

Reference ID: 11225929074_Blatecky

Reference ID: 11225929074_Blatecky

Submission Date and Time: 12/16/2019 10:17:58 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: Alan Blatecky - RTI International

Additional authors: None

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

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

Data, HPC, Networking, Cyberinfrastructure

Title of Response

Data Harmonization

Abstract

While most of the current CI efforts on supporting the use of data is focused on accessibility (including authorization) and interoperability, the much larger problem of having data that has undergone some sort of curation and begins to address harmonization issues has been largely ignored. What is required, is to use support and develop AI approaches to deal with messy data, data cleaning, data curation and

harmonization. Use of deep learning, machine learning, neural nets, GANs, and knowledge networks is the only practical way to achieve the scale and accuracy to address data curation and harmonization issues.

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.

While S&E research is being transformed by the increasing availability and scalability of computation and data, the rate of transformation of science and new discovery is being seriously hampered because of the huge difficulties associated with actually using data from multiple sources and disciplines. While most of the current CI efforts on supporting the use of data is focused on accessibility (including authorization) and interoperability, the much larger problem of having data that has undergone some sort of curation and begins to address harmonization issues has been largely ignored. What is especially ironic, is that the promise of big data is predicated on the ability of a researcher to use data from multiple sources; however, if the data is not adequately curated and harmonized, the data is useless and the promise remains hollow. For example, NIH has already generated, and will continue to generate, enormous amounts of data from each of the institutes. Because the data has not been adequately curated or harmonized, the data collections are not being fully used by researchers in related or distant fields to explore entirely new areas of research and discovery. The same can be said for data collections at DOE and NSF; the use of the collections is stymied because of the enormous amount of effort that must be done to make the data useful to support other research questions and approaches. Unfortunately, in spite of advances in technology, the problem is getting worse as the volume of data being generated is outstripping the research community's ability curate and harmonize this "new" data. Likewise, little is being done to curate and harmonize data that has already been collected which seriously limits its use for longitudinal studies or tracking change over time.

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?

At the heart of data-to-discovery and re-use of data is the ability to compare specific data in one data set with specific data in other sets. That is, the specific data being used in each collection needs have enough in common or understood enough that the data can be compared to other data. For example, if collection 1 has temperature data that has been collected by satellite and collection 2 has data that comes from sensors on ocean buoys, the two types of data can only be used for analysis if the two data points can be harmonized in some way. This may be through a set of simple error bars, use of time stamps, comparison to a pre-determined standard, or perhaps require details on the methods or specific instruments being used so that a number can be imputed for each data point so it can be used across multiple data sets. Without an understanding of the data itself including relationship(s) of the specific data points to other data being evaluated, resulting models, simulations or predictions will be invalid. This applies to all types of data including genomic, blood pressure, climate, environment, etc. While researchers routinely do this by hand for small data sets (they may use a short java or perl script to do this transformation or imputation), this approach does not scale for data collections that run into thousands, millions or billions of data. The level of funding required to do this with boutique tools and algorithms is not available; worse, even if the funding were available, the length time required to do the work makes it a non-starter. What is required, is to use support and develop AI approaches to deal with messy data, data cleaning, data curation and harmonization. Use of deep learning, machine learning, neural nets, GANs, and knowledge networks is the only practical way to achieve the scale and accuracy to address data curation and harmonization. The growth of new data is being driven by a range of advanced technologies and already outstrips the ability of those generating the data to deal with harmonization issues unless they also use AI technologies to clean and process the data in near-real time so that it can be widely and effectively used. For large existing data collections, use of these AI data technologies is most likely the only way that continued storage of the data can be justified. Lastly, it is also important to note that the focus of this effort cannot just be theoretical research, but it needs to be grounded in cyberinfrastructure as it has to also provide a service that is sustainable and production-oriented. It should be noted that this cannot be done by AI experts, but must be done by a multi-disciplinary team including domain experts (for each data collection being considered), statisticians, (to deal with error associated with each imputation as well as total error when the collections are used together), data scientists as well as computer science expertise. Until data collections have data that can be readily harmonized so that they can be combined with other data, the collections will continue to be only be useful sometime in the future and new science and discovery will be limited. Lastly, it is also important to note that the focus of this effort cannot just be theoretical research, but it needs to be grounded in cyberinfrastructure as it has to also provide a service that is sustainable and production-oriented. Unless these data collections have data that can be readily harmonized so that they can be combined with other data, the data collections will continue to be independent silos will have limited use or research value.

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.

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: 11225929074_Blatecky

Note, this approach assumes that efforts to support FAIR (findable, accessible, interoperable, reusable) principles is still required and critical.

-- End Submission --