

Reference ID: 11226555946\_Rominger

---

**Reference ID:** 11226555946\_Rominger

**Submission Date and Time:** 12/16/2019 1:39:28 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: Andrew Rominger - Santa Fe Institute

**Additional authors:** None

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

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

ecology; evolution; computational biology

#### **Title of Response**

Making data-intensive research more inclusive to foster open source solutions to model reproducibility and data synthesis

#### **Abstract**

Data-intensive research requires heavy educational investment that cannot be attained at far too many institutions. This challenge should be overcome with the creation of cyberinfrastructure programs that simultaneously advance research and achieve pedagogical goals. Increasing the diversity of data-

intensive researchers will help us grow as a community to find solutions to the challenges of making our data-intensive research reproducible and taking full advantage of the diversity of data available but currently not interoperable. Data-intensive models need to be well documented, tested, and open. New standards need to be developed to make this happen, and new incentives (for funding and publication) need to be put in place. Huge advances have been made in the hosting of open data in online repositories. The logical next step is to create innovations in how these repositories interoperate.

**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.

Three challenges confront progress in data-intensive research: 1) the diversity and inclusion of scientists engaged in the research; 2) the documentation, shareability, and reproducibility of data-intensive models; and 3) the interoperability of large, open data repositories. The use of data-intensive methods in research requires extensive training that continues to be difficult for all students to access, and yet the more scientists who participate, the better our research will be. Entering undergraduate students often come from low math literacy backgrounds (mathematics being key to data-intensive research) and enter into one of the majority of universities that do not offer accessible training in data science. And yet skills in data science can be tickets to success both inside and out of academia. Thus we need programs to both increase quantitative skills before university, and add curricula at the university level. In our Infrastructure Innovation for Biological Research (IIBR)-funded work we will design and deploy massively open online courses and in person workshops with open curricula. We elected these strategies because they can reach audiences at under-served institutions; however, we are still challenged as a community by a lack of resources and support at smaller, often minority-serving institutions, and in K-12 education. Data-intensive research fundamentally depends on models (mathematical, statistical, computational). With increasingly heterogeneous data, associated models are becoming more complex. Documenting these models for reproducibility is no longer achievable with traditional methods (e.g. journal articles). This problem is all the more dangerous because model outputs can become the data inputs of others, creating potential cascades of un-reproducibility. To overcome this obstacle we need a robust, discipline-agnostic model metadata language and rigorous standards (for funding and publication) of code and software open access, documentation, and testing. These standards need to be discipline-agnostic such that, for example, when modeled-climate predictions are used for model-based biodiversity forecasts, all assumptions and uncertainties can be made apparent and testable. Complex models are not only inspired by, but actually require heterogeneous data. The curation and querying of such data requires the integration of multiple open data repositories. For example, a rigorous analysis of biodiversity past, present, and future, might require interoperating data repositories from the geosciences, climatology, biological museum collections, and molecular genetics. Currently, different labs produce non-scalable custom scripts for

this task, but making heterogeneous data repositories easily interoperate should be a research priority for the cyberinfrastructure community.

**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?

Cyberinfrastructure (CI) development needs to produce tools that simultaneously advance research and education. Examples include the ABI Innovation-funded Wallace project (PI Anderson) and our own IIBR-funded Rules of Life Engine project. These projects deploy an accessible user interface that generates reproducible, well documented code and model metadata. Thus users can begin to self-learn coding from a graphical user interface. Creating open course content and curricula through meaningful broader impacts should also continue to be valued in CI proposals. However, these pedagogical resources will go under-used without more funding opportunities focused on data-intensive training at minority-serving universities. Such funding opportunities should also recognize the need to begin building inclusion before students even arrive at university by opening channels for K-12 training programs. Great advances have been made in requiring data to be open source, documented, and accessible. Similar pressure (via funding and publishing requirements) needs to be put on data-intensive models. The underlying code needs to be documented, tested, and made open. These requirements are more involved for models compared to data, particularly the process of documentation and testing, but are critical for validating results. The heavy overhead of model documentation and testing reaffirms the need for training in data-intensive research. Basic CI research also needs to be conducted to discover the most efficient and informative ways to document models. Core model metadata should be discipline-agnostic to facilitate trans-disciplinary research, but should also have model ontologies that can be customize to domain-specific modeling practices (e.g. climate models do not need to know about modeled speciation processes, but biodiversity models need to know about both climate and speciation). Open data repositories have contributed enormously to advancing CI. The next logical step is to make these repositories interoperate. A combination of several approaches will likely be necessary. First, the metadata associated with different repositories are often in different metadata languages; building dictionaries to translate homologous fields will bring us closer to interoperability. Second, we need open software that makes use of these dictionaries to synthesize heterogeneous data sources on-the-fly using well documented repository APIs. Third, we need link these data integration efforts with innovation in making and documenting complex models, and training more scientists to use

them. Sustaining demand for data repository interoperability will be key to catalyzing innovation the logical next step of not only hosting data, but supporting integration of those hosted data.

**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.

In summary, we need to broaden the community of data-intensive researchers by deliberately funding and deploying CI programs that simultaneously advance science and inclusion in that science. We need this larger, more diverse community to discover the complex models that can draw insight from heterogeneous data. We need these models to be open and well documented with an as of yet undiscovered modeling workflow enabled by a new, discipline-agnostic model metadata language. We need the heterogeneous data required by these models to be easy to synthesize from disparate repositories that do not yet interoperate but should.

-- End Submission --