UNIVERSITY OF MINNESOTA
Driven to Discover℠

March 17, 2020

Lisa Nichols
Office of Science and Technology Policy
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."[1]

- 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 (PUIDs): We agree that PUIDs 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.[2] Furthermore, a PUID 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.[3]

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 PUID, author, author PUID, author affiliation, author affiliation PUID, title, date published, source repository, source repository PUID, license, license PUID, abstract (of the data, not the related article), related publication, related publication PUID, 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.

---

[1] Research Data Alliance Term Definition Tool https://smw-rda.esc.rzg.mpg.de/index.php?title=Data_Repository

[2] 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.

[3] 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).[4] 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).[5]

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 prefered 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.[6]

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

[4] 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
[5] https://gdpr.eu/
[6] 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.[7] 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

[7] 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.