

Reference ID: 11226170198_McHenry

Reference ID: 11226170198_McHenry

Submission Date and Time: 12/16/2019 11:31:02 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: Kenton McHenry - University of Illinois Urbana-Champaign

Additional authors: Dan Katz, University of Illinois Urbana-Champaign; Colleen Bushell, University of Illinois Urbana-Champaign; Praveen Kumar, University of Illinois Urbana-Champaign; David LeBauer, University of Arizona Luigi Marini, University of Illinois Urbana-Champ

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

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

research software/cyberinfrastructure; geoscience; biology

Title of Response

Software Perspectives at the Intersection of Scientific Disciplines

Abstract

The NCSA Software Directorate responds to this RFI as an organizational body that supports a broad range of scientific research needs via technologies we build, adapt, and deploy, based on commonalities

we encounter across these observed needs. We first outline a set of emerging research challenges, particularly in areas of convergent research. From this, we point out common technology needs and gaps within the national data cyberinfrastructure, as well as growing needs for personnel to support these resources and to support the researchers that use them.

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.

From the vantage point of a supercomputing center supporting the computing, data, and software needs of numerous scientific domains, we have observed an increase in the number and diversity of disciplines using advanced computing/cyberinfrastructure, bringing new challenges as compared to traditional computing intensive disciplines such as physics. Here we describe five fields of research that present such challenges related to the use of data in transdisciplinary research. The first set of these can be referred to as end-to-end geoscience, a convergent area of work spanning geoscience itself and leveraging advances in areas such as phenomics, genomics, biology, ecology, machine learning, civil engineering, and agriculture, seeking discoveries that enable a number of societal impacts of relevance today. One specific field is that of ecological forecasting, or earthcasting. This area aims to model and predict aspects of ecology and the environment (plant/animal populations, resource availability, etc.), much like today's weather models, across multiple timescales, so today's decisions reflect and take into consideration the consequences and outcomes across years to decades, with results that have clear impacts to decision makers, society, and the economy, and are verifiable on the order of one or so years and are trusted by the public. A second field is phenomics which aims to measure the physical manifestations of genes by environment interactions, a task that benefits immensely from advances in machine learning, particularly in the areas of robotics and computer vision. Digital agriculture is a third field, leveraging a variety of data sources as well as machine learning advances, and aiming to optimize agricultural practices to produce more with fewer inputs and environmental impacts. Applications include precision agriculture, crop improvement, and prediction of yields and other ecosystem services. The fourth field is risk/hazard management. Because societal risk is increasing due to both global and local changes, these risks must be evaluated in context by bringing together event models (e.g., earthquakes, tsunamis, hurricanes, wildfires, tornadoes) with models of the natural (i.e., critical zone) and human-built regions, requiring the development of novel tools and workflows around machine learning/AI advances. The fifth field is materials science, where work to design novel materials leads to challenges in automatically capturing and curating instrument data and metadata. Common challenges for all five fields include the need for machine learning tools/workflows, data sharing/wrangling tools, and low cost elastic computational resources near datasets to support these tools.

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?

To address the challenges above, a number of CI components are needed. First and foremost is elastic compute in proximity to data storage centers, often referred to as the “cloud”. While many funding agencies and programs consider this to be a resource that should be provided commercially, and others think of it as just something suitable for a “testbed” resource, treating it in this way has a high cost to science. This type of resource provides flexible data analytics and pipelines, optimizations in the form of minimizing costly data movement by providing computational resources near data, on demand capabilities for aspects such as community data portals and scientific gateways, and overall user accessibility and usability. Such resources are of immense value to the bulk of scientific efforts, in addition to those mentioned above, work typically thought of as embarrassingly parallel though this is all that is needed by most researchers. Such resources and capabilities provided locally by non-profit organizations can be highly cost-effective: 3 to 5 times less expensive than commercial options when well-utilized (even without considering additional egress costs), with data ownership/movement being fully retained, and with the ability to interoperate with and burst to a commercial resource when needed (<https://doi.org/10.17226/21886>). This can enable more funds to be delivered to students, education, and science overall. Further, the notion that commercial options are more cost effective ignores the high cost of the software engineering that goes into all modern scientific work and the lock-in aspects of commercial resources in the form of egress and/or optimizations to computation costs by utilizing proprietary libraries, resulting in code that is costly to refactor if other resources are ever leveraged. Under the growing area of “hybrid cloud”, such local interoperable elastic resources based on familiar virtual machines/containers, with nearby data capabilities, are essential for the needs of most of science spanning image/satellite data analysis, LIDAR, the growing area of non-consumptive data analysis where direct access to the data must be restricted, and machine learning where a pre-trained model is applied to classify novel data, especially when a mix of capabilities are provided by such resources (e.g., both CPUs and GPUs for modern machine learning-based efforts). In terms of software CI, there are a number of common capabilities that continue to be needed and/or where improvement is needed, often in terms of usability. These capabilities include data gathering/ingesting/cleaning (sometimes called data wrangling), data management and sharing with support for flexible metadata and data labeling that can also support modern machine learning needs, user friendly visualization, readily accessible advanced resources along the lines of science gateways, and reproducibility/robust execution such as user friendly workflows. These needs, while typically requiring adaptation for specific disciplines, are at their core similar across disciplines. Thus, they can ideally be provided in a cross-

disciplinary yet tunable manner, which leads to improved sustainability and reduced cost. While options for such capabilities are emerging, many are too fine tuned for a specific discipline, not sufficiently interoperable, and not sufficiently accessible in terms of usability to the researchers who need them.

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.

The growing need for professional, non-student, long-term research software development support by the scientific community can not be understated, having boomed in the past decade with the need to continually develop, improve, support, sustain, and adapt software required by research as well as train those that use and contribute to it. This need is highlighted by the independent formation of pockets of these human resource capabilities internationally, recently coalescing into what is becoming known as the Research Software Engineer (RSE) movement. While for many, the need for RSEs as part of the modern academic/research ecosystem is now apparent, there are still many challenges in providing these human resources due to an academic reward/funding system that predates the digital age (e.g. <https://doi.org/10.5281/zenodo.3234083>). A better understanding of what defines an RSE, how they differ from other aspects within the academic environment, and the differing needs in terms of providing/retaining these resources is essential.

-- End Submission --