



March 6th, 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:

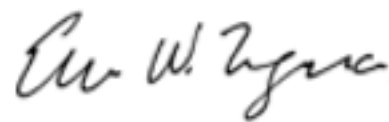
- To “assist investigators in identifying data repositories”, per this CFP, it is important that repositories document their own collection policies, clearly articulating their self-defined scope and use/reuse policies, including: (a) what does and does not meet the repository’s selection or inclusion 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 FAIR¹ 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.

¹ <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) 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.