Ecosystem dynamics of Western Pacific hydrothermal vent communities associated with polymetallic sulfide deposits

Data Management Plan

a. Data Generation Activities: The research products generated in the proposed effort include: i.) updated bathymetric maps and geological interpretations, ii.) biological samples; iii.) genetic and transcriptomic data; iv.) *in situ* and shoreside geochemical measurements; v.) photographic archive of ROV and shipboard activities; vi.) GIS products overlaying imagery and geochemical measurements; and vii.) physiological rate measurements. Shipboard data generation activities include the initial processing of ROV-based sonar data, CTD data, ISMS geochemical measurements, photographic surveys, and raw physiological measurements, as well as the associated sampling and station metadata. Each of these datasets will be additionally processed shoreside resulting in bathymetric maps with geological interpretations (i.), calibrated *in situ* geochemical and environmental datasets (iv.), photographic and GIS products (v.,vi.), and physiological rates (vii.). In addition, biological samples of microbial biomass and macro/megafaunal tissue, including preserved voucher specimens, will be collected shipboard (ii.).
Finally, datasets generated via shoreside activities will include measurements of water geochemistry via Seewald's gastight bottle samples (iv.) and sequencing of transcriptomes from experimental animals (iii.).
b. Roles and Responsibilities: Dr. Ferrini, who is a leader in marine geoscience data management, will function as the primary Data Manager and will ensure that data archiving and metadata standards and obligations are met. She will also be responsible for processing all bathymetric data and for depositing final geology/geophysical data products (including maps, grids, interpretations, and the results of quantitative bathymetric change analysis). Dr. Fisher will be responsible for photomosaic creation and for depositing all photographs, the photomosaics, and the integrated GIS products. Dr. Girguis will be responsible for the deposition of all geochemical data, as well as any microbial samples, which includes frozen material for future DNA/RNA/protein sequencing, as well as any resulting sequences that are generated during the analyses. Dr. Beinart will be responsible for the deposition of all animal data, including tissues, respiration data, sequencing data, voucher specimens, and any images. In the interest of fostering "open design" among the oceanographic community, any sampling technologies or methodologies that are developed during the course of this expedition will be made publicly available as well.
c. **In-project data management**: The data will be placed on common, easy to use web-based servers for rapid dissemination (e.g. Google Drive). 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 ecologists, marine microbiologists, geochemists and geophysicists, and even astrobiologists. 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.
Biological samples and data: Voucher specimens will be deposited in the Harvard Museum of Natural History's Collection (https://mcz.harvard.edu), which houses a broad Pacific deep-sea

collection. Additionally, all biological samples collected during the cruise will be sub-sampled and photodocumented for delivery to the Ocean Genome Legacy (OGL), a facility funded to archive biological material for any and all investigators who require the materials. All contextual data (e.g. sample collection site, weights, etc.) will be provided to the OGL and the National Science Foundation's Biological and Chemical Oceanography – Data Management Office (BCO-DMO).

Sequence data will be stored indefinitely on Girguis laboratory computers, as well as on the Harvard Research Computing Cluster, Odyssey, which offers over 2.5 Petabytes of raw storage for 16 its users and includes both daily checkpoints and off-site backups of stored data. Associated metadata for every sample will also be submitted to the BCO-DMO. Within two years of acquisition, assembled and annotated transcriptomic sequences will be made publicly available as well as archived through the genome repositories GenBank (NCBI) and the MG-RAST Metagenomics Analysis Server. Metadata associated with the transcriptomes will be stored on MG-RAST and the BCO-DMO. In addition, all data will be stored indefinitely on Girguis laboratory computers, as well as on Odyssey (see above for details). Physiological rate data collected during shipboard experiments will be made available, after publication, to other researchers upon request. This data will be stored indefinitely on Girguis laboratory computers, as well as on Odyssey. Final archive will be ensured by depositing this data into the BCO-DMO within two years of acquisition.

Photomosaic, GIS and bathymetric products: Bathymetric data and digital still images will be archived at the Marine Geoscience Data System (MGDS; www.marine-geo.org), part of a publicly accessible data facility funded by the National Science Foundation. Additionally, digital still images will be archived at PSU in the Fisher lab on our 24 terabyte internally redundant data storage system which is housed in a climate controlled room, in a separate building managed by the College of Science. The mosaics and GIS products will be stored at PSU similarly during analysis and deposited in the MGDS database within 2 years of acquisition of the raw data.

Geochemistry: Raw and calibrated geochemical data generated by the *in situ* mass spectrometer or shoreside measurements, with associated metadata, will be stored indefinitely on Girguis and Seewald laboratory computers. Final archive will be ensured by depositing this data into the MGDS and VENTDB (www.earthchem.org/ventdb) within two years of acquisition.

d. 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).

e. Data quality: 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. Dr. Ferrini has been a major proponent of ensuring that archived data are of the appropriate quality, and her inclusion in this proposal ensures that we will collectively hold to a high standard.

f. Status of Funding Support for Data Management: Each of the PIs have 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 nominal costs for these facilities to handle our data we will cover those costs from discretionary funds to ensure long-term data preservation and access.