

Reference ID: 11103983518\_Carstens

---

**Reference ID:** 11103983518\_Carstens

**Submission Date and Time:** 10/30/2019 8:17:42 AM

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: Bryan Carstens - The Ohio State University

**Additional authors:** None

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

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

Evolutionary biology

#### **Title of Response**

Cyberinfrastructure for specimen-based research.

#### **Abstract**

Cyber infrastructure challenges include the need for connection of data (environmental, genetic, behavioral, morphological) with the physical specimens that form the basis of organismal biology.

**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.

One challenge is created by specimen based research in the biological sciences. While there are an increasing number of databases for morphological, genetic, behavioral, or environmental data it can be difficult to connect specimens across these databases and access the relevant metadata. A second challenge relates to the analytical pipelines and scripts used in data analysis. These are often published as supplemental material by journals or on sites such as GitHub, but the permanence of these objects is questionable. Will researchers be able to find scripts published in 2010 in 2110? It's questionable. A third challenge relates to the databasing of raw sequence reads. While resources such as the Short Read Archive exist, it appears that many publications do not include access to the sequence offloads.

**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?

An investment in databasing software and hardware could address the first challenge. Museums house the physical specimens that enable organismal biology, but other types of data are hosted by entities ranging from non-profit groups to governmental agencies. A better integration of the databasing infrastructure with the museums would enable this integration, but require a tremendous investment of resources to make the back end available and accessible to institutions of every size. The second challenge could be addressed by creating a publicly supported code repository. If there was a considerable investment in museum cyber infrastructure, sequence offloads could be accessed with other specimen data.

**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.

*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: 11103983518\_Carstens

---

Any solutions must be global in their design and application. Particularly for the biodiversity sciences, cyber infrastructure needs are global because information sharing across borders and institutions is the best way to investigate species diversity.

-- End Submission --