

Reference ID: 11102301318_Morris

Reference ID: 11102301318_Morris

Submission Date and Time: 10/29/2019 3:53:22 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: George Morris - Dartmouth College

Additional authors: None

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

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

data science/information technology

Title of Response

Collections-based research

Abstract

In recent years, and with the support of NSF and other organizations, concrete advancements have been made in the infrastructure supporting digitization, aggregation, and sharing for online collections and specimen databases. Making best use of this collection infrastructure now requires expanding its scope to consider tools and processes for cross-disciplinary curation and use of data. I proposed NSF invests to:

(1) address the technology, expertise, and application gaps in today's research collections lifecycle; and (2) provide a modular, scalable framework to quickly build working collections that are easily publishable to extended members of the research and education community, initially focusing on projects and research audiences in under served communities. By establishing Collections Informatics capabilities through NSF solicitations, individuals and institutions working in a variety of disciplines (ranging from STEM fields to cultural heritage) will gain access to resources that facilitate the identification, selection, description, organization, curation, preservation, and presentation of digital and physical artifacts used in research, teaching, or outreach. Such projects can support the hiring of application specialists to develop workflows facilitating collections development and use. Such developers will create new open-source interfaces and storage systems to manage working collections assembled from available physical or digital. Finally, such investment will help researchers and educators develop collections-based grant applications and publish data to scholarly and general audiences.

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.

To test a theory about the relationships between the fourth metatarsal bone and the development of bipedalism in humans. Many data sources are available, though they are quite different: a paleoanthropology lab holds physical disarticulated skeletons, museums around the world hold thousands of skeletonized human remains in various states of physical preservation or digital imaging, and this particular study also requires MRIs and videos of human subjects walking. A way to digitize skeletal collections, capture locally-created MRI and video data, and ingest imagery from outside museums into a working collection. While the researcher still takes the lead on identifying appropriate specimens and contacting collection stewards, staff work with the researcher to develop ontologies for both basic descriptive metadata about the bones (age, geotagged origin, etc. mapped to Darwin Core terms) and specific metadata describing the aspects of the bones critical to the researcher's study (a custom ontology). NCIH then creates a new working collection of binary files to contain 3D imagery of the bones, describes it with a subset of the existing Darwin Core metadata form, adds a new metadata form to capture the researcher's custom ontology, and imports a generic research video module preconfigured with appropriate descriptions and relationships. While the researcher collects and analyzes data, staff continue to work with them to think about how to analyze and use the data in the working collection. In this case the analysis is visual and would be best accomplished by handling physical specimens, so staff facilitates transformation of the stored 3D imagery to a format suitable for 3D printing. The derivative printable files are also added to the working collection. When the researcher has examined all the data they find that yes, the shape of the fourth metatarsal bone significantly influences the gait of modern humans and, when applied to historic remains, probably tells us a great deal about how bipedalism developed over time. When the researcher publishes their research they

both export the entire data set for preservation in a data repository and send newly-created metadata files back to the source archives for their own records. Rather than taking this as a stopping point, staff and the researcher realize that they have all the components of an engaging classroom experience sitting in the working collection for on premise online learning.

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?

Our long-term vision is to develop and deploy a trustworthy collections informatics strategy for institutions of varying sizes and across diverse disciplines that drives innovative teaching, learning and research. To realize this vision, we have developed an agenda that helps build a new data-oriented resource with capabilities needed to access, manage, analyze, and present collections in a variety of contexts. In addition to supporting current collections-based research processes, we propose intervening in the current workflow at four key points that prepare collections for both better organization within disciplinary research and new interdisciplinary applications:

1. Planning – Though digitization processes are well developed, many researchers and institutions have collections of specimens that they have not digitized and shared. In some cases they may not have the resources to digitize their specimens or produce metadata assets describing them; in other cases, an institution holding the collection may not realize that it has research value in other disciplines or beyond the study for which the collection was assembled. Investing in expert staff to offer consulting services for collection owners to help assess the cross-disciplinary value of their collections, work on digitization plans, and identify sources of funding to support digitization.
2. Collection and Curation – NSF should develop a Working Collections Tool for researchers who want to select and curate assets sourced from both their own and external collections for use in a particular research study. Providing a tool designed for consumption of shared assets and addition of new descriptive metadata closes the reuse loop that has been left open by current collection aggregators. In addition to simplifying reuse by allowing individual assets from many sources to be merged into a single collection, NSF Working Collections Tool provides researchers an environment to organize heterogeneous assets, manage metadata, and track references in a flexible way that replaces the requirement for one-off databases.
3. Domain-Specific Analysis – While assets may have research value in different disciplines, actually reusing those assets for research is complicated when they were only cataloged with metadata tied to the assets' original use. NSF Working Collections Tool treats metadata as both modular and inherently faceted. NSF helps researchers conceptualize both basic descriptive ontologies meant for cross-disciplinary discovery

and additional metadata required to describe an asset in domain-specific terms. As many discipline-specific metadata records can be added as are required to describe the asset in all relevant research contexts. 4. Presentation and Publication – NSF supports researchers as they fulfill their obligations to release datasets for reproducibility and validation by making assets available through machine-searchable APIs and human-readable web interfaces. However, the collection-based research model also creates a potentially transformative opportunity to combine real research assets with didactic text for public presentation. Presentation to general audiences, as opposed to publication to expert audiences, is simplified by NCIH's integration of the Working Collections Tool with platforms such as Scalar and Omeka. Beyond providing technology, presentations are also supported by NSF funded learning designers and consultants who assist researchers in making their work clear and understandable to a variety of audiences from the general public to K-12 and higher education students and on to broader research and business communities.

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.

At a practical level, the problem is no longer a lack of information available to researchers; the problem is that there is so much information available that the data outstrips our ability to organize it. Even when a specific specimen is discoverable it can be difficult for researchers to pull that specimen out of massive online aggregators and into a local system where it can be utilized in their own research program. Any additional information or metadata added via this new research also has little chance of being linked back to either the aggregator or the original source collection; new knowledge created to support research ends up either buried in a disconnected dataset or is simply lost. Supporting interdisciplinary collections requires more than just new software, it requires rethinking cataloging practices and roles, developing new research workflows, consulting with disciplinary experts, and exploring the possibilities created by opening specialized repositories to general audiences—in short, reconsidering and expanding the definition of the full lifecycle of research data across the disciplines (Punzalan and Kriesberg 2017). Our proposal to create the Northeast Collections Informatics Hub (NCIH) engages a full lifecycle approach that meets the needs of both disciplinary and interdisciplinary investigators and ensures interconnectivity between collections, teaching, and research.

-- End Submission --