

Reference ID: 11226934943_Foster

Reference ID: 11226934943_Foster

Submission Date and Time: 12/16/2019 3:59:18 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: Ian Foster - University of Chicago

Additional authors: Rachana Ananthakrishnan (University of Chicago); Kyle Chard (University of Chicago), Vas Vasiliadis (University of Chicago)

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

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

Computer Science

Title of Response

Utility Data Services are a Prerequisite for Accelerated Discovery and Innovation

Abstract

Essentially all research today is data-intensive. However, the research data ecosystem remains primitive, with most researchers either lacking access to effective data management capabilities or being constrained by unscalable, costly, non-interoperable, and hard to maintain custom stovepipes. The

resulting cost to science is large in terms of both wasted effort and missed opportunities. To overcome this problem, NSF should establish a robust, sustainable foundation of broadly accessible and interoperable utility data services. These services should allow any and every researcher to find and access data regardless of location; share data securely and reliably across institutional silos; automate data manipulation tasks; and build discipline- and project-specific higher-value capabilities. They must have a sustainability model that can allow researchers, institutions, and projects to count on their persistence into the future. The feasibility and value of such foundational services is demonstrated by, for example, Globus, DataONE, and Figshare, each of which operates data services that have spurred innovation across US research institutions; the not-for-profit Globus and the for-profit Figshare have also demonstrated sustainability. The opportunity and challenge is now to extend these proven models to encompass more of the research data life cycle, from data generation to preservation.

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.

A primary objective of the recently updated National Strategic Computing Initiative is to “develop, broaden, and advance the Nation’s computational infrastructure and ecosystem” [1]. The scale of data-intensive research that will be enabled by future advances in the nation's computing ecosystem will make robust data management capabilities critical to realizing the full value of such advances. With exceptions (e.g., Globus), today's data management solutions are barely equipped to deal with current data volumes, and will almost certainly be incapable of supporting computation at exascale and beyond. A related set of challenges derive from advances in data-intensive instruments such as genome sequencers, cryo electron microscopes, and synchrotron light sources. Facilities and investigators face many hurdles in enabling effective use of the resulting data streams, due, in large part, to the lack of reliable infrastructure for describing, cataloging, and distributing experiment data reliably and securely. Further exacerbating the problem are proprietary and inflexible interfaces and storage systems on the instruments themselves, which make it difficult to build robust, repeatable data management solutions. Effective data-intensive research thus requires that investigators be able to collect, organize, analyze, and preserve the data that they generate--and then find, access, make sense of, and compute on both their own data and data generated by others. They must be able to perform these tasks easily, reliably, securely, automatically, and scalably. These foundational data management capabilities are, we believe, table stakes for all research endeavors. Meeting these needs requires solutions to the following cross-cutting challenges:

- Universal data access: Ability to efficiently access and operate on data irrespective of the underlying storage system implementation, interface, access protocol, security model, location, and network connectivity.
- Stovepipe identity and access management environments: While NSF-funded efforts (e.g., InCommon) provide a global trust fabric, they have not been universally adopted. A similar global authorization fabric, built on standard libraries (e.g., Globus Auth), is needed to enable interoperability between services.
- Duplicated components: The lack of easily reusable capabilities

result in an ecosystem of duplicated functionality (e.g., user and group management, search infrastructure, allocation management, notification services) with significant limitations (e.g., security and scalability concerns). - Scalable data pipelines: Increasingly complex research pipelines comprise distributed instruments, storage systems, computers, and researchers at multiple institutions. In order to keep pace with increasing data size and complexity, and to avoid human bottlenecks, new ways of automating all aspects of the data lifecycle are needed. [1] <https://www.whitehouse.gov/wp-content/uploads/2019/11/National-Strategic-Computing-Initiative-Update-2019.pdf>

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?

Few investigators have access to effective data solutions today. A major reason for this unfortunate situation is that, largely lacking foundational data services, individuals and projects are forced repeatedly to develop custom stovepipes, which are expensive to create and maintain, and furthermore tend to wither as funding streams end. The key to improving this situation is to embrace a utility model of data service delivery, to proactively invest in new data service utilities, and (see Q3 below) promote funding models that incentivize establishment and operation of data services.

a. Data management services as a utility Recent advances in cloud computing enable data management services to be operated as hybrid utilities, with management logic running on cloud computers; lightweight software agents deployed near storage, computers, or instruments that are to be managed; and investigators using simple APIs to interact with the resulting utility services. This approach has allowed Globus, for example, to deliver widely used data transfer and sharing services to more than 200,000 registered users at most US research institutions. The utility approach, when properly executed, can deliver powerful capabilities to large numbers of users at modest cost, due to economies of scale in service delivery. But few such utility services are available today. The challenge and opportunity is to deliver a full suite of data management services to every researcher as a utility. Much as many institutions provide investigators with compute and storage resources today, they should also provide access to data management services that allow researchers to make effective use of those compute and storage resources. We expand below on what capabilities we think those data management services should provide.

b. Critical data services In particular, we see the need to invest in the following areas, to ensure that they do not create friction or stand in the way of researchers doing their work (at scale):

- Data orchestration: secure, efficient movement of data among systems and institutions; broader access to shared data sets with appropriate security controls; easy-to-implement

automated data flows that allow an investigator to scale their work at the flip of a switch; full visibility into all of the above to ensure provenance and facilitate reproducibility. - Data discovery: indexing of metadata (currently stored in bespoke representations and conventions in files, file names, and associated files); enabling wide-area and collaborative discovery and management of data, irrespective of where it is stored, how it was created, or who it is used by. - Security and privacy: providing additional protections to maintain privacy, prevent disclosure of proprietary secrets, and ensure compliance with a growing list of laws and standards, as research data increasingly include sensitive elements. - Ubiquitous connectivity: uniform, intuitive interfaces to hybrid (cloud and on-premise) storage resources, as well as integration with storage management systems, given the growing adoption of cloud services by research institutions. Utility services should all support APIs, so as to also address another need highlighted in [1]: “the development and refinement of scientific gateways, portals, and associated workflow tools to enable more efficient approaches to solving challenging scientific and technical problems”. Development of such solutions has typically required specialized skills available only within a few large, and well-funded projects. A platform with well-supported APIs and toolkits will make it trivial to incorporate advanced data management functionality in bespoke solutions. Examples of services that offer open APIs include Globus, TACC Cloud API System (Tapis), DataONE, and Figshare.

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.

We focus our remarks here on the twin challenges of adoption and sustainability. While utility data services such as Globus, Figshare, and DataONE have been proven effective, there are many other data capabilities for which no service providers exist. This relative lack of effective data service utilities in contemporary academia represents a market failure: there is much need for such services, but few service providers, due to a lack of means of compensating service providers. NSF can overcome this problem by allowing investigators to request funding for the use of utility data services within research grants. The precise mechanism requires further investigation, but most importantly it must encourage funds to flow to services that are proven to be effective. Alternatively, the NSF could establish long-running programs to support system-wide data management, akin to efforts like XSEDE that provides system-wide access to computing CI. NSF can also promote data management utilities within research institutions. We observe widely differing opinions across institutions as to the importance of research data management infrastructure, as evidenced by the degree of institutional support provided for such infrastructure. That is, while few likely disagree with the importance of data management infrastructure in principle, we observe that, in practice, other non-research-related infrastructure needs (e.g., learning systems, applications for managing administrative operations) often seem to be prioritized for funding. The NSF has an important opportunity to promote leadership in this area via well-designed funding programs, much as it did for research networks in the past. For example, they might provide funding models that reward use, support education and workforce development programs to train the next generation of scientists to rely on such approaches, and discourage bespoke data service development.

Response to NSF 20-015, *Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research*

Reference ID: 11226934943_Foster

-- End Submission --