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
Heather Joseph, Executive Director, SPARC: heather@sparcopen.org>

Essential Characteristics for Repositories Managing Research Data Framework

Objective	Essential Characteristics
Discoverability of data	• High quality metadata (discipline-based or general metadata schema (e.g. Datacite or Dublin Core metadata) with an OAI-PMH feed • Repository has well documented APIs • Repository assigns a citable, persistent unique and universal identifier (PUID) that points to the landing page of the dataset[1] (even in cases where data is no longer available or data is not available for security purposes)
Equitable, free and ongoing access to data	• There is no cost to the user for accessing data once it is published • Repository ensures ongoing access to data for a publicly stated time frame • Repository has a contingency plan to ensure data are available and maintained during and after unforeseen events
Reuse of data	• Repository supports the use of machine readable licenses (e.g. Creative Commons Licenses) • Repository provides citable PUIDs[2]
Data integrity and authenticity	• Repository provides information about data provider(s) including contact information of the person(s) responsible for the data. • Repository provides a record of all changes to metadata and data in the repository • Repository provides documentation of its practices that prevent unauthorized access/manipulation of data
Quality assurance	• Repository undertakes basic curation of metadata and data[3] • Repository provides documentation about what curation processes are applied to the data and metadata

[1] Many existing repositories use Handles as persistent identifiers, so these should be admissible
[2] 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.
[3] As defined by the CORE Seal of Approval, basic level of curation involves brief checking and addition of basic metadata or documentation where needed.

Privacy of sensitive data (e.g. human subjects, etc.)	● 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 ● 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
Sustainability and preservation	● Repository (or organization that manages repository) has a long term plan for managing and funding the data repository ● Repository has a public data retention policy that defines the duration of time the data will be preserved and documentation about preservation practices
Other	● Repository has a contact point or helpdesk to assist data depositors and data users ● Repository provides documentation about the scope of data accepted into the repository

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.