

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:10:54

PAGE 1

REFERENCE NO: 257

This contribution was submitted to the National Science Foundation as part of the NSF CI 2030 planning activity through an NSF Request for Information, [https://www.nsf.gov/publications/pub\\_summ.jsp?ods\\_key=nsf17031](https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17031). Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

## Author Names & Affiliations

- Victoria Stodden - University of Illinois at Urbana Champaign

## Contact Email Address (for NSF use only)

(Hidden)

## Research Domain, discipline, and sub-discipline

Cyberinfrastructure and Data Science

## Title of Submission

CI is now central to NSF funded discoveries and would benefit from a specialized CI review process

## Abstract (maximum ~200 words).

The term Cyberinfrastructure can now be understood very broadly, since computational tools and infrastructure underlie nearly all scientific discoveries today and CI has evolved to be central (and crucial) to the scientific process. I suggest that this centrality be formally recognized within NSF by establishing a review process for CI projects that is customized to CI. Today CI projects are often reviewed as using research criteria, and budgets often support the two types of project without distinction. Review criteria for CI-oriented proposals should be different than for research proposals, for example including a recognition of the unique needs of the community they serve. CI proposals can end up competing within the same pool of funds as research proposals, to the detriment of both. The recent reduction of the MREFC eligibility threshold is a strong step to resolve these issues, and a community-driven reconsideration of appropriate review criteria would enable such projects to be better evaluated and better support scientific discovery at all levels of funding. It is important to note that CI is not a concept that uniquely attends to large scale or even mid-scale projects. It is just as important for enabling reliable and efficient discovery even at very small scales. The unique position of the Office of Advanced Cyberinfrastructure as serving all NSF directorates makes it well positioned to coordinate efforts toward the recognition of the importance of CI in discovery.

**Question 1** Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

The fundamental CI challenge facing computational research is the capture of relevant experimental details that can be stored, shared, and disseminated to the research community. This would enable computational reproducibility of published findings, allowing researchers to verify claims and extend and build on them. Today, such infrastructure generally does not exist, although some researchers are taking self-directed steps to do this in their research, for example using enterprise software tools such as docker or website services such as Github. These tools may be part of the solution but neither were developed with the needs of the scientific community in mind so there is an

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:10:54

PAGE 2

REFERENCE NO: 257

opportunity to meet the special needs of our community with our own CI investments. Additionally, many questions about CI-enabled discovery and how best to carry it out in the research context, are open. Sharing private data for example, testing codes that underly findings, developing standards for code, data, and workflows, and many others, are questions the research community is working to resolve today and are crucial to the development of a robust CI supporting scientific discovery.

**Question 2** Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

The typical computational research project uses a variety of disconnected tools to derive claims and findings: ranging from email and dropbox, to advanced computing resources at TACC and the University of Illinois at Urbana Champaign, for example. There is no streamlined way to capture all the computational steps that were taken in the course of generating findings. Additionally there is no commonly accepted way to disseminate these steps, even if they are captured. Repositories are emerging but a concerted effort to meet the repository needs for the different domains and communities is important for preservation of the scholarly record, broad and open access to scientific findings, and the independent reproducibility/verification of the findings themselves. At this point many researchers want to know \*how\* to enable computational reproducibility in their work, and \*where\* to share the results. We need the development of "experiment definition environments" that exist as cyberinfrastructure and provide an interface to compute and other resources used in the discovery process, without compromising the flexibility needed for creative scientific research. One could imagine, for example, a small set-aside per published article that permits a third party to certify the computational aspects of the claims in the article. Such a system would permit standards and tools to develop for relevant groups of users around particular problems and workflows.

**Question 3** Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Providing CI as a fundamental part of the research process is essential, and essential to that is the development of proposal review criteria specific to CI. There are also cultural challenges as the use of new tools can imply a change in existing research practices. Ensure a smooth and seamless transition - ease of use of CI - is essential. It is also important to recognize researcher efforts toward reproducibility that use of these tools can imply, and reward researchers taking these steps.

Enabling reproducibility in computationally-driven research is not obvious - different types of problems and research have different needs. Research is needed in the \*science of reproducibility\* to better understand the challenges facing the development of scientific CI and allow for effective usable solutions. It is important to involve all relevant stakeholders in these efforts, and important to ensure that scientific products, for example code, data, or other artifacts such as workflows, are accessible and usable by the research community (broadly understood) to facilitate research and cross-pollination of ideas and methodologies in research.

NSF is not well-positioned to take on long-term sustainability commitments of CI, however this gap needs to be filled in a responsible way. Funding for sustainability can come from a variety of different sources (for example institutions, external third parties) but the scientific community is best positioned to provide oversight of the development and maintenance of the scientific CI system and its products.

## Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:10:54

REFERENCE NO: 257

PAGE 3

---

---