

Reference ID: 11218908711_Williams

Reference ID: 11218908711_Williams

Submission Date and Time: 12/12/2019 9:05:46 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: John Williams - Department of Geography, University of Wisconsin-Madison

Additional authors: Jessica Blois, University of California-Merced; Gabriel J Bowen, University of Utah; Edward Davis, University of Oregon; Eric C. Grimm, University of Minnesota; Rob Guralnick, University of Florida; Steven M. Holland, University of Georgia; Eric C.

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

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

Archaeology; Geology; Global Change; Paleontology; Paleobiology and Paleoecology; Paleoclimatology; Stratigraphy

Title of Response

Building Open, Linked, and Community-Curated Data Ecosystems for the Past Earth-Life-Human System

Abstract

Understanding how biological and human systems responded to and affected past environmental change requires the assembly of a massively dispersed and heterogeneous site network of paleoenvironmental, paleoecological, and archaeological data, with many kinds of measurements generated by many researchers working across many disciplines and spatiotemporal domains. Their assembly and integrated understanding is beyond the reach of any single research team or discipline. In response, an open ecosystem is emerging of multiple community-curated data resources, each linked to networks of data generators, synthesizers, and stewards. Key needs and opportunities: 1) building data systems that seamlessly transfer data and metadata from point of collection in field or lab to community paleodata resources and analytical pipelines; 2) on-going integration of community paleodata resources with each other and domain-agnostic services such as DOI minting and ORCID identifiers; 3) full integration of data provenancing and microattribution into analytical and publication pipelines, to enhance reproducibility and better recognize data generators; 4) sustained support models for long-lived community data resources and software systems; 5) integrated training programs at the intersection of paleoscience and data-science, targeted to undergraduates, graduate students, and the current workforce.

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.

Our understanding of the interactions among past global climate changes, species and ecosystem dynamics, and the worldwide expansion of human societies is founded upon massively long-tail data: collected by many researchers from many disciplines (e.g. paleoclimatology, paleoecology, paleontology, Stratigraphy, genomics, geochemistry, geochronology, geography, archaeology), from many geographic locations, and representing many kinds of measurements. These data are used e.g. to test and improve the Earth system models used for 21st-century climate projections, assess the sensitivity of the Earth system to greenhouse gases and other climate forcings, study the responses of species distributions and diversity to novel and no-analog climates, understand the drivers of species extinctions, reconstruct the size and effects of a growing anthropogenic land use footprint, and study the processes governing societal resilience to past environmental change. Understanding these processes requires the on-going production, aggregation, and improvement of paleoclimatic, paleoecological, and archaeological data from sites distributed around the world and spanning timescales from the last few centuries to (for geo/bio data) many millions of years. Primary data collection usually requires labor-intensive fieldwork (to collect samples) and laboratory work (e.g. geochemical analyses, paleontological or archaeological morphometrics), while secondary data synthesis requires painstaking compilation of many records across many sites, QA/QC, data harmonization, and statistical modeling. This integrative research faces multiple cross-disciplinary challenges, all associated with paleodata being long-tail data for which the ‘last mile’ problem is paramount. 1) Current data pipelines from point of collection to community data resources are

inefficient and rely heavily upon manual labor, which hinders scalability and open data sharing by data generators. 2) The collective expertise needed to collect, assemble, use, and govern these data is distributed widely among scientists and disciplines. 3) Differing semantic and ontological frameworks within and among disciplines and community data resources. 4) Lack of data provenancing and microattribution, which hinders reproducibility and causes the intellectual work of primary data generators to be undervalued. 5) Need for informatics solutions to support active data stewardship and curation by experts, e.g. annotation services to support user-centered data corrections, augmentations, and interlinking. 6) A multi-decadal longevity of community paleodata resources, combined with increasing informatics complexity, that requires new models of long-term sustained support for community cyberinfrastructure resources and their support teams of developers with crossover expertise in the geosciences. 7) Scarcity of integrated training in the paleosciences and data sciences, with needs for undergraduate, graduate, and professional refresher training. 8) Better integration of disciplinary domain experts with library and information scientists to curate and preserve data and better integrate data with publishing workflows.

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?

In response to the above challenges, a grass-roots ecosystem is emerging of open, community-curated data resources, data standards, and open software, that is evolving to cover previously unsupported data types, interlink among resources, and create new analytical capacity. The overarching need for the next generation of Data-Oriented CI is to continue to build scientist-centered informatics systems and infrastructure capacity to support expert-curated data systems that support data contributions, curation, analysis, integration, and governance by widely dispersed networks of scientists. These scientific communities can be organized into three overlapping communities of data generators, data synthesizers, and data stewards, all of whom are contributing and receiving services from open community-curated data resources (CCDRs, <http://pastglobalchanges.org/products/pages-magazine/12705-26-2-build-harness-open-paleodata>). Specific recommendations include: 1) Supporting development of data systems that facilitate data and metadata tracking from point of collection in the field or laboratory to community data resources and analytical pipelines. 2) Improved support for data provenancing and microattribution via e.g. A) supporting free DOI minting services, or ongoing support for third-party DOI minting services like DataCite, B) removing artificial limits to citation counts in peer-reviewed publications, e.g. via Data Citation sections that would be citation-indexed and

could link to thousands of individual data records. 3) On-going support for community adoption and extension of common data standards and services such as ORCID, FundRef, IGSNs, DarwinCore, and LiPD. 4) Paleodata systems must continue to evolve from archival systems-of-record to systems of engagement, in which the primary emphasis is providing living, state-of-the field data, which are regularly updated by experts. This transformation in turn requires implementation and integration of services of user authentication, change-modification tracking, and data annotation. 5) Sustained support models for community-curated data resources with proven track records of close engagement with scientific communities of data generators, synthesizers, stewards, and memory institutions (libraries, archives, museums) as partners in stewardship. One source of conservatism in data generators is the risk of investing effort in open data systems that disappear after a few years. Moreover, few developers have the necessary crossover skills in geoscience and data science. Lack of sustained support for community data infrastructure leads to high turnover in these mission-critical research positions. 6) A new generation of training programs that operate at the intersection of the data sciences and traditional knowledge domains in the paleosciences targeted to at least three levels: 1) undergraduate, 2) graduate, and 3) refresher courses targeted to current professionals. A related need is to train peer reviewers and journal editors about the emerging state of the field and how to put FAIR principles into practice. These recommendations align well with the following NSF 10 Big Ideas: Growing Convergence Research; Future of Work at the Human-Technology Frontier; Rules of Life; and Harnessing the Data Revolution.

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.

Because paleodata are long-tail data generated by networks of scientists with distributed expertise, the critical challenges to building larger data ecosystems all lie at the intersection between informatics, social sciences, and user-centered design. The emerging community-curated data resources (CCDRs) are inherently socio-technical institutions (Williams et al. 2018, PAGES Magazine, p. 50) with close connections to their user communities of data generators, synthesizers, and stewards. Hence, all cyberinfrastructure development recommendations need to be closely integrated with workforce and governance needs. See our answer to Question 2 for a fuller set of recommendations.

-- End Submission --