

Reference ID: 11227038153_Crespi

Reference ID: 11227038153_Crespi

Submission Date and Time: 12/16/2019 4:40:49 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: Vincent Crespi - Penn State University

Additional authors: Co-author: Joan Redwing, Penn State; Venkat Gopalan and Kevin Dressler from Penn State also contributed to the discussion leading to this document.

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

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

Materials Research; Chemistry; Physics

Title of Response

Cyberinfrastructure for Capture and Correlation of Diverse Heterogeneous Materials Research Data

Abstract

Whereas astronomers share a single sky and life scientists a nominally unitary genome with base-pair (and other) variations, materials researchers work with hundreds of thousands of samples each with a unique synthetic history and its own nanoscale and mesoscale heterogeneities that often dominate the

resulting properties. New methods of data-intensive curation and analysis aligned to these heterogeneity challenges are needed to ask questions of this potentially transformative emerging data landscape. The new algorithms so developed will have potential application across diverse interdisciplinary fields that confront similar challenges of data heterogeneity, including subfields of life sciences, chemistry, and engineering.

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.

Materials research centers and facilities such as MRSECs, MIPs, etc. produce a large volume and variety of data from wide-ranging experimental and computational activities associated with materials synthesis, processing, characterization, theory and modeling. A key aspect of this data is its heterogeneity: sample quality is everything, so two samples of nominally the same material are actually quite different, and the available characterization data on each is necessarily incomplete, and incomplete in different ways. New multimodal measurement techniques such as 4D STEM and ultrafast spatiotemporal datasets vastly expand the data landscape, yet within the same highly heterogeneous landscape. Compounding these issues, most of this data is currently stored by individual investigators and institutions using a variety of formats and venues and thus is largely inaccessible to the broader community except through the limited window of traditional publications and point-to-point collaborations. A treasure trove of materials data thus exists in forms that are incompletely annotated, not easily accessible, difficult to locate, and currently impossible to correlate. Enabling and empowering researchers to organize and share their “hidden” data with the broader community in a useable form offers the potential to dramatically accelerate research productivity and make possible the asking of new research questions about materials data “at-scale” with full account of the diverse synthesis pathways that result in unique sample-by-sample material outcomes.

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?

These highly heterogeneous datasets require new approaches to collate and correlate data across different samples, different measurement modalities, different degrees of coverage, and different sample histories. Community buy-in to ensure cultural changes in research practice to facilitate ingestion of data from a diverse and broad user community presents additional opportunities and challenges. Community-wide data models must align to norms of behavior while also drawing the next generation of researchers to ask new questions and develop new data-intensive tools that exploit these opportunities. A new type of digital repository is required that can capture and curate diverse and heterogeneous experimental data, integrated with associated supportive theory and modeling (on complex energy surfaces, modes of spatiotemporal optical response, etc.), with modest and targeted human curation. Software tools are needed to search for useful patterns in these rich but heterogeneous datasets where the very object of study varies from sample to sample. Finally, an on-line collaborative research environment for data-driven science is needed to host the digital repositories, datasets and tools.

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.

Human resource development and community building will be key parts of attaining these goals.

-- End Submission --