs such as handles and ARKs are acceptable options for repositories.

Accession number using repositories are embedded in the biological sciences community, and so this is the only discipline in which we allow accession numbers as PIDs. The primary reason for this is that accession numbers are not consistently globally unique without knowing which repository these came from. We have are therefore working to implement the use of identifiers.org links across our journals where accession numbers are referenced<sup>1</sup>. An example of the identifiers.org implementation can be seen at the journal Scientific Data<sup>2</sup>.

*B. Long-term sustainability*

We think this is a very important characteristic, and regularly converse with repositories about their sustainability and contingency plans. We suggest it would be helpful to provide guidance on what a repository should do to evidence their long-term sustainability.

We suggest adding explicit guidance on what is expected regarding data preservation. At Springer Nature we suggest 'ten years after data publication' as a minimum, and 'ten years after the data were last accessed' as a gold standard for data preservation.

It's become clear that the license chosen for a dataset affects whether this can be archived in a US government funded archive. Our understanding is data with anything other than a Creative Commons waiver (CC0 ) cannot be preserved by the National Archives, and that a public domain designation (not a Creative Commons license) is often used by federal repositories and employees. We suggest mandating repositories to provide guidance on suitable licenses for long-term data preservation for their users.

### *C. Metadata*

We recommend differentiating between metadata enabling data reuse (which will differ between disciplines and communities) and metadata for data discovery. Metadata for facilitating data discovery should be machine readable and ideally would be in a common format across all repositories, e.g. schema.org. In cases where implementation of a standardized metadata schema is not possible, we suggest encouraging the use of common linked data formats for these metadata, e.g. JSON-LD.

We additionally suggest expanding this guidance to require terms of use (or licence) for each dataset should be clearly and consistently displayed on each dataset landing page. We consider this to be vital metadata enabling data reuse, but which we often see missing from dataset landing pages.

### *D. Curation & Quality Assurance*

We agree with this as written. Curation and quality assurance for its data holdings, is an important part of the service a repository should provide.

### *E. Access*

We suggest including the phrase 'sensitive data' to allow the guidance to be inclusive of non-clinical sensitive data (for example geolocation data of endangered species).

### *F. Free & Easy to Access and Reuse*

This is a very important characteristic, since clear documentation of data being in the public domain and available for reuse is often missing from many repositories, meaning that it is often unclear whether an accessible data may be reused or not, and for what purposes the data may be reused.

Attribution and receiving appropriate credit for data is an important consideration for researchers when depositing research data<sup>3</sup>. Attribution of sources is a key part of scholarly communication, since clarifying the provenance for one's research is essential for credibility. However, the perception that researchers may not receive due credit for sharing their data, negatively impacts the willingness of researchers to make their data available. We would welcome specific guidance for repositories on clarifying to their users whether attribution is required or recommended.

### *G. Reuse*

We recommend expanding this guidance to request repositories to provide clear guidance on how each dataset should be cited. This is essential for allowing data citation tracking, which will feed into data reuse tracking.

We recommend repositories should be explicitly guided to facilitate the provision of links from individual data holdings to other research outputs, such as scholarly articles. Clear linking between distinct but related research outputs is essential to facilitate reuse, thereby maximising the benefit of tax-payer funded research.

We also recommend that repositories be required to provide suitable data licences to facilitate maximal data reuse.

#### *H. Secure*

We recommend that the guidance be expanded to state that all documents demonstrating how a repository meet security criteria should be made accessible to all repository users.

#### *I. Privacy*

We recommend that the guidance be expanded to state that all documents demonstrating how a repository meet privacy criteria should be made accessible to all repository users.

#### *J. Common Format*

This characteristic will likely to be difficult for generalist repositories to meet, since they do not exert control over the data formats which a data depositor may use for their data. We suggest that a common data format is unlikely to be a useful characteristic.

However, and as stated above, we support the promotion of a common metadata schema for enabling data discovery.

#### *K. Provenance*

This is an important characteristic, since undocumented (silent) changes to a dataset after the dataset has been made public, are detrimental to the scientific record. In accordance with the Force 11 data citation principles<sup>4</sup>, data should be accorded the same importance as peer-reviewed literature. If data are to be considered as part of the scientific record, repositories should have clear policies (helping to set research expectations) with regard to published data associated with articles that are subsequently retracted or corrected.

**Comments on the proposed Additional Considerations for Repositories Storing Human Data (Even if De-Identified):**

In our experience, researchers occasionally leverage concern about participant privacy to inappropriately deny data sharing requests. We recommend that guidance on sensitive data management include clarification that legitimate concern about protecting research participants be balanced with the possibility of sharing desensitized and non-sensitive aspects of the dataset more openly.

*A. Fidelity to Consent*

We suggest adding explicit best practice guidance that a blank copy of the patient consent form should be included as part of the metadata record for human data. This would facilitate potential data reusers to fully understand the consent given by participants for the original data collecting study.

*B. Restricted Use Compliant*

This is an important consideration for human data. We feel it is appropriate for repositories to be tasked with restricting data access to unauthorized users. However preventing re-identification may not be possible for repositories, and as such including this aspect may limit the number of repositories willing to handle human data.

*C. Privacy*

We agree that repositories have a duty to assist researchers in preparing data to minimize inappropriate access.

*Characteristics D to I*

We applaud the OSTP for considering these important characteristics as being desirable for repositories handling human data.

**References**

1. Wimalaratne, S., Juty, N., Kunze, J. et al. Uniform resolution of compact identifiers for biomedical data. *Sci Data* 5, 180029 (2018). <https://doi.org/10.1038/sdata.2018.29>
2. On the road to robust data citation. *Sci Data* 5, 180095 (2018). <https://doi.org/10.1038/sdata.2018.95>
3. The State of Open Data 2019. *figshare* <https://doi.org/10.6084/m9.figshare.9980783.v2>
4. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 <https://doi.org/10.25490/a97f-egy>

## **OSTP RFC Response: Desirable Repository Characteristics**

**Submitted By: Duke University**

### **Contributors:**

#### **Duke University Libraries Research Data Working group**

(Co-Chair) Jennifer Darragh, Research Data Management Consultant

(Co-Chair) Sophia Lafferty-Hess, Research Data Management Consultant

Karen Barton, Biomedical Research Liaison Librarian

Ryan Denniston, Librarian for Public Policy and Political Science

Moira Downey, Digital Repository Content Analyst

Ciara Healy, Librarian for Psychology & Neuroscience, Mathematics, and Physics

Joel Herndon, Director, Center for Data and Visualization Sciences

Shadae Gatlin, Digital Repository Content Analyst

Alex Jakubow, Associate Director for Empirical Research and Data Support Services

Will Sexton, Head, Software Services

Lee Sorensen, Librarian for Visual Studies and Dance

### **The proposed use and application of the desirable characteristics (as described in the “Background” section above)**

We agree that the description of the rationale for the creation of desirable repository characteristics makes sense given the rapid proliferation of repositories being launched by commercial entities, funders, publishers, institutions and discipline groups. With the breadth of the data publishing and preservation landscape, it is useful that you have explicitly stated that these characteristics are not exhaustive, and are for guidance rather than any sort of official evaluation or endorsement. We would caution that guidance versus endorsement can be a difficult line to parse as researchers may assume “all or nothing” - as in, if a repository does not match every characteristic described it should not be used. If you are working towards a list of “minimally acceptable repository criteria,” it would be best to be explicit.

### **The appropriateness of the “Desirable Characteristics for All Data Repositories” (Section I) for data repositories that would store and provide access to data resulting from Federally supported research, considering:**

#### **SECTION 1:**

Generally, the characteristics are useful and map closely to other official certifications. They are relatively high level and include specific examples when necessary. They will also be helpful for repositories as they evaluate building in new features and curation processes.

#### D. CURATION & QUALITY ASSURANCE

We are glad to see that the term curation is listed explicitly, as it is an important part of making data meet the FAIR guiding principles. However, curation can be done at a very low-level (ensure appropriate discovery metadata is assigned) to a very high level (disclosure risk assessment, de-identification, code review, data harmonization, etc.). Researchers should be informed that not all repositories curate data the same way, and they will need to evaluate their needs versus what the repository offers.

#### C. METADATA

Levels also relate to metadata assignment. Some repositories offer more flexible metadata such as Dublin Core due to the variety of disciplines they support. Others may offer more granular metadata based on community or disciplinary standards due to the specific disciplines they support. Again, reminding the researcher to evaluate their needs versus what the repository offers is important.

#### H and I: SECURE AND PRIVACY

These two characteristics may be difficult for researchers to evaluate without more examples. As in, is it enough to know that a repository adheres to its parent organization's security policies and procedures? In instances where the repository is not stand-alone but hosted (which is often), security policies and procedures are at the hosting organization/enterprise level rather than the repository level. For example, Duke's Research Data Repository follows Duke University-wide system security policies. In addition, this type of information is not externally facing, and often for good reason (to avoid hacking). Explaining that researchers may need to specifically ask for security policies would be helpful along with some clarity regarding exactly what the researcher should confirm.

#### **ADDITIONAL CHARACTERISTICS**

We think that the RETENTION characteristic listed in Section II might be useful to include here, perhaps as part of B. Long-Term Sustainability. Does the repository have a formal retention or preservation policy? This is something we often recommend to researchers when they ask for our help in determining an appropriate repository solution.

#### E: ACCESS

It would be worthwhile to mention clear licensing and use terms here as many repositories support Creative Commons or other terms of use.



**Appropriateness of the characteristics listed in the “Additional Considerations for Repositories Storing Human Data (even if de-identified)” (Section II) delineated for repositories maintaining data generated from human samples or specimens, considering:**  
SECTION II

Overall, we think it is a good idea to have specific characteristics for the mitigated sharing of potentially sensitive research data. However, not all sensitive research data is about human subjects. It may be information under Export Control or for other ethical reasons (poaching, looting, etc.). It would be useful to make this clarification.

**E. DOWNLOAD CONTROL**

It should not be assumed that data would be downloaded if in a restricted-access repository. Some repositories may provide access through virtual machine environments (enclaves); others may send external encrypted media. This characteristic should be reworded to account for data delivery and access methods pursuant with data use agreements and security plan terms as required to protect data from unauthorized access. Perhaps a better term would be END USER ACCESS CONTROL or DATA DELIVERY MECHANISMS.

**F. CLEAR USE GUIDANCE**

It would be useful to include some examples here such as Data Use Agreement, Contract, Terms of Use, and Data Security or Data Management Plan.

**G. RETENTION GUIDELINES**

It is not immediately clear if the retention guidelines here refer to how long an end-user may access the data for or how long the repository will retain the data. If it is the former, being more explicit about how it pertains to the end user would be helpful. If it is the latter, this would be a useful item to include for all repositories in Section I., perhaps as part of B. Long-Term Sustainability.

**Considerations for any other repository characteristics which should be included to address the management and sharing of unique data types (e.g., special or rare datasets)**

Large datasets (100+ GB) pose a challenge to many repositories due to the complexities in data transfer (both upload and download) as well as with long-term storage and preservation. It would be helpful for funding agencies that know large-scale data will be generated from their funded projects to list repositories that readily accept data of this size.

As mentioned in our comments for Section II, it is important to note that there are other kinds of data that require protection and mitigated access but are not from “human subjects”

including data that may be subject to export control or for other ethical reasons to prevent crime (looting, vandalism, poaching).

### **The ability of existing repositories to meet the desirable characteristics**

Many repositories would be able to meet most, but not necessarily all of these characteristics - such as curation, security, privacy, common format and sustainability. Some repositories only offer self-deposit with very minimal curation. Having better definitions of what level of curation is at least minimally acceptable will help. It will also take additional work for repositories to ensure that they have public documentation on system security and privacy for researchers to easily access and understand. With regard to common format, many repositories will do their best to ensure that deposits are in open formats, however not all files are easily transformable due to the rapid pace of software development and highly specialized equipment use. Finally, for sustainability, this will vary between repositories and it is often unclear if long-term means “forever.” It would be helpful to provide clarification such as “sustainability that meets the terms of your funding agreement (i.e. 10 years post publication)” or that the “repository possesses a clear data retention policy.”

### **Consistency of the desirable characteristics with widely used criteria or certification schemes for certifying data repositories**

These characteristics seem to be particularly consistent with the CoreTrustSeal. This consistency is useful as it provides an avenue for repositories to transparently demonstrate that they meet these desirable characteristics and their overall trustworthiness even if they are not already certified or in the process of certification.

### **Any other topic which may be relevant for Federal agencies to consider in developing desirable characteristics for data repositories.**

How these characteristics will be used in practice is still an open question. Researchers will need help assessing these characteristics and reviewers of DMPs will need training to ensure they do not read these characteristics like a compliance checklist. It would be worthwhile to include information that encourages researchers/agencies to talk directly to a repository if they have questions about what characteristics are implemented or to seek help from their institution (from the libraries, research compliance or integrity office, IRB, etc.) in reviewing/identifying an appropriate repository.



March 17, 2020

Lisa Nichols  
Office of Science and Technology Policy  
[OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

Subject: University of Minnesota “RFC Response: Desirable Repository Characteristics”

Dear Dr. Nichols,

The University of Minnesota writes in response to the “Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research” posted January 17, 2020 as document 85 FR 3085 in the Federal Register. As a public land grant institution we strongly support federal agency policies to ensure that the results of federally funded research are properly stored in trusted data repositories that optimize the public’s ability to locate, manage, share, and re-use.

Our institution has a long history of providing stable, long-term repositories for research data; there are currently 15 data repositories hosted at the University of Minnesota (U of M) that are [listed in the re3data.org registry](#), including the [Clinical Data Repository](#) and the [Data Repository for the University of Minnesota \(DRUM\)](#). Our University has established infrastructure and support for research data sharing via the [Libraries’ Data Management Service](#), the [University Storage Council](#), the [Storage Champion Network](#), and governance via a [Research Data Management Policy](#). With input gathered by data repository managers and data service providers across campus, we would like to respond to the proposed characteristics.

### Background Section

- As representatives of a large multi-disciplinary university that works with dozens of funding agencies, we are happy to see improved consistency across agencies.
- We appreciate the use of the existing OMB circular A81 definition of data.
- While we appreciate any consistency with the Federal Data Strategy, we would like to better understand how (if) the proposed desirable characteristics will be included in the principles and actions of the strategy.
- While “data” is defined, a “data repository” is not well defined and could possibly be confused with a data catalog or data library (e.g., for physical specimens). Consider including a data repository definition such as “a type of repository where data, data

objects, and data collections are permanently stored, managed and made accessible.”<sup>1</sup>

- The intended uses of these characteristics is appropriate (guidance by federal agencies to help direct researchers, etc.) and the recommendations do not appear overly burdensome, rather, this is the norm for well-managed digital repositories.

## Section 1: Desirable Characteristics for All Data Repositories

**A. Persistent Unique Identifiers (PUIs):** We agree that PUIs for data is required. Also, to help ensure proper attribution, the repository should include a suggested citation for the dataset and have terms of use that require attribution back to the original researchers.<sup>2</sup> Furthermore, a PUI for the dataset should be accompanied by linked data (e.g., data with unique identifiers) to other contextual elements surrounding that dataset, including people, institutions, related publications, funders and the home repository.<sup>3</sup>

**B. Long-term sustainability:** We recommend a peer review system, similar to the CoreTrustSeal certification process, for data repositories to receive an independent peer-reviewed assessment of long-term sustainability.

**C. Metadata:** Sufficient metadata (when expressed in machine-readable formats) is a critical component for enabling the discovery, reuse, and citation of datasets. For data repositories, who may be serving a broad community of diverse disciplines, we recommend that OSTP present a recommended minimum set of metadata elements for repositories to adhere to. These should include: dataset PUI, author, author PUI, author affiliation, author affiliation PUI, title, date published, source repository, source repository PUI, license, license PUI, abstract (of the data, not the related article), related publication, related publication PUI, geographic coverage, temporal coverage, terms of use, level of openness (see Access).

**D. Curation & Quality Assurance:** We strongly agree that curation assistance is a key characteristic. Professional curators take many actions to ensure a dataset’s usefulness over time. For example, the University of Minnesota is the lead institution in the [Data Curation Network](#) and we train curators on applying CURATE steps to every dataset (Check, Understand, Request, Augment, Transform, and Evaluate for FAIRness). In addition the Data Repository for the U of M has eight data curators who help authors appropriately share their data for the repository.

**E. Access:** We agree.

---

<sup>1</sup> Research Data Alliance Term Definition Tool [https://smw-rda.esc.rzg.mpg.de/index.php?title=Data\\_Repository](https://smw-rda.esc.rzg.mpg.de/index.php?title=Data_Repository)

<sup>2</sup> Pierce, Heather H., Anurupa Dev, Emily Statham, and Barbara E. Bierer. "Credit data generators for data reuse." *Nature* **570**, 30-32 (2019). doi: 10.1038/d41586-019-01715-4.

<sup>3</sup> A recent conference (report yet to be released) expounds on this idea: "Implementing Effective Data Practices: A Conference on Collaborative Research Support, was held on December 11–12, 2019, in Washington, DC. <https://www.arl.org/implementing-effective-data-practices/>

**F. Free & Easy to Access and Reuse:** We suggest that repositories utilize standard licenses to enable the broadest possible reuse, such as CC0, when appropriate.

**G. Reuse:** Reuse is not an inherent quality of a repository since successful data reuse is dependent on many different trust factors related to the data itself (but also including repository reputation).<sup>4</sup> However this criteria speaks more about tracking reuse analytics. This criteria could be renamed or combined with PUIDs.

**H. Secure:** Data security standards governing a data repository should conform to all established federal and local laws. Citing these particular two standards does not pursue the highest level of data protection. For example, the [U of Minnesota Information Security policy](#) offers appendices that include 16 important security standards with detailed guidance.

**I. Privacy:** We are not clear whose privacy is referenced in this characteristic. Privacy of human subjects is addressed in Section 2. Does this characteristic refer to safeguarding the privacy of people who are accessing and downloading data from a repository? Also how does this characteristic take international standards for user privacy into account (e.g., GDPR).<sup>5</sup>

**J. Common Format:** The type of data formats that are submitted to a general repository can vary widely. For less common data formats, it may not be obvious what the standards-compliant format for that file is, nor possible to transform a particular file to a preferred format without specialized software. This characteristic suggests that the repository will be responsible for ensuring that data files are available in a standards-compliant format, but this may not be feasible for all instances. Instead, we recommend that repositories provide clear guidelines for preferred formats and how they will treat non-compliant formats in the long-term.<sup>6</sup>

**K. Provenance:** Provenance is an important characteristic of trusted data repositories and critical to maintaining and tracking the integrity and authenticity of data. One evolving feature of repositories is whether to make the detailed log-file public. This information of when the data were received into the repository, how long they remain in the curation process, detailed changes that were made (and by whom), and when they are released for public access may have an impact on scholarly metrics such as patents or citations. We ask, is “maintenance” enough, or should this information be made transparent for public use?

## Additional characteristics that should be included

**Preservation:** Repositories that actively monitor and take action to ensure the long term

---

<sup>4</sup> Yakel, E., Faniel, I.M. & Maiorana, Z.J. (2019). Virtuous and vicious circles in the data life-cycle. Information Research, 24(2), paper 821. Retrieved from <http://InformationR.net/ir/24-2/paper821.html>

<sup>5</sup> <https://gdpr.eu/>

<sup>6</sup> For example, see the preservation policy and format recommendations for the Data Repository for the University of Minnesota (DRUM), <https://conservancy.umn.edu/pages/policies/#preservation>.

preservation of data is a desirable characteristic. Evidence of this may be through the use of [PREMIS](#), the preservation metadata standard.

**Documentation:** Repositories that require adequate documentation describing the nature of the data at an appropriate level for reuse is a desirable characteristic. The repository should offer guidance and assistance prior to rejection for data that do not meet this criteria.

While structured metadata is often expressed in machine-readable formats, additional structured and/or unstructured “documentation” is often required to provide the level of detail needed for an individual to use and understand the data. Documentation can come in many forms such as a code book (a well-structured output file generated by a statistical software package), a lab notebook or lab manual (unstructured text detailing the methods, quality control measures, and other parameters of the data collection and processing), or a simple “readme” text file that provides core information about the dataset. Most data files are NOT self-describing and may include difficult to interpret codes, acronyms, symbols, blank/null cells, and other processing elements that have a direct impact on the interpretation and successful reuse of data. Therefore, data curators at the Data Repository for the University of Minnesota for example, often request additional documentation from the researcher or consult with them to create a readme file using [our template](#), prior to acceptance into a repository. Our policy requires that “Data must include adequate documentation describing the nature of the data at an appropriate level for purposes of reuse and discovery. All data receive curatorial review and data that are incomplete or not ready for reuse may not be accepted into the repository.”

**Clear Use Guidance and Retention Guidelines:** We would like to see these characteristics included in Section 1 and apply to all repositories.

## **Section 2: Additional Considerations for Repositories Storing Human Data (even if de-identified)**

**A. Fidelity to Consent:** We recommend “Restricts dataset access to appropriate uses **and audiences** consistent with original consent...” as consent forms typically restrict reuse of the data to certain contexts, but also to certain individuals (such as “only researchers will see the data,” which may preclude making the data available in a publicly accessible repository). Furthermore, for this characteristic to be implemented, the data repository must review a (blank) copy of the consent form to determine the appropriate level. It is often the case that researchers may have placed high restrictions on their data that limit sharing. Therefore it is very useful to have an IRB office associated that the repositories can turn to for expert guidance, as we do at the University of Minnesota.

**B. Restricted Use Compliant:** Research that shows how reidentification is possible is a valid

and important area of study.<sup>7</sup> Rather, repositories should restrict improper use of that information, for example, [DRUM Terms of Use policy](#) states “The user will not make any use of data to identify or otherwise infringe the privacy or confidentiality rights of individuals discovered inadvertently or intentionally in the data.”

**C. Privacy:** Privacy is not the same as security. Inappropriate access as described here is a security issue.

**D. Plan for Breach:** Some public access repositories may have deidentified human subjects data that are appropriate to share, and are publicly accessible and downloadable. This characteristic would not apply because a breach would be not possible when the data are publicly available for download.

**E. Download Control:** This may not apply to public access repositories that hold deidentified human subjects data.

**F. Clear Use Guidance and G. Retention Guidelines:** These characteristics should apply to all repositories.

**H. Violations:** This characteristic could be combined with the characteristic on restricted use compliance as addressing violation is a more reasonable expectation than prevention.

**I. Request Review:** How would this group interact or overlap with the IRB?

### **Additional characteristics that should be included**

We are happy that these guidelines go into detail for human subjects data. However, we recommend that this section be broadened to include other sensitive data types such as endangered species, protected sites, indigenous data sovereignty, and others.

Sincerely,



Lisa Johnston  
Director, Data Repository for the University of Minnesota (DRUM)  
University of Minnesota Libraries

---

<sup>7</sup> See for example, De Montjoye, Yves-Alexandre, Laura Radaelli, and Vivek Kumar Singh. "Unique in the shopping mall: On the reidentifiability of credit card metadata." *Science* 347, no. 6221 (2015): 536-539. <http://doi.org/10.1126/science.1256297>.

**From:** Lin Mohle <[mohlerlin@gmail.com](mailto:mohlerlin@gmail.com)>  
**Sent:** Tuesday, March 17, 2020 12:01 PM  
**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>  
**Subject:** [EXTERNAL] Data repositories comment

I agree, it sounds like a plan to eliminate data from consideration in environmental health decisions if the data were collected at a clinical trial or other study where the human consent form pledged confidentiality. They're doing it under the cover of "transparency".

They got away with taking down EPA data (saved by vigilant researchers around the country, thank you). What can we do to act on this?

Lin Mohler





American Society of Agronomy • Crop Science Society of America • Soil Science Society of America

5585 Guilford Road, Madison WI 53711-5801 • Tel. 608-273-8080 • Fax 608-273-2021

[www.agronomy.org](http://www.agronomy.org) • [www.crops.org](http://www.crops.org) • [www.soils.org](http://www.soils.org)

To: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)  
OSTP Chief of Staff, Sean C. Bonyun,  
Re: **RFC Response: Desirable Repository Characteristics**

Dear Mr. Bonyun,

The American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America represent more than 8,000 scientists in academia, industry, and government. We support more than 13,500 Certified Crop Advisers (CCA), and more than 700 Certified Professional Soil Scientists (CPSS). We remain fully supportive of open science initiatives that improve the accessibility and transparency of our sciences and thank you for the opportunity to provide comments on data repositories.

Our members are keenly interested in data repositories that ascribe to FAIR (Findability, Accessibility, Interoperability, Reusability) principles, but challenges and overly optimistic promises associated with the build-out of repositories for agricultural data have tempered enthusiasm. The Societies, therefore, urge OSTP to be thoughtful regarding burdens placed on researchers and the unrealized responsibilities of repositories that currently lack capacity and expertise to achieve OSTP's proposed characteristics.

#### **Researchers need federally supported tools and training to accelerate data reporting**

Most researchers are not funded, trained, or otherwise incentivized to annotate and organize their data and provide the necessary meta-data in a way that would maintain OSTP's proposed data repository characteristics and meet FAIR principles. Most of the proposed characteristics focus on managing data once it is in a repository, but our researcher's decade of experience has made clear that for data to be deposited, there is a critical need for data tools and workflows that enable researchers to take raw datasets and those in statistical formats and easily assemble them into formats that enable general reuse.

For example, federal agencies should hire specialists to create data extraction and upload wizards for automatic extraction, standardized formatting, and depositing of data directly from research equipment. These data specialists could work with research equipment designers and users to ensure that their products are equipped to deliver collected, calibrated data in FAIR format that can be user-verified and that include metadata on how logged data were verified, processed, and calibrated. The Societies also support a reasonable embargo period for data from uploaded but yet unpublished research (e.g. multi-season studies) so that researchers have time to conduct rigorous statistical analyses and submit manuscripts for peer review and subsequent publication.

Automatic upload of data from devices may afford an easier and more systematic path to data repository compliance. However, there are large amounts of data that cannot be automatically uploaded from most scientific equipment and is, therefore, manually recorded, sometimes with e-tablets and spreadsheets, other times with paper and pencil. We suggest that federal agencies offer training and workflow tools for researchers and students so that they understand ahead of time the

data repository requirements and FAIR principles so that these manually recorded data can more easily be transitioned to repositories.

### **Agriculture and natural resource researchers need a fully-supported data repository**

It is not enough to mandate principles for data repositories without fully supporting researcher participation and database functionality. For example, the U.S. Department of Agriculture's Agriculture Research Service (USDA ARS) supports the Ag Data Commons, but this data repository is too small and under-resourced to handle modern agriculture research datasets. The U.S. Department of Energy's Knowledge Discovery Framework (KDF), for example, is only open to data from biofuels research. The National Science Foundation's iPlant, now CyVerse, also has been proposed as an alternative federally-sponsored data repository, but it is not known among the agricultural research community nor has it invested in making its data FAIR, resulting in datasets with opaque identifiers and non-standard formats that are unusable to any but the researchers who deposited them.

Thought must be given to how federal repositories can be structured moving forward so that large and interdisciplinary datasets can be included, and this includes data created in conjunction with the private sector. For example, a member-scientist recently initiated a collaboration with an agricultural consultant who has assembled more than 530 million rows of data in a spreadsheet with nearly 100 geo-referenced traits for each row. No federally supported data repository is equipped or open to receive such a dataset, and no guidelines exist for how this data could be formatted to make it comply with FAIR principles. And yet, datasets resulting from public-private collaborations like this are the future of modern agriculture. Without investments in the work-flow tools that researchers need to get data into these repositories or the incentives to make uploaded data follow FAIR principles, progress will languish.

### **"Access" alone may not be enough to make data findable and usable**

The Societies are concerned that OSTP's proposed characteristics could potentially describe a "dark archive," where the data is there but not discoverable. Repositories must be readily searchable by commonly used search engines and data formats. Data should be linked to the publications, and vice versa. Inclusion of Persistent Unique Identifiers (PUIs), like a Digital Object Identifier (DOI), is an absolute necessity. Thought also must be given to versioning of data sets so that data accrued in multi-year studies and similar situations can be identified with absolute certainty.

### **Long-term sustainability and business models for repositories need to be defined**

As a public good, the Societies support federal funding of key data repositories to ensure long-term program sustainability. Preservation and curation practices should ascribe to the best management practices, including frequent file back-ups, strategic distribution, and disaster-recovery protocols. Business models should consider triaging data curation according to its use and apparent value, moving less-used and/or limited value datasets to less costly preservation systems. Library scientists, other preservation specialists, and the relevant research communities should work jointly to develop curation guidelines for data.

### **"Additional Considerations" for data privacy are needed for farm data**

Our scientists often depend on data collected on privately owned farmland. The requirement to make all such data public may deter these important studies, research that enables the scaling of research findings. For this reason, perhaps OSTP's proposed "Additional Conditions" should apply to on-farm data as well as human data. "Fidelity to Consent," "Restricted Use Compliant," "Privacy" and OSTP's other proposed characteristics for human data repositories may all apply to many on-farm research datasets

and their respective landowners. Also, the Societies suggest that OSTP include “confidentiality” to its list of privacy safeguards (II.C).

Again, we thank OSTP for providing our Societies the opportunity to comment on this important issue. Please feel free to contact me if you have questions.

Sincerely,

A handwritten signature in black ink, consisting of a stylized 'N' followed by a 'G' and a horizontal line extending to the right.

Nicholas J. Goeser, CEO  
American Society of Agronomy  
Crop Science Society of America  
Soil Science Society of America



OFFICE OF THE VICE PRESIDENT - RESEARCH AND INNOVATION

OFFICE OF THE PRESIDENT  
1111 Franklin Street, 11<sup>th</sup> Floor  
Oakland, California 94607-5200

March 17, 2020

Dr. Lisa Nichols  
Assistant Director for Academic Engagement  
Office of Science and Technology Policy  
Submitted via email: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

**RE: Docket ID [OSTP-2020-0001](#) Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research (RFC Response: Desirable Repository Characteristics)**

Dear Dr. Nichols:

I write on behalf of the University of California (UC) system with regard to the Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research issued on January 17, 2020.

The UC system is comprised of ten research-intensive campuses, six medical schools, and three affiliated U.S. Department of Energy national laboratories. As a system, UC receives approximately \$6 billion annually in extramural research awards and is the nation's largest academic recipient of federally funded research and other university-based projects. In 2018, UC received more than \$2.95 billion in federal agency research funding.

The UC strongly values open science and applauds this Notice issued by the Office of Science and Technology Policy (OSTP) to solicit feedback and recommendations on approaches for ensuring long-term stewardship of, and broad public access to, data resulting from federally funded research.

We appreciate that OSTP wants to decrease burdens on researchers by setting data repository standards for federal agencies to provide optimization and improved consistency across the federal government's repositories. While UC generally agrees with the OSTP's Draft Desirable Repository Characteristics, we believe that in order to make this policy a success, OSTP should further promote harmonization across agencies and departments and consider the cost to curating and preserving research data. These considerations, provided below, are in addition to the comments on specific aspects of the Draft Desirable Repository Characteristics.

## **Promoting Harmonization Across Federal Agencies and Departments**

UC recommends that OSTP promote harmonized regulatory guidance for data curation, preservation and sharing across all federal funding agencies. Currently, regulatory guidance varies between federal research agencies; this is due in part to the differences in types of data collected and varying practices between scientific disciplines. Improving regulatory alignment between agencies will help to improve greatly the overall research enterprise and reduce burdens and costs for both researchers and institutions.

One concrete way in which OSTP can promote harmonization is by encouraging the use of the Uniform Guidance OMB Circular A-81, section 200.315, definition of “research data,” referenced in this very Notice.<sup>1</sup> One recent example of how this definition is not used consistently is the issuance of NIH’s draft policy. The definition of “scientific data” as proposed by NIH expands beyond the Uniform Guidance definition by stating that scientific data includes recorded information that is “necessary to validate *and replicate* research findings” [emphasis added]. Aligning the definition of scientific data across all federal funding agencies will ensure proper management of scientific data and reduce confusion among the research community.

## **Cost to Research Data Curation and Preservation**

Universities, their research administrators, librarians, and technology specialists are in a good position to advise investigators as they curate and preserve their research data. However, the cost of data curation and preservation is huge, and cannot be fully borne by individual universities. We believe that federal funding agencies must increase their support for the expansions of local and disciplinary data storage capacity to meet the need to maintain data in a usable format.

These costs are, in fact, a critical component of disseminating research and ensuring research quality. Thus, data management, de-identification, curation and preservation costs should be allowed as a direct cost by research granting agencies. We note that the current 26% cap on indirect cost recovery constrains universities’ ability to pay for the infrastructure and additional resources necessary to ensure public access to research results, particularly for biological data collected from medical patients as patient data is understandably subject to strict confidentiality protocols.

## **Feedback on Desirable Characteristics for All Data Repositories**

We appreciate the comprehensive list of “Desirable Characteristics for All Data Repositories” (Section I) that provides information on how optimally to store and promote access to data resulting from federally-supported research.

---

<sup>1</sup> Uniform Guidance OMB Circular A-81, section 200.315 provides the following definition: “Research data means the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This “recorded” material excludes physical objects (e.g., laboratory samples). Research data also do not include: (i) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and (ii) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study.”

We particularly note that the recognition of datasets as a citable, authored sets of information will help protect the interests of individual stakeholders. In this regard, we suggest that data repositories provide each dataset with a citation that can be referenced in articles, resumes, etc. for both acknowledgement and for promoting reuse. To help ensure that citations and relationships between outputs (i.e. articles and related data) are indexed, repositories should send data and article relationships to DataCite, a central and open indexer for metadata.

We also note that the following attributes to the list of desirable repository characteristics should be included:

- *Persistent Unique Identifiers*: Repositories should support versioning of the Persistent Unique Identifiers like digital object identifiers, accession numbers, and others.
- *Metadata*: Repositories should implement best practices for standardized vocabularies in the metadata (such as, Crossref Funder Registry).
- *Free & Easy to Access and Reuse*: Repositories should be expected to implement Creative Commons licenses for published datasets.
- *Secure*: Adequate protection against security breach is important to protect the data from bad actors, both internal and external, to the US. Security measures would be different by area of science and should be set and evaluated by experts. We recommend that repositories provide documentation of its practices that prevent unauthorized access/manipulation of data.
- *Provenance*: Provenance tracking of datasets should be machine-readable.

Lastly, we strongly encourage the training for federal agency staff to act as partners with grantees and researchers in developing quality data and data management plans that include the elements laid out in the list of “Desirable Characteristics for All Data Repositories.” Data management is a constantly evolving field and agency partners should have the capacity to collaborate with researchers as data elements change.

### **Feedback on Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

The access to and sharing of human subjects-related data is governed by a complex, fragmented set of ethical and legal requirements. Frameworks for accommodating these data, at scale, have not been developed. UC appreciates that the draft acknowledges the importance of human subject protections without including a mandate for IRBs to verify or otherwise be a gatekeeper for data sharing and management. UC recommends that the OSTP work across federal funding agencies to provide guidance on appropriate ways to maintain sensitive data, use cases when providing access to others would be appropriate, and language about the need for repositories themselves to have in place mechanisms for preventing or discouraging re-identification of de-identified data. We also ask OSTP to consider issuing guidance on standards for uncontrolled access, de-identification, application of confidentiality policies, consequences of participant withdrawal and ability for a participant to decline data sharing, and how requirements such as the Health Insurance Portability and Accountability Act, the European Union General Data Protection Regulation and other data protection laws apply. This would not only decrease administrative burden on researchers and grantee institutions, but also promote the goal of long-term data maintenance and accessibility.

Thank you for the opportunity to comment on this important issue and we look forward to continued engagement on this issue as further policies and other guidance is developed.

Sincerely,

A handwritten signature in black ink that reads "Lourdes G. DeMattos". The signature is written in a cursive, flowing style.

Lourdes G. DeMattos  
Acting Executive Director  
Research Policy Analysis & Coordination  
Office of Research & Innovation

# Association of Research Libraries Comments on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research

March 17, 2020

## Introduction

The Association of Research Libraries (ARL) thanks the US Office of Science and Technology Policy (OSTP) for the opportunity to submit comments on desirable characteristics of repositories for managing and sharing data resulting from federally funded research. ARL is a nonprofit membership organization of 124 research libraries in the United States and Canada whose mission is to advance research, learning, and scholarly communication.

Our member libraries, which include academic libraries along with federal and large public libraries, manage data repositories and consult with researchers on deposit into disciplinary and/or agency repositories. Librarians also work with researchers to curate data for deposit. Research data stewardship—including curation, preservation, and development of tools for reuse—involves many different stakeholders, and OSTP’s guidelines to advance our shared understanding of repository characteristics are welcome. ARL recognizes the [excellent response of our colleagues](#) in the Confederation of Open Access Repositories (COAR) and SPARC to this request for information.

Just as OSTP recommends a common set of characteristics for data repositories, knowing there will be disciplinary and domain variation, ARL asks that OSTP consider harmonization of federal policies with respect to the definition of research data for sharing, as well as support for the cost of data curation and long-term preservation.

## I. Desirable Characteristics for All Data Repositories

ARL supports “Desirable Characteristics for All Data Repositories,” I-A through I-K, with the following additional recommendations and suggestions:

### A. Persistent Unique Identifiers

In order to deploy persistent unique identifiers (PUIDs) as a critical piece of infrastructure for provenance and replicability, ARL recommends that repositories:



- Embed digital asset versioning in PUIDs
- Include identifiers for people, organizations, data, and funding

### **B. Long-term Sustainability**

Research libraries seek accountability for both sustainability of the **software or repository platform** and the long-term sustainability of the **individual assets or data sets** within the repository. ARL recommends that data repositories:

- Develop long-term plans for funding and sustaining their infrastructures, and for documenting individual assets in accordance with public-data retention policies

### **C. Metadata**

In order to convey knowledge of data use terms, and to standardize where possible, ARL recommends that repositories include licensing and reuse terms in any metadata schema, and that OSTP:

- Direct generalist repositories that serve multiple disciplines to general purpose metadata standards, such as the [DataCite Metadata Schema](#)

### **D. Curation & Quality Assurance**

Data curation and quality assurance are critical for discoverability, long-term sustainability, and interoperability of assets in data repositories. These activities are also resource intensive. Research libraries expect the following:

- Curation is a partnership among data creators, curators, and repository managers, and that libraries are recognized as a source of broad expertise in this area.
- With targeted federal investment in university capacity, librarians and other experts can work with data creators to improve the quality of data sets before stewardship is transferred to a data repository, especially federal repositories.
- By partnering with national groups like the [Data Curation Network](#) that provide expertise not available locally as well as set standards for levels of curation, federal agencies can leverage distributed networks of knowledge.

### **E. Access**

In order to facilitate the broadest possible access to data, data repositories should:

- Ensure that data repositories are maximally open to machines as well as people, through user-friendly interfaces and open APIs
- Document access restrictions with reference to specific legal guidelines or ethical frameworks

### **F. Free & Easy Access and Reuse**

In order to ensure access and reuse, repositories should:

- Integrate and implement [Creative Commons](#) license terms for published data sets, and include clear disclosure of licensing terms in the metadata

### **G. Reuse**

In order to enhance discovery for reuse, repositories should:

- Include PUIDs, and machine-readable, standardized licenses, in citation metadata

### **H. Secure**

(Nothing to add.)

### **I. Privacy**

In recognition that some repositories exclusively collect data that will be made openly available, we ask OSTP to:

- Clarify that “In cases where the repository is collecting sensitive data, it will provide documentation related to the safeguards in place to protect data from access breaches.”

### **J. Common Format**

Providing access to data in a common format is dependent on the type of data that is provided to the repository. ARL recommends that:

- Transforming content that may be obsolete or content that may not have an open standard be excluded from this requirement

### **K. Provenance**

To further ensure clarity on provenance, ARL recommends that repositories:

- Implement versioned, machine-readable provenance tracking

## **Additional Characteristics Requested for All Repositories**

### **L. Retractions**

ARL recommends that repositories:

- Clearly indicate to potential data users if a data set is subject to a retraction

### **M. Open Source Platforms**

ARL recommends that repositories:

- Use open source tools and frameworks for repository development whenever possible

- Provide source code for the repository platform in a publicly auditable venue and preferably licensed with an open source license

## II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)

### A. Fidelity to Consent

- Ensure that appropriate systems are in place to confirm that data use is consistent with the original permission provided by the participants, even when the data is shared in a repository.
- For data sets with privacy concerns, a full data package will be required for consistency, including a copy of original consent forms, protocols, institutional review board (IRB) requirements, etc.

### B. Privacy

- Outline what security techniques to look for when evaluating a repository for storing human data.

## III. Additional Characteristics for Sharing of Human Subjects' Data

- Include documentation of the utility of the repository under various international privacy policies.
- Include documentation of the infrastructure in place to support the sharing of human data. Without such information, it is impossible for researchers to assess the appropriateness of the repository for their research.

Thank you for your consideration of these comments.

Sincerely,  
Mary Lee Kennedy  
Executive Director  
Association of Research Libraries

**About the Association of Research Libraries**

The Association of Research Libraries (ARL) is a nonprofit organization of [124 research libraries in the US and Canada](#) whose mission is to advance research, learning, and scholarly communication. The Association fosters the open exchange of ideas and expertise, promotes equity and diversity, and pursues advocacy and public policy efforts that reflect the values of the library, scholarly, and higher education communities. ARL forges partnerships and catalyzes the collective efforts of research libraries to enable knowledge creation and to achieve enduring and barrier-free access to information. ARL is on the web at [ARL.org](#).

###

**From:** Westra, Brian <[brian-westra@uiowa.edu](mailto:brian-westra@uiowa.edu)>  
**Sent:** Tuesday, March 17, 2020 2:25 PM  
**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>  
**Subject:** [EXTERNAL] RFC Response: Desirable Repository Characteristics

Responder: Brian Westra, University of Iowa  
Primary scientific discipline: Cross-disciplinary  
Role: Librarian – Data Services

The technical and human-mediated capacities of repositories are good indicators of the fit of a repository for a given need. The measures within each characteristic might be improved by outlining the technical, organizational, and human factors that will enable or support those objectives. For instance, alignment with FAIR principles is increasingly expressed as a goal in repository and data management communities. While FAIR has technical endpoints, with varying degrees of measurability, alignment with those principles is also dependent to a large degree on human mediated services and elements, such as curation, semantic frameworks, and the development and implementation of vocabularies that will evolve with research practices and scholarly communication.

Since the guidance is directed at three different potential audiences (repository developers and managers, researchers, and federal research funders), a glossary would contribute to shared understanding of terminology. The guidance should also note resources such as local experts in the data management, curation, preservation; repository communities; and online educational materials and resources that are already available. There is an established and growing community of data librarians and curators who can contribute to improvements in managing and sharing data. The intended audiences might benefit from generalizable use cases that demonstrate compliance with the guidelines, with the caveat that a use case should not be interpreted as a baseline for all repositories.

There is some likelihood that the guidance will be interpreted or used as a checklist, so the criteria that are included should be chosen with care. FAIR principles provide indicators across a spectrum of important characteristics, but unfortunately, casual references to FAIR within some communities are contributing to misapplication and misinformation. If FAIR is to be included in this guidance, it should be attached to technical metrics to mitigate this issue.

Similarly, CoreTrustSeal Trustworthy Data Repositories Requirements are a laudable goal, but official certification comes with a not insignificant cost in fees and staff time and effort. Some repositories may elect to instead do a self-audit using the criteria, and some repository managers might rely on existing certification and documented practices by others for the repository system they employ. Using official CoreTrustSeal certification as a baseline might unnecessarily penalize these repositories, even if they provide robust preservation and management of data, so it should be avoided at this time.

## I. Desirable Characteristics for All Data Repositories

**A. *Persistent Unique Identifiers:*** Assigns datasets a citable, persistent unique identifier (PUIID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (*e.g.*, of research progress), and research assessment (*e.g.*, identifying the outputs of Federally funded research). The PUIID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.

**B. *Long-term sustainability:*** Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.

Sustainability and preservation should be differentiated from each other and given separate coverage. The subset of characteristics outlined here are important considerations, although it may be difficult to quantify them in ways that can be used to evaluate and distinguish repositories.

It seems reasonable to expect repositories to provide descriptions of operational sustainability, including emergency planning and response capacity, as well as a sustainable business model. These qualities will generally need to be taken at face value. Requiring this information in an accessible online form can help address the need for clarity around repository infrastructure which is no longer supported or has a known end date. In the data management landscape there are many examples of systems that were developed as pilot projects, yet persist as digital ephemera with no capacity to offer a service at scale.

Preservation has been a component of data management guidance for quite some time, although data management plans often conflate preservation with storage and backup. The capacity for preservation of digital objects and their metadata should be an essential consideration in the selection of a repository. Therefore, the methods and practices of the repository should be documented through measures that are not too onerous but consistent within the repository community. As with other characteristics, the availability of local experts and repository and archival community-produced educational materials and guidelines should be noted.

**C. *Metadata:*** Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.

This brief statement notes the utility of metadata without getting attached to the specific approaches of different research domains and knowledge representation systems. The foundational importance of metadata to FAIR can also be noted in greater detail, and described in point D, Curation. The availability of additional guidance on domain-specific and general data

schema standards, and local or domain experts should be noted, with the caveat that vocabularies and schema are continually evolving and may be domain or repository-specific.

**D. *Curation & Quality Assurance*:** Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

Data curation services can make a considerable difference in the robustness of metadata, so this is a welcome criteria. However, the term “curation” has many different connotations, and should be more fully described. Similarly, data integrity, accuracy, and quality assurance are highly contextual terms, and invoking them as part of data curation services, without further delineation, has the potential to muddy the waters. A more explicit mapping of curation practices to these concepts, through a use case or matrix/diagram, could greatly improve how this point is interpreted and applied by data depositors and the other audiences for this guidance.

Not all repositories provide curation services, so using this as a required element, as outlined here, could potentially eliminate those options. In some cases, the repository workflow may require data depositors to address some of these elements from the outset.

Even among mediated deposit repositories, there can be a broad spectrum of curation interventions, from reviews of structured and unstructured metadata, to data structure, dataset organization, and selection of file formats. Since the nature of these services and their outcomes are so important, repositories should provide clear detailed descriptions of their data and metadata workflows and processes. These should include unambiguous and specific details about the curation services they provide, in addition to the costs, if any, as well as the expectations that are placed on data depositors.

**E. *Access*:** Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.

Sections E and F should be combined. Statement E is more inclusive of open access and restricted access cases, while statement F implies that all repositories should make all data available, which is not compatible with the “...legal and ethical limits...” modifier in statement E.

In keeping with FAIR and other good data sharing practices, this statement should be modified to include “metadata.” A basic tenet of FAIR is access to metadata records, while licensing, privacy and confidentiality, and embargo considerations may constrain access to the data itself.

Similarly, the provision of machine-readable metadata deserves stronger emphasis. Repositories should make metadata available and describe the methods for accessing it, whether through OAI-PMH alone, or also via well-documented APIs, if they exist.

**F. Free & Easy to Access and Reuse:** Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

**G. Reuse:** Enables tracking of data reuse (*e.g.*, through assignment of adequate metadata and PUID).

By noting metadata and PUID (which can be elements of data citations) for tracking data reuse, this statement seems to indicate that data citations would be used to track reuse of data. However, the collection of citation statistics is typically not a function of repository platforms, since those statistics would either need to be harvested from publication systems and citation indexes, or through self-reported registrations by users of repository systems.

A more realistic expectation for repositories would be that they provide standardized reporting of metadata/record accesses, views, and downloads of the data files themselves. These could be accomplished via open standards and protocols, which repositories should be expected to provide.

**H. Secure:** Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>) or the National Institute of Standards and Technology's 800-53 controls (<https://nvd.nist.gov/800-53>).

These standards may be informative to repository developers and managers but are probably much less useful to the typical data depositor. More user-friendly terminology would improve their usefulness as a criteria to data depositors. Security and data integrity would seem to be related issues as well and could be grouped together.

**I. Privacy:** Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.

These elements also seem to overlap with H, Secure. More detail should be provided on this point, since these characteristics might be assumed to apply only to human subjects and other restricted access/sensitive data and would therefore belong under II Additional Considerations... below.

This element could also reference the protection of privacy of users of the system. If it is to be applied in that way, that should be explained. In any case, repositories should be expected to provide open access to their user privacy policies.

**J. Common Format:** Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.

File formats are a critical consideration for the data depositor to ensure long-term reusability and preservation. The repository could influence or control the choice of file formats by data



depositors, but it can also provide guidance. This section should reference data management and curation guidelines, local experts, and generally accepted best practices that have been published by organizations such as the UK Data Service, Library of Congress, and numerous domain repositories.

**K. *Provenance*:** Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.

This seems like a core capacity of data repositories, though the use of 'record' instead of 'logfile' is preferred. This criteria could be grouped with H. Secure, and also corresponds with Integrity in section B.

## **II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

Repositories that provide access to and preservation of sensitive data would be expected to exhibit functionality that will protect the privacy and confidentiality of participants, through robust access controls, such as data use agreements and authentication and authorization measures.

**A. *Fidelity to Consent*:** Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).

This is outside of the purview or expertise that should be expected of a repository. I would assume this is a responsibility of the data depositor.

**B. *Restricted Use Compliant*:** Enforces submitters' data use restrictions, such as preventing reidentification or redistribution to unauthorized users.

A repository may have one or more processes for authenticating and authorizing users who agree to comply with a data use agreement. However, it seems unrealistic to expect the repository manager or personnel to be able to monitor what an authorized user does with the data, short of limiting their access to physically and technologically restricted data viewing options that could prevent unauthorized data transfers. Certainly, the repository should respond to abuses once it is aware, but it is

**C. *Privacy*:** Implements and provides documentation of security techniques appropriate for human subjects' data to protect from inappropriate access.

**D. *Plan for Breach*:** Has security measures that include a data breach response plan.

**E. *Download Control*:** Controls and audits access to and download of datasets.

F. *Clear Use Guidance*: Provides accompanying documentation describing restrictions on dataset access and use.

G. *Retention Guidelines*: Provides documentation on its guidelines for data retention.

Assuming this refers to the duration of data retention by the repository, this item should be moved to “I. Desirable Characteristics for All Data Repositories”. A plan or process for decision-making about de-accessioning data from the repository should be provided by the repository.

H. *Violations*: Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository.

I. *Request Review*: Has an established data access review or oversight group responsible for reviewing data use requests.

This should be combined with B, Restricted Use Compliant.

***CNRI Response to the OSTP Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research***

Submitted by Laurence Lannom, Vice-President  
Corporation for National Research Initiatives (CNRI)  
1895 Preston White Drive, #100  
Reston, Virginia 20191

CNRI is a not-for-profit organization formed in 1986 to undertake, foster, and promote research in the public interest.

March 17, 2020

CNRI applauds OSTP's continuing efforts, beginning with the 2013 memorandum on "Increasing Access to the Results of Federally Funded Scientific Research", to improve access to data and publications resulting from Federally funded R&D. The set of "Desirable Characteristics for All Data Repositories" listed in Section I, as well as the "Additional Considerations for Repositories Storing Human Data (Even if De- Identified)" described in Section II, are an excellent starting point for "Managing and Sharing Data Resulting From Federally Funded or Supported Research". While this set of characteristics, when implemented, will serve to reduce the burden for researchers, it may be insufficient to deal with the tsunami of data resulting from current research methods and programs; and this emerging requirement constitutes our primary response to the RFC. In this response, suggestions are provided by CNRI on how to enhance the performance of repositories for this purpose.

The remainder of this response focuses on the concept of digital objects, which are described more fully below. However, since the request from OSTP is focused on repositories, we feel it necessary to put the relationship between repositories and digital objects into perspective. A repository made accessible in the Internet may be viewed as providing a service that enables users to access digital objects that are accessible from that repository. However, since an individual digital object, when accessed, may provide or enable access to other digital objects contained within the digital object or obtained elsewhere, it has all the properties of a repository and is conceptually more general. Thus, rather than viewing a repository as providing a special kind of service apart from that provided by a digital object, one may view a repository as a digital object in and of itself – namely, one that contains other digital objects and is accessible to a network-based user community. It may even be a mobile program. In this way, every repository can be identified in exactly the same way that any other digital object is identified, every repository can have metadata about it, as do all digital objects, including especially, its identifier and other related information such as permissions and access rights, security measures, and privacy.

Managing and especially sharing data requires more than simply making data available. As embodied in the now widely accepted FAIR principles, data must not only be Findable and Accessible, but also Interoperable and Reusable. Fulfilling those last two requirements across

the wide swath of heterogeneous data currently filling data repositories requires an approach to simplifying and reducing the complexity of dealing with such heterogeneous data. How can researchers understand or reuse data which they had no part in creating and whose creators they may not know? We believe that the answer to that question is to move to a digital object approach to managing information. This approach, formally known as the Digital Object Architecture, started in the early 1990s at CNRI as an outgrowth of work by Robert Kahn and Vinton Cerf on mobile programs, and with funding from the U.S. Government.<sup>1</sup> After increasing adoption globally, the maintenance and evolution of this architecture in the public interest is now the mission of the non-profit DONA Foundation<sup>2</sup> founded by CNRI in 2014, and is based in Geneva, Switzerland. The Digital Object (DO) Architecture is also a primary consideration within the Research Data Alliance, an organization that was established with the backing of the US, EU and Australia in 2013.

The DO Architecture is a non-proprietary architecture based on the concept of a digital object (known as a “digital entity” for purposes of ITU-T Recommendation X.1255 (Sept. 2013)).<sup>3</sup> A key attribute of a digital object is its associated persistent unique identifier (PUID), referred to generally as a digital object identifier or handle, which provides a way for users or computer programs to reference or interact with a specific digital object unequivocally and with precision. In most cases today, data is treated as a collection of bits in multiple formats and structures with multiple names and identifiers and methods of access, with metadata in various degrees held closely with the data or located elsewhere. Accessing and acquiring such data typically requires an enormous effort to understand it sufficiently in order to combine with other data and/or to reuse or re-analyze it. The digital object approach raises the level of abstraction of data entities and treats them as logical single digital entities which can respond to a core set of transaction requests, including not only basic retrieval requests but also requests for extended operations, e.g., visualization or combination with similar data types. More generally, repositories should intrinsically support the execution of vetted operations on data contained within, and enable users to execute privileged operations, with adequate concern for individual privacy, rather than only enabling a request/response interaction pattern.

An open, simple, and powerful Digital Object Interface Protocol Specification (DOIP)<sup>4</sup> has been defined to show how these transactions can work over both short timeframes, and also after very longtime intervals, based on the use of PUIDs for actions to be taken (i.e., computational steps represented as digital objects) and also to designate the targets (i.e., digital objects) for those actions. Using DOIP as a front-end interface to information systems will create a data management environment which will radically reduce the amount of effort currently expended in managing scientific data and radically increase the amount of time and energy that scientists and data managers can devote to science instead of housekeeping tasks. Over the years, CNRI has made available in the public interest reference software implementations of the three main components of the DO Architecture for download at no charge.<sup>5</sup>

This overall approach has been described in a number of articles, a selection of which is referenced below <sup>6 7 8</sup>, has been adopted by the nascent Fair Digital Object Framework

movement<sup>9</sup>, has been shown to apply to specific domains<sup>10</sup>, and can be broadly illustrated by the following conceptual illustration:

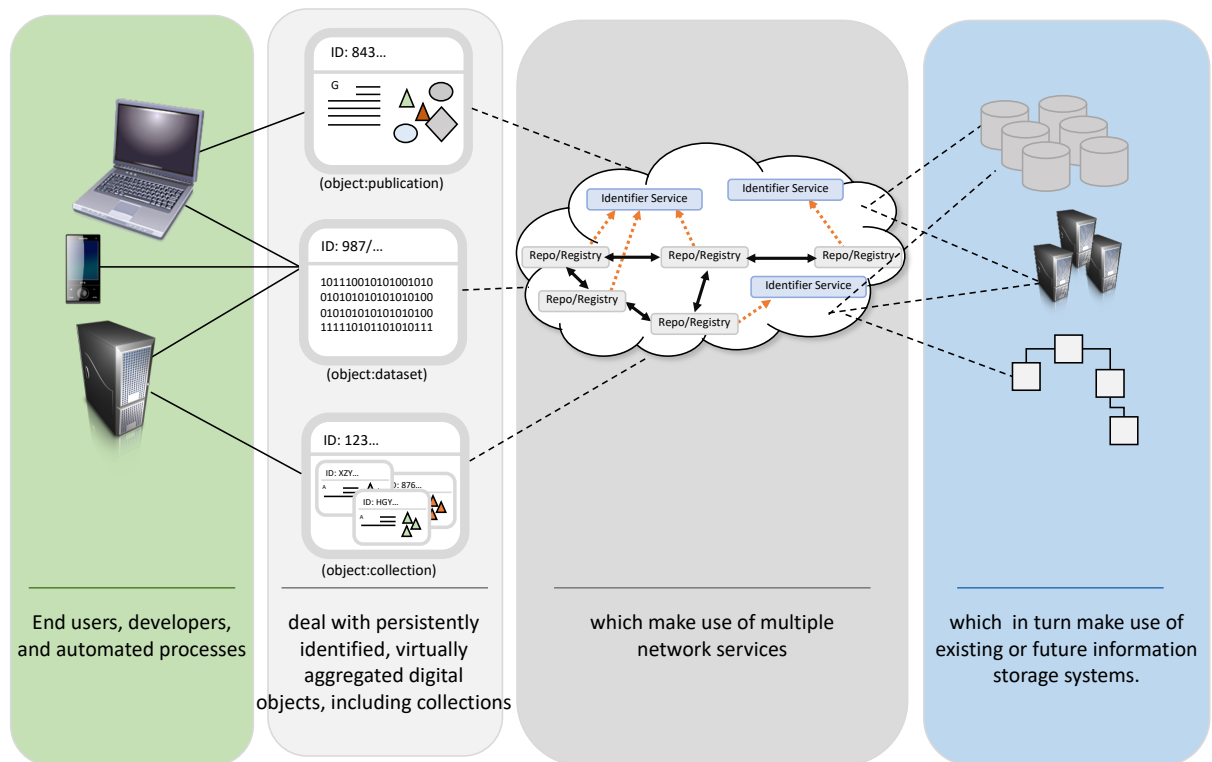


Figure 1. Digital Object Architecture enabling networked data and information interoperability and management across heterogeneous systems.

In addition to recommending that Federal efforts move in the direction of the vision described here, we realize that achieving this overall objective will involve a dedicated and persistent effort over a number of years.

Finally, we wish to comment on the term “interoperability”, which is often touted as a desirable characteristic of a given system, although interoperability becomes meaningful only within the context of a given set of systems and system attributes. The universe of data repositories, or digital objects more generally, should not appear to consumers as completely independent entities, but rather as interoperable entities that are part of a single infrastructure in much the same way that a collection of interoperable routers enables the Internet. Designing individual systems to behave and appear as part of a larger whole, in spite of the disparity in content, location, and governance, is an important characteristic for the collection of data repositories. This can be achieved via a series of design choices and practices that should include the following:

1. Reliance on a persistent unique identifier resolution system. The OSTP RFC references one example, the DOI system, which has been adopted within the publishing industry

for long-term stability, and which, in turn, relies on handle technology that is based on the DO Architecture.

2. Repositories should expose at least one interface (possibly among several agreed standards) that is based on semantic-free persistent unique identifiers, of which the DOIP Specification is one good example. This provision would create a means of withstanding technological change over the long haul.

CNRI has a few specific comments on the draft desirable characteristics:

Section I.A. *Persistent Unique Identifiers*: We agree that persistent resolvable identifiers are key to a coherent approach to managing the output of Federal research funding, and we agree that existing accession numbers used in well-known data repositories, e.g., GenBank, must be leveraged, but we encourage the use of PUIDs associated with digital objects, together with compliant registries and repositories, in connection with these accession numbers. Within a given community and using existing repositories these accession numbers will function just fine, but outside of those communities, and over time as repository access methods change, it will be important to add a level of indirection in the form of a general-purpose identifier resolution system, such that no particular up-to-date knowledge is required to access the data.

Section I.F. *Free & Easy to Access and Reuse*: Free & Easy Access and Reuse needs to be combined with detailed and complete understanding of the data that can be accessed. Without that ability, free and easy access may actually be counterproductive, as the reuse of misunderstood data will not yield anything useful, may waste time, and, in the end, may distort both the old and new research. In implementing the DO Architecture, it is assumed that the elements of each digital object are structured as <type, value> pairs, and the meaning of each type may be determined by resolving it using one or more type registries. Detailed and complete understanding of the data produced by others is ordinarily quite difficult to achieve, but could be greatly simplified by use of the DO Architecture, which is why we strongly recommend it. We also emphasize the need to focus on the key role of type information in the data along with type registries to enable global understanding of type information. In this approach, the details of the data are hidden behind explicit defined functions and operations, which can persist over changes in repositories and locations. The research community is inherently global at this point in time, and thus it is highly desirable that the mechanism(s) used to share and re-use digital information is/are able to work across international boundaries.

Section I.J. *Common Format*: Formats are inarguably important, but we must point out that simply knowing, for example, that a given data set is held in the form of a matrix with each cell a whole number and rows and columns explicitly labeled does not, by itself, give enough information for data re-use. For example, in the case of surface temperature, what instrument was used to record it and when was that instrument last calibrated? Or, if processing the results of a survey, what was the “skip pattern” used and how was the sample selected? There are an enormous number of variables involved in every research method and those variables must be bundled with the data and that bundling must be done in such a way that the data object itself carries the information needed to interpret and reuse the data. This is a critical contribution

enabled by computer-accessible metadata. The digital object approach described above provides such a framework. It is not a magic bullet, and may require considerable effort by researchers to implement; but, as has been the case with the evolving use of persistent unique identifiers over the last few decades, this approach will prove to be essential going forward.

---

<sup>1</sup> For background information on the DO Architecture, see “Blocks as digital entities: A standards Perspective,” <https://content.iospress.com/articles/information-services-and-use/isu180021> (2018).

<sup>2</sup> <https://www.dona.net/>

<sup>3</sup> ITU-T Recommendation X.1255, <https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=11951&lang=en>

<sup>4</sup> Digital Object Interface Protocol version 2.0 ( 2018), [https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec\\_1.pdf](https://www.dona.net/sites/default/files/2018-11/DOIPv2Spec_1.pdf)

<sup>5</sup> A recent CNRI software release based on components of the DO Architecture called **CORDRA** may assist in understanding the software implementations made available by CNRI. **CORDRA** is available at: <https://www.cordra.org/>

<sup>6</sup> Denning, Peter J. and Robert E. Kahn. "The Profession of IT: the Long Quest for Universal Information Access," *Communications of the ACM*, December 2010, Vol. 53, No. 32, pp. 34-36. <http://doi.org/10.1145/1859204.1859218>

<sup>7</sup> Kahn, Robert E. "The Architectural Evolution of the Internet". Corporation for National Research Initiatives, November 17, 2010, <http://hdl.handle.net/4263537/5044>

<sup>8</sup> Kahn, Robert E., Robert Wilensky, "A Framework for Distributed Digital Object Services". *International Journal on Digital Libraries*, (2006) 6(2): 115-123. <http://doi.org/10.1007/s00799-005-0128-x>. Reproduced by the International DOI Foundation with permission of the publisher [here](#). (First published by the authors May 13, 1995, "A Framework for Distributed Digital Object Services", <http://hdl.handle.net/4263537/5001>).

<sup>9</sup> <https://www.go-fair.org/today/fair-digital-framework/>

<sup>10</sup> L. Lannom, D. Koureas & A.R. Hardisty. FAIR data and services in biodiversity science and geoscience. *Data Intelligence* 2(200), 122–130. [http://doi.org/10.1162/dint\\_a\\_00034](http://doi.org/10.1162/dint_a_00034)

March 17, 2020

To: *OpenScience@ostp.eop.gov*

Subject Line: RFC Response: Desirable Repository Characteristics

Thank you for the opportunity to comment on the “**Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research**”. We are Indigenous and an allied scholar affiliated with the Global Indigenous Data Alliance (GIDA), the US Indigenous Data Sovereignty Network (USIDSN), the Te Mana Raraunga Maori Data Sovereignty Network (TMR), and ENRICH, the Equity for Indigenous Research and Innovation Coordinating Hub, a group of entities advocating for Indigenous rights and interests in Indigenous data and providing practical tools and mechanisms that support Indigenous control of Indigenous data. We have been working, along with a broad number of stakeholders, to advance changes to data policies and practices that enhance Indigenous control of data and enrich metadata. Much of this work is part of operationalizing and implementing the **CARE Principles for Indigenous Data Governance**: Collective benefit, Authority to control, Responsibility, Ethics ([gida-global.org/care](http://gida-global.org/care)). These Principles, developed and released in 2019, promote a new paradigm of responsibility, equity and transformative change in the production, research, collation, storage and distribution of Indigenous data. They currently set the international standard for rights and governance of Indigenous data.

The repository community has a significant role in the success of operationalizing the CARE Principles. The list of desired characteristics for scientific repositories is important and aligns well with the work we are doing. There are additional topics that should be included in the list.

1. **Development of new Provenance Standards for Indigenous data.** Indigenous data lacks clear and proper provenance. This affects how Indigenous data can be used now and into the future. With no standards, including metadata fields, that support Indigenous provenance, there is a real danger that Indigenous data within repositories will remain impoverished and unusable by Indigenous peoples and by collaborative researchers. Tracking full provenance enables possible reuse of existing datasets in new research. Full provenance is also important as it enables the original funders, communities, researchers and institutions that enabled the creation of any source dataset to have identity, attribution and rights of association where this is determined to be suitable and appropriate.
2. **Development of New Guidelines on the Collection of Indigenous Data in Federally Funded Research.** There are currently no Guidelines on the collection and storage of Indigenous data through federally funded research. This means that researchers have limited direction and support about ethical and responsible practices when collecting Indigenous data, and therefore also when depositing and storing Indigenous data in repositories. Moreover, repositories also have limited guidance in the care and management of Indigenous data. This has inevitable consequences for the future use and circulation of Indigenous data. These new guidelines need to address differentiated



privacy issues alongside ownership and control of Indigenous data. These Guidelines must follow current international standards for data and Indigenous data - namely -the FAIR principles and the CARE Principles for Indigenous Data Governance (gida-global.org.care).

- 3. Supporting Enhanced and Replicable Integrity in Research Practice.** To address barriers that historically have impeded ethical and responsible research practices, research repositories need to foster a culture of integrity and trustworthiness. Scientific discovery hinges on data analytics, but data systems are rife with biases and encumbrances that inhibit the ethical conduct of science. Indigenous data sovereignty draws on the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP) that reaffirms the rights of Indigenous Peoples to control data about their peoples, lands, and resources. Indigenous data governance enacts those rights through mechanisms grounded in Indigenous rights and interests that promote ethics and equity, while providing a framework for addressing deeper historical issues associated with barriers for underrepresented communities and knowledge systems. The ‘CARE Principles for Indigenous Data Governance’—Collective benefit, Authority to control, Responsibility, and Ethics—enhance and extend the ‘FAIR Principles’ for data findability and reuse—Findable, Accessible, Interoperable, Reusable—by centering equity and ethics as core guiding principles alongside those set out by FAIR. These concepts form a basis for normative standards for collective data rights that impact research agendas for data privacy, future use, reuse, and data stewardship. Implementation of the CARE Principles provides an International standard in exercising Indigenous rights for the governance of Indigenous data. Operationalizing the CARE Principles requires upholding tribal self-determination by requiring adherence to tribal codes, IRBs, guidelines, etc.; enacting repository policies for Indigenous data; and using tools such as metadata, labels, and collection notices to enhance transparency and integrity.
- 4. Clarification on the Limits of IP (Copyright and Patents) for Indigenous control of Indigenous data.** The current IP system treats Indigenous interests in different ways. Historically it has promoted Indigenous culture and relevant collected data to be open and available to all. This approach has led to the disclosure of valuable and secret cultural information, the widespread appropriation of Indigenous knowledge and cultural forms, and the derogatory treatment of Indigenous culture through a failure to appreciate and respect nuances in forms of sharing and use of knowledge. These problems extend into the sciences and Indigenous data. Copyright and patent law continue to exclude Indigenous interests, and this means that license agreements, or other control mechanisms, tend to be unfairly biased against Indigenous interests. We recommend that clarification on these limits of the law are made for all those researchers working with and collecting Indigenous data in order to make Indigenous rights clear and to support informed decision-making at every level of the research process.
- 5. Access to reliable and supported training that addresses Indigenous data governance.** There is currently no supported or reliable training offered to researchers around Indigenous data governance. Training creates the opportunity for increased knowledge around Indigenous data governance and the possibility of the extension of best

practices for Indigenous data. Directed training in specific science and research areas - for instance, genomic sciences, health sciences, environmental sciences - can support better engagement in the collection of Indigenous data, including increased reliability for using Indigenous data owing to proper attention to issues of provenance. Training through webinars can be an effective means for increasing researcher knowledge and supporting Indigenous community engagement with researchers. Training can also help build trust between historically unequal parties in the research process.

**6. Promotion and Adoption of tools that support the application of the CARE Principles - the TK (Traditional Knowledge) and BC (Biocultural) Labels and Notices System**

The TK and BC Labels and Notices System has been developed to support Indigenous interests in the documentation of Indigenous knowledge and in the production of Indigenous data, especially in contexts of governance, decision-making, provenance and control. Within this system, the TK and BC Notices have been designed as specific tools for researchers to help promote transparency and integrity in the collection and engagement with Indigenous data. For instance, the TK & BC Notice system allows a researcher to fix a Notice to specific data *as additional metadata* when they know or have reason to believe that there are specific, or underlying Indigenous interests that will need attention and Indigenous engagement into the future. As a distinct mechanism for both researchers and for data repositories, these Notices allow researchers to apply CARE Principles in their research practice. We recommend that Federally Funded Research makes recommendations to researchers to use these tools in a similar way to how these tools have been recommended by Mellon and other federal funded research in the social sciences when researchers are addressing the rights, ethics and data sections of their grant applications. See <https://www.youtube.com/watch?v=s18DaM6TXHE>

In addition to the specific desirable characteristics indicated above, GIDA, USIDSN, TMR, ENRICH, and other entities have been collaborating with scientific and research repositories to define and develop leading practices. We hope that these draft guidelines recognize and complement this effort and that as leading practices continue to develop are sufficiently adaptable.

Thank you again for the opportunity to provide comments.

Best regards,

Stephanie Russo Carroll  
Assistant Professor, Public Health, University of Arizona  
Associate Director, Native Nations Institute, University of Arizona  
Chair, Global Indigenous Data Alliance  
Co-Founder, US Indigenous Data Sovereignty Network  
Implementation Team, ENRICH-Equity for Indigenous Research and Innovation Coordinating Hub

Maui Hudson

Associate Professor, Faculty of Maori and Indigenous Studies, University of Waikato

Co-Founder, Global Indigenous Data Alliance

Co-Founder, Te Mana Raraunga Maori Data Sovereignty Network

Co-Founder, Biocultural Labels Initiative

Co-Director, ENRICH- Equity for Indigenous Research and Innovation Coordinating Hub

Jane Anderson

Associate Professor, Anthropology and Program in Museum Studies, New York University

Affiliated Professor, Engelberg Center on Innovation, Law and Policy, School of Law, New York University

Director, Local Contexts: The TK Labels and Notice System

Co-Founder, Biocultural Labels Initiative

Co-Director, ENRICH- Equity for Indigenous Research and Innovation Coordinating Hub

**From:** Cole, Stanley R. (LARC-D3) <[stanley.r.cole@nasa.gov](mailto:stanley.r.cole@nasa.gov)>  
**Sent:** Tuesday, March 17, 2020 4:29 PM  
**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>  
**Cc:** Rivers, H Kevin (LARC-D3) <[h.kevin.rivers@nasa.gov](mailto:h.kevin.rivers@nasa.gov)>; Kilgore, William A. (LARC-D3) <[william.a.kilgore@nasa.gov](mailto:william.a.kilgore@nasa.gov)>; Mark, Michael I. (LARC-B2) <[michael.i.mark@nasa.gov](mailto:michael.i.mark@nasa.gov)>  
**Subject:** Response to Request for Information

The following are comments on the OSTP “Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research.”

(Notice issued March 5, 2020, p. 12950 of Federal Register.)

These comments are provided by the Research Directorate at the NASA Langley Research Center.

Stan

---

Stanley R. Cole  
Chief Operations Officer  
Research Directorate  
Cell: (757) 303-4011  
NASA Langley Research Center  
Mail Stop 41  
Hampton, VA 23681



The approach outlined looks pretty comprehensive and would be of a benefit to better retain and standardize Federally Funded research data. It would be nice if there was a way to enforce these characteristics (that is likely outside the realistic scope of this subject document).

Some specific comments on the individual clauses:

1. Clause D. Curation & Quality Assurance – Recommend incorporation of the concept of mandatory data back-up/data preservation.
2. Clause E. Access – Recommend including a facet of access control. Data should be open in the interest of sharing, but access at a minimum should be controlled to understand user statistics and mitigate potential malicious data manipulation. Access control would likely be through some sort of a free account tied to an email (similar to how NSPIRES is set up). This would also aid in metrics pertaining to Clause G “Reuse” and is closely related to Clause H (but does beg inclusion here).

3. G. Reuse – from Gov't-Industry Data Exchange Program (GIDEP) experience, this can be a double edged sword if not accomplished in the correct manner. You do not want this to be obtrusive to the user (or they just won't use the system). Recommend wording similar to "Enables tracking of data reuse in a manner unobtrusive or transparent to the user." Users hate filling out utilization reports.
4. Recommend adding verbiage to having an established and followed Data Retention Schedule. Some data will simply not be needed to be held indefinitely and should be disposed per established guidelines. Without a retention schedule, you could have a situation of IT-Intensive repositories of little or no practical use.

It is important that format compatibility is taken into account for long-term electronic storage. There are two issues:

- 1) Binary software for reading data stops working after a few generations of hardware and OS evolution, and
- 2) Storage medium also evolves and becomes obsolete.

In addition to archiving and providing data-read software (if applicable) with each database, one needs to provide the uncompiled code and a data file format document which describes in detail the organization of the file, including headers, data precision (12-bit, 16-bit, etc.), big versus little endian, and so forth.

Corinna Turbes, Policy Manager  
[corinna.turbes@datacoalition.org](mailto:corinna.turbes@datacoalition.org)  
202-573-7975

March 17, 2020

Subcommittee on Open Science of the National Science and Technology Council's  
Committee on Science  
OpenScience@ostp.eop.gov

Re: RFC Response: Desirable Repository Characteristics

Dear Subcommittee Members,

The Data Coalition is America's premier voice on data policy. As a membership-based business association, the Data Coalition advocates for responsible policies to make government data high-quality, accessible, and usable. Our work unites data communities that focus on data science, management, evaluation, statistics, and technology in companies, nonprofit organizations, and academia.

The Data Coalition members have long supported transparency for government information, which is often facilitated with improved data standards and access mechanisms. The Data Coalition appreciates the effort made by the subcommittee of the White House Office of Science and Technology Policy (OSTP) to align with the principles and practices outlined in the Federal Data Strategy. We hope that the strategy, as well as the implementation of the Foundations for Evidence-Based Policymaking Act of 2018 (Evidence Act), will continue to bolster your efforts to improve access to data and publications resulting from federally-funded research.

While overall the draft characteristics cover a range of topic, the Data Coalition offers the following brief comments for further consideration as the characteristics are revised and finalized:

**1. Data Repository Guidance Must Align with National Secure Data Service Planning.** While we applaud the inquiry and development of a strategy from OSTP, the Data Coalition strongly encourages the subcommittee to plan its work and subsequent publications to align with the forthcoming Advisory Committee on Data for Evidence

Building, established by the Evidence Act. This committee is explicitly tasked to provide recommendations to OMB on, among other things, how to facilitate and coordinate improved data sharing across agencies to build upon the unanimous recommendations from the U.S. Commission on Evidence-Based Policymaking. As you know, the Evidence Commission recommended the establishment of a National Secure Data Service. The successful implementation of such a service would hinge on the capabilities embodied in data repositories established across the federal government. Thus, based on this new law and requirement we strongly encourage the subcommittee to align with this line of work as well.

With this in mind, we also recommend the “Draft Desirable Characteristics” specify that the tool is relevant for data sharing under the Evidence Act and to support the development of any federal data service.

**2. Additional Considerations for Confidential Business Information.** While the “Draft Desirable Characteristics” include reference to “human data” there is no apparent recognition for other data that might be collected under a pledge of confidentiality or require appropriate protections. The Data Coalition recommends the characteristics under Section II of the characteristics that other units of analysis should also be privy to these considerations, including confidential business information. Other units of analysis could include household units and the like.

**3. Use of Privacy.** In discussing data privacy the Data Coalition recommends envisioning the variety of potential mechanisms that can support responsible protections, and supports the inclusion of consideration for “administrative, technical, and physical safeguards.” Throughout the draft characteristics, the term privacy is written to suggest only access limitations, without regard to other mechanisms to protecting confidentiality, including output privacy and disclosure avoidance which could be made more explicit within the stated definition

**4. Alignment with the OPEN Government Data Act.** Title II of the Evidence Act, the OPEN Government Data Act, establishes expectations related to data inventories. The characteristics defined by OSTP for the repositories should minimally align with those core legal requirements for federal systems, including the availability of information about data sensitivity as an aspect of a dataset’s metadata.

**5. Sustainability and preservation.** Data repositories and their underlying infrastructure require long-term plans for managing and funding the repository. This should generally include a publicly-available and transparent policy that documents

preservation practices. Some members of the Data Coalition, such as Digital Science, have especially unique and relevant expertise that may be helpful for OSTP's further revisions on the characteristics regarding this topic.

**6. Define data repository.** While the draft outlines characteristics of a repository, the document and future policy guidance would benefit from clearly explicating the definition of a repository for which the guidance is intended to apply.

**7. Confidentiality Characteristics.** In general, the Data Coalition recommends that in coordination with OMB, an additional section to the guidance could be included that highlights characteristics relevant for data collected under a pledge of confidentiality. In the federal government, such pledges are required for use of the privacy framework in the Confidential Information Protection and Statistical Efficiency Act of 2018 (Title III of the Evidence Act). The U.S. Commission on Evidence-Based Policymaking previously unanimously recognized this legal framework as foundational for envisioning the future of government data sharing and management. The draft characteristics could meaningfully and productively benefit from relating these concepts more clearly.

**8. Accessibility and Machine-Readability.** Data repositories should generally have accessibility for non-sensitive data and metadata, enabling humans of all abilities and machines to understand and read information consistent with requirements in the OPEN Government Data Act.

Thank you for the opportunity to submit comments on this important set of draft characteristics. If you or your staff have any questions about the Data Coalition's comments, please contact Corinna Turbes at [corinna.turbes@datacoalition.org](mailto:corinna.turbes@datacoalition.org).

Respectfully,  
Corinna Turbes



17 March 2020



**To:** OpenScience@ostp.eop.gov

**Subject:** RFC Response: Desirable Repository Characteristics

Thank you for the opportunity to comment on the “**Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research**”. We are affiliated with the American Geophysical Union (AGU), a society representing the international Earth and space science community. The AGU community, along with a broad number of stakeholders, has been working on changes in practices and policies requiring published research to have supporting data and software made available through an appropriate repository, preferably a domain-specific repository when available while also recognizing the importance of general and institutional repositories. This work is building on the **Enabling FAIR Data [1]** project, funded by Arnold Ventures and subsequent work by the National Science Foundation, that established a Commitment Statement [2] as well as Author Guidelines [3] adopted by most of the major publishers. This effort advocates a change in scientific culture towards data sharing and making data open and FAIR (findable, accessible, interoperable, reusable) [4]. AGU has also recently updated its Data Position Statement [5] affirming the importance of repositories for preserving scientific data and other relevant evidence and the need to follow leading practices...to ensure data is “**processed, shared, and used ethically, and is available, preserved, documented, and fairly credited.**”

The repository community has a significant role in the success of the goals defined in the Enabling FAIR Data project. The list of desired characteristics for scientific repositories in the Draft is important and aligns well with the work we are doing with AGU and our community. There are additional topics that should be included in the list:

1. **Linking of research products:** Research efforts create digital research products including publications, data, software, and more. Few platforms support all possible formats, and commonly, different repositories are required for proper management and preservation even within single projects. These products need to be linked through persistent identifiers, grant ID’s, and researcher ID’s (e.g. ORCID). This would allow for robust discovery and transparency. Search platforms need to be trained to find and index different types of products especially if they have a persistent identifier.
2. **Confidential access to datasets during publication peer review:** In order to evaluate research, peer reviewers need access to data. Repositories make this possible by either publishing the data prior to the paper being published, by providing a temporary “share link”, or by providing secure “embargoed” access to reviewers only. As the publishing community more closely integrates with scientific repositories, this support to publication peer review is necessary. Single-blind and double-blind peer review requires this access to be confidential. Further, once the journal

publication related to a dataset is publicly available, publication of the dataset should be linked and/or triggered automatically. At this time this process is manually coordinated (at best).

3. **Software as context for data.** We learn more about research data by looking at the related software (e.g. code, workflow, models). Software sharing is necessary to better understand the data and the associated research. Repositories that support software preservation are important along with relevant links to related data and publication.
4. **Clear licensing to support data and software reuse:** Researchers who seek data or software for use in new research need to understand the usage license. There are different leading practices for software and data research products.
5. **Versioning and provenance:** Data analyses associated with a research grant commonly require existing datasets to be reprocessed and/or integrated with other datasets. The process of preparing data for this integration and possible transformation into more useful formats results in valuable derived data products that can be of significance to the broader community. Versioning also occurs when data is reprocessed to create better versions of a dataset as issues are found, or new understanding emerges. Establishing links from original datasets to derived data products is important to be able to unambiguously track the full provenance and possible reuse of existing datasets in new research. Full provenance is also important as it enables the original funders, researchers and institutions that enabled the creation of any source dataset to have identity and attribution.
6. **Funding and archive support for domain repositories:** Many domain repositories originate from a need by the community and tend to be funded with grants having time frames of 2-3 years. This short time frame limits long-term planning and the ability to evolve capabilities to meet desired characteristics like those listed in this draft. Adequate funding allows for incorporating needs from researchers and provides data products that are easier to discover and use. Making government-funded archives available to these domain repositories provides long-term preservation for data that are part of the scientific record, protecting them from loss of funding. This is one of the most significant and important challenges [4]. Leading repositories should have long-term support such that researchers and publishers will be confident that data supporting peer-reviewed research will be available. When policy recommends a future date to purge data, the data landing page with descriptive metadata will be maintained to identify that the data existed information about the data.
7. **Machine readable and actionable:** As required by the FAIR Data Principles [6], repositories need to provide both human and machine capable methods for discovery, access, and action. Researchers will benefit from using tools to help locate relevant data. Machine ability to read and take action will enable efficiencies for researchers, thereby allowing more time for analysis.

In addition to the specific desirable characteristics indicated above, societies such as AGU and the broader publishing community have been collaborating with scientific repositories to define and develop

leading practices. We hope that these draft guidelines recognize and complement this effort and that as leading practices continue to develop are sufficiently adaptable.

Thank you again for the opportunity to provide comments.

Best regards,

Shelley Stall



Sr. Director for Data Leadership  
American Geophysical Union

Brooks Hanson



Executive Vice President, Science  
American Geophysical Union

[1] <https://copdess.org/enabling-fair-data-project/>

[2] <https://copdess.org/enabling-fair-data-project/commitment-statement-in-the-earth-space-and-environmental-sciences/>

[3] <https://copdess.org/enabling-fair-data-project/author-guidelines/>

[4] Stall, S, et al. (2019), Make scientific data FAIR, *Nature* **570**, 27-29 (2019) doi: [10.1038/d41586-019-01720-7](https://doi.org/10.1038/d41586-019-01720-7)

[5] [https://www.agu.org/Share-and-Advocate/Share/Polycymakers/Position-Statements/Position\\_Data](https://www.agu.org/Share-and-Advocate/Share/Polycymakers/Position-Statements/Position_Data)

[6] 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>



March 17, 2020

Lisa Nichols

Office of Science and Technology Policy

[OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

Dr. Nichols,

We, JHU Data Services (<https://dataservices.library.jhu.edu>) submit this response based on our experience over the last 7 years with supporting institutional data sharing and archiving within our JHU Data Archive (<https://archive.data.jhu.edu>). Overall, we feel the characteristics listed within this draft are the right ones to include and would not remove any of them. However, we feel that it would be helpful to the OSTP for us to provide more specific feedback on a few of the features below, in particular those associated with human subject data. We full endorse the response submitted by the Data Curation Network (DCN) and supplement that response with additional comments here.

#### **I. Desirable Characteristics for All Data Repositories**

- **Curation and quality assurance.** We agree that it is important to have quality assurance as a feature of repositories. If data quality review can have a process similar to peer review for journal articles, it will assure the data quality. It is a problem how to accomplish that. Seeking out help from library professionals could be one way to find experts in data curation.
- **Reuse (G):** Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).
  - This characteristic echoes characteristics A (Persistent Unique Identifiers) and C (Metadata). We believe this could be expanded to include more detail on what data repositories can do to enable tracking of data reuse, for example, having recommended citations or gathering download and usage statistics.
  - Repositories can track how many people have downloaded a dataset and if possible, the location of them. This will provide valuable information for researchers to track the impact of their work.

- Repositories can also provide a citation for the user. If the user uses the citation in a paper, it can be tracked like journal citation.
- **Data repository evaluation.** The ways in which federal agencies might use these characteristics include “identifying specific repositories that a federal agency might designate for use for particular types of research data” and “evaluating data management plans that propose to deposit research data in a repository that is not operated by a federal agency.” However, the document does not explain how data repositories will be evaluated or to what lengths a reviewer might go to evaluate a repository. We wonder whether suitable repositories will be disregarded for not fitting these characteristics at first glance, for example, institutional repositories that have expert curation, track provenance and meet accepted security criteria, but do not readily publicize this information.

## **II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

A. *Fidelity to Consent:* Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).

In data repositories with full open access policies that do not register or restrict who can download files, there is not a direct technological restriction on releases of datasets that do not have appropriate consent or adequate de-identification of data. Our institutional repository, [JHU Data Archive](#), screens datasets for remaining risk; however, we ultimately rely heavily on researchers’ compliance with a deposit agreement. The submitter is also responsible for abiding by IRB agreements and other compliance requirements of their institution and funders. Researchers from professional settings are generally aware of consequences of violating subject consent and privacy requirements. A researcher's home institution also shares responsibility for breeches, potentially. A gray area, not often tested to our knowledge, is how institutional repositories, both within academic institutions and as independent organizations, share in the consequences of violations by facilitating distribution of data that violates consent or privacy disclosure. Clear deposit agreements and data use agreements can at least document formal responsibilities of researchers, repository, and the academic institution. Repositories, however, should ideally anticipate the informal 'backlash' to each of these stakeholders should repositories inadvertently facilitate submitter's non-compliant data releases. OSTP could recommend that acceptable data repositories have adequately discussed and documented their direct responsibilities for disclosure protection in relation to submitters and their home institution's compliance offices. At Johns Hopkins Data Archive, we request that all submitters of human subject data submit sample consent forms for our review of permission to share data, including de-identified and we refer questionable cases to IRB. OSTP might offer similar recommendations to repositories, such as encouraging IRBs to provide sample consent forms that acknowledge data sharing.

- Fidelity to Consent: This is an important one to check. If study participants do not consent to share data (even de-identified ones) with the public, researchers should not share in a data repository. But usually when it is time to upload data to a repository, it is too late to re-write the consent form. If the repository can provide some sample consent forms for researchers to use, it will help them with getting consent from participants of their studies.
- Consent forms: Not sure if “requiring consent forms to be deposited along with the data” is implied by the bullets of “fidelity to consent” or “clear use guidance”, but I think it might be worth stating that it should be strongly recommended if not required.

B. *Restricted Use Compliant*: Enforces submitters' data use restrictions, such as preventing reidentification or redistribution to unauthorized users.

Open access repositories such as the JHU Data Archive rely on researchers to prepare fully de-identified datasets. These are the conditions specified in our deposit agreement. Repositories should consider specifying the particular criteria for de-identification, which for public access datasets should, in most cases, meet HIPAA's "expert/statistical determination" level. This level requires full removal of direct and quasi-identifiers, the latter using adequate statistical anonymization techniques as required applied by those with sufficient expertise in techniques. Also, data should be reviewed for remaining risk by those with comparable expertise. We find that most researchers are not familiar with the techniques. These are considerations for OSTP regarding whether human subject data can be recommended for open access repositories, in the context of open access as the preferred means of sharing data by open access policies. We consider full de-identification to be achievable for open access and do not recommend restricting that option. We would recommend, however, that OSTP and funders strongly encourage data repositories to a) institute protocols for screening data for remaining privacy disclosure risk, b) provide training to personnel so that screening is done with sufficient expertise, and c) facilitate training of researchers or other assistance for preparing fully de-identified datasets where open access may be feasible. Repository managers and compliance offices can recommend restricted access when full de-identification is not feasible. The challenge is that techniques and training for data de-identification and disclosure risk screening are not readily available either for researchers or repository data curators. There is also not a strong consensus among academic open access repositories on what levels of de-identification and screening standards should apply. OSTP and the affiliated funding offices could participate in facilitating resources and training for better skills and policy for de-identifying data for open accesses.

C. *Privacy*: Implements and provides documentation of security techniques appropriate for human subjects' data to protect from inappropriate access.

This is a good best practice for data repositories, and may not currently be uniformly applied. Our repository deposit agreement does not currently require submitters to include documentation of their de-identification techniques, but our School of Medicine IRB and compliance offices are requiring de-identification protocols to be submitted and reviewed. We maintain our documentation of screening datasets, but OSTP might consider encouraging documentation retention more formally.

*D. Plan for Breach:* Has security measures that include a data breach response plan.

JHU Data Archive may be typical of other academic data repositories in that we would rely on our academic compliance offices in the event of a breach. What may be less common is that repositories have fully worked out with their compliance offices the possible scenarios and responses of breach from publicly released data. OSTP might help encourage such planning, and ideally help share community standards for response plans. Repositories that are more independently operated outside of academic institutions may need to develop comparable compliance responses internally.

*E. Download Control:* Controls and audits access to and download of datasets.

Control and audit of downloads of datasets is not common to many data repositories, including the JHU Data Archive. For JHU, this is currently a policy decision for the terms of open access. OSTP should clarify this stipulation as a feature of repositories that do promise restricted access and human subject data that is not fully de-identified. OSTP, however, cannot uniformly require this for open access repositories, including those that allow de-identified human subject data. The problem mentioned above, however, is that repositories should recognize the significant challenge of screening deposits for public access at HIPAA's "expert/statistical determination" criteria.

*F. Clear Use Guidance:* Provides accompanying documentation describing restrictions on dataset access and use.

We agree that this is an important stipulation for open access repositories that admit human subject data. Developing appropriate oversight and data screening; however, can be a significant challenge to repositories. This is especially so for smaller repositories and at the level of staffing and training data curators in best practice screening techniques. Neither the privacy industry nor the data repository community has developed methods or readily available resources and training to help repositories meet these standards in a timely manner. Although Johns Hopkins Data Archive has screening procedures in place, we are still working with our compliance offices and other institutions in the academic data curation community to improve procedures and policy. OSTP could encourage community development of these standards. OSTP should be aware that many repositories currently in operation are not meeting these standards.

*G. Retention Guidelines:* Provides documentation on its guidelines for data retention.

H. *Violations*: Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository.

I. *Request Review*: Has an established data access review or oversight group responsible for reviewing data use requests.

In the context of open access data repositories, including our JHU Data Archive, there are currently no logged data use requests, which is typical for open access policies. Users of the data are asked to abide by terms of use stipulated on each dataset. Restricted access repositories should have such procedures, and at JHU, collaborative use of restricted data includes data use agreements and is handled by Research Administration and IRB offices. OSTP may need to clarify that this stipulation may apply differently to open access repositories but should require that repositories can do the "front-end" assurance that only fully de-identified data are released, meeting HIPAA expert/statistical determination criteria.

Sincerely,

JHU Data Services:

Mara Blake, Manager of Data Services

Chen Chiu, Data Management Consultant

David Fearon, Data Management Consultant

Betsy Gunia, Data Management Consultant

Marley Kalt, Data Management Consultant





Office of Science and Technology Policy  
Dr. Lisa Nichols  
[OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

March 17, 2020

Subject: RFC Response: “Desirable Repository Characteristics”

Dear Dr. Nichols,

The Council on Governmental Relations (COGR) is an association of 188 research universities and affiliated academic medical centers and independent research institutes. COGR concerns itself with the impact of federal regulations, policies, and practices on the performance of research conducted at its member institutions.

We appreciate the opportunity to respond to the Office of Science and Technology Policy (OSTP) Request for Public Comment (RFC) on “Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research.” COGR recognizes the value that data repositories provide to the public and our nation’s scientists. Among other benefits, access to data can give researchers new ways of looking at old problems and a path to new discoveries.

In addition to specific comments on some of the data repository characteristics, we are also responding to the general principles and larger context related to data repositories and their characteristics.

As a starting point, it is critical that definitions applicable to data standards and repository characteristics are clear and consistent, developed through consultation across academic and administrative disciplines. The stakeholders are quite varied, including repository managers, researchers, and funders, each of which bring their own interpretations of terminology. A clear set of definitions across these groups is absolutely essential.

The practices, policies, and guidelines that will emerge from the desired set of characteristics should provide a clear vision of the future, while also accommodating an evolving landscape. A successful data repository will ensure that data receive proper technical and scientific governance, both when deposited and while being maintained and curated.

To this end, funding agencies can lead by example, while minimizing the workload and unfunded mandates placed on grantees, by curating and maintaining data centrally where possible. This would also facilitate inclusion of and access to data currently housed in agency repositories and would allow agencies to ensure standard metadata, quality, longevity, sustainability, accessibility, and security on a discipline-by-discipline basis.

Although we recommend that agencies centralize data repositories to the degree possible, doing so will be best achieved with input from relevant scientific and disciplinary communities. Discipline-specific context is essential in determining short and long-term uses, replicability, and transparency. In doing so, it is also important to take into consideration which additional disciplinary communities are likely to find the data useful. In a time where science is a team activity and interdisciplinarity is a goal across areas of research, the community that is interested in the data is not always the same one that builds the data set, especially in applied fields such as biology or information technology. Examples of successful partnering led by federal agencies in the past include genomic and high energy physics data.

Absent centralized agency resources, smaller locally developed repositories are likely to proliferate as a way to cheaply, quickly, and easily meet the letter of the requirements, creating redundancies at a small scale and challenges to the FAIR principles. In addition, such repositories are often developed by individual PIs and risk being neglected after the end of the project. Though specifying the desirable characteristics that OSTP proposes may help nudge researchers toward more robust methods for data storage and maintenance, it may also make projects seem overwhelming and untenable, particularly for a single PI without technical and curation expertise or support.

The goals of data preservation and sharing affect not only the repositories, but the entire life cycle of data in research and creative endeavors. Data creation, curation, analysis, sharing, and preservation are all connected and intertwined, along with the progress of knowledge creation and the careers of researchers. To this end, it is important to consider that the requirement to share data through a repository will create administrative and scientific workload in all aspects of the way research data creation is performed. As we consider the characteristics of repositories, we also need to consider that data deposition should be simple and straightforward to minimize the administrative burden for researchers; that governance of the repositories should include representatives from agencies, researchers, and research administrators to ensure standard practices; and that any new processes, guidance, or policies should be extensively tested by active users of the repositories before being scaled up, to ensure both stability and functionality, and that benefits exceed associated costs.

Beyond the administrative barriers, the cost of maintaining a data repository, including, for example, credentialing (such as ISO standards), may not be insignificant. In cases where the government is not curating and maintaining a repository itself, it would be appropriate for the government to find a way to cover the associated costs. This will be critical if the government intends to successfully drive greater development and use of distributed data repositories.

Finally, additional consideration should be given to ways in which researchers, data curators, data collectors, and data stewards can be recognized for contributing to the shared goal of transparently managed and shared data in research. This can include required attribution at different stages of the data life cycle such as attribution to the data collector when data is used by a third party, attribution to the data curators, and citation of the repository used in publications at very least.

We also have specific responses to some of the data repository characteristics (excerpts from the RFC are italicized).

***B. Long-term sustainability:*** *Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.*

Long-term sustainability of data for research is important for discipline-specific studies for reproducibility purposes but does not come without substantial costs and risks to an institution that if exposed (e.g., patient data), may cause irreparable harm. Further analysis of long-term data preservation should be vetted by both funders and institutions as research progresses. We recommend that both funders and the research community further analyze studies that warrant long-term preservation.

In addition, there is ample confusion on the definition of long-term. The answer will vary not only by discipline, but by perspective, for example, a researcher versus a librarian. A successful data repository will ensure that data sets are given appropriate life cycles both from a technological and a scientific perspective, beyond the principal that re-use of data is valuable. Appropriate considerations for the definition of long-term include:

- At what point is the data obsolete?
- At what point does the format of the data set make it unusable?
- What then should happen to the data set and who is responsible for those actions?
- Who pays for the support of the long-term plan?

***G. Reuse:*** *Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).*

Defining “adequate metadata” can be complex, time-consuming, and costly. Appropriate assignment will also need to accommodate the ability to maintain a link from the dataset and its metadata to its primary source. For example, if the source were proven to suffer from fabrication, falsification, or plagiarism, such a link would allow the flawed data to be removed. Similarly, if the raw data were re-analyzed leading to different conclusions, such a link would be helpful.

***H. Secure:*** *Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization’s ISO 27001 (<https://www.iso.org/iso/iec-27001-informationsecurity.html>) or the National Institute of Standards and Technology’s 800–53 controls (<https://nvd.nist.gov/800-53>).*

Adequate protection against security breach is important to protect the data from bad actors, both internal and external. This should be connected to the applicable U.S. security measures as these will vary by area of science. Such security measures should also be determined and regularly evaluated by experts and maintained by the federal government.

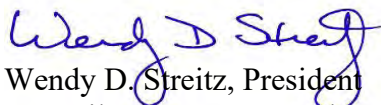
## *II. Additional Considerations for Repositories Storing Human Subjects Data (Even if De-Identified)*

We appreciate OSTP's recognition of protections for scientific data generated from humans or human biospecimens and as we shared with NIH when they requested similar feedback, we ask that OSTP explicitly acknowledge the role of the Institutional Review Board (IRB) in ensuring that such plans are appropriately disclosed in informed consent materials. OSTP may want to consider the existing NIH Genomic Data Sharing (GDS) Policy and related guidance as a model, as it provides a framework for IRB considerations such as risks associated with data sharing and evaluation of informed consent, including identification of circumstances where informed consent may not adequately address data sharing. There must be consistency between the plan and the informed consent obtained from human participants.

We also ask OSTP to consider issuing guidance on standards for dealing with uncontrolled access, de-identification, application of confidentiality policies, consequences of participant withdrawal or election to decline data sharing, and addressing requirements such as the Health Insurance Portability and Accountability Act, the European Economic Area's General Data Protection Regulation and other data protection laws, especially as the data could ultimately be used for commercial purposes through uncontrolled access.

In closing, we ask that OSTP continue to work with stakeholders with the goal of arriving at achievable standards for improving public access to data while minimizing the associated costs and burdens.

Sincerely,



Wendy D. Streitz, President  
Council on Governmental Relations (COGR)

[www.cogr.edu](http://www.cogr.edu)

Subject: RFC Response: Desirable Repository Characteristics

From: Jennifer Doty, Research Data Librarian, on behalf of Emory University Libraries

Discipline: research libraries support faculty, staff, and student researchers across all disciplines

Comments on DRAFT: Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded or Supported Research

**I. Desirable Characteristics for All Data Repositories**

- A. *Persistent Unique Identifiers*: No comment
- B. *Long-term sustainability*: How would long-term sustainability be reconciled with point F (*Free & Easy to Access and Reuse*)? For any repository providing access to data free of charge, another funding stream should be identified to have a sustained existence (i.e. for the repository to exist beyond the start-up period often funded by soft money).
- C. *Metadata*: No comment
- D. *Curation & Quality Assurance*: No comment
- E. *Access*: No comment
- F. *Free & Easy to Access and Reuse*: See comment for B (*Long-term sustainability*)
- G. *Reuse*: We recommend changing this to “Citation & Reuse Tracking” since it is focused on tracking reuse of datasets, which is typically achieved with proper citation and use of persistent unique identifiers.
- H. *Secure*: No comment
- I. *Privacy*: No comment
- J. *Common Format*:
- K. *Provenance*: No comment

Not included in section I characteristics:

- *Linking to Publications*: providing a direct link to related publications (e.g. journal articles, book chapters, working papers, etc.) is essential to understanding the data, and is a common practice amongst established data repositories.
- *Software and Other Computational Environment Information*: this could be included as its own characteristic, or *Metadata* could be more robustly defined to include this information and related parameters for using data.

**II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

- A. *Fidelity to Consent*: We are concerned that current consent language used for some human subject research is already too restrictive, and would prevent appropriate sharing of deidentified data from the initiation of a project.
- B. *Restricted Use Compliant*: “...preventing reidentification or redistribution to unauthorized users”—this phrasing is problematic due to the nature of

disclosure risk and the impossibility of preventing **all** future reidentification of data. We recommend restating this characteristic.

- C. *Privacy*: No comment
- D. *Plan for Breach*: No comment
- E. *Download Control*: No comment
- F. *Clear Use Guidance*: No comment
- G. *Retention Guidelines*: Retention by whom? If it is the end-user of the data, that should be explicitly stated. If it's by the data repository administrators, this characteristic should also be included in section I.
- H. *Violations*: No comment
- I. *Request Review*: No comment

Not included in section II characteristics:

- *Compliance with Policies*: we expected to see an explicit characteristic stating that any repository must comply with federal, state, and institutional policies governing the collection, retention, and dissemination of human subject data.

## **LTER IMC Response to OSTP Desirable Repository Characteristics RFC**

To: OpenScience@ostp.eop.gov

Subject Line: RFC Response: Desirable Repository Characteristics

Date: March 13, 2020

Dear Subcommittee on Open Science (SOS),

Thank you for the opportunity to comment on the document “Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research”. The responses provided here reflect the collective input of the Information Management Committee (IMC) of the Long Term Ecological Research (LTER) network and its data repository, the Environmental Data Initiative (EDI). LTER is a program supported by the National Science Foundation (NSF), consisting currently of over 2000 researchers at 28 sites (but has included a total of 32 sites over its history) who apply long-term observations, experiments, and modeling to understand how ecological systems function over decades. Providing open, accessible, well-documented data has been a cornerstone of LTER science and policy since its inception in 1980. EDI, also supported by NSF, houses meticulously curated data from the LTER, as well as from other ecological research projects, and is grounded in data and metadata practices and software systems originally developed by the LTER IMC. EDI and LTER are signatories of the Enabling FAIR Data (<https://copdess.org/enabling-fair-data-project/>) principles.

The LTER Network and EDI strongly support OSTP’s effort to identify and amplify best practices for data repositories, and applaud the broad principles outlined in the draft document. We draw on our deep experience in the area of ecological data curation to suggest a few areas that may benefit from additional detail to avoid a repository landscape that may fulfill requirements of data publishing but nonetheless render those data difficult to discover and reuse.

### **Comments to document sections**

#### *Section I. Paragraph B. Long-term sustainability*

Comment: Long-term sustainability—as the draft policy notes—is critical and repositories must be responsible for developing long-term plans and contingencies. But there are real costs associated with repository operations and the potential sources of support are limited. As the research-data-publication ecosystem evolves and develops appropriate support mechanisms, repositories

will experience significant and costly disruption if agencies are unable to support gradual transitions between funding models. It is important to consider repositories as both homes for -- and sources of -- data. Marketing new repositories to dispersed audiences of data users entails significant costs, so there is value in balancing stability with innovation.

#### Section I. Paragraph C. *Metadata*

Comment: Efforts to encourage data publishing and reuse are laudable, and we are witnessing tremendous successes on these fronts. However, inadequately documented data can be rendered unusable or contribute to misuse. The policy should provide clarity and guidance regarding what it means to provide sufficient metadata. Guidelines should speak to documentation at multiple scales from the level of the study (e.g., methods employed to collect data) to attributes and units of measured variables to spatial and temporal characteristics. Such clarity, in our experience, can optimize the needed balance between the effort required to document data and the value of the data for reuse.

#### Section I. Paragraph G. *Reuse*

Comment: Metrics are essential to understanding patterns-of-use for improving tools and approaches going forward, and maximizing returns on investments. Metadata and PUIDs are important metrics in these regards but may not be sufficient. A better understanding of organizations, investigators, or other users who are accessing data is essential to ensure that data repositories have details needed to document the data life cycle from submission to reuse. Repositories and the agencies that fund them should consider additional approaches to cataloging data reuse, such as by requiring and collecting the ORCID (<https://orcid.org/>) identifier of data requesters, and by supporting clear guidelines and requirements for data citation.

#### Section I. Paragraph J. *Common Format*

Comment: Standards should be sufficient for machine readability, enabling data discovery and easing migration into current or future repositories and analytical systems.

#### Section I. Paragraph K. *Provenance*



Comment: A record of provenance, as described here, is essential and achievable for original datasets from a single study. Although perhaps beyond the scope of this RFC, and also relating to “Paragraph G. *Reuse*,” the role of repositories is expanding to accommodate synthetic data derived from multiple studies with multiple authors. In such situations, provenance tracking becomes more complex, but is just as important for assigning credit and maintaining the integrity of data citation.

#### Recommended Additional Characteristic for Inclusion

Comment: One aspect left out of these guidelines is the area of workforce development. A repository is only as good as the data and metadata being submitted. Significant consideration of personnel requirements is vital, as substantial cost is involved in training and data curation.

The policies adopted by the OSTP have important ramifications, and we thank you again for the opportunity to provide comments on the document.

Sincerely,

Stevan Earl, Co-chair LTER IMC  
Information Manager, Central Arizona–Phoenix LTER  
Data Manager, Global Institute of Sustainability, Arizona State University

Corinna Gries, Co-PI EDI  
University of Wisconsin, Madison

Suzanne Remillard, Co-chair LTER IMC  
Information Manager, Andrews Forest LTER, Oregon State University

Mark Servilla, Co-PI EDI  
University of New Mexico

Special thanks to these contributors:

Dan Bahaiddin  
Renée Brown  
Marty Downs  
Margaret O'Brien

**Name:** Raleigh L. Martin, Ph.D.

**Organizational Affiliation:** Self

**Primary Scientific Discipline:** Geosciences (Physical Sciences)

**Role:** Researcher / Administrator (Policy)

**Date:** March 17, 2020

**Re:** “Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research”

**Response:**

This RFC response is my personal opinion alone, and it does not reflect the official opinion of my current or previous employers. Nevertheless, these comments build on years of experience working as a research geoscientist, as the co-leader of a small disciplinary data repository, and as a AAAS Science & Technology Fellow hosted in the Directorate for Geosciences at the U.S. National Science Foundation (NSF), where I focused on improving data sharing by NSF grantees in the geosciences.

*The single most important factor in designing data repositories for federally-funded research is **people**.* Even if a repository builds on the best-available technology, adopts the most widely agreed-upon data and metadata standards, and is backed by the strictest funder and publisher data policies, it will fail without buy-in from all the people along the full research data lifecycle. Researchers need the training, support, and incentives to guide them in spending a portion of their valuable work time appropriately curating and sharing their data. Journal reviewers and editors need clear policies and review processes to enforce publication data sharing mandates. Federal agency program officers and grant reviewers likewise need robust policies and practices for reviewing proposal data management plans (DMPs) and ensuring that grantees follow through on data sharing promises. And the technologists who develop and manage data repository resources need to maintain a close dialogue with all stakeholders in the research data lifecycle, to ensure that repositories meet user needs and to advocate for the provisioning of sufficient resources for long-term sustainment of valuable data resources.

The research world is diverse, and therefore it is hard to specify in advance what the specific “desirable characteristics” might be for any particular research data repository. Even precise-sounding principles like “FAIR” (Findable, Accessible, Interoperable, Reusable) are at best aspirational, and such principles will be implemented in widely different ways depending on research community needs. For example, the success of the U.S. Global Change Research Program in providing FAIR data underlying National Climate Assessments reflects a centralized federal agency mission and the dedicated data curation work of federal staff and contractors. In contrast, the research of university scientists supported by extramural federal grants is highly heterogeneous, and overly centralized federal data mandates and repositories could actually serve to stifle the creativity and individuality of such research work.

Therefore, **my primary recommendation** is that federally-supported data repositories be designed and managed through a consultative and iterative process that brings together the full range of stakeholders involved in the research data lifecycle. In addition to the stakeholders above, such consultations should include educators and citizen scientists who may not traditionally engage in academic research, but who stand to benefit greatly by increasing the openness of federally-supported data and knowledge generation. One successful example of such a consultative process is the recently-completed “Enabling FAIR Data Project,” led by the American Geophysical Union (AGU) and supported by the Laura and John Arnold Foundation.<sup>1</sup> This project significantly advanced the conversation on research data sharing by bringing together data repository managers, infrastructure providers, scientific publishers, and other stakeholders for sustained conversations on coordinating research data best practices. The work was hard and the progress was slow. But such sustained conversations and hard work are essential to the development of research data resources that successfully align with the actual data needs and practices of the scientific community and the public.

<sup>1</sup> Stall, S., et al. (2018), Advancing FAIR data in Earth, space, and environmental science, *Eos*, 99, <https://doi.org/10.1029/2018EO109301>. Published on 05 November 2018.

Date: March 6, 2020

To: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

From: University Libraries, University of Nebraska–Lincoln

Subject: Response to “Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research”

The University of Nebraska–Lincoln (UNL) serves as the flagship and land grant institution of the University of Nebraska system and is classified within the "Carnegie R1: Doctoral Universities: Highest Research Activity" category. The University Libraries at UNL is a campus entity that directly supports scholars in their research and data needs. In response to the "Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research" (<https://www.federalregister.gov/documents/2020/01/17/2020-00689/request-for-public-comment-on-draft-desirable-characteristics-of-repositories-for-managing-and>), the following comments are provided:

In response to I.A., *Persistent Unique Identifiers*: Persistent identifiers should be applied to metadata records for datasets as well as to the datasets themselves. The addition of persistent identifiers for metadata records describing datasets is especially important in cases where data are not or cannot be considered "open" or may not be perpetually retained. Metadata records with a persistent identifier can be found by researchers, demonstrating that such data exist or existed for research purposes, including reducing redundancy.

In response to I.B., *Long-term sustainability*: Long-term plans for managing data should address data that are not intended to be retained in perpetuity; preservation of data should not be understood to mean perpetual retention in all cases. Context-specific planning with regard to the data lifecycle should be incorporated into long-term retention and preservation.

In response to I.D., *Curation & Quality Assurance*: The provision of data curation and quality assurance requires significant expertise and resources, and at present there are not consistent, sustainable models for providing this expertise or for funding it. The Data Curation Network (<https://datacurationnetwork.org/>) has made steps in this regard. However, curation and quality assurance require a level of resources and expertise that most institutions are not in a position to fund and staff adequately, particularly at public institutions and other organizations seeking to serve the common good.

In response to I.E. *Access*: If researchers are not working with sensitive data, there should be an expectation that it will be open. The current language of the draft may not state this directly enough. Likewise, the draft lacks substance on what it means to provide equitable access, and we recommend more specific guidance and thinking about this element of the *Access* section in particular. What are the implications for how people package their data? For example, in areas with inadequate internet speed and reliability, downloads of large datasets may not be possible or practical.

In response to I.G., *Reuse*: Communities of practice and/or disciplinary societies should be encouraged to determine appropriate metadata schema (e.g., the ecology and evolutionary biology scientists and DARWIN Core), to facilitate reuse. While application of the FAIR guidelines will vary in extent, communities of practice and researchers should have a baseline for what constitutes "FAIR enough" for discovery and reuse. With regard to the data repositories themselves, repositories should have mechanisms in place to provide transparent analytics, allowing researchers to see if their data are being accessed and downloaded. Repositories should also support easy citation of data.

In response to I.K., *Provenance*: Consideration should be given to changes introduced or made by systems as well as those that may stem from such phenomena as data decay, in addition to those introduced directly by users.

More broadly, the characteristics seem specific to data that can be represented in discrete captures. The guidelines do not reflect the complexities of scale and volume, complexities at the heart of a growing number of federally-funded projects. Likewise, the characteristics do not appear to consider data that are generated or affected continuously and in real-time, which might be considered data streams, rather than data sets.

We also return to a point made above about the resources--financial, human, technological--necessary to support the sharing of data from federally-funded research. If the OSTP is prepared to frame characteristics for the repositories of this data, then there should be a commitment as well to the resources necessary to undertake this repository and related work. The generation of the data is supported by federal dollars, and typically the federal funding for a research project stops short of supporting the level of curation, maintenance, and preservation called for in these draft guidelines. There is a significant financial cost for institutions to maintain repositories and to maintain compliance with the OTSP characteristics as described in the draft. Federal support should aid in maintaining current technological infrastructure to ensure free and equitable access according to the proposed guidelines. This challenge is not likely to be solved through one-time funds for such work associated with specific grants, in part because the work of preservation is active and ongoing. Many institutions do not have the capacity or funding to address all of their researchers' needs for data management and sharing, especially in cases of extremely large collections and/or sensitive data. One possibility is that federal funders should plan, establish, and subsidize repositories capable of ingesting and preserving data from funded projects (e.g., NLM's/NCBI's GenBank).

Filed by:

Leslie Delserone (life sciences, data librarian), University Libraries, University of Nebraska–Lincoln

Casey Hoeve (librarian), University Libraries, University of Nebraska–Lincoln

Elizabeth Lorang (humanities, librarian), University Libraries, University of Nebraska–Lincoln

Claire Stewart (administrator), University Libraries, University of Nebraska–Lincoln

## **Response to OSTP RFC - 85 FR 3085 - Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research**

*Melissa Haendel, Monica Munoz-Torres, Nomi Harris, and Chris Mungall, on behalf of the members of the Monarch Initiative.*

On behalf of the Monarch Initiative, we provide the following response to the Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research.

The Monarch Initiative (<https://monarchinitiative.org>), funded by the NIH, semantically integrates information from many public biomedical data repositories related to genes, variants, genotypes, phenotypes, and diseases in a variety of species. We develop algorithms and tools to identify animal models of human disease through phenotypic similarity for differential diagnostics and to facilitate translational research and mechanistic discovery. We are thus very familiar with the challenges and requirements of both utilizing and creating repositories for managing and sharing research data. Collectively, our team members cover a range of roles, including biomedical science researcher, bioinformatician, data scientist, standards developer, research repository manager, library/information scientist, and data curator. We focus on the life sciences.

Formidable challenges are involved in making scientific data repositories usable and useful in both the short term and the long term. Some of the challenges can be summarized as taking steps to make data findable, accessible, interoperable and reusable (FAIR). This includes clear and consistent ways to record data provenance, including the use of persistent identifiers; processes for data QA/QC; consistent and versioned analysis processes and pipelines; links to community-developed ontologies, etc. While the idea of FAIR has created community awareness that is to be applauded, more needs to be done to make data truly reusable. Policy recommendations should focus on the important, specific, and enforceable practices that truly make resources more Findable, Accessible, Interoperable, and Reusable. We refer the reader to our earlier FAIR-TLC response to RFI NOT-OD-16-133 Metrics to Assess Value of Biomedical Digital Repositories. <https://doi.org/10.5281/zenodo.203295>.

Making data FAIR requires a substantial amount of effort; one that is not always rewarded by funding agencies. Making data repositories sustainable presents additional challenges. For example, data formats change over time, analytical methods and pipelines evolve or become obsolete, measurement modalities are replaced. In sciences such as biology, with rapidly evolving knowledge, techniques and measurement modalities, preserving data in a useful form for more than a few years is a major undertaking. The Biden Cancer Moonshot report "*The Enhanced Data Sharing Working Group Recommendation: The Cancer Data Ecosystem*", which we contributed to, explained that to maximize the utility of the data assets to accelerate progress towards improving cancer outcomes, we must consider all of the dependencies of the ecosystem of data, software, standards, and people - over time. Creating such interoperability requires an enormity of coordination and standardization, both in terms of the data itself, but also the standards and tools required to ensure data interoperability.

<https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel/enhanced-data-sharing-working-group-report.pdf>

We note that it would be a good idea to define exactly what is meant by "Managing" and "Sharing" data. We feel that "Data Management" should be an integral and iterative part of ongoing research, and that this critical role requires specialized skills to ensure that shared data is useful and reusable where possible. Regarding "Sharing", there are some specifics that should also be mentioned, such

as access mechanisms (API, data downloads, etc.), archival plans and persistence (DOIs or other persistent identifiers), standards and data harmonization (checklists, models, ontologies), and licensing (use rights, flow-through terms). We address some of these in our comments below.

## Persistent Unique Identifiers (PUIDs)

The appropriate provisioning and use of persistent identifiers are key in making data FAIR. Monarch team members and colleagues have published recommendations for identifier best practices (<https://doi.org/10.1371/journal.pbio.2001414>) based on our extensive efforts to reuse data from public knowledgebases and databases.

We support the use of resolvable PUIDs to identify and access datasets. We recommend that there should be a standard way to resolve the PUID to a machine-readable description of the datasets, and that where appropriate, dataset values use PUIDs for data items rather than relying solely on names or symbols which may be ambiguous. For example, in a dataset where each row represents observations about a gene (such as its expression under particular conditions), a standard identifier for that gene should be included in addition to the gene symbol. The identifier should follow community standards for disambiguating identifiers, such as those described in our 2017 publication, <https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2001414>.

## Metadata, Provenance, Interoperability, and Common Formats

**Metadata and Provenance.** Data that describe the data, formatted in a standardized way, are critical for making datasets findable and usable in the short and long term. These metadata are also critical for provenance and attribution, two key attributes required for maximizing data reuse and persistence. Other standards for metadata in wide use include numerous W3C standards and [schema.org](http://schema.org).

**Interoperability.** We emphasize the necessity of the use of ontologies and standard terminologies for structuring and annotating data. Controlled vocabularies or ontologies can radically reduce ambiguity while retaining human-readability. For the life sciences, the Open Bio Ontologies (OBO) Foundry project (<http://obofoundry.org>) provides a collection of community-developed ontologies for use in standardizing data. Within Monarch, we have been able to demonstrate the ability to reuse data from many sources by leveraging interoperability using ontologies. Monarch's integrated data from across the phylogenetic spectrum of model organisms is now used to help diagnose rare disease patients. In translational science, this work has been foundational in demonstrating the impact of data reusability using semantics for interoperability.

**Common formats.** It is not relevant to enumerate the numerous common formats here; however, we provide an example to illustrate the impact that a common format can have. Phenopackets is an exchange standard developed within the GA4GH; its goal is to be able to exchange, in a non-lossy manner, phenotype data about an individual from within Electronic Health Records, Journals, databases, patient registries, and clinical laboratories. Similar to FASTA for sequence data, a common format is critical to support the myriad of uses and innovation. However, developing a common format or standard often can take years of iterative community coordination, implementation, evaluation, and adoption.

## Curation & Quality Assurance

One of the most under-valued and least understood activities in populating and maintaining a data repository is curation. The quality and consistency of the data going into any given repository directly impacts the utility and long term preservation of the data. Curation requires robustly defined SOPs, standards, terminologies, identifier strategies, scientific validation, and forward migration tools in

order to sustain quality of the resource. While automated strategies can be used in combination with curation to enhance efficiency, it is almost never the case that they can fully replace the specific expertise required to finalize high quality data for reuse. Finally, sharing of data curation protocols and best practices is also not often done, although it is highly useful for downstream users of the data. We refer the reader to the International Society for Biocuration (<https://www.biocuration.org/>) for a variety of expert areas in curation, as well as this manuscript (<https://doi.org/10.1371/journal.pcbi.1006906>) detailing 10 best practice rules for biocuration:

- Tip 1. Know the subject area and assemble a team of experts
- Tip 2. Clearly define the intended use of the curated data
- Tip 3. Automate as much curation as possible
- Tip 4. Share your data in a standard structure
- Tip 5. Use ontologies and persistent identifiers to annotate your data
- Tip 6. Develop robust curation guidelines that include provenance and attribution
- Tip 7. Curate early, stay cozy with the data
- Tip 8. Commit to maintaining data
- Tip 9. Learn basic programming for ad hoc data wrangling
- Tip 10. Persist the data and provide it in multiple formats

## Access

**Diverse data access mechanisms.** Where practical, the resource should provide the option to download all data via one or more well documented mechanisms. Evidence of dissemination mechanisms that enable the community to use knowledge and data in innovative and reproducible ways should be obvious. Recommendations:

1. **Dumps:** Whole database dumps are available (where appropriate)
2. **Query:** Query interfaces or Mart-style exports, where possible
3. **Downloads:** Slices of the database and individual records can be downloaded (e.g. as JSON/XML/tab delimited, etc.)
4. **API:** Application Programming Interface (API) for the data exists

**Well structured and provisioned APIs.** If the resource provides an API, the following implementation guidelines are recommended. Direct database endpoints (e.g. MySQL, SPARQL etc) can be valuable; however, expertise in using these varies. Therefore, it is important to also wrap these with an API wherever possible. A summary of important REST principles is below; see also SSI REST best practices here

(<https://www.software.ac.uk/blog/2016-10-06-top-tips-creating-web-services>).

Recommendations:

1. **RESTful:** Follow RESTful API pattern
2. **JSON:** Return JSON or JSON-LD if possible, TSV if not
3. **Retrieval:**
  - a. Allow retrieval of a single record by using its identifier
  - b. Allow batch retrieval of a list of data entities using a list of identifiers
4. **Paging:** Provide a query interface to return matching data entities with paging support
5. **Versioned:**
  - a. Provide versioned URL pattern for future API changes
  - b. Document policies for change management
6. **Uptime:**
  - a. Provide an API uptime report (third-party services are available to reduce the implementation burden)



## 7. Access:

- a. Grant access requests (e.g. new accounts or API keys) promptly and efficiently
- b. Grant write access to trusted partners to make contributions, corrections, suggestions to records

## Data Licensing and Reuse

Not all data resources are free to use, derive, and redistribute, even if they are publicly funded and seemingly publicly available. We believe that there needs to be better awareness of the impacts of data license choices among both resource providers and government agencies. Moreover, few databases produce just data; most also produce software source code, algorithms, and applications. There should be licenses explicitly covering each of these products. We have created a rubric for assessing licenses of data held within data repositories and have applied this rubric at <http://reusabledata.org/> and published a manuscript (<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0213090>).

### Recommendations:

1. **Documented:** Explicit data use terms (ideally formal licenses) should be defined by the resource providers and easy to find
2. **Clear:**
  - a. At a minimum, licenses/data use agreements must be clear and easy to understand. A variety of specific examples of data use/reuse conditions should be included.
  - b. Licenses should not require negotiation and licenses themselves should be legally redistributable without engaging legal counsel
3. **Minimally restrictive:** The licenses and/or data use agreements should explicitly permit downstream data reuse, derivation, and re-dissemination
4. **Standard licenses.** We note that considerations for data are significantly different than those for software and they must be considered separately (see this blog for example - <http://lu.is/blog/2016/09/14/copyleft-and-data-databases-as-poor-subject/>).
  - a. **Standard data license:** For data, ideally CC0.
  - b. **Standard software license:** For software, ideally Apache version 2. Note that software license choices are the subject of much community discussion especially regarding “copy-left” approaches and there are other valid standard options available (such as GPLv2, GPLv3, AGPLv3, etc.)
5. **Contactable:** There should be an appropriate person available for contact with questions about licensure; this person’s contact information should be easy to find
6. **Transparent about flowthrough implications.** If others’ data is redistributed, clarity about the licensing implications of the redistribution is critically important. Concrete metrics:
  - a. Documentation about which source resources/data, if any, come with flowthrough implications
  - b. Links to the original licenses/data use terms of all redistributed content. It is currently commonplace that such terms do not exist; in such cases, it should be clearly stated that license/terms could not be found.
  - c. If specific authorization has been obtained for redistribution
  - d. Flowthrough implications are especially important for downstream data integrators. If we truly want to maximize data reuse, we must make it easy to redistribute freely.

March 17, 2020

**Response to: *Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research (OSTP)***

We thank the Office of Science and Technology Policy for the opportunity to respond to the Request For Public Comment released on January 17th, 2020 on the Federal Register.

Like our partners and sponsors in the federal government, we at Stanford University are committed to FAIR access and preservation of research data, and the benefits to the public that such practices make possible. In addition to the numerous discipline-specific repositories that our faculty use, including federally funded repositories operated by NIH and others, Stanford has its own institutional repository, the Stanford Digital Repository (SDR) which has been in constant operation since 2006, and is mentioned by name in the AAU-APLU Public Access Working Group Report and Recommendations (November 2017), along with Stanford's recommended best practices in research data management. As professional researchers, librarians, preservationists, and technologists, we look forward to engaging with OSTP on this matter. Your proposed framework has many laudable elements.

**Summary:**

Though we respond directly to most characteristics, below, there are two high level items that we would call to your attention first as top priority that should be addressed.

1. Cost implications
2. Harmonizing (security) standards

**Cost**

Regarding cost (**Section F**), the RFC states that a repository should "make datasets and their metadata accessible free of charge..." While a laudable goal that is consonant with the public interest, we encourage OSTP to examine how this policy would impact repository operators in practical terms. Research datasets are becoming increasingly large, reaching the petascale in some disciplines. We live in an inflationary universe of data storage. For research universities at the cutting edge of science, we cannot predict how data storage needs will expand further still, as our researchers are constantly inventing new techniques and areas of inquiry.

A common strategy for dealing with the explosion of data that we are currently witnessing is to rely on cold storage from cloud providers, as one of the only scalable routes forward. Part of the reason for this approach is its affordability: cold storage is cheap...so long as the data is left alone. As soon as it is requested, and read, there is an incremental, and potentially non-trivial, cost involved for data egress. If a repository is expected to provide such access for all research data to anyone who requests it, that may come at an unsustainable cost, and some allowance for metering, throttling or charge-backs (at cost) may become essential.

**Recommendation:** Metadata for all datasets should be available free of charge in a timely manner after submission. Data sets should also be available free of charge, but may incur a recall charge from a data storage service, if applicable, that may be passed through to the requestor.

## Security

Regarding information security standards (**Section H**), the RFC mentions two extant standards for information security, ISO 27001 and NIST 800-53. These are worthwhile, but heavyweight, examples that raise concerns for compliance. In particular, the effort and investment required to certify ISO (for example) compliance may be overwhelming to viable repositories that do not carry PHI, high-risk or other sensitive data. Further, in an institutional environment, information security is an enterprise concern which requires a layered approach, and cannot be described in the context of a single system. For example, data center physical security, network security and account management for the institution--and increasingly for its IT service suppliers--all must factor in. Finally, transmission to and analysis in associated compute environments--with rapidly changing architectures and needs that may evolve more quickly than security standards--need to be factored in as an essential component of the (eco)system surrounding a data repository.

Rather than selecting a general purpose systems standard for information security, there may be advantages to naming a third applicable standard in OSTP guidance language that is more domain specific, such as **CoreTrust Seal** (<https://www.coretrustseal.org>). CoreTrust Seal names 16 distinct requirements, including security; many of the other requirements of the CoreTrust Seal are in alignment with additional areas that OSTP addresses in this RFC (metadata, expertise, integrity, authenticity, etc.), and it has been endorsed by the Research Data Alliance to harmonize international requirements standards for digital repositories. (<https://www.rda-alliance.org/rda-coretrustseal-adoption-story-across-domains-and-regions>)

## Response to Other Proposed Criteria

Regarding the other desirable characteristics proposed (**Section A-E,G,I-K**), we endorse these principles as necessary for providing FAIR access to research data.

Section A, proposing *persistent unique identifiers*, is necessary to ensure that each dataset is findable not only in the present but into the future, with a consistent way to track versions, and record changes, and link to associated datasets. While the creation of these persistent unique identifiers is not sufficient to track versions and record changes as those tasks and processes require additional infrastructure and metadata, datasets must be uniquely and persistently trackable in order for these other tasks and processes to be possible.

We heartily endorse Section B, *Long-term sustainability*. Only repositories with a preservation mandate, and viable means for providing long-term access (even when no longer commercially viable) should be considered meritorious homes for federally funded research data. Ideally,

this would also include long-term direct access to the datasets themselves, though it may require such long-term access be provided through means that have costs associated (see above discussion of Section F). Datasets must be preserved in such a way that they are robust to catastrophic loss, but we recognize that providing access to

data after such an event may have costs associated that should not be the sole responsibility of the preservation host (i.e., repository).

Section C addresses the need to preserve the necessary and sufficient information about each dataset required to enable its reuse "using a schema that is standard to the community the repository serves." Many discipline-specific communities have established standards and metadata schema which they have determined to be necessary to enable reuse of the research generated in their fields. While interdisciplinary research calls on work from many different fields, the specialized metadata schema used by these diverse fields often have little interoperability with each other. In order to maximize discovery of datasets, particularly for interdisciplinary research and across repositories, we recognize the need to adopt universal standards (e.g., Dublin Core, DataCite, DDI, Schema.org, etc.) in order to facilitate metadata interoperability across repositories in the name of FAIRness. Ideally, the standards will be built in a way that allows for discipline-specific schema to be subsets of broader and more general schema that support interdisciplinary metadata interoperability.

With regard to Section D, Curation and Quality Assurance, we note that institutional repositories are particularly well-suited to address the needs to provide expert human guidance through the research data lifecycle, in particular for the workflows around data curation and quality assurance, for many domains, including for research data that does not have a strong, established discipline-specific repository.

Section E addresses the need to have comprehensible and transparent access policies that are tailored to meet the needs of each dataset. Like in sections B and C above, long term plans to ensure updated, transparent and standardized (when possible) access policies are required. Access requirements for high-risk data in particular are subject to change. Further, for E it should be stressed that machine-actionable access is highly desirable for data repositories; humans may, but should not have to, access data directly via a download link.

Section G seems to be subsumed by the requirements of Sections A and C, and would benefit from further expansion or distinction. One potential point to support tracking of data reuse is the ability to capture and relate PUIDs for *all relevant entities associated with the research data*: identifiers for researchers, their institutions, shared equipment, associated articles, etc. The repository should be able to capture not just the research data itself, but also the context around it. Data repositories should further enable tracking of reuse by a.) instrumenting downloads or "hits", and b.) providing preferred citation formats as a feature.

Section J addresses the technical requirements for equitable access. To the extent possible, and without losing data, datasets should be reformatted into non-proprietary and openly accessible standards. Data in proprietary and commercially inaccessible formats, while nominally "accessible," do not provide meaningful access to those without the means to access the proprietary tools. We recognize that this will require curatorial assistance and that such assistance will require skilled labor.

Finally, Section K addresses the need to preserve the entire research data lifecycle of each dataset, a "lab notebook for a dataset" to adopt a metaphor of sorts. This often takes the form of a rich codebook or other long-form narrative description such as a laboratory notebook; and if the intent of this requirement is to ensure that information about: 1) how the data were generated, 2) how the data were collected, 3) where the data were collected, 4) what rights and obligations are created/promised through the collection process, and 5) how the data are modified at each step 6) with all of the above events in a time-auditable log. Further, the data

repository should also be capable of taking a deposit of any software essential to generating or interpreting the research data in a persistent and secure way (i.e., not just a link to an external, potentially ephemeral source code repository).

Respectfully submitted on behalf of Stanford University,



Mimi Calter  
Deputy University Librarian  
Stanford University

### Works Cited

AAU-APLU Public Access Working Group Report and Recommendations

<https://www.aau.edu/sites/default/files/AAU-Files/Key-Issues/Intellectual-Property/Public-Open-Access/AAU-APLU-Public-Access-Working-Group-Report.pdf>

Accessed on: 3/3/2020

Stanford Minimum Security Standards

<https://uit.stanford.edu/guide/securitystandards> . Accessed on: 3/2/2020

FIPS Pub 199: Standards for Security Categorization of Federal Information and Information Systems.

<https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.199.pdf>

Accessed on: 3/5/2020

CoreTrustSeal Trustworthy Data Repositories Requirements: Extended Guidance 2020–2022.

<https://zenodo.org/record/3632533> . Accessed on: 3/5/2020



March 17, 2020

Lisa Nichols  
Assistant Director for Academic Engagement  
Office of Science and Technology Policy  
Executive Office of the President  
1650 Pennsylvania Ave NW  
Washington, DC 20504

Re: Draft Desirable Characteristics of Repositories for Managing and Sharing Data  
Resulting From Federally Funded Research

Dear Dr. Nichols,

On behalf of the American Educational Research Association (AERA), thank you for the opportunity to comment on the Office of Science and Technology Policy (OSTP) request for comments on a draft set of desirable characteristics of data repositories used to locate, manage, share, and use data resulting from Federally funded research.

AERA is the major national scientific association of 25,000 faculty, researchers, graduate students, and other distinguished professionals dedicated to advancing scientific knowledge about education, encouraging scholarly inquiry related to education, and promoting the use of research to improve education and serve the public good. AERA has long been committed to data sharing as set forth in Standards 14.06 (a) – (f) of the *AERA Code of Ethics* (<https://doi.org/10.3102/0013189X11410403>) and in the *Standards for Reporting on Empirical Social Science Research in AERA Publications* (<http://journals.sagepub.com/doi/pdf/10.3102/0013189X035006033>) as well as in many prior statements to OSTP, federal science agencies, and the Academies.

AERA has encouraged education researchers to benefit from the use of data repositories to facilitate data sharing in safe, secure, discoverable, accessible, preservable, and citable form. In a major, longstanding initiative supported by the National Science Foundation, AERA established an education research data repository with the Inter-University Consortium for Political and Social Research (ICPSR) with an emphasis on providing technical assistance to NSF awardees with projects with potential for multi-users (DRL-0941014). With NSF continued support, we have continued this work with ICPSR widening our scope so that dissertation and early career scholars would be mentored in data sharing and data management of their research and locate their data and data-related products at the AERA-ICPSR repository upon publication of their work (DRL-1749275).

1430 K Street, NW • Washington, DC 20005 • (202) 238-3200

Facsimile (202) 238-3250 • <http://www.aera.net>

In our work, AERA has given considerable thought to what an expert, experienced, and high-quality data repository can provide for primary researchers as well as for data users to support sound and respectful data use. As a matter of AERA policy since 2015, authors in AERA journals are encouraged to archive their article-related data in public or restricted access form at ICPSR. AERA also has established a dedicated archive at *Open ICPSR* for each of our seven journals so that article-related data and data products (code, manuals or field guides) can be shared as part of the publication process. The editors of AERA's open access journal, *AERA Open*, lead among our editors in working proactively with authors to do so as the default process.

We appreciate the attention of the OSTP Subcommittee on Open Science to develop characteristics of data repositories as part of encouraging the sharing of data resulting from federally-funded research. As federal agencies continue to develop and refine data sharing and management policies, providing federal research grantees with guidance on trusted repositories is important in the responsible storage, use, and sharing of data. We also welcome the additional consideration of ensuring the safe storage and use of human subjects data in repositories. This additional consideration is particularly important in education research with the involvement of students, teachers, and parents to inform understanding how school and community contexts influence teaching and learning, as well as to improve educational outcomes.

AERA is pleased to see that the Subcommittee on Open Science's efforts to ensure that the draft desirable characteristics of repositories align with the FAIR (findable, accessible, interoperable, and reusable) principles. We see the final version of these characteristics as a valuable resource not only for federal agencies that have built their own repositories, but also for grantees seeking a repository to store data that is appropriate to their research.

### **Response to I. Desirable Characteristics for All Data Repositories**

Overall, we strongly support the draft desirable characteristics listed for all data repositories. As part of an AERA workshop on data sharing in the education and learning sciences supported by the National Science Foundation (DUE-1656866 and DUE1745569), representatives from data repositories highlighted several shared responsibilities and activities to facilitate data sharing and data use with appropriate protections, including:

- Making data easily discoverable and actionable
- Ensuring the sustainability and durability of repositories
- Preserving data in easily reusable formats
- Setting security, confidentiality, and privacy standards<sup>1</sup>

---

<sup>1</sup> Levine, F. J., Rosich, K. J., Nielsen, N., Talbot, C. (forthcoming, 2020). *Data sharing and research transparency at the article publishing stage: A workshop report*. Washington, DC: American Educational Research Association.

All of the desirable characteristics listed in this draft proposal reflect standards that long-standing data repositories, such as ICPSR, have in place. We wish to emphasize the importance not just of data but of archival attention to materials related to data documentation and data procedures that enhance the scientific value and impact of the research. Also, an archive's attention to the inclusion of complete meta data that can help with discoverability of archived data and the linkages to other research is important.

AERA wants to particularly applaud the inclusion of a persistent unique identifier as one of the desirable characteristics. Researchers conducting replication studies or examining new research questions with extant data need to cite those data as scientific contributions in their own right (not just cite the articles that report on those data) to further cumulative knowledge, make more transparent the foundations of their research, and provide attribution to the federal agencies and other funding sources that make possible the data and the findings upon which they rely or link.

We also support under the characteristic of "access" that repositories can provide broad, equitable, and maximally open access with appropriate privacy and confidentiality protections. Ensuring multiple levels of access is particularly important in managing human subjects data, as addressed in the response on those draft desirable data repository characteristics. For decades, archives like ICPSR and federal statistical agencies like the National Center for Education Statistics (NCES) have provided public-use data and restricted-use data under license agreements and have been attentive to testing the possibilities of deductive disclosure and levels of potential risk (based on the research issues, the populations under study, and the potential for reidentification). As advanced technologies have evolved and the potential for inadvertent disclosure has increased, consideration of tiered access and the rationale to locate data in trusted, secure, and knowledgeable repositories are even more important not to limit use but to have the capabilities to determine the conditions for sound and safe access and use.

## **Response to II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

We appreciate the attention to the storing and potential use of data involving humans in research. Several federal laws and regulations, including the Federal Policy for the Protection of Human Subjects known as the 'Common Rule' (45 CFR 46, Subpart A) and the Family Education Rights and Privacy Act (FERPA), provide language on the informed consent process and ensuring the confidentiality of data in education records. We welcome the inclusion of technical practices to protect the privacy of research subjects, such as a plan to address potential data breaches and review data requests.

We also encourage OSTP to review policies of agencies and data repositories that include human data related to survey data collection, storage, and subsequent use that can inform additional guidance. As one example, in its survey programs, NCES makes sensitive data and those with a greater risk of deidentification available only in restricted-use form. NCES is experienced and rigorous in its practices. Authorized users are subject to the laws, regulations, and penalties that apply to the NCES use of



confidential data of up to \$250,000 and six months in jail. The NCES Statistical Standards Program monitors the licensing process and inspections. The NCES website also has extensive materials on data access to public-use and restricted-use data, including a Restricted-Use Data Procedure Manual (NCES 2007 at <http://nces.ed.gov/pubs96/96860rev.pdf>).

As a second example, Databrary, a data repository that offers the ability to store and share data from video recordings, ensures that data submitted in the repository adheres to guidelines regarding consent and that any personally identifiable information in recordings is not compromised. The resources include templates for video data release for research participants; best practices for authorized users of data that include password generation, configuration of computers (e.g., disabling automatic log-in); and working with Institutional Review Boards on informed consent protocols. Databrary has benefitted from support from the National Science Foundation and the National Institute of Child Health and Human Development in building this high-value repository.

The mention of funding and support for data repositories merits attention. Repositories that meet the characteristics that we seek and can be certified as operating at the highest levels serve an important mediating and educative function for science in securing, managing, and enhancing data assets; preserving and making data accessible; keeping up with use possibilities and opportunities; and serving data providers and users beyond what individual investigators can do for themselves. We applaud OSTP giving attention to desirable characteristics and see also considerations of how to enable the best to be an important next step.

Thank you once again for the opportunity to comment. Please do not hesitate to call upon AERA if we can be helpful in informing the development of the desirable characteristics for data repositories.

Sincerely,

A handwritten signature in cursive script, appearing to read 'Felice J. Levine'.

Felice J. Levine, PhD  
Executive Director  
[flevine@aera.net](mailto:flevine@aera.net)  
202-238-3201



Tracy K. Teal, PhD  
*Dryad*  
Durham, NC 27702  
Phone: 530.341.3230  
E-mail: [tkteal@datadryad.org](mailto:tkteal@datadryad.org)  
URL: <http://datadryad.org>

Dr. Lisa Nichols  
Assistant Director for Academic Engagement  
Office of Science and Technology Policy  
Submitted via email: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

**RE: Docket ID OSTP-2020-0001 Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research (RFC Response: Desirable Repository Characteristics)**

March 17, 2020

Dear Dr. Nichols:

I write on behalf of Dryad with regard to the Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research issued on January 17, 2020.

The Dryad Digital Repository is a curated resource that makes research data discoverable, freely reusable, and citable. Dryad is a nonprofit organization committed to its mission of providing the infrastructure for, and promoting the re-use of, research data. It originated from an initiative among a group of leading journals and scientific societies to adopt a joint data archiving policy (JDAP) for their publications, and the recognition that open, easy-to-use, not-for-profit, community-governed data infrastructure was needed to support such a policy. These remain our guiding principles.

Dryad is a leader in data curation and data publishing, and has strategic partnerships with the California Digital Library and Zenodo. For the last ten years, Dryad has focused primarily on research data, supporting a CC0 license and manually curating each incoming dataset. Dryad has been broadly adopted by the biological research community and has grown across all research data fields. More than 100,000 authors have deposited 32,000+ data packages containing a total of more than 90,000 data files. The data were associated with peer-reviewed publications in 900+ different journals and 2,100+ institutions.

Dryad values open scholarship with its vision of a world where research data is openly available, integrated with the scholarly literature, and routinely re-used to create knowledge. We appreciate the Notice issued by the Office of Science and Technology Policy (OSTP) to solicit feedback and recommendations on approaches for ensuring long-term stewardship of, and broad public access to, data resulting from federally funded research.

While Dryad generally agrees with the OSTP's Draft Desirable Repository Characteristics, we believe that in order to broadly support and further data sharing and re-use, OSTP should consider the cost to curating and preserving research data and its relationship to equal access, repositories integration in the peer review process and support for software preservation and citation. These considerations, provided below, are in addition to the comments on specific aspects of the Draft Desirable Repository Characteristics.

## **Cost to research data curation and preservation affecting equal access**

As an existing general repository, we are aware of the costs associated with data curation, preservation and storage. As datasets increase in size and complexity, costs in staff time, software maintenance and storage associated with curation and preservation have continued to increase, even while being able to increase efficiency through updated systems. In our work with university research administrators, librarians, technology specialists and individual researchers, we see that the cost cannot be borne by individual universities or researchers, nor does this mechanism provide equitable access to sharing data or complying with federal requirements.

The goals around data sharing are that all data can be shared, not just data where researchers or institutions have the available resources to appropriately prepare data for sharing, cover data deposition costs and support the infrastructure for data transfer. Relying on individual or institutional level support for data sharing runs the risk of prioritizing data from privileged institutions or individuals in our society, in conflict with the goals of democratizing data for people to share and have access to data to address the questions that are important to them and their communities and advance all areas of science and society.

We support mechanisms to directly support researchers' data sharing in areas where they are under-resourced or support to general repositories that complements their business models and gives them the additional capacity to handle datasets from all researchers.

## **Peer review access to research datasets**

Dataset deposition to repositories is largely tied to article publication. Because of this, it is essential that repositories support blind-peer review access to research datasets. By providing a link for the journal office and reviewers to access the data, before curation or publication of the dataset, repositories can ensure that datasets are a part of the peer review process. Without this step, data can be left out from the review process and issues may arise post-article publication. To promote full transparency and best practices for open science, repositories should accommodate the peer review process and be able time-release and restrict access to reviewers.

## **Support for software preservation and citation**

Reproducible and transparent data publishing practices rely on software and underlying code used to analyze data. Repositories should support software preservation and citation within the data repository platform or through partnerships with software repositories.

## **Feedback on Desirable Characteristics for All Data Repositories**

We appreciate the comprehensive list of "Desirable Characteristics for All Data Repositories" (Section I) and the recognition of datasets as research products that have impact for researchers and the research community.

Supporting the recommendations of the University of California Office of the President and providing additional recommendations, we note that the following attributes to the list of desirable repository characteristics should be included:

- *Persistent Unique Identifiers*: Repositories should support versioning of the Persistent Unique Identifiers like digital object identifiers, accession numbers, and others. Additionally, to aid in the discoverability, transparency and re-use of datasets, repositories should support linking related works through persistent identifiers. For instance, to help ensure that citations and relationships between outputs are indexed, repositories should send data and article relationships to DataCite,

a central and open indexers for metadata

- *Long-term sustainability*: The organization sponsoring the repository should have a governance or leadership model that reflects the community and allows for decision-making aligned with the continued open access and sustainability of the data.
- *Metadata*: Repositories should implement best practices for standardized vocabularies in the metadata (such as, Crossref Funder Registry).
- *Curation & Quality Assurance*: Repositories should provide clarity on their dataset requirements and levels of curation.
- *Access*: Repositories user interfaces for deposition and access to the data should comply with federal accessibility requirements.
- *Free & Easy to Access and Reuse*: Repositories should be expected to implement Creative Commons licenses for published datasets.
- *Provenance*: Provenance tracking of datasets should be machine readable.

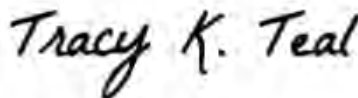
We also strongly encourage support for training of federal agency staff, researchers, librarians and technical specialists in helping to create, maintain and provide oversight for data that complies with FAIR standards. The stewardship, deposition and re-use of high quality data is an endeavor that requires all stakeholders to be educated and involved and partnerships across researchers, libraries and federal agencies.

#### **Feedback on Additional Considerations for Repositories Storing Human Data (Even if De-Identified)**

The access to and sharing of human subjects-related data is governed by a complex, fragmented set of ethical and legal requirements. Frameworks for accommodating these data, at scale, have not been developed. Dryad recommends that the OSTP work across federal funding agencies to provide guidance on appropriate ways to maintain sensitive data, so that general repositories, which are unlikely to meet these standards, can provide researchers a standard set of recommendations on appropriate repositories or resources for human data.

Thank you for the opportunity to comment on this important issue, and we look forward to continued engagement and discussion as further policies and other guidance is developed.

Sincerely,

A handwritten signature in black ink that reads "Tracy K. Teal". The signature is written in a cursive, slightly slanted style.

Dr. Tracy K. Teal  
Executive Director  
Dryad

**From:** Simon Hodson <[simon@codata.org](mailto:simon@codata.org)>

**Sent:** Wednesday, March 18, 2020 7:54 AM

**To:** Open Science <[OpenScience@OSTP.eop.gov](mailto:OpenScience@OSTP.eop.gov)>

**Subject:** Confinement en France! RE: [EXTERNAL] "RFC Response: Desirable Repository Characteristics"

As a necessary step to reduce the further spread of the SARS-CoV-2 virus, [the French government has required everyone in the territory to stay at home, only going out as briefly as possible for specific, justifiable reasons](#). Preparation for this state of affairs has interfered with my work for the last 24 hours. I will now be working from home for a number of weeks. I will adapt as quickly as possible to these circumstances. Please understand if there is a delay in responding during this period.

**Pending my reply, the following may be of interest:**

CODATA President, Barend Mons: 'World View' Opinion Piece in Nature: ['Invest 5% of research funds in ensuring data are reusable'](#)

[Webinar: Sustainable and Resilient Urban Ecologies – Possible lessons from recent Australian Bushfires](#) - Theresa Anderson, Associate Professor, Ethics for AI, 31 March, 11:00 UTC

SAVE THE DATE! [International FAIR Convergence Symposium and CODATA General Assembly](#), 22-24 October 2020, Paris, France

[February 2020 Publications](#) in the [CODATA Data Science Journal](#)

**Stay in touch with CODATA:**

- Find out what's happening! [Sign up to the CODATA International News List](#)
- Looking for training and career opportunities in data science and data stewardship? Sign up to the [CODATA community-run data science training and careers list](#)
- Follow us on social media! [Twitter](#) - [Facebook](#) - [LinkedIn](#) - [Instagram](#)

--

---

Dr Simon Hodson | Executive Director CODATA | [>http://www.codata.org<](http://www.codata.org)

E-Mail: [simon@codata.org](mailto:simon@codata.org) | Twitter: @simonhodson99 | Skype: simonhodson99

Tel (Office): +33 1 45 25 04 96 | Tel (Cell): +33 6 86 30 42 59

CODATA (Committee on Data of the International Science Council), 5 rue Auguste Vacquerie, 75016 Paris,  
FRANCE

Response to an OSTP notice on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research (3/5/2020)

The University of Florida (UF) Libraries working in collaboration with UF Research Computing and other campus partners developed a collaborative response to the [OSTP](#) Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research.

Data that is findable, accessible, reusable, and interoperable requires good data management. Good data management requires collaboration between many stakeholders to support the infrastructure and resources needed to enable sustainable good data management long term.

The [CoreTrustSeal](#) Trustworthy Data Repositories Requirements 2020-2022, Coalition for Publishing Data in the Earth and Space Sciences ([COPDESS](#)) Enabling FAIR Data – FAQs, FAIRsharing Collaboration with DataCite and Publishers: Data Repository Selection, Criteria That Matter<sup>1</sup>, FAIR Guiding Principles<sup>2</sup>, and Confederation of Open Access Repositories (COAR)<sup>3</sup> were referenced in the development of the comments.

The University of Florida and George A. Smathers Libraries supports “Desirable Characteristics for All Data Repositories,” I-A through I-K, with the following comments and recommendations:

## I. Desirable Characteristics for All Data Repositories

- A. Persistent Unique Identifiers:** Assigns datasets a citable, persistent unique identifier (PUIID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (e.g., of research progress), and research assessment (e.g., identifying the outputs of Federally funded research). The PUIID points to a persistent landing page that remains accessible even if the dataset is de-accessioned or no longer available.
  - 1. [Datasets PUIID should be agnostic, platform-independent, embedded, and support versioning.](#)
  - 2. [Facilitate PUIID for funders, institutions, researchers, and data.](#)
- B. Long-term sustainability:** Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.
  - 1. [Develop multi-stakeholders’ collaboration that support capacity, infrastructure, and resources with embedded contingency planning and continuous data management policy review for long-term sustainability over time.](#)
- C. Metadata:** Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.
  - 1. [Support generalist repositories \(e.g. \[NIH Generalist Repositories\]\(#\)\) with best practices, recommendations, and use cases from multiple disciplines to better support institutions in meeting funding agencies’ evolving data mandates.](#)

---

<sup>1</sup> McQuilton et al. (2019). FAIRsharing Collaboration with DataCite and Publishers: Data Repository Selection, Criteria That Matter. Retrieved from <https://osf.io/n9qj7/>.

<sup>2</sup> FAIR. (2016). FAIR Principles. Retrieved from <https://www.go-fair.org/fair-principles/>.

<sup>3</sup> COAR. (2020). What should be the essential baseline practices for repositories that manage research data? Retrieved from <http://tinyurl.com/r3m6cgu>.

- D. Curation & Quality Assurance:** Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.
1. The management of data throughout the data curation lifecycle involves the collaboration of key stakeholders. The stakeholders include the University, Office of Research, Research Compliance Office, Information Technology Department, Researchers, Academic Units, and the Library<sup>4</sup>. Strategic alliances between and across these stakeholders enable support, creation, operational and tactical<sup>5</sup> data curation throughout the data lifecycle at scale. A stable data repository infrastructure (e.g. OAIS<sup>6</sup>) for data creators, stewards, and users is key.
- E. Access:** Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.
1. This goal requires investments for APIs, software, and tools development that enable the infrastructure to support FAIR, open science, and reproducibility.
- F. Free & Easy to Access and Reuse:** Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.
1. Articulate different levels of access (e.g. open, embargo, closed) as appropriate.
- G. Reuse:** Enables tracking of data reuse (e.g., through assignment of adequate metadata and PUID).
1. Include machine-readable licenses, citation metadata and PUID for data reuse.
- H. Secure:** Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>) or the National Institute of Standards and Technology's 800-53 controls (<https://nvd.nist.gov/800-53>).
1. Include support for the Securing American Science and Technology Act of 2019. See Appendix 1.
- I. Privacy:** Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.
1. Clarify distinction in safe guards for open data, secure, and sensitive that may require offline or secure computing environment for select data according to regulatory frameworks.
- J. Common Format:** Allows datasets and metadata to be downloaded, accessed, or exported from the repository in a standards-compliant, and preferably non-proprietary, format.
1. Articulate responsibility of data creators and data owners to adhere to respective best practice, guidelines, and standards for common format. Recommend templates for QC/QA to ensure data input, integrity, and standards-compliance.
- K. Provenance:** Maintains a detailed log file of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.

---

<sup>4</sup> Erway, R. (2013). Starting the Conversation: University-wide Research Data Management Policy. Retrieved from <http://tinyurl.com/tjzlrk8>.

<sup>5</sup> UNSW. (2019). Research Data Governance & Materials Handling Policy. Retrieved from <http://tinyurl.com/s787ro9>.

<sup>6</sup> OAIS. (nd). Consortium of European Social Science Data Archives (cessda). Tutorial: OAIS. Retrieved from <http://tinyurl.com/sgre27w>.



1. Update “log file of changes” to record of changes. Maintain detailed versioning of changes to datasets, date, user, and data representation for accurate information.

## II. Additional Considerations for Repositories Storing Human Data (Even if De-Identified)

Stakeholders must ensure the management of sensitive data adheres to regulatory frameworks. Senior stakeholders must ensure the institutions and researchers have the capacity, infrastructure, and resources to support the management of sensitive, especially unfunded mandates. Thus, investment in infrastructure for secure computing environment (e.g. [UF Resvault](#)) can enable management of sensitive data before sharing via a data repository. The repository may offer restricted or closed access for some data that require access restriction. A desirable repository must include various levels of access commensurate with data type.

- A. **Fidelity to Consent:** Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).
  1. Agreed. Restrict access contingent on credentials and data type.
- B. **Restricted Use Compliant:** Enforces submitters' data use restrictions, such as preventing reidentification or redistribution to unauthorized users.
  1. Ensure Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule for PHI datasets.
  2. Articulate the Two methods to achieve de-identification in accordance with the HIPAA Privacy rule: Expert Determination; Safe Harbor<sup>7</sup>
- C. **Privacy:** Implements and provides documentation of security techniques appropriate for human subjects' data to protect from inappropriate access.
  1. See B.1 and B.2.
- D. **Plan for Breach:** Has security measures that include a data breach response plan.
  1. Include data deletion and/or shut down of functions during data breach.
- E. **Download Control:** Controls and audits access to and download of datasets.
  1. Flags abuse, high use, or misuse with warning notifications leading to restriction.
- F. **Clear Use Guidance:** Provides accompanying documentation describing restrictions on dataset access and use.
  1. See A.1.
- G. **Retention Guidelines:** Provides documentation on its guidelines for data retention.
  1. Articulate mandatory and regulatory data retention guidelines and stewardship.
- H. **Violations:** Has plans for addressing violations of terms-of-use by users and data mismanagement by the repository.
  1. Articulate policy for data use violations terms-of-use with informed consent.
- I. **Request Review:** Has an established data access review or oversight group responsible for reviewing data use requests.
  1. Ensure a systematic review for data use requests with two-levels of review.

The Appendix 1 includes an attempt to map comments (note three additional comments not previously included) to OSTP Response to FAIR Principles to Core Trust Seal for ease of reference to demonstrate overlap.

---

<sup>7</sup> U.S. Department of Health & Human Services (HHS). (2020). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Retrieved from <http://tinyurl.com/yxm3yo34>.

**Appendix 1: Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded or Support Research - Aligning Comments to OSTP Request for Public Comments to Core Trust Seal to FAIR**

ID	Comments	OSTP RFC	FAIR Guiding Principles	CoreTrustSeal
1	Assign unique id, digital object identifier (DOI), for all datasets upon deposit with version control feature.	<b>Persistent Unique Identifiers</b>	F1, F3	R13
2	Enable multi-stakeholders support across organizations for long-term sustainability of capacities, infrastructure, and resources (e.g. CERN, OpenAire, and European Commission’s support of <a href="#">Zenodo</a> ).	<b>Long-term sustainability</b>		R1, R3, R5, R6, R7, R8, R9, R10, R11, R12, R15
3	Promote appropriate metadata standard with multiple export options.	<b>Metadata</b>	F1, F2, F3, F4, A1, A2,	R4, R14
4	Provide support for curation expertise across disciplines (e.g. ARL-CARL Task Force on Research Data Services).	<b>Curation and Quality Assurance</b>	F1, F2, F3, F4	R4, R5, R6, R7, R8, R9, R11
5	Provide different levels of access (e.g. Open Access, Embargoed, Restricted, and Closed Access). Support <a href="#">OAI-PMH</a> via API; <a href="#">OSTI OAI Repository Manual</a> .	<b>Access</b>	A1, A1.1, A1.2, A2	R3, R14
6	Provide resources to support <a href="#">open science</a> (e.g. <a href="#">EOSC</a> ) and data sharing agreements (e.g. <a href="#">USGS</a> ).	<b>Free and Easy to Access and Reuse</b>	A1.1, R1, R1.1, R1.2, R1.3	R3, R14
7	PUID, citation metadata, human and machine-readable licenses	<b>Reuse</b>	R1, R1.1, R1.2, R1.3	R14
8	“ <a href="#">Support</a> the <a href="#">Securing American Science and Technology Act of 2019</a> , which would require federal research and security agencies to coordinate in an effort to better safeguard federally funded research from foreign influence, attacks, and theft.” - <a href="#">APLU</a>	<b>Secure</b>		R5, R16
9	Adhere to established responsible code of conduct and data protection.	<b>Privacy</b>		R4
10	Promote best practices/standards, <a href="#">repository schema</a> , and <a href="#">TRAC Metrics</a> .	<b>Common Format</b>	I1, I2, I3, R1.3	R8
11	Ensure copyright/IP compliance	<b>Provenance</b>	R1.1, R1.2, R1.3	R2
12	<i>Repository type: Institutional, <a href="#">Generalist</a>, or Discipline-specific.</i>			
13	<i>Repository status: development, production.</i>			
14	<i>Repository certification: level of fitness; trustworthiness; icons ( e.g. <a href="#">re3data</a>)</i>			



March 18, 2020

National Science and Technology Council  
Office of Science and Technology Policy  
The White House  
[OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

RE: RFC Response: Desirable Repository Characteristics

Thank you for the opportunity to comment on Desirable Repository Characteristics. The NSC Alliance is a non-profit scientific organization whose members include natural history museums, botanical gardens, and other scientific collections, the people who make and care for these collections, and all who use scientific collections for research and education that benefits science and the public.

The North American museum community has committed to digitizing data from their collections and making these data publicly accessible through data portals (national and international). A primary broader goal is improved access to data and publications resulting from federally funded research. These efforts have received federal support, particularly from the National Science Foundation. Importantly digital repositories continue to expand as the community continues to gather and create new types of data that must be linked to the original specimens for maximum scientific benefit and use.

Millions of specimens are now digitized and even more millions are left to digitize. Long-term access to these data requires new cyber-infrastructure associated with data storage, attribution, and access for the broad network of scientific collections across the country, now and into the future. Federal repositories are part of this network. The benefits are a time series of biological diversity data for use in research and conservation far into the future. These data are critical for the nation's bioeconomy and events such as the current COVID-19 pandemic are examples where effectively collected data in these repositories can be essential for understanding and addressing events involving pathogens with animal to human interactions.

**Desirable Repository Characteristics**

Below are characteristics that are essential for governmental repositories of biodiversity data.

- 1) **Regional distribution of facilities.** The federal government should support, with input from the museum community, a strong regional network of repositories for data.
- 2) **Necessary staffing with experts in science and collection management (digital and traditional).** Recent decisions by the United States Geological Survey to cut long-term support for curatorial and collection management positions in the Biological Survey Unit that were stationed onsite at the Smithsonian's National Museum of Natural History are contrary to what is needed to build and sustain a data repository. Agencies across the government should be adding positions to support the development and maintenance of data repositories.
- 3) **Funding to continue to digitize collections.** This has been highlighted in several recent reports and articles including the Biological Collections Network (BCoN) report: *Extending U.S. Biodiversity Collections to Promote Research and Education*. The most-speciose groups of non-bacterial organisms, insects and invertebrates are still poorly digitized because they present significant challenges by virtue of physical size, but also in the variety of manners in which they are preserved. What the scientific community needs with respect to biodiversity repositories are digital solutions allowing the assembly of and common access to the best and most comprehensive information about each specimen. This allows comparison and contrast of specimens by distributed networks of experts who can develop taxonomies and phylogenetic trees. These results can then be incorporated into an ever-expanding number of other research questions. This will open the door for research in many other scientific fields from ecology to neurobiology and epidemiology. Cyber-infrastructure for improving data attribution and connectivity is evolving, but the challenges associated with this necessarily distributed network of networks need to be overcome with visionary cyber-informatic approaches. Massive collections-based datasets of the evolutionary relationships of all biodiversity will be the source of an endless set of critical information about the interactions with humans and the rest of biodiversity now and into the future.
- 4) **A mandate to continue to collect.** The numerous recent reports focus on the need to digitize specimens already in collections because they can be studied to address many questions about what is happening now, but the value of these specimens coupled with new, extended specimens provides essential specimen time series to help address issues in every environment in this country far into the future.
- 5) **Outfitting for modern capacity for long-term preservation of extended specimens.** The new specimens that will be added to national repositories, government and otherwise require the necessary facilities for long-term preservation, that also provide access.
- 6) **Funding for in-house research.** Repositories for specimens and biodiversity data must be studied and experts directly associated with these collections are essential in this network, because of their understanding of specific repositories.
- 7) **Funding for effective in-house information technology expertise and outside partnerships.** There will be important technological advances in this digital age, and the museum community needs to have the in-house expertise necessary to take advantages of these advances. These experts will provide the cyber-infrastructure necessary to make the growing digital datasets and to effectively partner with entities outside our community on implementation of new solutions.

- 8) **Funding for data storage.** Data storage is a primary need for all such repositories and the needs are growing with more data-rich specimens and the need to archive imagery such as ct-scans and other new types of data.
- 9) **Connectivity via networks of similar repositories.** The crux of success for biodiversity repositories lies in the ability to create networks or networks across the country. This is because the only way to effectively monitor and understand our national biodiversity across different geographic scales. Biodiversity must be monitored locally and there must be quality repositories for the specimens and associated data that are gathered.
- 10) **Funds to develop and implement community-wide standards for data attribution.** The ability for the network of networks to function requires the development of mechanisms to insure tracking of data so that newly gathered data from specimens wherever it may be gathered is re-associated with the voucher specimens from which the data came.

Effectively expanding the digitization of data in repositories for all biodiversity from whales to microbes is essential for understanding environmental change effects ecosystems and regions through time. Cyber-infrastructure commitments to collections-based data are a national and global imperative required for to understand a world with rapidly changing climates. There is a somewhat analogous cyber-infrastructure to what is needed, which is the already existing databases that comprise GenBank at NCBI. These databases are constantly growing, data comes from government, public, and private sources. The data are stored and made publicly available in a stable system that continues to evolve and expand (from DNA sequences to whole genomes). With respect to biodiversity, DNA data archived in GenBank are most valuable when tied to phenotypic data such as those associated with extended specimens.

As a final characteristic, the government agencies should engage the museum community at large (private and academic museums across the country) more effectively. This includes providing continued financial support and interaction with respect to strategic initiatives across the nation's collections. As an example, the Government's Inter-agency Working Group on Scientific Collections is about to present a decadal report on Government collections and this committee has no non-governmental representation. To the broader museum community, this presents the potential for decisions that may affect our institutions in all manner of ways without adequate input from community experts outside of the government.

Thank you for your consideration of these issues. Please do not hesitate to contact me at [jbates@fieldmuseum.org](mailto:jbates@fieldmuseum.org) if the NSC Alliance can provide additional information.

John Bates, Ph.D.  
President  
Natural Science Collections Alliance

Date: March 17, 2020

Document #: 2020-00689

Re: RFC Response: Desirable Repository Characteristics

Submitted on behalf of: American Association of Physics Teachers  
American Crystallographic Association  
American Meteorological Society  
American Physical Society

---

### **Desirable Characteristics for All Data Repositories**

The characteristics included in this RFC are all worthy of being included. Furthermore, it is good that the draft characteristics do not over-specify the definitions of identifiers, APIs, metadata, discovery services, etc. as standards. In addition, the likely ramifications of technology change over time, and the need for experimentation to determine what standards best fit the market should be acknowledged.

We appreciate the fact that the characteristics listed are neither “intended to be an exhaustive set of design features for data repositories,” nor to be used “to assess, evaluate, or certify the acceptability of a specific data repository, unless otherwise specified for a particular agency program, initiative, or funding opportunity.” The characteristics and considerations enumerated below are similarly non-exhaustive and nonmandatory.

### **Additional characteristics that should be included**

Interoperability - Repositories should work towards utilizing common storage and reading formats and should shy away from using proprietary, and difficult to interconvert, storage and reading formats. Beyond storage, data and metadata should use a formal, accessible, shared, and broadly applicable language for knowledge representation, such that data and metadata are ready to be combined with other datasets by both computer systems and humans.

Removal – Repositories should have clear rules for removal of data and recordkeeping regarding removal of data. Additionally, removal of data should be considered as a final resort. Considerations regarding the removal of data should include the scientific process, scientific integrity and legal matters (fraud and privacy, for example). Specifically, to help ensure the integrity of the scientific record, it is important to mark retractions and the reasons for retraction, but there is likely merit in keeping the data set available. This can be particularly relevant in cases where others have incorporated the retracted data into their research.

Repository Governance – Repositories should have effective and relevant governing and advisory bodies that ensure alignment with the needs of the scientific enterprise as a whole and those of particular disciplines.

National Security – Research and data that is intended to be published and shared widely should only be censored through traditional classification. If preserving national security requires that certain data be censored, repositories should not be subject to a new method of restriction.

### **Other topics relevant for Federal agencies to consider in developing desirable characteristics for data repositories**

Researcher Incentives – One of the major challenges with regard to data management is participation by researchers and/or authors. Depositing data in repositories can be a significant burden on researchers. In some cases, scholarly publishers assist, facilitate or deposit data related to journal articles. However,

it should be recognized that systematic deposition by publishers on behalf of authors is complex and comes with real and non-negligible costs. Scientific societies and publishers can be part of the solution to these issues, but these activities must be adequately funded to ensure that they are sustainable over the long term. Societies can also help by raising awareness, supporting data submissions and facilitating dialogue between and among interested parties.

Cost-benefit Analyses - Scientists place great value in rigor; that is, the strict application of the scientific method to ensure unbiased and well-controlled experimental design, methodology, analysis, interpretation and reporting of results. As part of this, scientists regularly present or share their work so that others can examine procedures and learn about results. In this same vein, repositories should regularly be evaluated for their contributions to the field of science and for their cost-benefit ratios.

Metrics of Success - Good metrics should be actionable and drive successful behavior. Metrics could be established by funding agencies, be government-wide, or be left to the individual repositories. Some standardization of metrics across fields and repositories is desirable. Flexibility should be a priority as needs, technology and market demands change.

Definition of Data - A larger conversation around the definition of data may be warranted. The definition given in OMB Circular A-81, section 200.315, "Research data means the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communication with colleagues." is a sufficient baseline. More data may have been collected during the course of an experiment that could be of interest. For example, null results may have been recorded and not submitted for publication. Researchers should be able to deposit all appropriate data associated with their research. However, there should be careful consideration of the burdens associated with submitting and storing related, null or ancillary data.

#### **Additional Considerations for Repositories Storing Human Data**

For repositories storing human data, repository guidelines should also address Health Insurance Portability and Accountability Act (HIPPA) requirements and documentation of review by an Institutional Review Board (IRB).

March 17, 2020

Lisa Nichols  
Office of Science and Technology Policy  
[OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

Subject: RFC Response: Desirable Repository Characteristics

Dear Dr. Nichols,

The Data Curation Network (DCN) thanks you for the opportunity to respond to the “Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting From Federally Funded Research” posted January 17, 2020 as document 85 FR 3085 in the Federal Register.

The [Data Curation Network](#) is a collaboration of 10 academic and general data repositories that share data curator expertise to overcome common challenges. Partner repositories include Cornell University’s eCommons, Dryad Digital Repository, Duke University’s Research Data Repository, Johns Hopkins Data Archive, University of Illinois Data Bank, University of Michigan’s Deep Blue Data, Data Repository for the University of Minnesota (DRUM), New York University’s (NYU) Faculty Digital Archive, Pennsylvania State University’s ScholarSphere, and Washington University in St. Louis’s Open Scholarship.

The DCN generally agrees with all of the desired characteristics, and rather than echo what [SPARC and CORE have expertly framed in their response](#), we would like to drill down on two characteristics in greater detail:

- D. Curation & Quality Assurance
- F. Free & Easy to Access and Reuse.

**D. Curation & Quality Assurance: Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.**

1. “Curation” is a term that may have multiple meanings depending on context and perspective. We recommend that the OSTP adopt a clear definition of “curation” to aid both researchers and repositories in understanding and adhering to



## DATA CURATION NETWORK

expectations in managing and sharing data. Our preferred definition of curation is the activity of managing and promoting the use of data from their point of creation to ensure that they are fit for contemporary purpose and available for discovery and reuse.<sup>1</sup>

2. Our research has shown that researchers view themselves as playing a key role in providing curation and quality assurance for their data, often starting when the data are created.<sup>2</sup> As such, data repository curators must bring in multiple perspectives, including the originating author, when providing additional curation and quality assurance services.
3. Curators employed by data repositories should be recognized as trained professionals who draw from an educational foundation in digital archives grounded in subject matter expertise. For example, [our data curators in the DCN](#) often have a PhD in a discipline combined with a terminal degree in library information science (e.g., MLS or MLIS) and supplement this with ongoing professional development in digital curation practice (e.g., SAA digital archives specialist certification<sup>3</sup>).
4. Not every general or multi-disciplinary data repository can hire an expert for the wide variety of data types and discipline-specific data formats that we receive (such as spatial data, code, databases, chemical spectra, 3D images, and genomic sequencing data). Therefore, the Data Curation Network, in addition to establishing a shared staffing model among our partner repositories, also created a platform for others to share expertise through Data Curation Primers.<sup>4</sup> These freely available tools are interactive, living documents that detail a specific subject, disciplinary area or curation task and that can be used as a reference to curate research data. Primers published by teams of experts include:

---

<sup>1</sup> CoreTrustSeal Standards and Certification Board. (2019, November 20). CoreTrustSeal Trustworthy Data Repositories Requirements: Glossary 2020–2022 (Version v02\_00-2020-2022). Zenodo. <http://doi.org/10.5281/zenodo.3632563>.

<sup>2</sup> Johnston, L.R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R. and Stewart, C., 2018. How Important is Data Curation? Gaps and Opportunities for Academic Libraries. *Journal of Librarianship and Scholarly Communication*, 6(1), p.eP2198. DOI: <http://doi.org/10.7710/2162-3309.2198>

<sup>3</sup> Society of American Archivists. Digital Archives Specialist (DAS) Curriculum and Certificate Program. <https://www2.archivists.org/prof-education/das>

<sup>4</sup> <https://datacurationnetwork.org/resources/data-curation-primers/>

## DATA CURATION NETWORK

- a. [Acrobat PDF](#) Primer
  - b. [ATLAS.ti](#) Primer
  - c. [Confocal Microscopy Image](#) Primer
  - d. [Geodatabase](#) Primer
  - e. [GeoJSON](#) Primer
  - f. [Jupyter Notebooks](#) Primer
  - g. [Microsoft Access](#) Primer
  - h. [Microsoft Excel](#) Primer
  - i. [netCDF](#) Primer and [Tutorial using an NCAR dataset](#)
  - j. [SPSS](#) Primer
  - k. [STL](#) Primer
  - l. [R](#) Primer
  - m. [Tableau](#) Primer
  - n. [WordPress.com](#) Primer
5. Curation should protect the chain of custody of a dataset and ensure authenticity of the data. Therefore, we recommend that data repositories strive for transparency in the curation actions taken both generally as well as the specific curation actions taken for an individual dataset. Such transparency would benefit data depositors when selecting a repository, as well as data consumers, when determining whether to use the data. For example, members of the data curation network take generalized actions for all data sets, called the Data Curation Network CURATED steps<sup>5</sup>, as well as specific actions that are detailed in a curation log. The CURATED steps include (briefly):
- a. **Check** - Create an inventory of the files and review received metadata
  - b. **Understand** - Run the data/code, read documentation, assess for QA/QC red flags
  - c. **Request** - Work with the author to address any missing information or changes needed
  - d. **Augment** - Enhance metadata for discoverability and contextualize data with appropriate linkages (e.g., PUID for paper or published code, etc.)
  - e. **Transform** - Convert files to non-proprietary formats, if appropriate
  - f. **Evaluate** - Review overall data package for FAIRness
  - g. **Document** - Record all curation activities in a log file
6. Levels of curation vary from repository to repository (see Table 1). Based on an examination of the work we do in the Data Curation Network, we recommend that federally funded research be shared in data repositories that practice **Enhanced curation** to ensure that data sets are complete and understandable to someone with similar qualifications and in formats that allow for long-term use.

---

<sup>5</sup> Data Curation Network (2018). "Checklist of CURATED Steps Performed by the Data Curation Network." <https://datacurationnetwork.org/resources/resources-2/>.

## DATA CURATION NETWORK

Table 1: CoreTrustSeal Levels of curation mapped to descriptions provided by the Data Curation Network

<i>Level of Curation</i>	<i>Description and Examples</i>
A. Content distributed as deposited	<p>Data sets are accepted into the repository with no curator intervention.</p> <p>e.g. FigShare, Zenodo, OpenICPSR, many institutional data repositories.</p>
B. Basic curation – e.g., brief checking, addition of basic metadata or documentation	<p>Basic curation is often applied at the metadata record level. Descriptive metadata, such as keywords using a controlled vocabulary, are reviewed, verified, and/or added to improve discoverability.</p> <p>e.g., Springer (\$), Mendeley Data (\$), some institutional data repositories.</p>
C. Enhanced curation – e.g., conversion to new formats, enhancement of documentation	<p>Enhanced curation is often applied at the file-level where data files are checked for completeness and documentation is reviewed and/or enhanced to be understandable by someone with similar qualifications as the data creator.</p> <p>e.g., Data Curation Network institutions. We follow the CURATE(D) steps in order to apply enhanced curation for a wide variety of data types.</p>
D. Data-level curation – as in C above, but with additional editing of deposited data for accuracy	<p>Data-level curation is often applied by a subject matter expert who reviews the contents of the files in a process analogous to peer-review. This deeper level of curation may involve quality control, harmonization to increase interoperability with other data sets, and domain-specific metadata augmentation.</p> <p>e.g., many domain repositories such as Protein Data Bank, ICPSR, DBGap, GenBank</p>

## DATA CURATION NETWORK

**F. Free & Easy to Access and Reuse: Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.**

1. We agree that data should be available to the user without cost. However, there are significant costs attached to providing long-term discovery, access, curation, preservation, and stewardship for data. As data repository managers, we ask that OSTP address this criteria by recognizing how data repositories fund these services.
  - a. In many cases, academic data repositories are supported through federal grants via indirect costs as well as through state and tuition funds for general operating costs.
  - b. If sufficient funding is not available from the federal agencies or publishers who mandate the deposit of data into repositories, the repository may need to charge reasonable, cost-recovery fees to researchers depositing their data to cover operating expenses. For example, the Dryad Digital Repository includes a \$120 deposit fee for authors, which may be covered by a researcher's institution via annual membership.

Sincerely, representatives of the members in the Data Curation Network

Lisa Johnston, University of Minnesota

Kathryn Wissel, New York University

Elizabeth Hull, Dryad

Mara Blake, Johns Hopkins University

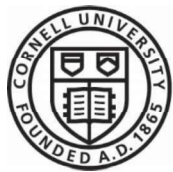
Cynthia Hudson Vitale, Pennsylvania State University

Joel Herndon, Duke University

Hoa Luong, University of Illinois

Wendy Kozlowski, Cornell University Library

Jake Carlson, University of Michigan



Cornell University

March 6, 2020

Dr. Kelvin Droegemeier  
Director  
Office of Science and Technology Policy  
Executive Office of the President  
Eisenhower Executive Office Building  
1650 Pennsylvania Avenue  
Washington, DC 20504

**Submitted online to: [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov).**

Dear Dr. Droegemeier,

On behalf of Cornell University and Weill Cornell Medicine, we write to provide comments on the US Office of Science Technology and Policy (OSTP) “Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research.” We applaud this effort and appreciate the opportunity to contribute our thoughts and expertise.

Cornell University is a world-class research institution known for the breadth and rigor of its curricula, and an academic culture dedicated to preparing students to be well-educated and well-rounded citizens of the world. Its faculty, staff and students believe in the critical importance of knowledge—both theoretical and applied—as a means of improving the human condition and solving the world’s problems. With campuses in Ithaca, New York and New York City, including Weill Cornell Medicine, and a location in Doha, Qatar, Cornell is a private Ivy League research university and the land-grant institution of New York State. Weill Cornell Medicine is committed to excellence in patient care, scientific discovery and the education of future physicians in New York City and around the world. The doctors and scientists of Weill Cornell Medicine — faculty from Weill Cornell Medical College, Weill Cornell Graduate School of Medical Sciences, and Weill Cornell Physician Organization—are engaged in clinical care and cutting-edge research that connects patients to the latest treatment innovations and prevention strategies.

Robust data repository guidelines will help researchers to make sound decisions about the repositories they use, present agency staff with a uniform set of considerations in evaluating the suitability of repositories for data resulting from the research that they fund, and help inform the management of repositories used for research data.

We offer the following general comments on the draft:

- Our position is that well-run general-purpose repositories, such as those operated by campus IT organizations and academic libraries, are a viable option for providing access to research data. Where established disciplinary or federal repositories do not exist, or the research data in question do not fall within the collecting scope of these repositories, a general-purpose repository may well be the *only* option available to federally-funded researchers. The final “desirable characteristics” should not preclude the use of such repositories by researchers.
- While we think not requiring formal certification is the right decision, we do wonder how repositories can demonstrate alignment with the desirable characteristics, and how researchers and federal staff will determine a repository’s fitness for purpose. More specific guidance on both fronts would be helpful.
- As noted in the draft, the document will need to be updated. The draft would benefit from a more specific plan for review and update, given the rapidly changing research and information technology landscape.
- It is not always clear when a criterion is something a repository should be capable of doing, and whether it can or must publicly demonstrate the same. For example, repositories may record complete provenance information with respect to changes to metadata, and file uploads and replacements, but may or may not display this information publicly, making compliance with a provenance requirement difficult to ascertain or demonstrate.
- Consideration should be given to the fact that data access requests could be abused or misused and institutions should be protected from harassment by individuals or entities that make infinite and/or unreasonable requests that institutions have legitimate reasons to deny. We strongly suggest that repositories that are responsibly managed are protected from extreme misinterpretations of the data access regulations.
- Finally, we recommend aligning as much as possible with existing and emerging data repository evaluation criteria, in order to avoid conflicting recommendations. One such example is “Data Repository Selection: Criteria That Matter” (<https://doi.org/10.17605/OSF.IO/N9QJ7>), developed collaboratively by the Research Data Alliance FAIRsharing working group, DataCite, and publishers.

We offer the following feedback on Section I, *Desirable Characteristics for All Data Repositories*:

- A. *Persistent Unique Identifiers*: PUIDs are key to both long-term access and tracking as well as for citability of the datasets. We are happy to see that the guidance is not prescriptive to a particular type of PUID, but that key functionality and good practice are highlighted.
- B. *Long-term sustainability*: This is a very broad and ambitious set of recommendations combined into a single criterion. We suggest unbundling them. This particular criterion might, at its core, be about the long-term sustainability of the repository itself – that is, the commitment of its parent institution and any succession plans that are in place. Specific requirements for the long-term retention of data sets, if that is what is meant, should be specified separately. Additional clarity regarding the degree to which this would include point-of-contact/custody chain issues relying on

long-term validity of contact details of the individual reasonable party for a given dataset would be helpful.

- C. *Metadata*: We suggest explicitly acknowledging that established, general purpose schema may be used when appropriate, particularly when metadata and data are accompanied by robust supplementary documentation that facilitates understanding and reuse. We also recommend that the National Library of Medicine or similar federal entity invest in the development of discipline-specific metadata standards creation as an ongoing priority. A minimal mandatory subset of metadata should also be included in the definition.
- D. *Curation & Quality Assurance*: We strongly agree on the importance of curation for increased value of datasets and related metadata. We suggest emphasizing the tangible improvements that result from curatorial activities, such as reproducibility and reusability. We also encourage a statement in this section (or in the section on provenance) that recommends a record of curatorial actions performed on a dataset, and inclusion of that information in the provenance documentation.
- E. *Access*: In addition to privacy and confidentiality issues, there may also be intellectual property, patent, or commercialization reasons to restrict access.
- F. *Free and Easy to Access and Reuse*: As written, this section is difficult to differentiate from “Access.” We suggest the wording be changed to more clearly or expressly allow the possibility that there will be costs incurred in providing access, and that sharing at not more than the cost of distribution is acceptable. Federal agencies should provide funding through their granting mechanisms for research to charge the cost of data retrieved from long-term storage. We strongly encourage the addition to this section that when sharing, terms of reuse should be clearly communicated.
- G. *Reuse*: Issues related to reuse are also addressed in the PUID and long-term sustainability characteristics, and as written, this section could be eliminated. Consideration should be given to the DataCite construct for tracking re-use, as this will create a virtuous cycle to incentivize depositing data.
- H. *Security*: We feel implementing and documenting NIST 800-53 controls is generally untenable/burdensome for most institutions (even for low or moderate impact data), as this standard typically applies to Federally classified information, which, by nature, would never be publicly released. Alternatives suggested below are more flexible and would allow organizations to choose the framework that best fits their program. Given that, we suggest the following wording changes:
  - a. For Restricted Use or Private datasets, provides documentation of meeting relevant criteria for security to prevent unauthorized access or release of data, such as the criteria described in the International Standards Organization's ISO 27001 (<https://www.iso.org/isoiec-27001-information-security.html>), the National Institute of Standards and

Technology's 800-171 controls for Controlled Unclassified Data (<https://csrc.nist.gov/CSRC/media/Publications/sp/800-171/rev-1/archive/2018-02-20/documents/NIST.SP.800-171r1-20180220.pdf>), or Center for Internet Security's Top 20 (<https://www.cisecurity.org/controls/cis-controls-list/>).

- I. *Privacy*: There are unique considerations that must be taken into account with privacy, especially for data privacy in evolving fields of research such as genetics and genomics.
- J. *Common Format*: It is important to acknowledge that common formats are not necessarily well-established for all data or all disciplines, and it is unclear what is meant by “standards-compliant.” We suggest a focus on meeting and communicating relevant standards as appropriate and, when standards do not exist, using non-proprietary formats to mitigate obsolescence and facilitate reuse over time.
- K. *Provenance*: We suggest not just maintenance of provenance records, but that repositories specifically provide mechanisms or actively enable recording of changes to datasets and metadata. We suggest you replace the word “logfile” with “documentation” to be more inclusive of the variety of data repository infrastructure and workflows in use. For example, there are a number of standards for digital versioning/logging of changes to text/code, while “git” is used for the open source community. Versioning of raw/numerical data is significantly more complicated.

We offer the following feedback on Section II, *Additional Considerations for Repositories Storing Human Data (Even if De-Identified)*:

- A. *Fidelity to Consent*: We strongly support a human subjects data repository that maintains IRB approval and data use agreements, and provisions access only in accordance with such consent agreements. We are very concerned, however, about the costs associated with inventing a new method for maintaining a scalable repository of consents linked to the data and access controls. Additional clarity is needed regarding the use of consented data in research repositories, as recent clarification attempts have only further contributed to confusion. In particular, there is a pressing need for simplified paths to “broad consent” and a path through overlapping and contradictory cross agency and cross government regulations. Unified guidance in this instance would also be beneficial.
- B. *Restricted Use Compliant*: We endorse a secure electronic human subjects data repository, actively managed by data custodians who capably restrict access to authorized users based on data use agreements. Perpetual use agreements should be avoided, but subject to review if established. As a rule, data should be retractable in the event that the status of a user or use-case changes. UKBioBank procedures could be used as an example of application for data access, use, and custody.
- C. *Privacy*: We strongly support a secure data enclave that is auditable and ensures that each access to human subjects data is fully documented. The enclave construct has limitations, however, and shared access and controls should be considered.



- D. *Plan for Breach*: It is unclear if this section is in reference to existing requirements or if this would be a new requirement. Regardless, we would caution against the creation of a new mandate that conflicts with existing mandates.
- E. *Download Control*: Access should not be assumed by this regulation. Liability for misuse should fall to the user; data providers should be appropriately indemnified by the user if they follow appropriate data access controls before granting a download.
- F. *Clear Use Guidance*: Clinical data should be shared only if explicitly allowed by broad and explicit consent of the patients involved. That said, new broader consent regulation would be helpful to allow for a hypothetical data collection with limited identifications for use in qualified research.
- G. *Retention Guidelines*: We are unsure how to interpret this section, as the community currently lacks clarity on existing data retention guidelines. It would be helpful to clarify whether this is intended to simply mean that a repository should have a retention policy as opposed to adhering to some other standards.
- H. *Violations*: Institutions can self-police to the extent that they have oversight of their own personnel and policies. Institutions, however, cannot be enforcement agencies for people with whom they are required to share but have no oversight over. We believe this is the responsibility of the NIH and other government agencies.
- I. *Request Review*: We support this kind of dual oversight for explicit human protections and data management. Additionally, we support the inclusion of patient advocates in some portion of the oversight process, though these advocates should be required to have some training in research methods, the ethical conduct of research, and principles of patient advocacy. We recommend the NIH *All of Us* Research Program data access review process as a model for requesting review.

Cornell University and Weill Cornell Medicine would like to thank you again for undertaking this review in such a thoughtful manner and especially for soliciting input from stakeholders. We would be pleased to work with you going forward. If you have any questions or would like to additional information or elaboration on anything contained in this letter, please contact Alessia Daniele, Associate Director for Federal Relations at Weill Cornell Medicine, at [ald2035@med.cornell.edu](mailto:ald2035@med.cornell.edu) or via phone at 646-962-485.

Sincerely,

  
Gerald Beasley (Mar 6, 2020)

**Gerald R. Beasley**  
Carl A. Kroch University Librarian  
Cornell University



**Curtis L. Cole, MD, FACP**  
Chief Information Officer  
Assistant Vice Provost for Information Technology  
Frances and John L. Loeb Associate Professor of  
Libraries and Information Technology  
Associate Professor of Clinical Medicine, Healthcare  
Policy and Research  
Weill Cornell Medicine



# AMERICAN SOCIETY OF HEMATOLOGY

2021 L Street, NW, Suite 900, Washington, DC 20036 **ph** 202.776.0544 **fax** 202.776.0545 **e-mail** ASH@hematology.org

March 10, 2020

## 2020

### President

Stephanie Lee, MD, MPH  
Fred Hutchinson Cancer Research Center  
1100 Fairview Avenue N, D5-200  
PO Box 19024  
Seattle, WA 98109  
Phone 206-667-5160

### President-Elect

Jane N. Winter, MD  
Northwestern University  
Robert H. Lurie Comprehensive Cancer Center  
676 N. Saint Clair Street, Suite 850  
Chicago, IL 60611

### Vice President

Martin Tallman, MD  
Memorial Sloan-Kettering Cancer Center  
1275 York Avenue  
Howard Building 718  
New York, NY 10065  
Phone 212-639-3842

### Secretary

Robert Brodsky, MD  
Johns Hopkins University  
Ross Building, Room 1025  
720 Rutland Avenue  
Baltimore, MD 21205  
Phone 410-502-2546

### Treasurer

Mark Crowther, MD  
McMaster University  
50 Charlton Avenue East  
Room L-301  
Hamilton, ON L8N-4A6  
Canada  
Phone 1-905-521-8024

### Councillors

Alison Loren, MD, MS  
Bob Lowenberg, MD  
Belinda Avalos, MD  
John Byrd, MD  
Cynthia Dunbar, MD  
Arnold Ganser, MD  
Agnes Lee, MD, MSc, FRCPC  
Joseph Mikhael, MD, FRCPC, Med

### Executive Director

Martha Liggett, Esq.

Lisa Nichols, PhD

Assistant Director for Academic Engagement

National Science and Technology Council Subcommittee on Open Science

Office of Science and Technology Policy

1650 Pennsylvania Avenue, NW

Washington, DC 20504

Comments submitted online to [OpenScience@ostp.eop.gov](mailto:OpenScience@ostp.eop.gov)

RE: Request for Public Comment on Draft Desirable Characteristics of Repositories for Managing and Sharing Data Resulting from Federally Funded Research [FR Doc. 2020-00689]

Dr. Nichols:

The American Society of Hematology (ASH) appreciates the opportunity to provide comments on the Draft Desirable Characteristics of Repositories to Consider for Managing and Sharing Data Resulting from Federally Funded or Supported Research, as proposed by the Subcommittee on Open Science (SOS) of the National Science and Technology Council's Committee on Science. ASH's journals, *Blood* and *Blood Advances*, currently mandate that datasets be accessible by reviewers and editors at the time of paper submission and must be publicly available as of the date of publication. Given ASH's journal policy with respect to data sharing, we support the effort by the SOS to improve the consistency of information that Federal Agencies provide to scientists on the long-term preservation of data resulting from Federally funded research, along with the effort to improve and support the discoverability, management, and sharing of data.

ASH represents more than 18,000 clinicians and scientists worldwide, who are committed to the study and treatment of blood and blood-related diseases. These disorders encompass malignant hematologic disorders such as leukemia, lymphoma, and multiple myeloma, as well as non-malignant conditions such as sickle cell disease, thalassemia, bone marrow failure, venous thromboembolism, and hemophilia. In addition, hematologists are pioneers in demonstrating the potential of treating various hematologic diseases and continue to be innovators in the field of stem cell biology, regenerative medicine, transfusion medicine, and gene therapy. ASH membership is comprised of basic, translational, and clinical scientists, as well as physicians providing care to patients.

In general, the draft desirable characteristics as proposed in both sections I and II of the proposal are reasonable, and ASH appreciates that "Federal agencies would not plan to use these characteristics to assess, evaluate, or certify the acceptability of a specific data repository" since different public sharing solutions may be needed given the various types of research data. If access to data generated by Federally funded research is done appropriately, it will enhance research transparency and accuracy, as well as foster the

reproducibility and reliability of the data. More importantly, it will provide an opportunity to analyze data in new ways that might further enhance scientific discovery and promote collaborative interactions.

The Society has a few suggestions and questions for OSTP to consider as it drafts its final set of characteristics that Federal funding agencies can use when issuing guidance around data sharing and management. First, the draft characteristics in both Sections I and II currently do not address who is responsible for maintaining, updating and disbursing the data once it is deposited and ensuring that use is compliant with requirements (the government repository, the non-government repository, or the investigator/institution). For example, in the case of restricted use agreements and protected health information, how will Federal agencies monitor whether the investigator/institution complies? Second, the Society is concerned that the RFC did not address where the resources will come from to collate, share, and store all these data in a manner compliant with a new policy. Federal agencies should allow for grant dollars to be used to comply with data sharing and storage policies. Please also consider applying any policy changes to newly funded studies and not to ongoing projects that were not designed to deposit data with a compliant repository and do not have budgets to support this work. Third, it would be helpful to clarify that this repository guidance is intended for deposition of primary data collected by federally funded investigators and does not apply to data collected by others, i.e., an investigator, federally funded or not, should never deposit someone else's data. The Society also offers our comments on the following draft data characteristics:

D. Curation & Quality Assurance: Provides, or has a mechanism for others to provide, expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

- This category would be better titled “Data Quality.” We also recommend deleting “expert” which is vague and adding instead, “... provide, training for data capture methods, data architecture designed to maximize data quality, automated or manual data cleaning, and/or conduct internal data validity evaluation.”

E. Access: Provides broad, equitable, and maximally open access to datasets, as appropriate, consistent with legal and ethical limits required to maintain privacy and confidentiality.

- ASH recommends rephrasing this description as follows: “Provides broad and equitable access to datasets, as appropriate and consistent with legal and ethical limits required to maintain privacy and confidentiality.” We recommend deleting “maximally open access” since access might be determined by considerations beyond legal and ethical limits required to maintain privacy and confidentiality, such as the requestor (e.g., a foreign government), planned use of the data (e.g. scientifically questionable projects), or prioritization (in the case of limited resources). The requirements under *Request Review* suggest that there will be some review of the request before data are released.

F. Free & Easy to Access and Reuse: Makes datasets and their metadata accessible free of charge in a timely manner after submission and with broadest possible terms of reuse or documented as being in the public domain.

- ASH recommends deleting “free of charge” as this prohibits any cost sharing in providing the data. As noted above, ASH's journals, *Blood* and *Blood Advances*, currently mandate that datasets be accessible by reviewers and editors at the time of paper submission and must be publicly available as of the date of publication. We would also like to highlight the need to ensure Federal funding includes costs involved with making datasets and their metadata accessible, if the intent is to make data free of charge to others. Costs may include creation of the dataset, documentation, review and approval procedures, data storage and access, data security procedures and data maintenance. Requiring that datasets be available free of charge shifts substantial costs and responsibilities to the investigators and data depositors rather than the users. Many datasets will not be of interest to other investigators, and ASH hopes that data access plans can be calibrated to the likelihood of use. A model where data repositories can recoup some of their costs from users provides a sustainable model

that unburdens investigators while providing the best practices in repository management in an efficient manner. This is especially important if long-term sustainability is required beyond the funding period.

- ASH recommends deletion of “and with broadest possible terms of reuse or documented as being in the public domain” as repositories may wish to reserve the right to review data requests to make sure they meet certain criteria.
- Who will make sure that analyses are correct, and attribution is given? Disclaimers should be required if the primary team is not involved in the reanalysis.

The Society also has questions and suggestions related to Section II, Additional Considerations for Repositories Storing Human Data, as follows:

A. Fidelity to Consent: Restricts dataset access to appropriate uses consistent with original consent (such as for use only within the context of research on a specific disease or condition).

- Restricting access to appropriate use consistent with original consent is critical. Who bears responsibility to ensure that data requestors’ use of the data is consistent with the original consent? These terms are usually delineated in Data Use Agreements or through IRB review, but it does not appear that these oversight mechanisms will be used with data requestors.

C. Privacy: Implements and provides documentation of security techniques appropriate for human subjects’ data to protect from inappropriate access.

- Protection of clinical trial data as it relates to increasing the inclusion of minority patients in clinical trials is especially important. A conclusion from a recently published paper (PMC2990341) about biospecimen repositories was that “Minority blood donors are less likely to participate in biospecimen repositories than Caucasians, though other variables also influence participation. The reluctance of minority donors to participate in repositories may result in a reduced number of biospecimens available for study and a decreased ability to definitely answer specific research questions in these populations.” Mistrust of data use was discussed in the study as a reason for a lack of participation in biospecimen repositories; similarly, ASH is concerned that this same mistrust could be the reason for low participation rates of minority patients in clinical trials focused on therapies that would benefit patients with sickle cell disease or multiple myeloma. It is not just important that repositories of clinical trial data are protected from unauthorized users (i.e., law enforcement), but equally if not more important that these communities trust that their data are protected. The desirable characteristics might also include a description outlining who authorized and unauthorized users are.

Again, ASH appreciates the opportunity to provide comments on desirable characteristics for managing and sharing data from federally funded or supported research and remains available for consultation as the National Science and Technology Council Subcommittee further refines the characteristics. We also call your attention to [comments](#) submitted on January 15, 2020, in response to the National Institute of Health’s Data Sharing and Management Policy, for further reference. Please use Suzanne Leous, ASH Chief Policy Officer, as your point of contact at [sleous@hematology.org](mailto:sleous@hematology.org) or 202-292-0258, if you require additional information from the Society on this matter.

Sincerely,



Stephanie Lee, MD, MPH  
President