Data Management Plan

Data Policy Compliance: The goal of our data management plan is to ensure that all data and products are archived, shared, and accessible for the long term. Here, we present a data management plan that is compliant with the NSF Division of Ocean Sciences Sample and Data Policy (NSF 17-037). We will cooperate and collaborate with all appropriate data repositories, particularly the Biological and Chemical Oceanography Data Management Office (BCO-DMO; http://www.bco-dmo.org/), and adhere to the overall data sharing philosophy (section II). Our cruise data will be archived at the Marine Geoscience Data System (MGDS; www.marine-geo.org) as well as the Rolling Deck 2 Repository (R2R, http://www.rvdata.us).

Types of data*:* The project will ***produce*** five types of data:

(1) vent fluid and seawater samples
(2) metagenomic sequencing data
(3) cell and virus count data
(4) Scripts/code for bioinformatics data analysis
(5) Curricular materials in the form of syllabi and laboratory protocols

Additionally, the project will ***collect*** publicly available data derived from the OOI Regional Cabled Array in the form of chemistry data collected by the RAS instrument deployed at Tiny Towers in the International District at Axial Seamount. All OOI data are made publicly available through the OOI Data portal; the Regional Cabled Array has a project-specific data portal at interactiveoceans.washington.edu.

Data Standards: We will follow the metadata guides and references produced by the Marine Metadata Interoperability project (marinemetadata.org). All genomic sample collection will conform to the Genomic Standards Consortium's (GSC) minimum information about a genome sequence (MIGS) or the minimum information about a metagenome sequence (MIMS). We will also follow the principles articulated by Goldstein et al. (2003, Chemical Geology 202, 1-4) regarding open reporting of analytical metadata. Calibration standards, standard reference materials, internal standards used to quantify analytical precision, and other items analyzed as part of good analytical practice will be fully described in relevant publications. Ultimately, the PIs are responsible for the appropriate level of QA/QC that is consistent with best practices in their respective field. In some cases, there is no established protocol, and as such the PIs will endeavor to provide the data in a manner that enables other investigators to easily access the final calibrated data, as well as replicate the conditions under which the data were collected. All code generated by the project will be annotated and shared publicly via a lab GitHub page (https://github.com/carleton-spacehogs). Curricular materials will be shared on the Carleton College web servers and/or informal science learning web servers (such as *howtosmile*) when appropriate.

Data Access and Sharing: We will make all data publicly available no later than two years after the data are collected, per NSF policy. All data will be provided to BCO-DMO and linked to core cruise metadata in MGDS and R2R. Raw and assembled and annotated sequences will be made publicly available as well as archived through the genome repositories GenBank (NCBI; http://www.ncbi.nlm.nih.gov/) and Joint Genome Institute Integrated Microbial Genomes (JGI/IMG; http://img.jgi.doe.gov/) server. Both of these databases are federally supported and stable and can be linked through BCO-DMO. We will use a master sample/metadata spreadsheet and database that will allow us to rapidly cross-reference samples to facilitate of our publication efforts. This will also assist others who might be interested in these samples or data, including marine microbiologists and geochemists. Moreover, if data products are yielded by third parties integral to the successful completion of the proposed project, we will ensure that these too are made publicly available along the same timeline, via BCO-DMO.

Data Archiving: As noted above, all data will be deposited into publicly accessible archives, including BCO-DMO, JGI IMG, the OOI Data portal, and NCBI. Final archive will be ensured by depositing this data into the BCO-DMO within two years of acquisition, and where stable links to data housed in other repositories (i.e., NCBI) can be made. The PI will upload all cruise data and sequencing data, and will take primary responsibility for complete archiving. For data storage, the PI owns a total of 70TB of usable space on a new NetApp shelf (24 6TB drives), physically located on campus and dedicated to academic/research. All data created by this project will reside on this machine. All data generated by proposal research will be archived. The NetApp server is regularly backed up via Snapshot technology. The PI and appropriate technical staff will be responsible for the management of the data, including physical storage, archiving, and submission to public databases. Management of the NetApp server is done by members of the ITS staff and the computer science department technical director in consultation with the PI. The PI grants authorization to access to this server. The PI is also the responsible party for artifacts released online via repositories such as GitHub. Associated metadata for every sample will also be submitted to BCO-DMO. To the extent possible, all products will also be described in peer-reviewed literature, to ensure public dissemination and long-term accessibility beyond the scope of this grant.

Policies for access and sharing*:* The following specific steps will be taken to protect privacy, confidentiality, and security of the data:

- Access to the server used to store the project data will require both explicit authorizations from the PI as well as password identification by the user.
- Authorization to access the server storing the project data will only be granted to personnel that have a legitimate need to access it.
- No data with any identifying information will be provided to anyone without explicit authorization and a legitimate need to access the data.
- Appropriate technical measures to ensure security of the system will be taken. This includes, but is not limited to: restrictive file/directory permissions, security measures for databases behind firewalls, user authentication, patches updated regularly, etc.
- Any real data collected will be stored and shared in accordance with regulations set forth by the institution providing the data.

Metadata and Documentation: All associated metadata will be deposited along with the data in the appropriate databases (see previous for descriptions of public and private archives planned for each dataset).

Status of Funding Support for Data Management: The PI has access to institutional data storage facilities and experience managing the types of data we will generate and do not require additional funding for this activity. Federally-funded data facilities exist for each of the data products we propose to generate. If there are nominal costs for these facilities to handle our data we will cover those costs from the awarded funds to ensure long-term data preservation and access.