

Reference ID: 11226658347_Voorhees

Reference ID: 11226658347_Voorhees

Submission Date and Time: 12/16/2019 2:16:27 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: Peter Voorhees - MaRDaC Co-Chair, Northwestern University

Additional authors: Bill Mahoney, ASM International; Eric Toberer, Colorado School of Mines; Ian Foster, University of Chicago; Gil Gallegos, New Mexico Highlands University; Jeffrey Rickman, Lehigh University; K. Rajan, State University of New York at Buffalo; Apurva Met

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

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

Materials Science and Engineering

Title of Response

Towards an Interoperable, FAIR Compliant, Sustainable Materials Data Ecosystem

Abstract

We live in an era of "abundant data" of such value that data has been called the "new oil". In the materials domain, much of our data remains dispersed and isolated and significant CI advances are

needed to leverage our data for major scientific and engineering challenges. Creation, adaptation, and improvement of materials data repositories are essential to establishing a sustainable CI ecosystem for materials data with a focus on interoperability and computable representations of experimental and computational data. Implementing FAIR (Findable, Accessible, Interoperable, Reusable) data principles will enable the materials community to capitalize on the nation's investment in producing materials data and the great promise of artificial-intelligence-directed materials discovery. If this effort is successful, the result will be the ability to produce novel materials that address society's greatest challenges and to deploy them at a greatly reduced cost. This response is from the Materials Research Data Council (MaRDaC), a committee with broad representation in the materials community.

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.

Data-driven research has the unique ability to cut across science and engineering, especially in the areas of chemistry and chemical engineering, and materials science and engineering, leveraging advances in statistics and computer science. Significant opportunity exists in materials research for the curation of data and associated tools, application of emerging methods in machine learning and artificial intelligence, leading to accelerated materials design, deployment and integration into manufacturing processes. These efforts have the potential to translate directly into industrial job creation and to reduce or quantify the materials-related risks that new startup companies face when developing new technologies. Several reports have highlighted the importance of open and secure cyberinfrastructure in enabling data-driven acceleration in materials science and engineering: 2018 National Academies report *Open Science by Design*, NASA's 2018 report on *Vision 2040*, 2019 DOE Basic Research Needs for *Scientific Machine Learning Core Technologies for Artificial Intelligence*. These reports emphasize the critical role of high-quality, interoperable data in fueling the next revolution in energy sciences including development and application of novel materials. The key challenges in data-based approaches for materials science and engineering are:

1. The ability to rapidly and easily obtain data from controlled experimental and simulation conditions with documentation of known bounds on statistical variations;
2. Curation of large amounts of sparse, isolated data as well as annotation and management of increasingly large datasets generated by state-of-the-art imaging devices and computations;
3. Creation of sufficiently consistent, hierarchical and flexible metadata standards across the spectrum of highly heterogeneous materials data distributed over diverse properties;
4. Creation of common formats and interoperability among the parallel cyberinfrastructure developments in multiple, geographically distributed centers;
5. Reconciliation of well curated data from different sources;
6. Balance of the cost of security and/or IP against the benefit generated from wider availability of data;
7. Integration of physics-based approaches with machine learning and data-based approaches;
8. The need to have both education and research in balance for the infrastructure to be of value to the community;
9. The need to address the incentives and social

engineering of data contributions from a broad spectrum of PIs, including credit attribution and neutral database management. 10. Harnessing the power of AI-driven laboratories to explore large search spaces of potentially promising materials in ways that contribute to a broader pool of knowledge. 11.

To train materials scientists on the proper use of these databases and to provide background on machine learning methodologies.

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?

The Materials Genome Initiative has been central in establishing a number of disparate data resources and computational tools across the materials domain. Many resources are focused on computational materials science, where simulations provide predictions of specific material properties covering a broad range of possible design spaces. Some resources encompass long tail experimental data where the uniqueness of every experiment and heterogeneity of data creates challenges to data curation and robust standardized vocabularies. Computational tools that use these data, such as AI-guided materials discovery, are also being developed. Despite these important first steps, the vast majority of materials data remains isolated and outside of shareable resources, and interoperability and reuse of materials data has not yet been achieved. The huge value in sharing the vast amount of orphaned materials data is an enormous lost opportunity for the nation. Now is an opportune time to draw lessons from successful efforts within and outside our field to enhance and expand the CI supporting materials research that will usher in a new era where the materials required to tackle society's greatest challenges are discovered and fielded at an unprecedented pace. To ensure that the investment in the MGI achieves its full potential, it is essential to provide resources to create robust and easily accessible platforms, to support ongoing evolution and maintenance of the frameworks, and to create tools that enable data interoperability and widespread reuse. No single CI system can be developed to hold and provide ready access to the wealth of materials data, and therefore an array of federated, interoperable systems is the optimal goal for a materials data ecosystem. To realize the overarching goal of a robust, interoperable materials data ecosystem, the CI needs to enable creation and implementation of high-level common vocabularies and metadata standards, including standards for appropriate statistical representation of uncertainty, as well as robust APIs for capturing data flow provenance and analysis. At the same time, developed CI must emphasize usability design principles to minimize barriers to broad adoption. Tools that enable simple and automated data capture and curation must be developed across the spectrum, spanning data acquisition, computable representation of experiments, models, and

algorithms. Development of incentives for robust data archiving in the established repository ecosystem can be a successful approach. Data security and IP concerns must be addressed in a uniform manner. Creation, adaptation, and improvement of materials data repositories are essential to establishing a sustainable CI ecosystem for materials data with a focus on interoperability and computable representations of experimental data, and computational and theoretical models. This foundation will provide a rich substrate for the materials community writ large to contribute innovative tools and datasets for research and industrial design. The ready access to large quantities of data under specific conditions enabled by the interconnected materials CI will enable applications of machine learning via accurate interpolation of data in hyper-dimensional surfaces. Implementing FAIR (Findable, Accessible, Interoperable, Reusable) data principles will enable the materials community to capitalize on the nation's investment in producing materials data and the great promise of AI directed materials discovery. The net result will be the ability to produce novel materials that address society's greatest challenges and to deploy them at a greatly reduced cost.

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 live in an era of “abundant data” of such value that data has been called the “new oil”. In the materials domain, much of our data remains dispersed and isolated and significant CI advances are needed to leverage our data for major scientific and engineering challenges. NSF and DOE sit at the center of materials data production, strategically positioned to drive the goal of maximizing data value. We identify three important foci:

- Changing the culture. Data sharing and publication following best practices needs to change from being the exception to the norm. This requires both CI and a change in culture. The community is rapidly coalescing into a grassroots network of data resource stakeholders to develop and adopt data standards and tools. Functional CI development should align with these community developments through partnerships and programs.
- Sustainability. It remains a challenge to support CI resources for research data in the long-term, beyond the initial grants used to create them. Increased investment at the federal level to maximize the impact of materials CI infrastructure should be similar to investment in large-scale user facilities.
- Workforce development/learning. Realizing the potential of data science in materials science and engineering requires educating new generations of scientists and engineers with a working knowledge of data science and re-training our existing workforce. NSF should encourage the creation and sharing of learning modules and workshop materials for students and working professionals with an emphasis on democratizing access to data and facilitating access to the breadth of materials CI resources. NSF is uniquely positioned to construct a solicitation that incorporates the grand vision of a holistic CI for materials science and engineering and moves the research community and industry towards an interoperable, FAIR compliant, sustainable materials data ecosystem.

-- End Submission --

Response to NSF 20-015, *Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research*

Reference ID: «Respondent_ID»_«Primary_Last»
