

Reference ID: 11225723903_Maginn

Reference ID: 11225723903_Maginn

Submission Date and Time: 12/16/2019 9:10:40 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: Edward Maginn - University of Notre Dame

Additional authors: None

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

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

chemical engineering; thermodynamics; molecular dynamics; Monte Carlo; molecular simulation

Title of Response

Obtaining Data in Chemical Sciences

Abstract

How can people in the chemical sciences work with data scientists to retrieve the large amounts of legacy data that are "out there"?

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.

One of the big challenges people in my community (chemical sciences) face is that there is a huge amount of data out there in the form of databases, recent publications (that are available electronically) and older publications. How do you extract that information and get it into useful forms? How do you assess the quality of it? How do you handle missing or incomplete data? My sense is that some of the problems seem "mundane" to computer/data scientists, but they are absolutely crucial to enabling advances in machine learning and simulation for materials discovery, and my community often does not have the background to solve these problems.

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?

Much of the data I referred to in Question 1 is contained in databases (Beilstein, NIST, DIPPR, as well as corporate). Most are proprietary or pseudo-proprietary, and each has its own format and structure. Thus access to the data is a huge issue and even if one has access, custom tools are needed to extract the proper information. The same problems apply when looking to extract data from technical publications. Format, naming conventions, etc. vary widely. So access to data, developing tools that can extract the data and then (potentially) publishing the data in ways that researchers can use it is a multi-faceted problem that needs to be addressed. I am especially worried about making the data available - who "owns" it and can it be made freely available?

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.

-- 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»
