

Reference ID: 11227080534\_de La Beaujardiere

---

**Reference ID:** 11227080534\_de La Beaujardiere

**Submission Date and Time:** 12/16/2019 4:58:36 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: Jeff de La Beaujardiere - National Center for Atmospheric Research (NCAR)

**Additional authors:** Seth McGinnis, NCAR; Brian Bonnlander, NCAR; John Clyne, NCAR

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

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

Big data; data management; data visualization; climate model downscaling

#### **Title of Response**

NCAR CISL response to NSF RFI on Data-Focused Cyberinfrastructure

#### **Abstract**

The National Center for Atmospheric Research (NCAR) Computational and Information Systems Laboratory (CISL) believes that data-focused cyberinfrastructure must be supported for the long-term rather than exist only during the lifetime of grant, and must include high-volume, persistent, publicly-accessible storage; scalable computing resources that can directly access the data; ability to find, launch,

reuse, and modify well-written and well-document analysis tools; improved data standardization and metadata; and training opportunities on how to use the infrastructure, write appropriate code, and prepare data for data-intensive research.

**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.

(a) Ability to reproduce other researchers' analyses, and to build on previous analyses. (b) Ability to prototype locally on small data, and then leverage Cloud or HPC for large-scale analysis. (c) Support for analysis on multiple large datasets that are not co-located on same storage infrastructure. (d) Growing disparity between compute (numerical modeling) and storage capability. (e) Lack of homogeneity/interoperability/standardization among datasets. Increasing level of computing literacy required to do modern science. Recent graduates not sufficiently trained yet; later-career researchers falling behind the technology curve.

**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?

(a) Storage for medium- or long-term hosting of large datasets on which results were based. Ability to compute directly on data in that storage area to reproduce or build on earlier results. Access to the data should be public as early as possible, and certainly by the time of first publication based on the data. (b) Reusable workflows and notebooks. Reliable and free way to host and quickly launch Jupyter Notebooks or other workflow tools. There should be an ability to run workflows/notebooks locally (w/smaller data) and then scale up in the Cloud, with minimal reconfiguration. (c) Better ability to move data as needed; ability to move/merge only results; investment in data lakes to serve as temporary locations to store data during analyses. (d) Strong enforcement of metadata, consistency, and standardization in NSF-funded data, with tools that data producers can use to verify their own data. Also independent (3rd-party) automated checkers. Process to be discussed in data management plan, and final payment on grant not made until checks passed. (e) Semantic mappings and other

transformations to allow inter-disciplinary studies using pools of data that are each internally consistent but are different between disciplines. (f) Workforce development, as well as additional investment/improvement in the tools.

**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.

(a) Besides CI, significant investments in labor to do data cleanup, data management, improved interoperability among datasets, fixing models to create more standardized output. (b) Build on what has previously been funded by NSF. Don't keep issuing new solicitations to re-invent everything -- need to get beyond CS R&D for its own sake. (c) Enable ongoing infrastructure, not temporary things that are killed at the end of the grant. (d) Instead of building and operating NSF-specific storage and computing resources, consider partnering with commercial cloud vendor(s) to obtain needed capacity at affordable prices, possibly including unlimited access to a designated pool of resources at a predefined annual cost. (e) Pervasive problem of poor-quality software, and of people needing to re-invent code that was already written but is either not discoverable, too hard to use, or not good enough. Continue activities such as Cyberinfrastructure for Sustained Scientific Innovation (CSSI). Consider a software curation/recommendation/approval process, focusing on a tight set of reliable, non-overlapping code bases. Also, give credit for good, reusable open-source software (as much credit/respect as for a published paper in CV).

-- End Submission --