

**Response to the NASA Public Access Plan for Increasing Access to the Results of
NASA-Supported Research**

Dr. Rebecca Ringuette, NASA Goddard Heliophysics Digital Resource Library

Dr. Brian Thomas, NASA Goddard Heliophysics Digital Resource Library

Dr. Robert Candey, NASA Goddard Space Physics Data Facility

Increasing Research Transparency and Findability through an
Infrastructure Designed for Efficient Validation

The overall goal of the NASA Public Access Plan is for the community to be able to build upon previous work for more efficient and capable science, but in an efficient way to avoid overburdening the infrastructure and funding resources. There are restrictions both in policy and technology on what is currently possible and what must be done, but there are some improvements in both the policy and implementation approaches that we provide comment on here.

The **current infrastructure situation is increasingly inefficient**, with publishers attempting to archive data and software, modeling centers attempting to serve data, and researchers resorting to generalist repositories to avoid the mess and the extended timeline needed for metadata curation. **These practices are impeding the stated goals of discoverability, accessibility, and usability, and provide an increasingly insurmountable barrier for validation efforts.** Datasets and software are increasingly hard to find and discover due to use of generalist repositories and the use of publisher-provided mechanisms, making truly connected search interfaces nearly impossible to develop. When datasets and softwares are available, many do not even have DOIs due to the lack of the application of related best practices of those attempting to serve them, short-circuiting efforts to improve citations of and access to these artifacts. Even basic instructions on how to use many artifacts do not exist since curation of these artifacts is not performed by generalist repositories such as Zenodo, and is not performed to the domain-specific standards by the publishers hosting the artifact. Archive websites are difficult to navigate, especially outside of Earth Science, simply because the archives are attempting to design them themselves, when there are a plethora of much more capable and efficient commercial entities to be engaged. Expecting these various entities to apply best practices in the full spectrum of situations and resources is too much to ask, and is resulting in an increasingly unnavigable jungle of resources and waste of funds (e.g. p. 24 of the ISTPNext Workshop Report https://bit.ly/ISTPNext_report). As a result, validation of publications is an increasingly insurmountable task in the current infrastructure landscape. If the goals of the NASA Public Access Plan are to be realized within a reasonable timeframe and the funding available, then **a more coordinated and organized approach to infrastructure is needed.**

In short, **let each infrastructure entity do what they do best, and distribute responsibilities accordingly.** Publishers should provide peer-review services, archives should curate and serve research artifacts (e.g. data and software), modeling centers should install and run models for the community, and researchers should describe their research. Commercial services should be strongly considered for needed capabilities not within the skillset of the entity. In a more efficient infrastructure landscape, publishers provide peer-review services, repositories curate and serve research artifacts to both the community and the publishers as needed, and modeling centers are engaged for questions related to their hosted models and run those models for the community. Publishers and modeling centers would not attempt to curate and serve data and software, and archives would not attempt to perform complex model runs for the community. **A more efficient infrastructure is one in which crossover of responsibility is minimized and collaboration is coordinated, resulting in a more capable infrastructure to support validation efforts.**

Incorporating validation into the peer-review process can be streamlined in such an infrastructure as follows. The researcher collects all the research artifacts they think is needed for someone else to repeat the work, such as data and software, according to general rubrics provided by a joint effort between the publication journal and domain-specific repositories. They submit these items to the selected domain-specific repository in parallel with the submission of the publication to the publisher. The author works with the repository to complete metadata curation tasks as the publisher performs their tasks. When temporary links to the artifacts are available, they are shared privately with the publisher to be shared with the peer-reviewer(s). The first main stage of the peer review process after the initial checks are completed is for the reviewer to reasonably reproduce the work using these artifacts. Notes from the peer-reviewer describing how the work was validated are added to the artifacts' metadata. Once the reviewer successfully validates the publication, they then review the remaining portions of the publication as usual (e.g. text clarifications and other items). Once both portions of the peer-review process are completed, the publisher waits for the repository to notify them of successful completion of metadata curation in accordance with the repository requirements and any imposed by the peer-reviewer before accepting the publication. In this process, the researcher and the peer-reviewer can be in close contact while the work is being validated, and notes on how to reproduce the publication are included in the metadata, increasing its usefulness and trustworthiness. Incorporation of validation into the peer-review process in such a manner creates useful collaborations between infrastructure entities such as domain-specific repositories and publishers, distributes tasks according to demonstrated skill, directly connects a publication to all relevant artifacts, and improves the quality, usefulness, and trustworthiness of our publications.

We can imagine several improvements to this workflow and infrastructure in general as technology advances. Such improvements could include support of a shared cloud environment for the peer-reviewer and researcher to validate the work, which could then be linked from the publication for public use as restrictions allow, using AI technology to aid in metadata creation and re-curation, creating a searchable network of modeling centers hosted by the modeling community with models installed in the cloud and made executable from a notebook, and improvements in repository websites and search technologies. Peer-reviewers could also be recognized for their validation efforts by being included in the artifact metadata or even the publication as a contributor.

We ask for the agencies, including NASA, to coordinate validation efforts and distribute infrastructure responsibilities in such a cost-effective and distributed fashion to take advantage of the demonstrated experience across our resources, to increase our efficiency, and to keep costs down as prioritized in the Public Access Plan and related policies. Intelligent coordination and collaboration will decrease our costs, but must be led and encouraged by the agencies, particularly the needed redistribution of responsibility between publishers, archives, and modeling centers.

An important caveat to such a redistribution is that the domain-specific **repositories** chosen to host NASA-funded research artifacts **must have demonstrated ability** to create DOIs or the

most relevant persistent identifiers for their hosted research artifacts, provide a searchable interface for and immediate access to those artifacts, and preserve similar artifacts. This request builds upon the requirement for repositories to align with the Desirable Characteristics of Data Repositories by extending its application to all repositories hosting research artifacts, including software, and additionally requiring a search interface where users can search for hosted artifacts either by using the artifact's persistent identifier or with other keywords from its metadata. Ideally, the repository would also provide machine search and access to its hosted artifacts. We note that generalist repositories do not align with the Desirable Characteristics of Data Repositories policy since they do not provide curation, and so should be excluded from hosting NASA-funded research artifacts. A lack of such features in a repository already hosting NASA-funded artifacts would be considered a motivator for a funding request to add the needed capabilities, given a plan to efficiently use the funds and assuming those hosted resources fall within the repository's demonstrated expertise, or a phasing out of the hosting responsibilities of the artifact type(s) in question and transfer of those artifacts to a more appropriate repository. We also note that such requirements do not exclude new repositories or related resources from forming, but simply require that new entities demonstrate capability and permanence before being designated as an acceptable repository for archiving NASA-funded research artifacts.

A middle ground in the generalist repository discussion could be that generalist repositories could be used *if* done in coordination with a domain-specific repository so that metadata curation still occurs. However, the domain-specific repository needs to be the long-term destination of research artifacts to preserve those artifacts, including the related DOIs, and avoid preventable loss of resources. This is particularly relevant to the current trend for software archiving. We are submitting a separate RFI focused on infrastructure needed for software archiving, and leave further thoughts on that topic for that response.

In the meantime, validation studies in each science directorate can be funded through the ROSES open science AOs to provide initial groundwork for validation efforts, ranging from validation studies limited to the same domain as the publication for the most basic needs to similar studies performed by research groups in different science directorates than the publication topic for multi-disciplinary uses. Long-term studies of the technology needed for validation of example publication types (e.g. workflows requiring large datasets) should also be performed to understand what reasonable goals should be in long-term preservation efforts. We expect such studies to be useful to both the publishers and archives as we work towards understanding how to validate research, and can be patterned off of replicability efforts in other sciences led by the Center for Open Science.

As a final note on this topic, the policy uses the words validation and reproduction interchangeably, which is incorrect. **The steps and conclusions of a given publication can be validated, or confirmed to be reasonable, without completely reproducing every step.** In many cases, such as for publications with complex modeling efforts, completely reproducing a given publication is not feasible in a reasonable time frame and with a reasonable effort. However, requiring that all publications be validated through a peer-review process is a vital step towards increasing our research efficiency and building trust in our science. As implied

earlier, validation will require a larger breadth of research artifacts to be archived, including all analysis codes and a number of datasets, both intermediate and fully processed, which will add cost to the archiving effort. However, if a peer-reviewer can reproduce enough of the publication to validate it and leave notes for others on that process, then the community will have a higher trust in that result and can more easily build upon that work. We agree on the desired effect, but **reproducibility as a goal is not feasible or necessary. Validation is.** Please correct the wording.

Thank you for your attention.