

Reference ID: 11228633757\_Pfeiffer

---

**Reference ID:** 11228633757\_Pfeiffer

**Submission Date and Time:** 12/17/2019 9:04:04 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: Nicole Pfeiffer - Center for Open Science

**Additional authors:** Brian Nosek, UVA, COS; David Mellor, COS

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

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

Mechanical engineering, Materials science, Product Development, Agile software

#### **Title of Response**

Cross-disciplinary data sharing limitations are both technical and social

#### **Abstract**

The challenges for cross-disciplinary data sharing are both technical and social. Enhancements to research tool infrastructure will enable the metadata collection, but requires research communities to define their disciplinary schemas. Using those schemas, along with incentives, to build a FAIR workflow that is mapped across disciplines will provide the path.

**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.

Interdisciplinary data generation and sharing requires improving FAIRness of shared data, with more examples from individual researchers. Discipline specific metadata schemas will go a long way to improving FAIRness of the data being generated, however additional efforts are needed to standardize the schemas and map them together into a heterogeneous schema for cross-disciplinary uses. The current model uses centralized discipline-specific repositories, which are limited by quality metadata and metadata mapping for achieving FAIR cross-disciplinary data sharing. Further efforts that could assist this work are visibility sharing behaviors to create new norms (e.g., using badging to increase visibility) and growing a community of practice that provides examples for colleagues to emulate. That is, the challenge is not just technical, it is also social. Supporting both will be productive.

**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?

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? CI services that would be most impactful would enhance existing tools researchers use for sharing data. Enhancing these services could advance establishment of common, standardized discipline-specific metadata schemas to be widely available and easily adopted by researchers. Building in incentives for this level of data stewardship will encourage uptake and wide

adoption. The immediate first step is community building activities among the disciplines to unite and align schemas for detailed data descriptions. Some disciplines have made headway into this, while others need examples and the support for meetings and other events to gather research communities and foster their metadata development. A few examples of a research communities with established discipline-specific metadata schema are: Brain Imaging Data (<https://bids-specification.readthedocs.io/>), Frictionless data (<https://frictionlessdata.io/specs/tabular-data-package/>), and Psychology Data ([https://docs.google.com/document/d/1u8o5jnWk0lqp\\_J06PTu5NjBfVsdoPbBhstht6W0fFp0/edit#heading=h.1m6sa5bs9j0e](https://docs.google.com/document/d/1u8o5jnWk0lqp_J06PTu5NjBfVsdoPbBhstht6W0fFp0/edit#heading=h.1m6sa5bs9j0e)). Once these schemas are developed within research communities, connecting communities can work on mapping them to one another using a common language to enable cross-disciplinary research efforts. Here again, support of action-focused community meetings will be necessary to build the mapping and promote its use. There is also need for CI to support easy to use interfaces for researchers to upload their data and apply discipline specific metadata, with a backend mapping of that metadata to a cross-disciplinary schema. This includes easy discovery of the data being shared accompanied by the discipline specific metadata and available cross-disciplinary metadata to support research consumers. Ideally the interfaces would allow application of multiple metadata schemas to the data at once to support robust metadata in multiple schemas. Data and metadata should be easy to export and download with standard protocols so transfer and portability of data is not an obstacle to reuse.

**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.

Sustainability is a persistent challenge for data-intensive infrastructures that support science as a public good. If the community is motivated to treat the infrastructure and its hosted data as public goods, then it needs to address the fundraising challenges for infrastructure providers. Calls for proposals focusing on new, innovative infrastructure are great, but without accompanying support mechanisms for sustenance and incremental improvement of core infrastructures the foundation of the scientific infrastructure will remain fragile and prone to defaulting to business models that monetize on the access and use of data.

-- End Submission --