

Reference ID: 11227192168_Sterner

Reference ID: 11227192168_Sterner

Submission Date and Time: 12/16/2019 5:50:05 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: Beckett Sterner - Arizona State University

Additional authors: Nico Franz, Arizona State University; Edward Gilbert, Arizona State University

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

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

Systematic biology; biodiversity informatics; philosophy of biology

Title of Response

Decentralized but globally coordinated data science

Abstract

Semantically-aware decentralized versioning for scientific datasets is an emerging need across multiple NSF directorates. An emerging custom services model for data analysis, while better able to accommodate local communities of practice in a collaborative process of integrating scientific knowledge with decision-making, risks fragmentation of the resource pool of scientific knowledge

overall. Reliable aggregation of pooled scientific data must be engineered and revised over time to ensure continued coordination of new data contributions and modifications to existing datasets. This sets up a critical challenge for decentralized approaches to data aggregation: how to engineer the capacity for competing hypotheses and distributed curation work without losing the connectivity required for global data sharing? A semantically-aware decentralized versioning system for scientific knowledge products, especially data, is critical to realizing greater impacts for scientific knowledge on decision-making through the custom service model.

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.

A leading hypothesis for the primary cause of widespread and continued disconnect between scientific knowledge and societal decision-making has been the scarcity of ready-made knowledge for decision-makers to access and apply. A common response has been to pursue a loading dock model for producing scientific knowledge for decision-makers whereby synthesis researchers generate standardized knowledge products that are collected and made available in a single centralized location such as an online repository. Standardization is crucial to the loading dock model's appeal because it putatively helps scientists achieve new efficiencies of scale by leveraging automation based on globalized formats and production workflows. However, growing evidence shows that making an abundance of off-the-shelf scientific knowledge is insufficient to cause a fundamental transformation in the knowledge-action gap. Other obstacles that are at least as important include decision-makers' trust for the science and scientists involved, the substantial work required to customize and augment current knowledge to address local contexts, and the limited reach of stable consensus knowledge in many areas of great relevance to society. While the loading dock model is valuable in appropriate circumstances, it no longer viable as a general answer to the full challenge of incorporating scientific knowledge into decision-making to achieve better outcomes. An emerging alternative, a "custom service" model, emphasizes cyberinfrastructure built to enable sustained interactions among a community of practice with overlapping interests oriented toward addressing a shared decision, research problem, or domain subject. The custom service model is oriented toward growing adoption of collaborative approaches to integrating science into decision-making where all stakeholders participate in the process of developing scientific knowledge and infrastructure. Instead of reflecting the authority of a centralized consensus-making body, cyberinfrastructure on the custom service model therefore facilitates the growth and sustainability of a decentralized but still coordinated ecosystem for data, models, software, and other scientific knowledge products. Cyberinfrastructure design therefore prioritizes a broader set of functions than in the loading dock model, and achieving efficiencies of scale on a global level is one concern but doesn't override other aims. For example, cyberinfrastructure should provide affordances for decentralized governance over pooled knowledge resources at a community of practice scale rather than entrenching globally centralized control. Also critical is leveraging computing power to

accommodate customized development of best practices within communities of practice through enabling local experimentation, constructive competition, and specialization to maximize relevance for situated decision-making or research challenges.

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 custom service model, while better able to accommodate local communities of practice in a collaborative process of integrating scientific data with decision-making, risks fragmentation of the resource pool of scientific knowledge overall. Reliable aggregation of pooled scientific data must be engineered and revised over time to ensure continued coordination of new data contributions and modifications to existing datasets. This sets up a critical challenge for decentralized approaches to data aggregation: how to engineer the capacity for competing hypotheses and distributed curation work without losing the connectivity required for global data sharing? A semantically-aware decentralized versioning system for scientific knowledge products, especially data, is critical to realizing greater impacts for scientific knowledge on decision-making through the custom service model. The Git model for decentralized version control provides a powerful and successful foundation for realizing decentralized but globally coordinated data services (Loeliger and McCullough 2012). Perhaps best known through its implementation by GitHub, the Git model allows a group of collaborators to create parallel versions (“forking”) of a shared reference standard (the “master”) and edit these versions locally before merging the edits with the reference standard (via a “pull request”), which may itself have changed in the meantime. Similarly, local versions can be updated with changes from the reference standard (a “push”) by reconciling edits to the local and reference versions. Adopting the Git model for a project comes with implicit governance decisions about who has the ability to create local versions, request and approve changes to the reference standard, and push updates from the standard to local versions. Contributors to a collaborative project will generally form a community of practice, and the appropriate governance strategy within a community can vary from highly centralized to highly decentralized; indeed, communities often evolve over time as they grow or change identity (Shaikh and Henfridsson 2017). Existing Git implementations, however, have key limitations for applications to semantically annotated scientific knowledge products, especially for data. One critical limitation of current implementations is that while they can track line edits to documents, they do not evaluate the semantic implications of those edits. It is therefore essential to extend the basic Git model to incorporate semantically-aware conflict detection and reconciliation between datasets with different

metadata classification systems (Arndt et al. 2019). A second limitation is that semantically-aware data reconciliation needs to be possible across multiple reference standards rather than solely with respect to local versions of a single reference standard. A setting where biologists, for example, maintain multiple, partially conflicting species checklists (e.g. as has been the case with birds for decades) is most analogous to handling decentralized versioning across multiple Git projects (i.e. multiple “master” versions), each of which has its own local versions. Aligning concepts rather than text documents (i.e. the meanings of metadata terms rather than the documents specifying them) is therefore doubly essential for accurate and machine-automated data aggregation across parallel metadata systems.

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.

Semantically-aware decentralized versioning for scientific datasets is an emerging need across multiple NSF directorates. While our own expertise is species occurrence data for biodiversity science, we’ve observed parallel challenges for colleagues in anthropology and archaeology seeking to synthesize survey data collected across geographic subdistricts worldwide and archaeological collections from different sites and research teams. The challenge also extends to citizen science organizations such as iNaturalist and NGOs such as NatureServe that rely on data classifications (e.g. biological taxonomies) that scientists regularly update in a decentralized fashion with minimal versioning information. Right now scientists in these different fields are not communicating and identifying shared cyberinfrastructure needs for robust and agile data analysis, so a shared process to articulate and align infrastructure design specifications would have high impact.

-- End Submission --