Data management plan

Data Policy Compliance: The goal of our data management plan is to insure 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 11-060). 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 at sea sampling efforts and experiments will be linked to Dr. Geoff Wheat's research cruise supported by MG&G, and all data from that cruise 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). Finally, because some components of this work are directly leveraged with the NSF Science and Technology Center C-DEBI, we will insure that the C-DEBI data management plan is followed where appropriate as well.

A) Data Types Generated During this Effort. The research data and products generated in the proposed effort include 1) crustal fluid and seawater samples; 2) geochemical measurements; 3) microbial activity measurements; 4) microscope images; and 5) genomic and transcriptomic sequencing data. Each of these datasets will be processed, resulting in calibrated geochemical and biological datasets.

B) 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.

C) 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, geochemists and geophysicists. 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.

D) Data Archiving. As noted above, all data will be deposited into publicly accessible archives, such as BCO-DMO, JGI IMG, 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. Additionally, raw and processed sequence, geochemical, and imaging data will be stored indefinitely on Huber, Pearson, and Girguis laboratory computers, as well as on the Harvard Research Computing Cluster, Odyssey, which offers over 2.5 Petabytes of raw storage for its users and includes both daily checkpoints and off-site backups of stored data. At MBL, long-term storage systems consist of three server and storage arrays with a total usable space of approximately 230T. These systems provide over 150T of tape-backed "permanent" storage. Associated metadata for every sample will also be submitted to the BCO-DMO. For data without national repositories, such as microbial activity and microscopic image data, we will work with BCO-DMO to make this data archived as well, which they are willing to do. 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.

E) 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).

G) 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 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.