

Reference ID: 11226878755\_Towns

---

**Reference ID:** 11226878755\_Towns

**Submission Date and Time:** 12/16/2019 3:37:22 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: John Towns - University of Illinois at Champaign-Urbana

**Additional authors:** Larry Di Girolamo, Atmospheric Sciences, UIUC; Dan Katz, NCSA, UIUC; Kevin Leicht, Sociology, UIUC; Andre Schleife, Materials Science and Engineering, UIUC; Ester Soriano, NCSA, UIUC

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

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

astronomy; astrophysics; earth sciences; climate modeling; materials science; environmental sciences

#### **Title of Response**

Data Infrastructure and Services to enable multi-disciplinary Data Sharing and Analyses

#### **Abstract**

We see the need for a shared facility partnered with other facilities, optimized for exploitation of data from a wide range of sources which produce more and more data and serve larger and larger communities, where world-class science requires access to products from multiple instruments. Cross-

cutting nearly all domains is the need to allow a large number of individual investigators to focus on their research topics by relying on the capabilities and expertise of a facility providing an appropriate level of data and technology support. Part of the decision to create the facility includes ensuring that it is well-governed by community input and participation, and to provide ongoing and agile support based on a staff with the requisite professional skills.

**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.

Earth science and astronomy have and are building instruments that will produce many petabytes of data. The data needed to extract the most advanced science comes from multiple instruments, which have independent observing programs, and sometimes from multiple agencies, including international partners. The scale of instrumentation and the reliance of larger communities on fewer but more massive observational datasets is growing. Candidate instrumentation from WFIRST, LSST, Euclid, CMB-S4, MAIA, etc., form a vast body of observations that are foundational to leading-edge observational science. The ultimate purpose of these massive observational programs is to extract knowledge and insight from the interrelated data. A big challenge in materials science is gathering and homogenizing data from a large number of small experimental or theoretical research groups. Gathering data and obtaining related metadata from researchers is difficult and currently not standard in the field. Centralized infrastructures, development and dissemination of metadata schemas, and ontologies to build connections between different experiments as well as experiment and theory is necessary. Achieving this successfully across synthesis, characterization, and theory may enable the goal of designing materials for specific applications and bears great potential for tremendous societal impacts, including energy, health, consumer electronics, and many more. Further, UIUC seeks to develop a distinctive form of medical education with an engineering emphasis. Part of that emphasis needs to develop ways to tap environmental and atmospheric hazards in close-to-real-time in the immediate environments of patients. Further, the school needs to train physicians and health policy researchers in the uses of, and advances in, remote sensing technologies to assess the quality of air, water, soils, and the built environments surrounding patient populations. These data can be combined with data from law enforcement agencies, the Census, and other government agencies to provide a much better accounting of the natural and human environments in which patient populations live. The ability to convert these data into real time assessments of exposure to hazards would represent an invaluable tool for policy analysts studying health and illness and ultimately for physicians seeking timely diagnosis of patient ailments. Currently, accessing this data is daunting for most research teams and requires access to a substantial facility and at-scale data science and engineering techniques to fully exploit these data. In practice, only a few researchers are able to achieve this combination of facilities and skills. A path forward is needed to enable exploitation of datasets of this size.

**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?

Cross-cutting all of these is the need to allow a large number of individual investigators to focus on their research topics by relying on the capabilities and expertise of a facility providing an appropriate level of data and technology support. In order to effectively support a range of domains the data ecosystem must evolve to ensure that the needs of the research community for highly heterogeneous data can be supported in order to enable researchers to productively exploit massive amounts of data to obtain knowledge and insights. Addressing these challenges requires data infrastructure. We see the need for a shared facility partnered with other facilities and optimized for exploitation of a growing collection of data which will serve larger and larger communities. To accelerate science and benefit decision-makers and the economy by supporting a large number of projects in this new data environment, we see the need for a shared facility that 1) provides for co-location of data with computational resources for analysis; 2) provides for reuse of data by multiple independent investigators, reducing redundant copies which inevitably would be supported by the agencies; 3) provides for application and dissemination of data science and data engineering expertise; and 4) working with the communities, advances the state of the art relevant to the facility, jointly and economically benefiting the scientific domains that it serves. Part of the decision to create the facility includes ensuring that it is well-governed by community input and participation, and to provide ongoing and agile support based on a staff with the requisite professional skills. The essential characteristics of a facility are that it is 1) agile and well-governed by the community; 2) independent of, while collaborating with, multiple projects that act as data sources; and 3) sufficiently large to command economies of scale, both in provisioning and expertise. This facility and associated infrastructure must be developed and maintained over time by professionals with knowledge of best practices. These professionals, such as Research Software Engineers (RSEs), must be supported with suitable university career paths and role-specific evaluation metrics used for hiring and promotion. While these roles are now supported by some universities, NSF policies must recognize and enable them, such as allowing full-time staff to be funded on projects, and by working with reviewers to help them balance the return against the cost of staff versus students and/or postdocs. Professionals at the intersection of various disciplines and data science are needed with domain knowledge in both fields, in order to build efficient and useful metadata schemas. Domain scientists need to be trained accordingly and this effort is necessarily highly cross-disciplinary. Infrastructure is also needed to store experimental and simulation data and make it accessible to all

researchers. Training infrastructure is needed to train the broad community to contribute their data/metadata and to use the data accessible on the infrastructure. Finally, while there is significant and valuable discussion of the workforce development efforts, they focus on the technical skills of staff. There must also be a focus on the development of the management and leadership skills of these professionals in order to create a complete and effective workforce.

**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.

Data infrastructure needs to both be as general as possible but also specific to disciplinary-needs. In order to design such infrastructure, disciplinary needs must be clearly understood. NSF should support a process organized and executed by representatives of the disciplines. Once needs are understood across a number of disciplines, a clearer understanding of common needs along the differentiated needs for various communities can be better understood and inform a strategy that leverages shared infrastructure and services to the greatest extent possible allowing greater focus for development of differentiated services. A number of communities, e.g. materials science, are not used to sharing their raw data or using other's raw data, aside from specific teams of collaborations. Establishing this culture, overcoming reservations, justifying the additional effort of proper curation and sharing, and related problems will be instrumental to make such efforts successful. Further, CI resources and staff dedicated to the creation and maintenance of various data with the goal of creating and implementing "point and click" access to curated data and software tools for researchers and practitioners as well as policy researchers and analysts to evaluate how, for example, changing immediate patient environments contribute to health outcomes over wide geographic space is sorely needed. Researchers in need of access to CI often find themselves requesting allocations from different providers to marshal the necessary capabilities. Allocation processes are frequently variations on a common high-level approach. A single allocation process across, a single point of entry for users and consumers of data, and a consolidated mechanism to manage allocated services, would allow researchers more productive access to CI services; enable a more streamlined means to share data; and make it possible for CI providers to tailor allocations to researchers' needs, and to assign and manage resources more efficiently.

-- End Submission --