

Reference ID: 11226376516\_Servilla

---

**Reference ID:** 11226376516\_Servilla

**Submission Date and Time:** 12/16/2019 12:36:53 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: Mark Servilla - University of New Mexico

**Additional authors:** Corinna Gries; University of Wisconsin; cgries@wisc.edu

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

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

Biology; ecological and environmental informatics, computer science; Servilla Limnology; ecological and environmental informatics; Gries

### **Title of Response**

Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research in the Ecological and Environmental Sciences: A Repository Perspective

### **Abstract**

Data-focused cyberinfrastructure needed to support future data-intensive science and engineering research in the ecological and environmental sciences is changing rapidly. Requests from the research

community to publish and archive large data (100's of Gigabytes to Terabytes) and numerous data sets (millions of individual data files) is becoming more common. Existing repository infrastructure cannot easily accommodate this class of data and must be re-tooled with future cyberinfrastructure demands in mind. In addition, workforce training must guide the users of repositories to better take advantage and strategize techniques for the management of such 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.

One data-intensive research challenge that is now becoming apparent for programs/organizations that publish and archive data (i.e., data repositories) is the recent explosion and subsequent management of large volume data or sets of data. Historically, scientific data in the environmental and ecological domain have been small in volume (MegaByte scale) and number (100's), with its own set of challenges caused by their high degree of variability in structure, semantics and sampling approaches. More often today, scientists are asking to archive very large (100's of Gigabytes to Terabytes) data files or to archive sets of data consisting of millions of data files so that they may be published for broader use. These data are now becoming more common as new technology increases both the fidelity and frequency of data collection from sensors or allows the operation of complex and data-intensive models, and they are becoming more ubiquitous across all sub-disciplines of environmental and ecological sciences.

**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 management and disposition of large volume (multi-TeraByte) data from across research domains is now emerging as a critical need for environmental and ecological data repositories who are looked upon as the final archive for such data. New research technologies and methodologies are producing vast quantities of data, including imagery (both still and moving) from drone and other capture platforms, mapping and geographical information systems (including oceanic bathymetry), genes and genetic variants and phenotypes, and numerical models. Research programs are now accumulating vast

amounts of heterogeneous data that span scientific specialties. Many of these programs prefer to publish and keep data archived within a common repository, thereby simplifying the data life-cycle process. Unfortunately, thematic repositories with mass storage do not host data outside of their own area of specialization. This leaves projects searching for more general data repositories that can accommodate large volumes of varying data. If a repository accepts large volume data, there are two specific technical barriers to the management of such data: (1) general Internet capacity makes the initial deposition of such data difficult, but reuse of large volume data to downstream consumers is almost impossible; current techniques still rely on old-school “sneakernet” transfers and (2) storage capacity of NSF-funded projects is being exceeded at extraordinary rates as new research is producing more data than can be reasonably managed. The issue related to Internet capacity is not new, however transport technology is not necessarily keeping up with the demand for data access. Repositories must now begin to consider offering computational hosting so that data-intensive research can occur at the location of the data storage rather than moving data to the researcher’s location. The second issue, storage capacity, can only be addressed through better planning and forecasting of expected data volume, along with the need for better design of repository architectures so that scalability can be accomplished seamlessly and cost effectively. Perhaps storage models should adopt the approach of the commercial cloud where common bulk-storage may be partitioned for domain-specific needs. Although large storage requirements may be resolved in a domain agnostic approach, many other aspects or current data publishing needs are still better handled within the community. The environmental and ecological research community in particular is still in a ‘data mobilization’ phase, i.e., a phase of culture change to more openness and transparency in the research process. This culture change is best driven and supported from within the community. Hence, data curation, awareness and trust building, training, general process development, and response to improvement suggestions currently still need to happen within communities. A major obstacle to efficient meta-analyses in environmental and ecological research is the almost non existence of standardized sampling methods and semantics for describing the data. This is another area of extensive work the may benefit from common tools, but will have to be conducted within the community of experts and data curators educated in those fields.

**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.

Workforce training, or more specific, data-user training (both producers and consumers of data) must be considered as part of the development of any repository infrastructure. Such training will ensure that evolution of technology meets the needs of users and users may be trained on optimal strategies for packaging and exploiting data held within the repository. Data repositories and their users share a symbiotic relationship, where each must build upon the technological needs and knowledge of the other. To fully integrate data publishing into the research process, workforce training needs to start much earlier, however. Only if well integrated into introductory science classes can a real shift in research practices be expected.

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: 11226376516\_Servilla

---

-- End Submission --