

Reference ID: 11226548477\_Jennewein

---

**Reference ID:** 11226548477\_Jennewein

**Submission Date and Time:** 12/16/2019 1:36:54 PM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. 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>).

*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 publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

**Consent answer:** I consent to NSF's use and display of the submitted information.

### **Author Names & Affiliations**

Submitting author: Douglas Jennewein - Arizona State University

**Additional authors:** Christopher Kurtz, ASU; Gil Speyer, ASU

**Contact Email Address** (for NSF use only): (hidden)

### **Research domain(s), discipline(s)/sub-discipline(s)**

Campus Cyberinfrastructure; Research Software Engineering; Systems Architecture

### **Title of Response**

Broadening campus support for data-enabled discovery

### **Abstract**

Arizona State University's response to this NSF Request for Information describes grand challenges in the areas of data acquisition, long term data storage, and the federal funding landscape. We recommend new advancements in storage technology such as non-POIX storage systems; industry partnerships, leveraging existing expertise is big data stewardship and analysis; and new funding

programs, complementing efforts with CC\*, DIBBS/CSSI, and MRI. Finally, we discuss the additional need for workforce development through data science training and education, and expanded Cyberinfrastructure facilitation roles on each campus.

**Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s).** Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

Computational approaches for research are broadly applied today due to their effectiveness and the widespread availability of advanced computing resources for analysis. With the surge of data-driven approaches to complement classical simulation, the need for accessible data resources is clear. Many of Arizona State University's (ASU) faculty conduct computational work exclusively, and most faculty use computational resources frequently, with campus cyberinfrastructure a necessary and enabling element of their success. Challenge 1: Acquisition. However, the available data is growing more quickly than storage capacity, while campus technology budgets are stagnant, resulting in prohibitively complicated data movement solutions, often with data being moved by hand with physical media such as USB drives. This complexity is compounded when data is located in multiple places. The data acquisition challenges we see at ASU mirror those found around the country: 1) Data that is acquired remotely from multiple repositories or instrumentation; and 2) Data that is generated onsite by simulations or acquired by campus instrumentation. Both must be transferred to ASU systems for analysis, publication, and archival. This requires sufficient network and data cyberinfrastructure to acquire, store, and archive the data, as well as the computational capacity to analyze it in a timely manner. Challenge 2: Long-term storage as a service. Compared to commercial IT, the research IT market is vanishingly small, with vendor solutions often not well aligned with our needs. Having teams of engineers to "roll our own" infrastructure is not feasible across the range of campuses needed to affect real national change. Though recent NSF efforts such as the Open Storage Network are a step in the right direction, there is no XSEDE- or NERSC-like entity for general purpose research storage across the lifetime of a project. There is not yet a system that has done for data management what systems like Globus and iRODS have done for data movement. The GridFTP protocol and the concept of rules-oriented object storage predate both technologies, but Globus and iRODS have made them vastly more usable and ubiquitous. Additionally, research data is increasingly being hosted "free" in cloud providers (such as Amazon's AWS) with users "encouraged" (forced by necessity) to use those cloud providers, which is expensive, especially long-term. Additionally many institutions subsidize capital purchases but not cloud (OpEx vs CapEx). Challenge 3: Funding. We have found that current external funding opportunities are no longer well aligned with the pace and nature of cyberinfrastructure needed at ASU, especially in the areas of data storage, archival, and stewardship; data acquisition, staging, and movement; and in-network computing. Specifically, the Major Research Instrumentation (MRI) program's limited submission mechanism and the Campus Cyberinfrastructure (CC\*) program's long standing focus on networking and some aspects of computing make them poorly suited to fund data-specific infrastructure. By developing a funding

mechanism for campus storage and data management, whether through existing programs like CC\* or DIBBS/CSSI or through a new initiative, NSF will be able to do for data what it has successfully done for advanced networking. Many challenges such a funding program could address are disciplinary-agnostic, including: data classification and qualification; authentication and federation; data literacy; and data movement.

**Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s).** Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

Recent NSF investments in the Big Data Innovation Hubs, FABRIC, the Open Storage Network, and Open Science Grid are exploring how some of these challenges can be addressed by: 1) scaling capacity and accessibility of storage and high speed networks (FABRIC, OSN), 2) scaling and distributing expertise (Big Data Innovation Hubs), and 3) scaling out distributed computing capacity (OSG). However, other approaches are necessary as storage and network capacity have always lagged behind data growth. Many approaches have been explored to meet the challenge of increasing data volume, velocity, and variety, including 1) traditional data movement to position data adjacent to advanced computing systems; 2) containerization allowing more portable computation; 3) employing in-network computing to analyze data in motion; and 4) automated data qualification, classification, and organization to sift out only actionable data. **Advancements Needed:** 1) Non-posix storage solutions Today's hierarchical POSIX filesystems (first developed in the 1980s) are becoming less suited to storing large data sets. Development of non-POSIX, non-filesystem storage methodologies will be necessary to scale data sets to Exabytes and beyond, with metadata that is not bound to a particular storage mechanism. 2) Industry partnerships As NSF has done with efforts like Exploring Clouds for the Acceleration of Science, NSF should explore industry partnerships to inform and develop a national cyberinfrastructure for big data. From social networking to finance to commerce, industry is managing enormous amounts of data already. While with these partnerships come issues of data privacy and security, ECAS has successfully demonstrated the effectiveness of leveraging industry expertise and capabilities in the cloud computing space. Future NSF efforts around data-driven computation and discovery should do the same. Rather than invest in the development of entirely new Cyberinfrastructure, NSF can leverage significant industry experience and expertise in managing large amounts of data and in "storage as a service" partnerships. 3) Campus infrastructure support In addition to architectural efforts and industry partnerships, the NSF funding landscape itself must evolve to better develop and enable data-centric research, establishing data infrastructure as a first class citizen in the campus Cyberinfrastructure

ecosystem. Consistent and significant emphasis and investment from NSF through the Campus Cyberinfrastructure (CC\*) program has transformed the campus discussion about advanced networking, giving concepts like the Science DMZ architecture traction across campus technology organizations. Similarly, recent CC\* solicitations have included a computing track, making funding transformational campus computing feasible outside the realm of the often more restrictive Major Research Instrumentation (MRI) program. However, aside from a single data storage track in the CC\* program in 2016, there has been no similar effort to enable campus data Cyberinfrastructure. Existing NSF efforts towards funding data infrastructure such as the DIBBS and CSSI programs have been too few to be effective across the breadth of campuses that CC\* has been able to affect.

**Question 3 (maximum 300 words) – Other considerations.** Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

It is important to note that most commonly accepted definitions of Cyberinfrastructure include people. There is a broadening talent gap characterized by a shortage of data scientists to analyze data compounded by a lack of expertise in how to move and manage it in a secure and timely manner. These gaps will be bridged only with a renewed focus on workforce development through training and education. Expanded campus Cyberinfrastructure meeting the needs of today's data-enabled science, engineering, and health research also requires the development of a new class of computational and data scientists. A recent update from the National Strategic Computing Initiative noted that the need for a larger workforce of interdisciplinary cyberinfrastructure practitioners is dire, with deficits not only in higher education but high school and middle school as well. This will also require the adoption or expansion of an emerging "bridging" role at each campus: that of the Cyberinfrastructure facilitator. Unfortunately, even on campuses where these positions exist, they often lack a well-defined and stable career path. However, drawing upon initiatives like its Data Science for Undergraduates program NSF could develop new funding programs to develop not only Data Science majors, minors, and faculty, but also a broad basic understanding of Data Science throughout academia. Similarly, building upon the success of ACI-REF and the Campus Champions, NSF could fund workshops and co-located events to stimulate the development of a culture of Data Science and engineering awareness within existing communities of practice.

-- End Submission --