

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:20:51

PAGE 1

REFERENCE NO: 260

This contribution was submitted to the National Science Foundation as part of the NSF CI 2030 planning activity through an NSF Request for Information, https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17031. 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>).

Author Names & Affiliations

- Matthew Collins - University of Florida
- Jennifer Hammock - Smithsonian Institution
- Jorrit Poelen - Manylabs
- Anne Thessen - Ronin Institute
- Alexander Thompson - University of Florida

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Data aggregation, data management, biodiversity data

Title of Submission

Data Management Plans Aren't Enough to Ensure Usable Data

Abstract (maximum ~200 words).

We applaud NSF's commitment to foster effective management and dissemination of research data. It is indeed a challenge to ensure that infrastructure and process requirements and their guidelines stay up to date in an era of rapid advances in technology and technique. Herein we address contemporary expectations for findability, quality, and interoperability of data products and strategies by which NSF funded projects can meet them. This is a critical issue for data openness and will dramatically impact our ability as a research community to fully leverage published research products. Data may be freely available online but if it is not easy to discover it, get it to interoperate with existing data, and adapt it for re-use, especially in a time of increasingly urgent global scale research questions, then it is not fully open.

Two changes to existing data management plan expectations can go far towards addressing our concerns:

1. Data produced in projects should have all core concepts mapped, identified, and linked to authoritative data products in the domain
2. Data repositories need a computation component that can inspect, enforce, enhance, and make discoverable such relationships between data products

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:20:51

PAGE 2

REFERENCE NO: 260

Contemporary data science has demonstrated the power of well structured linked data to rapidly discover anomalies and trends in vast data sets. This impressive and occasionally creepy superpower has been demonstrated in realms like social media and e-commerce, where data is born not only digital, but highly structured, and well linked to other relevant records.

The native shape of scientific data are also highly structured and linked, but most of this is not recorded in the biodiversity domain. This is overwhelmingly prevalent, and understandable, in historic data sets that are not born digital, but even modern data sets with rich metadata are often described only with human readable text. Finding related data sets, much less individual records from different data sets that are relevant to each other, requires the human effort of scholarship. Given the knowledge that we have in hand today, a scholar with an infinite supply of time could discover preliminary reports of an unfolding epidemic, records of wildlife vectors, properties of the pathogen, historic records of earlier outbreaks, concurrent climate, demographic and health and human services data from the earlier outbreaks and the present, and then access lessons learned and apply them to the current scenario. Or a machine could do it, at great speed, if only the data were structured so that the relationships among the records could be identified. Currently, for the most part, they are not.

The potential insights and rapid advances that structured scientific data could yield are significant, but their scope and scale are difficult to estimate. We will know more when we have begun to leverage the methods described below.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Data management plans aren't enough, and in some cases the desire to faithfully comply with the requirements of the data management plan are counter productive. Even in the best of current scenarios, when all of the products of a research project end up freely available in repositories with their appropriate metadata, we end up just building slightly larger silos of data. Some of these silos are very large indeed, like DataONE or data.gov, and many are very narrow in scope, like iDigBio or NCBI, but they are still just passive holders of data. It is often difficult to relate any of the data in any one product to the data in another product. This isn't really anyone's fault, it is simply a natural outgrowth of the fact that most data products are designed to be self contained, often so that it is easy to deposit the data in institutional and remote repositories.

We think that the research data community can do better. Specifically, we think that it is possible to change the underlying assumption that grant funded data products are stand alone items. Data products are subject to network effects; their utility increases with each new connection between them. NSF should expect its grantees to produce data that fit together in the broader context of our existing databases just as written research is expected to be placed in the context of the existing literature.

This starts with the data producers. NSF should require that research data sets use globally unique identifiers for concepts from other data sources, published ontologies, and existing vocabularies customary in the domain. Examples from the biodiversity domain might include the Open Tree of Life for taxonomy, USGS data sets such as the hydrologic units for geography or ontologies such as ENVO for expressing environment information. To facilitate this, it may be necessary for directorates to build a listing of existing large scale global or national data products that are relevant to their domain and explicitly require linking to them.

It continues with the repositories. The places where scientific data is deposited by default like DataONE or data.gov provide few resources for providing visibility into and linking between data sets. What is needed is investment in repositories capable of representing and building relationships between their data sets both from explicitly provided data and metadata and use of recognized ontologies, as well as from inference methods such as deep inspection of data to locate identifiers, terms, and relationships not explicitly provided by the original researcher.

Further, static repositories of data are prone to becoming outdated and not usable for current research. To address this we need live repositories where data is treated as a living thing. Its creation, sharing, and use should be tied together in active flows and made continuous. Having data be living provides an opportunity for continuous review and improvement instead of the only review (and often a cursory review at that) being done at the time of publication by a reviewer who likely does not actually use the data. This improvement process would eliminate the drawbacks of both of the two competing views of data publication: delaying data publication until full review denies access to current data; and publishing data quickly that may not be well quality controlled results in errors propagating into other's

research.

Requirements to use persistent and resolvable identifiers to do the linkages and new repositories that are able to maintain these links would lead to an uptick in interest in linking data sets, and hopefully linked data services, much like the original data management plan requirements have helped spur interest in institutional and domain repositories.

It is not enough to simply warehouse or archive data. Information needs to be machine-readable, linked, and maintainable to be usable for science. The National Science Foundation should encourage the development of research cyberinfrastructure that facilitates these needs.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Continuous maintenance of data requires a continuous source of funds for work. We propose a grant program to support quality assurance, quality control, and interpretation of data sets. Such efforts have direct parallels to study replication efforts in other domains such as medicine and the social sciences. The products of this work can be represented as augmentations to the existing data and also be considered uses of the existing data which will demonstrate the value of the original data publisher's work. New citation and usage metrics can be developed to reward people who both produce and enhance relevant and highly related data to provide documentation of the impact of this work to funding agencies besides NSF.

Unlike the perhaps antagonistic tone that reproducing work like a medical trial can take, improving the quality of a data set should be considered a positive event for both the original publisher and the improver. After all, there would be nothing to improve if the original publisher had not made their contribution available to the whole scientific community.

Another potentially useful mechanism would be to borrow the model of the RCN program to create something like a Data Coordination Network to allow existing data providers to self-organize and establish their own ADBC-hub like organizations where they have identified the need for infrastructure and community leadership to produce an integrated data product. These DCNs could be focused on the typically one time efforts of standardization and planning with the goal of enabling each participant to sustainably and continuously produce their component of the integrated data product once the DCN grant is done. These DCNs could also be encouraged to rely on existing large scale cyber-infrastructures such as XSEDE for computation or DataONE for storage by incorporating dedicated sustainable funding models into the proposal process, like a startup-scale XSEDE allocation that persists as long as the data product needs it too.

All the linking, inference, annotation, and general processing of data described here requires a workforce skilled in the practice and theory of modern computation. NSF already has existing workforce development programs for informatics capacity building and we view them as a prerequisite for our next generation of scientists to be able to move beyond single-data set, and even single-domain, focused science and into the inter-related data world that we are advocating.

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 publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."
-