

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-03-27 15:27:03

PAGE 1

REFERENCE NO: 196

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](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

- John Ayers - Vanderbilt University

## Contact Email Address (for NSF use only)

(Hidden)

## Research Domain, discipline, and sub-discipline

Geochemistry

## Title of Submission

Expanded use of online geoscience databases to improve data accessibility and scientific reproducibility

## Abstract (maximum ~200 words).

While most of the sciences have made concerted efforts to make data accessible online, enabling the use of data mining and large-scale data synthesis, some geoscience disciplines have lagged behind. NSF must institute incentives for uploading results from NSF-funded data collection to databases, and ensure the sustainability of online databases, to maximize the utility of the data. All data should be spatially referenced to enable use in GIS, and sample data should be cross-referenced with SESAR, the System for Earth Sample Registration. Workshops may help geoscientists reach agreement on standard data formats and the organizations that will host and maintain the databases.

**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.

Currently there is a patchwork of geoscience databases that are sometimes hard to hunt down; many Geoscience subdisciplines lack any database resources. Others are woefully incomplete and becoming obsolete. For example, I recently tried to download sediment chemical composition data for Bangladesh from SedDB. It returned only 9 samples, all from a single reference. SedDB also reports that it has been static since 2014 and will not be updated until further notice. I personally know that NSF has funded projects that have collected sediment composition information on hundreds if not thousands of samples from Bangladesh. My research group is interested in identifying areas where sediments have high arsenic content, which may result in produced rice with high As. NSF must provide incentives for research groups to upload results from NSF-funded data collection to databases, and to ensure the sustainability of databases, to maximize data utility.

NSF should first fund a workshop for interested parties to come up with a plan for developing a database. Participants should consider

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-03-27 15:27:03

PAGE 2

REFERENCE NO: 196

---

standard formats and metadata requirements. They should also explore the potential for third parties such as Mendeley Data <https://data.mendeley.com/> and Elsevier Research Data <https://www.elsevier.com/about/open-science/research-data> for making online data accessible. The working group should include representatives from geoscience publishers, especially the societies such as AGU and GSA. Perhaps they can come to an agreement on a centralized portal, and require that future submissions include the posting of datasets online to improve reproducibility.

**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).

To ensure the sustainability of such online resources, organizations rather than individuals or research groups should be chosen as database hosts. Organizations could include societies such as AGU, or Institutes such as Woods Hole or Scripps. In addition to funding the creation and implementation, NSF would also have to guarantee long-term funding for maintenance. The operations could be cost-effective if data submission forms make it easy for researchers to submit their data to databases. NSF should require grantees to submit all data acquired by funded projects to be submitted to the appropriate online database. Each database should include citation information in response to queries; derived papers should cite the original data source, giving researchers another incentive to make their data available online. Each query could be embedded in a permalink that would allow journal article readers to easily download the dataset from the database, making it easier for readers to verify claims and thereby enhancing scientific reproducibility. NSF should maintain a portal that provides access to the database resources it funds. This would make the databases more visible, increasing their usage and making it easy for NSF to track usage and demonstrate the cost-effectiveness of their investments. Popular query permalinks could be included on the main web page for a database.

## 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>)."
-