

Reference ID: 11224742879\_Hersam

---

**Reference ID:** 11224742879\_Hersam

**Submission Date and Time:** 12/16/2019 12:24:56 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: Mark Hersam - Northwestern University

**Additional authors:** James Rondinelli, Northwestern University

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

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

Materials Science and Engineering

#### **Title of Response**

CyberInfrastructure Recommendations from the Northwestern University MRSEC

#### **Abstract**

CyberInfrastructure (CI) has the potential to positively impact materials research, especially in multi-investigator efforts such as the Northwestern University Materials Research Science and Engineering Center (MRSEC). A key challenge to successful implementation of CI in materials research is engagement of suitably interdisciplinary teams that not only include materials scientists but also

statisticians and computer scientists. In addition, materials preparation, characterization, and properties all have endless conditions, combinations, and permutations, which present further challenges for constructing uniform data ontologies. Since the accepted source of reliable data is in the published literature, journals will need to allow data to be extracted and incorporated into standardized databases. Once databases are populated with reliable data, researchers will then need to be incentivized to develop data tools and codes since this type of work generally does not result in the high-impact publications that lead to career advancement. Strategic investments by the federal government that create suitable incentives thus have the potential to accelerate the development of CI by leading researchers in the materials field.

**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.

Current and emerging data-driven methods include natural language processing and Gaussian process Bayesian optimization aimed at delivering intelligent synthesis, which goes beyond standard design of experiment. Materials predictions are ever more common, and increased emphasis on laboratory realization is necessary for transformative impact. This effort will involve a convergence of expertise (domain and non-domain scientists). At the same time, there is an increase in “abuse” of data-driven methods (especially machine learning). Sometimes researchers use these methods as a numerical fitting tool, and quite often they will be able to find a suitable working model. Peer-reviewed journals cannot always prevent these studies from being spread to the research community since papers submitted to materials science journals will typically be reviewed by materials scientists as opposed to statisticians and computer scientists. Therefore, cross-disciplinary reviewers are needed. There should be increased emphasis on applicability/generalizability of data-driven models -- whether a new model/method works or not depends on its applicability to solve real-world problems. Molecular mechanics is currently experiencing a renaissance because new, high-accuracy interatomic potentials can be created using machine learning techniques (e.g., neural networks, kernel ridge regression, etc.) from large data sets of first principles data. This renaissance has not only been spurred by development of machine learning techniques but also the computational affordability of the thousands of required first principles calculations. With accurate interatomic potentials, molecular dynamics can be reliably used to study chemical reactions with quantitative accuracy, whereas qualitative results were the previous norm in the field. For reproducibility, thousands of atomic structures with energies and forces need to be shared. For example, projects like Openbabel, ASE, and Pymatgen allow for easy operability of atomic data formats. The other component is how to share interatomic potentials where there may be hundreds of coefficients. The openKIM project aims to achieve this by sharing them as implementations in C++, C, and Fortran modules so that they can be seamlessly swapped. The next step is creating a large database to permit sharing of large first principles datasets for training data. This

database would need a large initial investment in funds and effort to accommodate the data set sizes, number of formats, and the number of different first principles methods used in their creation.

**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?

Materials research is a difficult field for data reuse. Materials preparation, characterization, and properties all have endless conditions, combinations, and permutations, which makes them difficult to construct data ontologies. Additionally, the cost of reproducing many physical samples is low enough that there is little incentive for data reuse. These difficulties have slowed the creation and use of large materials databases and might be insurmountable. Since the field follows publications as the accepted authority on materials data, automating the extraction of data from publications may be the best solution. Modern natural language processing is likely the answer to this problem and has been pursued by a few investigators in the US, but is highly limited and presents challenges in specificity. For example, extracting melting points is relatively straightforward, but ensuring that these data are all at a certain pressure or with the same processing history can be difficult. The other major challenge is coordinating with journals to permit extraction of data that is usually forbidden by their terms of service. Journals will need to standardize how they will permit natural language processing of articles for this method to become widely adopted and consistent. Similar issues with experimental data continue to pose a barrier to collaboration between experiment and theory/computation groups. There is still no standardized (cross-journal) way to submit/retrieve/aggregate large datasets. There are plenty of existing databases whose creators would have valuable insight into creating this kind of CI, meaning a competition or initiative like MGI would likely make headway. Wiley is pushing in this direction with their new platform Manuscripts (<https://www.manuscripts.io/>) that allows you to collaborate in writing a manuscript but also include mathematical equations, code, and data and execute figures live with a behind-the-scenes Jupyter computational notebook integration. In practice, CI has typically been built in two ways. The first is a small project, which through a series of custodianship, grows to wider use. These small projects often struggle to expand beyond a single research group. If it makes it past this barrier, the next hurdle occurs when the project requires more funding than the original research group can spare from their normal activities. From here, the difficulty in finding project-specific funding is the barrier to becoming a major part of the CI of a field. Incentives for PhD students to pursue this work are low, thus necessitating a dedicated pool of experts. The second way of building CI is to secure a large amount of initial funds specific to the goal. In this context, the Materials Project is an excellent example.

Developing tools or code to make a better workflow takes away from effort on publications and funding, which are the main incentives in academia. Funding large projects with strong leadership may circumvent this issue. Another solution would be to incentivize CI development at a more local level. This funding model has been successfully implemented in the freelance programming markets.

**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.

There is typically a large gap between techniques used in academia and industrial applications. It would be helpful to encourage more communication between researchers in academia and industry (e.g., through co-op programs or internships). Since a growing number of students are seeking non-academic routes (e.g., industry, government, etc.), practical work experience during the PhD training would improve career prospects. For example, a growing number of materials research trainees are interested in supply chain optimization (LLamasoft) and data science (Facebook and Amazon) positions, because the learned quantitative skills and hierarchical materials systems level approach to problem solving can be leveraged in these fields. With all scientific software projects (databases included), the ingredients for success are the same as with open source software development: 1. Continuous funding since most fields do not reward software with high profile publications; 2. Leadership driving long term developments targets and enforcing code quality and documentation standards; 3. Community building and feedback otherwise the project fades into obscurity while not meeting the changing needs of the field.

-- End Submission --