

Reference ID: 11227054073\_Bart

---

**Reference ID:** 11227054073\_Bart

**Submission Date and Time:** 12/16/2019 4:47:56 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: Henry Bart - Tulane University

**Additional authors:** Yasin Bakış Tulane University Biodiversity Research Institute ybakis@tulane.edu

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

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

Biology; Biodiversity Informatics; Computer Science; Oceanography; Geospatial Science; Meteorology; Environmental Science

#### **Title of Response**

Cyberinfrastructure for Global Oceanic Ecosystem Forecasting

#### **Abstract**

Oceans are the main life support system on Planet Earth and protecting this system is of vital interest to humankind. However, our understanding of the structure, function, and interdependencies of the living and environmental components of oceanic ecosystems is limited because the myriad databases of

oceanographic information have not coalesced into an integrated and robust cyberinfrastructure ecosystem that can drive new thinking and transformative discoveries for sustaining Earth's oceanic ecosystems. Integrating databases of marine biodiversity information (e.g., the Ocean Biogeographic Information System), oceanic physical and environmental data (e.g., World Ocean Data) and global weather and climate data (e.g., World Weather Records Clearing house), will permit scientists to relate biodiversity trends to environmental and climatic patterns in different marine ecosystems and predictively model how marine ecosystems will respond to warming oceans, fishing related removal of apex predators, and expansion of marine protected areas to name just a few research areas that could be explored. We propose to unite relevant data on global ocean ecosystems into a robust CI/web interface that will permit scientists to search, filter, map and download integrated ocean data, including summary statistics, reports and graphs that will enable scientists to perform Data-Intensive Oceanographic research.

**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.

Oceans encompass the Earth's largest ecosystems and its most important life-support systems. They make up 97 percent of the Earth's water. Oceans also play an important role reducing climate change impacts by absorbing carbon dioxide from the atmosphere. The diversity and productivity of the world's oceans is a vital interest for humankind. Our economy and our survival require healthy oceans. Knowledge and understanding of the host of services healthy oceans perform for humans and Earth ecosystems broadly is limited, but could be enhanced by integrating data across the myriad data portals and databases of ocean information. A host of ocean-related data portals have come online in recent years (e.g., FishNet2, OBIS, GOOS, ODIS, EOSDIS NASA's Earth Observing System Data and Information System, etc) which provide data such as specimens of fish and other marine organisms in biodiversity collections, observational biodiversity records, and physical factors such as temperature, salinity, sea surface topography, ocean circulation, winds, heat exchange between the ocean and atmosphere, interaction of solar radiation with the ocean and so on. These platforms serve a wide array of interdisciplinary research at the boundaries of Biology, Geography, Chemistry, Meteorology, and Oceanography (we need to mention about cross disciplinary aspects of the projects in this section). Existing ocean data portals serve the needs of scientists focusing on topics such as assessing/sustaining stocks of commercially important marine species, climate change impact on ocean currents, impacts of ocean acidification on marine organisms, conservation of imperiled marine species, and documenting currently unrecognized marine biodiversity. Existing portals also offer many benefits to the public, such as increasing public understanding of ocean resources and marine environments past and present. However, the lack of integration of data in these systems complicates the process of posing questions across the different data platforms. Questions that could be answered if existing ocean data portals were integrated and made interoperable: o What are the consequences of climate change on

marine food webs and areas of high species diversity such as coral reefs? o How do seemingly disparate environmental factors such as ocean strontium ion concentration and bicarbonate ion concentration interact to produce sea shells of marine organisms, and how can this relationship be used to moderate the impacts of ocean acidification? o How would doubling of Marine Protected Areas do to reverse declines of stocks of commercially important marine fish species and where should the new reserves be created?

**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?

The number of available sources of oceanic data, and the abundance, spatial, temporal, and bathymetric extents of records is quite satisfactory. However, the task of discovering and gathering data from the various platforms and formatting the data for analysis can be challenging. There is a need for a system that presents data to researchers in an integrated form and facilitates the process of data analysis, by providing summary statistics, modeling tools, forecasts with explanatory charts, graphs and maps for visualizing data. Such a system would benefit not only scientists, but decision makers or even the general public. We propose to build a "system of systems" cyberinfrastructure for merging from disparate ocean data platforms based on temporal, geospatial and bathymetric overlap of the data. This will allow scientists to seamlessly gather data from any of these overlapping zones for analysis. An information retrieval server for big data analysis will be built for this purpose to collect the data from data providers, to store the data and make the data publicly accessible with a search interface. Accomplishing this will require a Relational High Performance Database system that will provide a fast querying of the harvested datasets on a parallel computing platform with fastest high-end server drives, large memory units and multiple cores with fast CPUs. The information system will harvest data from the data providers periodically and update its data cache, provide its users with the most up-to-date information. An information retrieval server for big data analysis will be built for this purpose to collect the data from data providers, to store the data, make the data publicly accessible, with a search interface and perform defined analyses along with production of reports, maps, charts and graphs. The information retrieval system will provide all types of ocean related data to be filtered by the user through a web interface and/or R package and create downloadable search results. Stable standard reference formats will be used for interoperability – such as Darwin Core. The analysis part of the system will create automated reports, generate maps and graphs based on user's search criteria. The interactive maps will be generated in the web interface by using a JavaScript based search interface

showing the accumulation of the data based on selected criteria. Similarly, graphs and charts will be generated within the web interface in the statistics page based on selected criteria. The data collected from all those different providers will create very large files for download and that will be computationally costly to analyze. Therefore, an R package will be developed as a second tool to practice searches, creating reports, graphs and maps and additionally performing certain analysis. One can simply install the R package on to their workstations/high performance computers or their desktop/laptop computers and perform several different types of analysis that are included in the package from simple statistics to Big Data analysis. Finally retrieve the results as reports, spreadsheets, data files or image or vector images on to their local machines.

**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.

The project requires composition/collaboration of several institutions/organizations that are willing to provide data for the platform. In addition to TUBRI, we would invite relevant scientists representing OBIS (TUBRI is a member of the OBIS Steering Group), the World Registry of Marine Species, NOAA, NASA, the U.S. Geological Service, and from equivalent international scientific organization to participate in the project. Administrative officials of these organizations would be invited to serve on a steering committee to provide project oversight, advise the scientific participants on project management issues and emerging scientific agendas that should be tracked and responded to. A technical committee would also be required for advice on designing, maintaining and sustaining the cyberinfrastructure. Finally, a committee of public outreach and education professionals would be engaged to design education, outreach and public dissemination resources, and for communicating scientific results on popular media platforms.

-- End Submission --