COGR

Council On Governmental Relations

An Association of Research Institutions

Office of Science and Technology Policy March 17, 2020
Dr. Lisa Nichols
OpenScience@ostp.eop.gov

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/isoiec-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

Wendy D. Streitz, President
Council on Governmental Relations (COGR)
www.cogr.edu