

Reference ID: 11227078986\_Elbert

---

**Reference ID:** 11227078986\_Elbert

**Submission Date and Time:** 12/16/2019 4:57:57 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: David Elbert - Johns Hopkins University

**Additional authors:** Tyrel McQueen, Johns Hopkins University

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

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

Materials Science and Engineering (NSF-DMR)

#### **Title of Response**

A Role for Centers of Excellence in Development and Delivery of Data Focused Cyberinfrastructure

#### **Abstract**

User-centric, efficient, flexible cyberinfrastructure is required for broad adoption and maximal impact of high-value data across the scientific enterprise. A focus on human information, expertise, and services is central to success; there must be bridges between data science, implementation, and domain expertise to build truly seamless data interoperability and the skill to apply it. Development of Centers of

Excellence for data-driven science and related cyberinfrastructure development can be an important part of the strategic vision for data-focused cyberinfrastructure.

**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.

Fundamental challenges in fields as diverse as materials science, turbulence, astronomy, oncology, and life sciences are marked by a convergence around data. To meet these science needs, CI faces its own challenge to connect complex services consistently and sustainably, and to provide services in forms readily consumed and utilized by a broad swath of scientific disciplines. CI development needs to embrace the integration of many, varied datasets and resources to enable faster and more transparent use of existing data sources, and create value added aggregations. A focus on data context with self-documenting data models allows a scalable, adaptable architecture that can provide end-to-end value. Although data-driven science is in principle domain agnostic, some fields have already advanced significantly further than others. Astronomy, for example, stands out for developing agreement on data description and format; sharing of data models and tools; large public datasets; and NSF sponsored development of a collaborative computational and data management research environment (SciServer, an NSF DIBB). As other fields adapt the astronomy model for accessible and interoperable data, applications of data-intensive techniques are expanding. Relative to astronomy, many communities are still in need of their data “watershed” moment – where the value add of data clearly enables science not possible via older approaches. This is fundamentally a culture and people challenge: how does one seed a sufficient embrace of data to engender these defining successes?

**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?

Of the many lessons learned in the last decade, two of the most important are 1) bringing small, seemingly unrelated data together creates new value, and 2) simple interfaces are more powerful than complex ones. A user-centric, efficient, and flexible cyberinfrastructure (CI) ecosystem depends on

human information, expertise, and services for maximal productivity. An effective CI spanning disciplines with domain independent tools and services requires particular attention to bridging data science, implementation, and domain experts. Seamless data interoperability can fuel creative collaboration and empower transformative science, but it requires a holistic, nimble CI able to support unexpected findings and approaches. CI development in the next decade should include a focus on: 1. building a shared, connectable data ecosystem to make full federation a reality 2. adopting automated FAIR data 3. producing sustainable repositories and associated facilities 4. an end-to-end data stewardship philosophy to maximize the return on investment in data. An important part of achieving these goals can be centers of excellence that combine technical expertise with the structures and people needed to bridge the data and domain expertise gap. Such centers can act like inverted user facilities where data experts bridge outwards to the domain users to accelerate projects and application. The location of the data and the computation will certainly change in the coming decades, but the need for seamless connection, simple interoperability, and people to bridge the domains will remain vital to success.

**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 need a model for data and CI sustainability that is consistent with the goals enumerated in our response to Question 2. The cost of curating data and associated infrastructure is estimated to be small compared to the price of producing the data or the value derived from the data; at the same time, the additional burden of curation of data is often perceived as a productivity sink that is not worth the small marginal time and money investment. Data success stories and community networks can make inroads into this latter problem, however, a long-term, stable model for high value data still needs investment. A possibility is to combine private endowment with a government component, much like the archival model of the Smithsonian Institutions. Education and workforce development also will be central to the success of data-driven research in any domain. To aid those efforts, community structures with open development of standards and practices should be incentivized. Centers of excellence for data-intensive CI should also provide services for curricular advance. An additional focus on training and collaboration with midcareer scientists can be especially effective at accelerating adoption of data-driven science.

-- End Submission --