

Reference ID: 11227098570_Bates

Reference ID: 11227098570_Bates

Submission Date and Time: 12/16/2019 5:07:50 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 Bates - Natural Science Collections Alliance

Additional authors: None

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

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

Biology; natural science collections; biodiversity; ornithology; evolution

Title of Response

Cyber-infrastructure and the extended specimen concept for biodiversity research

Abstract

The North American museum community has committed to digitizing data from their collections and making these data accessible through public data portals. They have had support from NSF in this endeavor, but work is just beginning because the community continues to gather and generate new types of data that must be linked to the original specimens for maximum scientific benefit and use.

Millions of specimens have now been digitized, but millions more remain. Long-term access to these data requires new cyber-infrastructure associated with data storage, attribution, and access for the broad network of scientific collections across the country, now and into the future. The benefit is time series biological diversity data that support novel research, development, and education that promotes economic growth while also improving public health, environmental stewardship, and our national security.

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.

Digitized collections data are used in ever-expanding ways to address research questions in the biological sciences. One recent example of the novel research value of these collections data begins with new work documenting a loss of 3 billion individual birds across over 500 species over the last 40 years; based on randomized continent-wide road survey data. Forty years of digitized specimens (>70,000 specimens) from data documenting migrating birds killed from window strikes are archived in collections at The Field Museum of Natural History. These data have shown that body size has gotten smaller in a subset of 52 of these species (changes likely related to changing climate). This research used data on large samples sizes from a digitized collection to address a broad pattern of change in birds. These data from this vouchered time series are still available for additional and different research. The numbers of vertebrate samples in the nation's collections are dwarfed by the herbarium specimens (mostly digitized) and the insect and invertebrate specimens (mostly undigitized). These collections are important resources that document global biodiversity. The example above typifies densely sampled museum collections useful for questions that go far beyond their ongoing value for systematics and taxonomy. Effectively expanding the digitization of data in collections for all biodiversity from whales to microbes is essential for understanding how environmental change effects ecosystems and regions through time. Cyber-infrastructure commitments to collections-based data are a national and global imperative required to understand a world with rapidly changing climate. There is a somewhat analogous cyber-infrastructure to what is needed, which is the already existing databases that comprise GenBank at the National Center for Biotechnology Informatics. If we want to do science related to the relationship between genotype and phenotypes in changing environments such as the Arctic (two of NSF's 10 big ideas), then it is essential to more effectively develop and maintain the cyber-infrastructure associated with these digitized collections data.

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?

As recently outlined in the Biological Collections Network (BCoN) report: Extending U.S. Biodiversity Collections to Promote Research and Education, the cyber-infrastructure of the collections-based community must develop to effectively serve diverse and distributed biodiversity collections across the country. This is different than astronomy where a new telescope can be developed to provide a point source to gather massive amounts of new data. No new telescope can be built and then pointed at the earth to collect biodiversity data, biodiversity occurs in all corners of the planet and the expertise to document and describe it is similarly dispersed as are the necessary collections of specimens for this research. An additional reason for this decentralization is the continued need to train new diverse generations of local and regional experts across the country. The most-speciose groups of non-bacterial organisms, insects and invertebrates are still poorly digitized because they present significant challenges by virtue of physical size and variety of manners in which they are preserved. What the scientific community needs with respect to collections are digital solutions allowing the assembly of and common access to the best and most comprehensive information about each specimen, so that it can be effectively compared and contrasted with related specimens by a distributed network of experts who can develop taxonomies and phylogenetic trees, which can then be incorporated into an ever-expanding number of other research questions. This will open the door for research in many other scientific fields from ecology to neurobiology. Cyber-infrastructure for improving data attribution and connectivity is improving, but the challenges associated with this necessarily distributed network of networks need to be overcome with visionary cyber-informatic approaches. Massive collections-based data sets of the evolutionary relationships of all biodiversity will be the source of an essentially endless set of forms that are interacting with humans and the rest of biodiversity now and into the future.

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.

As mentioned above, the museum community is a network of networks, subject to the budget issues facing all academic institutions. Centralization of visionary cyber-infrastructure for museum collection databases can help unify these networks to provide bigger data sets for study, it also can serve to help establish the value of and need for local expertise, collections, and training. iDigBio is a developed model that demonstrates this can work, but long-term success depends on stability. There is a huge opportunity to diversify the workforce necessary to undertake the effort to digitize and curate collections-based biodiversity data, but the solution must be one that allows for a distributed network to serve and receive attribution for data, which also being able to receive feedback to curate the data.

Response to NSF 20-015, *Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research*

Reference ID: 11227098570_Bates

-- End Submission --