

Reference ID: 11227061924_Bhatt

Reference ID: 11227061924_Bhatt

Submission Date and Time: 12/16/2019 4:50:46 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: Asti Bhatt - SRI International

Additional authors: Roger Varney (SRI); Todd Valentic (SRI); Ashton Reimer (SRI); Elizabeth Kendall (SRI); Leslie Lamarche (SRI)

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

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

Space Physics and Aeronomy

Title of Response

Cyberinfrastructure Challenges for Space Physics and Aeronomy

Abstract

There are many challenges with the existing space physics and aeronomy (SPA) cyberinfrastructure (CI). Presently, individual projects are individually responsible for storing and distributing the data that they collect, resulting in data being available in many formats from a wide variety of platforms with differing

quality and uncertainty as to what happens to the data when a project ends. NSF should develop its own, FAIR compliant CI framework, capable of handling continuous data from a variety of instruments, while maintaining standards in data format and quality. In designing this framework, NSF should consider lessons learned from NASA's approach to handling data and solicit input from facility/instrument operators who generate large quantities of data. Facilities should be accountable for their progress in data distribution as part of their annual reports. A centralized data infrastructure should continually work to migrate towards modern storage technology, ensuring that data and software maintain their usefulness long past the lifetime of a project.

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.

The Decadal Survey for Solar and Space Physics identified as one of the science challenges to “[d]etermine and identify the causes for long-term (multi-decadal) changes in the [atmosphere-ionosphere-magnetosphere (AIM)] system.” Addressing this goal requires analysis of data across multiple 11-year solar cycles. Long-term changes in the AIM system are being driven by a combination of long-term trends in solar activity, secular changes to the geomagnetic field, and human activities such as greenhouse gas emissions and other emissions from rockets. Disentangling the multiple sources of long-term change requires preserving long-term data sets, including data sets from historical facilities that are no longer operating, and ensuring those data continue to be accessible and analyzable. The science challenges in the Decadal Survey also emphasize the highly coupled nature of the AIM system and the need to study it across global, regional, and local scales. Addressing these questions of coupling between different regions and between different scales necessitates combining data from heterogeneous sources. Working with highly varied data can be very challenging and requires standardized treatments of uncertainties and quality flags across data providers. Combining data from multiple sources also requires those data to be accessible from centralized sources in standardized formats. Standardized formats refer to usage of human and machine readable non-proprietary open-source file formats such as hdf5 and cdf, with similar classes of instruments adhering to a common file structure. For example, magnetometers should all use the same file structure, but should not have the same file structure as all-sky cameras.

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 space physics field needs extensive CI improvements in order to conduct our research effectively. Currently we lack standardized data products for cross-disciplinary science. Data providers should provide a higher degree of transparency in their data processing and quality control, including careful documentation of assumptions and models used during analysis. Reproducibility should be emphasized, and all data should contain metadata that documents exactly which version of the analysis code was used to produce it. Furthermore, there needs to be a way to track the integrity and chain of custody of data products from the instruments through every step of analysis. Currently individual instrument operators are each operating their own data management and distribution plans, with varying levels of quality. The NSF needs to establish a centralized data infrastructure that has the ability to ingest and distribute vetted, version-controlled data and associated analysis software from all current and past NSF-supported work. This will require continuous maintenance that needs to be fully funded by NSF for both technology and human resources. This also requires the ability to continuously adapt the infrastructure to changing storage technologies.

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 NSF should regularly engage the community to establish needed data and software infrastructure. The current system of evaluating data management plans in proposals is inadequate. Projects that involve the creation and distribution of data should not only be required to submit a data management plan in the proposal, but they should be evaluated on their data management performance throughout the award. The current format of annual reports does not track data management performance effectively. The current academic environment provides little incentive for creating scalable and usable software and data systems. Creating a better CI landscape would also significantly improve experiences for students. If data formats were standardized and data were accessible and discoverable through a centralized source with a user-friendly interface, students would be able to learn to work with data more quickly. Lowering the barrier to entry to data analysis could also enable a larger number of undergraduate students to work with real data earlier in their academic development. A well-designed user experience for a centralized data infrastructure should be created in collaboration with experts in the field of user experience design. Industry is far more advanced than academia in the design of functional user interfaces, and that expertise should be exploited.

-- End Submission --