To: Martin Halbert
         2415 Eisenhower Ave.
         Alexandria, VA 22314.

From: Association of Public and Land-grant Universities (APLU): Contact – Kacy Redd, kredd@aplu.org
         Association of American Universities (AAU): Contact – Kate Hudson, kate.hudson@aau.edu
         Council on Governmental Relations (COGR): Contact – Krystal Toups, ktoups@cogr.edu

Date: January 19, 2024

RE: RFI Response: NSF Public Access 2.0

On behalf of our organizations representing the higher education and research community, we write to comment on NSF's Public Access Plan 2.0. We appreciate that NSF has been deeply committed to engaging with the community throughout the development of its public access policies and plans. It was with support from NSF (NSF#1837847 and #1939279) and the National Institutes of Health that APLU and AAU were able to host workshops with researchers, senior research officers, librarians, chief information officers, and organizations supporting increasing public access to research. Those community discussions inform our responses below.

While there are several positive contributions in the Plan that will advance public access to research data and publications, we are concerned by the lack of clarity for sharing scientific data for juried conference proceedings and the definitions of scientific data and metadata implied in NSF's Plan. We thank you in advance for your careful review of our comments, and we urge NSF to improve the Plan accordingly. Please know we are eager to be a resource as the process moves forward. Public access to federally funded research is crucial for rigorous science, discovery, and reproducibility, and research universities are committed to sharing the results of their research whenever appropriate.

Appreciation for NSF's flexibility in where and how to publish and recognition that public access requires investment
We appreciate that NSF's Plan provides flexibility in where researchers publish and that submitting the peer-reviewed and accepted manuscripts is allowed in the Plan. We note that in journals with the broadest reach and impact, investigators might be prohibited from sharing the accepted version by licensing agreements. If the investigators or their institution do not have the funds to cover article processing charges to enable the version of record to be available publicly, they may be barred from publishing in these journals. This may perpetuate inequities in the system. Addressing policies and funding to facilitate

public access, especially for under-resourced researchers and institutions, is essential to democratize knowledge and to ensure equity among researchers in the sharing of science.

We appreciate NSF's acknowledgment that ensuring equity for sharing information will "require significant investments in data cyberinfrastructure and training, particularly for the benefits of public access to reach local, public or marginalized communities." We also support the statement that "this approach also relies on NSF continuing to provide adequate resources to repositories to ensure appropriate data preservation and access over time, per Sec. 3.B.iii of this public access plan, as well as the availability of domain-specific expertise to curate meaningful datasets within scholarly communities." Providing adequate support and resources for public access is critical to ensure researchers from all institutions can participate equitably. Additional unfunded mandates on researchers risk further excluding those with fewer resources.

Significant concern with the expanded definition of scientific data in the Plan
The Glossary of the NSF's Plan defines Scientific Data as:
 "Any electronically stored information, including:
- Observations, calibrations, coefficients, documentation, algorithms and any ancillary information.
- Information needed to validate the scientific conclusions of peer-reviewed publications. This includes data underlying figures, maps and tables.
- Material captured under the heading' research data'"

The inclusion of "observations, calibrations, coefficients, documentation, algorithms and any ancillary information" will significantly impact the burden on the researcher to comply with sharing scientific data. This language goes beyond the requirements of the 2022 OSTP memo and current regulations. Such a new standard for 'scientific data' poses additional burdens on researchers and potentially encompasses data that has not been peer-reviewed. This creates uncertainty for researchers in determining when and what data needs to be shared, which may negatively impact research productivity and exclude smaller institutions with fewer resources to comply. We strongly recommend aligning with OSTP's definition and removing "observations, calibrations, coefficients, documentation, algorithms and any ancillary information" from the final definition of "scientific data" used in the NSF's final public access plan.

We support the text included in Section 3.C. *Researcher Responsibilities for Managing and Sharing Data Arising From NSF Awards* of the *Plan* that requires "Within these limits, beginning in 2025, NSF will require that researchers deposit scientific data — as defined in the 2022 OSTP memo and consistent with the norms in the scientific community of the funded researcher— undergirding peer-reviewed publications deposited in NSF-PAR in data repositories, as noted in Section 3.B.i." Usage of the definition within the 2022 OSTP memo will help harmonize requirements across federal agencies.

Lack of clarity in the definition of metadata and how it is referred to in NSF's Plan
Metadata is commonly referenced in two ways in public access discussions. First, metadata is used to mean data that helps index and identify a data set (e.g., digital object identifiers (DOIs) to identify research data, Open Researcher and Contributor (ORCID) IDs to identify researchers, and funder grant numbers). These persistent identifiers (PIDs) are machine-readable. Additionally, the term metadata is used to encompass all

the information about a data set that makes it reproducible and reusable – things such as methodology and observational variables. This end of the spectrum for many disciplinary fields does not have standards for what should be included or how it should be noted to standards. This would make machine-readability very challenging. These concepts of metadata are, of course, on a continuum. We urge NSF to be consistent about what is meant by metadata within the NSF Plan. Clear and precise definitions are needed to help the research community understand requirements and comply with regulations. We urge NSF to clarify metadata definitions and requirements.

We recommend defining metadata by requiring well-defined PIDs. APLU, the Association of Research Libraries, the California Digital Library, and AAU released a [report](#) in 2019 with recommendations for data practices supporting an open research ecosystem (NSF #1945938). Through those discussions, we came to a consensus on five PIDs that would help ensure that research data is consistent with FAIR principles and are machine-readable. These were:
1. DOIs to identify research data, as well as publications and other outputs;
2. ORCID IDs to identify researchers;
3. Research Organization Registry (ROR) IDs to identify research organization affiliations;
4. Crossref Funder Registry IDs to identifier research funders; and
5. Crossref Grant IDs to identify grants.

Clarity required for including juried conference proceedings in the public access requirement
NSF's Plan expands the current requirement to share peer-reviewed journal articles to include juried conference proceedings. However, requiring data sharing for conference proceedings may discourage researchers from presenting emerging findings at conferences, which has the potential to delay sharing of new knowledge. Conference proceedings, even juried conference proceedings, are more akin to pre-prints than they are to peer-reviewed publications. For these reasons, we recommend clarifying that data sharing is encouraged but not required for juried conference proceedings.

The Plan also indicates that NSF is considering extending public access requirements to *non-juried* conference proceedings. Conferences are a venue to share emergent results to get feedback from colleagues. It is a valuable resource for researchers. Requiring researchers to share their conference presentations more broadly before traditional peer review may create a chilling effect on sharing at conferences. We therefore do not support extending public access requirements to *non-juried* conference proceedings.

Burdensome expansion to require software sharing
The NSF's consideration of extending public access requirements to software, as mentioned in Section 3.A.i of the Plan, merits careful consideration due to potential impacts on researchers and institutions. While the OSTP 2022 memorandum's aim for broader access is laudable, this expansion raises concerns for several reasons. First, sharing software and code varies significantly across academic communities. Open-source and open-access practices flourish in some disciplines, while others lack established standards or even definitions of "software." This heterogeneity complicates any potential one-size-fits-all approach to sharing.

The term "software" could encompass a vast spectrum within this policy, ranging from an Excel formula to complex lab-developed code with embedded proprietary licenses. Defining which and how such diverse research products should be shared requires clear guidance and consideration of different levels of complexity and intellectual property concerns. The community will need more guidance about which and how this heterogenetic class of research product must be shared.

Some lab-developed software might incorporate licensed proprietary packages essential for its functionality. Extracting or sharing such code could violate license agreements and pose significant challenges. Additionally, identifying the provenance of code, especially older codebases vital for ongoing research, can be difficult. Researchers should not be obligated to share software that they do not own.

We would be happy to work with NSF to better understand the challenges and opportunities with sharing software and code and to help advance a culture of software sharing within and across disciplines.

Recommend that institutional repositories be identified as appropriate to house scientific data
Many institutions have made great investments in institutional repositories to archive research data. This resource is often at no-cost or low-cost to the researcher and may be the only repository available to them, if their discipline does not have an appropriate disciplinary repository. We recommend that NSF's Plan specifically identify that these institutional repositories may be appropriate for archiving research data as they do with disciplinary repositories. We also recommend replacing "disciplinary repositories" with "disciplinary or institutional repositories" or "appropriate repositories."

Recommend that NSF encourage the inclusion of adequate funds for public access in research grant funding
Another concern for the research community is the lack of clarity on the ability to use grant funds to comply with public access requirements, both in publication and repositories, and to meet the significant costs associated with data management. While OSTP policy permits "reasonable" publication costs to be direct charged by researchers to grants, this does not reflect the reality of the significant costs now to be shouldered by researchers and their institutions. Greater clarity and guidance are crucially needed as to which costs are considered allowable and which would require support from sources outside of the research grant. For example, guidance is needed for when research outputs exceed the funding allocated in their grant or publication comes after the end of an award due to delays in the peer review and revision process. That said, it is essential that NSF support adequate funding within grants to ensure public access. Insufficient funds could hinder researchers from disseminating results in their preferred journals or repositories, thereby impacting visibility and potentially marginalizing those from emerging research institutions.

To make progress on the overarching goals of public access, we urge NSF to more clearly define allowable costs associated with compliance as well as explore options to 1) permit pre-payment of publication expenses, 2) allow institutions to retain allocated publication funds post-award, and 3) provide supplementary funding for publication costs.