

Reference ID: 11227045917\_Elbert

---

**Reference ID:** 11227045917\_Elbert

**Submission Date and Time:** 12/16/2019 4:44:01 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 - Platform for the Accelerated Realization, Analysis, and Discovery of Interface Materials (PARADIM), a NSF Materials Innovation Platform (Cornell-Johns Hopkins-Clark Atlanta Collaboration)

**Additional authors:** Tyrel McQueen, Platform for the Accelerated Realization, Analysis, and Discovery of Interface Materials (PARADIM), a NSF Materials Innovation Platform (Cornell-Johns Hopkins-Clark Atlanta Collaboration)

**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**

Interdisciplinary Data-Focused Cyberinfrastructure Needs and Opportunities In Materials Science and Engineering

## Abstract

Centers of data production, including mid-scale infrastructure like MIPS, should play a central role in the integrated development and deployment of data-centered cyberinfrastructure. Such facilities touch the entire materials workflow providing opportunities to create the tight, functional connections required for seamless adoption of data-centric research in a complex, interdisciplinary domain. A national vision for domain-agnostic cyberinfrastructure will be best served by focusing on bridging infrastructure components and data science expertise through collaboration, co-development, and high quality training across application domains.

**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.

The importance of advancing data infrastructure arises from the confluence of a revolution in our ability to capture, store, and process data with the transformative new ways science is proceeding by focusing on data. The materials domain provides a particularly exciting opportunity for this Fourth Paradigm of Science and Engineering. Development and deployment of novel materials spans disciplines, has a profound impact on sectors of broad societal importance, and provides a crucible for creation of cyberinfrastructure advances needed across the research landscape. Materials research encompasses condensed matter physics, chemistry, optoelectronics, polymer science, metallurgy, and more. Required experimental data spans imaging, spectroscopy, and physical property measurements used in every NSF Directorate. Modeling relies on computational tools from atomistics to hydrocodes. The field is on the cusp of disruptive change in the design of novel quantum materials, catalysts, superconductors, optoelectronics, and polymers, with associated transformative applications for society in the energy, transportation, security, computation, cybersecurity and related sectors. Specific challenges and drivers include:

1. Breakthroughs in discovery and deployment of new materials: Materials research is evolving from serendipitous discovery to an iterative, Materials-By-Design loop incorporating multiscale characterization, theory-based computational prediction, and synthetic control of material structure and related properties. Data-centric approaches, especially machine learning (ML), promise accelerated design and discovery for every major science driver in materials science and engineering.
2. Understanding synthesis-structure-property relationships Establishing the connections between synthesis and processing, micro-/nano-structures, and emergent physical properties is a grand challenge. The development of new synthetic techniques and additive manufacturing opens up novel possibilities for controlling structure and properties. Developing data-driven approaches will enable a more complete understanding of these fundamental relationships and lead to the creation of models that enable tailoring material functionality. Such work will place high-value on developing understandable, invertible ML models as we proceed along the path to put every atom in its place, scalably.
3. Creating Seamless, Interoperable Data Structures and Infrastructure: The materials

domain includes big, highly varied, complex data produced in centralized and decentralized fashions. Materials research workflows include complex steps with myriad hidden variables that directly impact outcomes. Developing and integrating functional, implementable semantic frameworks for capturing all relevant information for instrumental and human workflows is a central challenge to realizing the potential for data-intensive science in the materials domain. This challenge more than any other highlights the importance of attending to human networking, collaboration, and training as an integral part of cyberinfrastructure development.

**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?

Reports from TMS (2017), the National Academies (2018), NASA (2018) and DOE (2019) affirm the importance and need for open science and data in the materials domain just as in other disciplines. Indeed, the arguably greatest, through-going challenge across disciplines is development of a comprehensive, sustainable, nimble, and seamless cyberinfrastructure (CI) encompassing computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people. NSF and other agencies have made critical investments in high-quality, high-volume experimental data production through user facilities and instrumentation. NSF has made similarly important investments in CI in the form of computational resources, software development, and high-speed networking all of which have made major impacts across disciplines. Despite these advances, however, data resources remain largely sequestered at the facilities or left to individual investigators who focus on the originally planned use of the information to solve the problem that drove its collection. The cost of advanced facilities and investigator time make data our most under-utilized resource. Materials discovery and development has an inherently complex workflow with end-to-end challenges. The high-volume, heterogeneous data encountered in a workflow connects physical objects (samples) with activities that modify them or create new objects (synthesis, processing, characterization, and computation). To meet these challenges, we need: CI services and capabilities that describe and work with data across the workflow; integrated FAIR data across the entire workflow, and consistent machine-based approaches for speed. That does not mean we need a single semantic description, but we do need commonalities in data description that allows mapping or translation so we can automate discovery and use of data across the domain. In a domain marked by such varied tools and practitioners a principle for broad success is seamless integration of data streams with shared, transferable services. This enables efficient, impactful education advances and workforce development.

Such an approach will be most successful by focusing on necessary but not excess metadata and APIs. Adopting a minimum-viable-product approach as thin layers that are easy to connect has been shown in other fields to be most efficient; we should avoid building more impressive platforms that fail to be widely adopted. To create tools and services with broad impact we need to connect people at the beginning of the projects rather than try to promote created tools later. Community standards and goals lead to thin, nimble development that can be adopted easily and quickly by any platform. Indeed, creation of community structures and bridge people are critical to bake-in connections across the data ecosystem. An example of such connection is shared development of data platforms across multiple mid-scale or large-scale facilities. Our platform, the PARADIM Materials Innovation Platform (MIP) has begun working with the 2DCC MIP at Penn State to connect our data visions and implementations. The immediate impact of such a relationship is a better infrastructure for our users. A more important, long-term impact of the relationship is deeper thinking about data use and delivery creation of shared data for long-term use by investigators beyond our immediate users. We suggest extending this kind of thinking to link different types of facilities by sharing data expertise and infrastructure development across experimental, computational, and synthesis/processing centers. This type of designed resource sharing during development is envisioned as the most effective way to build common access for data discovery, publishing, and interoperability. Our users move between these facilities so our infrastructure development should also move between them. As noted earlier, materials data problems are a particularly good place to invest because we span disciplines and types of facilities making developed tools inherently domain agnostic.

**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.

Community buy-in to shared resources and sustainability remain the biggest challenges for CI and facilities. The value of materials data requires appropriate mechanisms for crediting data and data analysis sources, while the inherent long-term value of data demands development of sustainable models for repositories, shared APIs, and semantic standards. CI development should, therefore, engage the community, encourage development of tools and services that allow partnerships with archival specialists, and encourage adoption of agreed upon certification and management frameworks (e.g. CoreTrustSeal). A CI made up of trustworthy repositories and open tools in a distributed, connected network model provides the most sustainable model for the tools and services created. The 2019 TMS Report on the Materials Genome Initiative Workforce is comprehensive view of learning and workforce development directions in the materials domain. The report recommendations suggest materials research data production sites (large- and mid-level facilities) partner with industry, National Labs, and academia to produce and deliver CI centered workshops, short courses, curricular innovation, and hackathons. Such partnerships put training directly in the loop with CI development which refines the scope and applicability of the CI itself while also cultivating deeper understanding in the expanding workforce and engendering community buy-in. Ideally, CI development should be coupled to curricular integration in undergraduate and graduate education. NSF sponsored university user facilities can help

lead curricular innovation at their home institutions. They are already experienced in delivery of summer programs, but CI development should expand that role to empower the democratization of science that can be one of the most exciting outcomes of a holistic CI and broader data sharing. With open resources, students and investigators from any institution can participate in modern, data-driven research. Summer institutes focusing on professional development of faculty at non R-1 institutions can create meaningful ways to generate broader impacts.

-- End Submission --