

Reference ID: 11225724551_Hemphill

Reference ID: 11225724551_Hemphill

Submission Date and Time: 12/16/2019 9:10:55 AM

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: Libby Hemphill - Inter-university Consortium for Political and Social Research, University of Michigan

Additional authors: on behalf of ICPSR

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

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

data curation; data reuse; information science; applied machine learning

Title of Response

Elements of transformative cyberinfrastructure to support groundbreaking, transparent, and reproducible science

Abstract

Improving shared human and technical resources that facilitate data documentation, sharing, dissemination, and preservation will dramatically impact research by reducing redundant effort, making

data and metadata more broadly accessible, and making science more easily reproducible. Challenges around data sharing, standards, nondesigned data, and computational overhead can be addressed by supporting archives in developing software and expertise to support researchers at each stage of the data lifecycle.

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.

We have identified five pressing challenges for data-intensive research: Data sharing and access—PIs often don't think about sharing their data until late in the research cycle; they lack funding for preparation, storage, and dissemination; PIs aren't trained to prepare data for reuse and leaving the burden on them is inefficient and means best practices in data curation are not followed. For example, PIs may provide the tabular data they used in a specific regression analysis but do not explain missing values or label values in categorical variables. They may also not be aware of archives that can preserve their data in perpetuity and manage access to those resources, freeing PIs from building websites (that may not be maintained in perpetuity) and fielding requests. Bespoke standards and data storage—much of the data used for research has multiple incompatible standards and lacks interoperability; neither robust, widely used standards across the research data lifecycle nor accessible tools for making diverse types of data interoperable yet exist for the social sciences. Duplicated efforts in nondesigned data use—researchers duplicate one another's efforts to collect and manage nondesigned data such as administrative data and social media data. For instance, because social media platforms restrict both access and sharing, individual researchers must collect and store data separately when a single shared source would probably be better. It would reduce computational and storage costs and facilitate access; the computational overhead needed to collect, manage, and analyze nondesigned data creates inequalities of access. Computational overhead—working with data at scale requires computational skills and resources that are not evenly or widely distributed. People with strong skills in data analysis may not have skills in acquisition or maintenance (e.g., server administration), and asking them to develop those skills or manage the infrastructure is an inefficient use of their time and expertise. Collaboration across institutions—research institutions increasingly provide research computing to their employees, and many have secured access to cloud services through agreements that are between a single institution and a cloud provider. These services, whether on-premises or in the cloud, are often governed by policies, whether in contracts or institutional policies, that restrict access to employees of the institution. These restrictions, and the costs associated with the computation mentioned above, make it hard to establish and maintain resources that are accessible by collaborators from more than one institution.

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?

To address the urgent need for new modes of access, confidentiality protection, methodological approaches, and tools so that research using a variety of data types meets accepted scientific standards, we suggest NSF support the development of three essential data-related CI capabilities: Expand existing archive capabilities to include the entire research lifecycle: registries of documentation related to research design, hypotheses, data management, and confidentiality protection. A framework for describing different lifecycle events in research management (e.g., research registry documents, analysis code associated with a specific research publication) would address the challenge of bespoke standards and data sharing. Archives such as ICPSR that currently host data could expand their infrastructure to accommodate documentation and research software code to connect these resources with the data they leverage and produce. The system should also enable researchers to send their materials to journals, sponsors, and archives to achieve transparency, reproducibility, and preservation goals. Build software to facilitate harmonizing data and generating appropriate metadata to assure findability, accessibility, interoperability, and reusability (FAIRness) of research data. We recognize that researchers are often not well-equipped to prepare their data to meet FAIR data standards. NSF could encourage and financially support researchers' partnerships with archives to achieve FAIRness. New developments in automated curation software could also facilitate the curation and sharing of high quality, discoverable, and reusable data and significantly reduce the costs of preparing data and metadata. Software could employ a human-in-the-loop approach to guide researchers through common harmonization and documentation steps, training researchers in best practices in data management in the process. Develop cloud-based platforms for analyzing confidential or large, complex, social science data that leverage advances in collaborative scientific computing and security. We propose a suite of computational templates that are designed to support common research approaches using shared datasets. These templates reduce the computational and system administration overhead that accompanies data-intensive research. Templates will include specifications for both hardware and software for various analysis activities. Hardware specifications will vary from bare-bones single-server configurations to high-performing server clusters designed to meet advanced computing needs. Users will be able to indicate which datasets they want to use, what kinds of analysis they want to conduct, and which statistical software they wish to use, and the system will generate a computing environment that contains the necessary hardware and software. Such a system also allows the data to stay in a single location and facilitates monitoring and auditing of access, thus reducing the risk of leak or breach.

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.

While technical infrastructure will be key to addressing these challenges, the human aspects of cyberinfrastructure will also require attention. For instance, data-intensive research generates increasing demand for professionals such as data curators, privacy experts, and data security managers. The computational and ethical landscapes around privacy are evolving, and both individual researchers and collaborative research efforts need support to meet their goals while protecting individuals from harm. Consortium funding models like ICPSR's are a means to build sustainable CI that includes technical and human resources while ensuring shared risks, benefits, and governance. Institution-level commitments made through consortia also help ensure that developments in one area (e.g. social sciences) are leveraged and connected to others (e.g., physics, geography).

-- End Submission --