NATIONAL SCIENCE FOUNDATION
2415 EISENHOWER AVENUE
ALEXANDRIA, VIRGINIA 22314

NSF 20-015

Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research

October 22, 2019

Dear Colleagues:

Twenty-first century science and engineering (S&E) research is being transformed by the increasing availability, diversity and scales of computation and data. NSF recently outlined a *holistic vision for an agile, integrated, robust, trustworthy and sustainable cyberinfrastructure (CI) ecosystem to drive new thinking and transformative discoveries in all areas of S&E research and education.* This vision was developed following extensive inputs from the community and encompasses the spectrum of CI resources, capabilities and services; the need for continuous innovation in CI and the translation of this innovation; the need for close collaboration with the S&E community to couple cycles of innovation across disciplines; the need to increasingly focus on CI usability and on CI training and workforce development; and the need to balance continuity and stability with approaches that promptly address new challenges and opportunities in an era of disruptive technologies and changing S&E needs.

In particular, NSF notes that digital data is playing an increasingly central role in all areas of S&E research, resulting, in part, from the dramatic growth in the scale and complexity of a variety of digital sources, from experimental and observational instruments, to computation and simulation, to discipline-specific digital repositories. Consequently, fostering S&E-driven, robust, reliable and scalable data-focused CI is a key component of NSF's vision referenced above. Such CI must be designed to flexibly accommodate both existing and future disciplinary and multi-disciplinary data sources and to provide essential capabilities and services that enable new and evolving integrative and cross-disciplinary S&E efforts to translate data from those sources to knowledge and discovery.

This Request for Information (RFI) invites the community to provide input to NSF on specific data-intensive S&E research questions and challenges and the essential data-related CI services and capabilities needed to publish, discover, transport, manage and process data in

secure, performant and scalable ways to enable that data-intensive research. Recognizing that data-oriented CI and services exist in many S&E disciplinary domains, NSF is particularly interested in understanding how broader **cross-disciplinary and domain-agnostic solutions** can be devised and implemented, along with the structural, functional and performance characteristics such cross-disciplinary solutions must possess. Such new CI services and capabilities should allow for seamless data integration and interoperability; support existing S&E drivers, users and usage modes; and foster the initiation of future modes of discovery. While no one technical solution will likely be able address the expansive S&E research enterprise that NSF supports, NSF is interested in understanding how different data-related CI solutions might support heterogenous ensembles of data-intensive disciplines - owing, for instance, to common requirements due to similarities in data set sizes, types and utilization workflows, or to novel shared goals for cross-disciplinary data integration and discovery. Note that NSF is especially interested in responses that build on existing and future data sources (including repositories) and address services for publishing, discovery, access, management and processing of the data.

The community input received from this RFI will inform refinement of NSF's CI investment strategy and planning of future NSF funding opportunities. Responses from individuals, organizations, as well as groups or collaborative networks are welcome.

INSTRUCTIONS TO SUBMITTERS

NSF invites both individuals and groups of individuals to provide their inputs via the online submission form (link below). The submission form requires the following information[1]:

- Contact person name and affiliation.
- Valid contact email address.
- Additional author name(s) and affiliation(s).
- Research domain(s), discipline(s)/sub-discipline(s) of the author(s).
- Title of the response
- Abstract (maximum 200 words) summarizing the response.
- 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.*
- 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?

- 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.*
- Checkbox to consent to NSF's use and display of the submitted information, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode). *NSF anticipates making submissions publicly accessible through a website*[2].

To respond to this RFI, please use the official submission form available at **https://www.surveymonkey.com/r/NSFDataCIRFI**.

Submission deadline. Contributions must be received on or before 5:00 PM Eastern time on December 16, 2019.

For questions concerning this RFI and submission of input, please contact Stefan Robila, Amy Walton and William Miller, NSF Office of Advanced Cyberinfrastructure, nsfdatacirfi@nsf.gov.

Sincerely,

Erwin Gianchandani
Acting Assistant Director
Computer and Information Science and Engineering

Manish Parashar
Office Director
Office of Advanced Cyberinfrastructure

---

[1] The valid OMB control number for this collection is 3145-0215. The time required to complete this information collection is estimated to be approximately 30 minutes per response.

[2] Submissions are expected to be professional in tone and address subject matter

relevant to this RFI. NSF reserves the right to not post or otherwise not consider any response not meeting this expectation.