

Reference ID: 11226500100_Berman

Reference ID: 11226500100_Berman

Submission Date and Time: 12/16/2019 1:19:29 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: Helen Berman - Rutgers University

Additional authors: None

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

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

Structural Bioinformatics

Title of Response

The importance of standards development

Abstract

It should be common practice to archive data that are the basis for scientific publications. To do this well, data standards are required as are the existence of repositories for data.

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.

I am currently involved in two projects that require integration of data across disciplines. The first project involves archiving integrative structural models where data come a variety of methods including x-ray crystallography, NMR 3DEM, FRET, mass spectroscopy etc. Computational methods are used to determine the structures of large macromolecular machines. In order to make the data available to the broader community a pipeline is being created to collect, validate, and archive these data. In a second project we are tackling the grand challenge of creating a spatial temporal model of an entire human cell. Facile data exchange is key to the success of this effort.

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 problems that we are addressing would be vastly simplified if there were well developed standards for all the methods we are using. Well defined standards would make the job of data exchange far simpler. The existence of well run repositories for data would also make the job much easier.

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.

The mind set of the community needs to change so that that data archiving is a prerequisite for publication. This has been true of the structural biology community but not others. It should also be common practice for workflows to be captured so that experiments could be reproduced

-- End Submission --