To: OpenScience@ostp.eop.gov
OSTP Chief of Staff, Sean C. Bonyun,
Re: **RFC Response: Desirable Repository Characteristics**

Dear Mr. Bonyun,

The American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America represent more than 8,000 scientists in academia, industry, and government. We support more than 13,500 Certified Crop Advisers (CCA), and more than 700 Certified Professional Soil Scientists (CPSS). We remain fully supportive of open science initiatives that improve the accessibility and transparency of our sciences and thank you for the opportunity to provide comments on data repositories.

Our members are keenly interested in data repositories that ascribe to FAIR (Findability, Accessibility, Interoperability, Reusability) principles, but challenges and overly optimistic promises associated with the build-out of repositories for agricultural data have tempered enthusiasm. The Societies, therefore, urge OSTP to be thoughtful regarding burdens placed on researchers and the unrealized responsibilities of repositories that currently lack capacity and expertise to achieve OSTP's proposed characteristics.

Researchers need federally supported tools and training to accelerate data reporting
Most researchers are not funded, trained, or otherwise incentivized to annotate and organize their data and provide the necessary meta-data in a way that would maintain OSTP's proposed data repository characteristics and meet FAIR principles. Most of the proposed characteristics focus on managing data once it is in a repository, but our researcher's decade of experience has made clear that for data to be deposited, there is a critical need for data tools and workflows that enable researchers to take raw datasets and those in statistical formats and easily assemble them into formats that enable general reuse.

For example, federal agencies should hire specialists to create data extraction and upload wizards for automatic extraction, standardized formatting, and depositing of data directly from research equipment. These data specialists could work with research equipment designers and users to ensure that their products are equipped to deliver collected, calibrated data in FAIR format that can be user-verified and that include metadata on how logged data were verified, processed, and calibrated. The Societies also support a reasonable embargo period for data from uploaded but yet unpublished research (e.g. multi-season studies) so that researchers have time to conduct rigorous statistical analyses and submit manuscripts for peer review and subsequent publication.

Automatic upload of data from devices may afford an easier and more systematic path to data repository compliance. However, there are large amounts of data that cannot be automatically uploaded from most scientific equipment and is, therefore, manually recorded, sometimes with e-tablets and spreadsheets, other times with paper and pencil. We suggest that federal agencies offer training and workflow tools for researchers and students so that they understand ahead of time the data repository requirements and FAIR principles so that these manually recorded data can more easily be transitioned to repositories.

Agriculture and natural resource researchers need a fully-supported data repository

It is not enough to mandate principles for data repositories without fully supporting researcher participation and database functionality. For example, the U.S. Department of Agriculture's Agriculture Research Service (USDA ARS) supports the Ag Data Commons, but this data repository is too small and under-resourced to handle modern agriculture research datasets. The U.S. Department of Energy's Knowledge Discovery Framework (KDF), for example, is only open to data from biofuels research. The National Science Foundation's iPlant, now CyVerse, also has been proposed as an alternative federally-sponsored data repository, but it is not known among the agricultural research community nor has it invested in making its data FAIR, resulting in datasets with opaque identifiers and non-standard formats that are unusable to any but the researchers who deposited them.

Thought must be given to how federal repositories can be structured moving forward so that large and interdisciplinary datasets can be included, and this includes data created in conjunction with the private sector. For example, a member-scientist recently initiated a collaboration with an agricultural consultant who has assembled more than 530 million rows of data in a spreadsheet with nearly 100 geo-referenced traits for each row. No federally supported data repository is equipped or open to receive such a dataset, and no guidelines exist for how this data could be formatted to make it comply with FAIR principles. And yet, datasets resulting from public-private collaborations like this are the future of modern agriculture. Without investments in the work-flow tools that researchers need to get data into these repositories or the incentives to make uploaded data follow FAIR principles, progress will languish.

"Access" alone may not be enough to make data findable and usable

The Societies are concerned that OSTP's proposed characteristics could potentially describe a "dark archive," where the data is there but not discoverable. Repositories must be readily searchable by commonly used search engines and data formats. Data should be linked to the publications, and vice versa. Inclusion of Persistent Unique Identifiers (PUIDs), like a Digital Object Identifier (DOI), is an absolute necessity. Thought also must be given to versioning of data sets so that data accrued in multi-year studies and similar situations can be identified with absolute certainty.

Long-term sustainability and business models for repositories need to be defined

As a public good, the Societies support federal funding of key data repositories to ensure long-term program sustainability. Preservation and curation practices should ascribe to the best management practices, including frequent file back-ups, strategic distribution, and disaster-recovery protocols. Business models should consider triaging data curation according to its use and apparent value, moving less-used and/or limited value datasets to less costly preservation systems. Library scientists, other preservation specialists, and the relevant research communities should work jointly to develop curation guidelines for data.

"Additional Considerations" for data privacy are needed for farm data

Our scientists often depend on data collected on privately owned farmland. The requirement to make all such data public may deter these important studies, research that enables the scaling of research findings. For this reason, perhaps OSTP's proposed "Additional Conditions" should apply to on-farm data as well as human data. "Fidelity to Consent," "Restricted Use Compliant," "Privacy" and OSTP's other proposed characteristics for human data repositories may all apply to many on-farm research datasets and their respective landowners. Also, the Societies suggest that OSTP include “confidentiality” to its list of privacy safeguards (II.C).

Again, we thank OSTP for providing our Societies the opportunity to comment on this important issue. Please feel free to contact me if you have questions.

Sincerely,

Nicholas J. Goeser, CEO
American Society of Agronomy
Crop Science Society of America
Soil Science Society of America