Data management plan

Peter R. Girguis (OEB, Harvard Univ.)

DIMENSIONS: COLLABORATIVE RESEARCH: The phylogenetic and functional diversity of extracellular electron transfer across all three domains of life

Data Policy Compliance: The research products generated in the proposed effort include 1) geochemical measurements of samples pre- and post-incubation; 2) genetic, genomic, and transcriptomic data from electroactive community experiments; and 3) physiological rate measurements. Here, we present a data management plan that is compliant with the NSF Sample and Data Policy (NSF 11-060).

Briefly, we will cooperate and collaborate with all appropriate data repositories and the Biological and Chemical Oceanography Data Management Office (BCO-DMO; www.bco-dmo.org) as appropriate, and we will adhere to the overall data sharing philosophy (section II). The details on the use of these repositories are presented below.

A) Samples and Data Collected During This Effort: (1) metadata on the collection of the samples used in this project, such as latitude and longitude, temperature, depth as appropriate, height above or below grade (for terrestrial samples), solar irradiance as appropriate, pH, and all the associated geochemical measurements outlined in the proposal; **(2)** experimental treatment data including electrode material and surface areas, electrode potential, and current density; **(3)** 16S and 18S rRNA gene sequences, as well as metagenomic and metatranscriptomic sequences, derived via next generation sequencing; **(4)** potential novel microbial isolates from the bioelectrochemical incubations; and **(5)** frozen and preserved samples resulting from these experiments, including inocula, colonized electrodes, and geochemical samples.

B) Roles and Responsibilities: The PI and CoIs named herein are each responsible for the deposition of any and all microbial 16S and 18S rRNA gene sequences, experimental treatment data and metadata, and geochemical data that result from these efforts. PI Girguis is responsible for overseeing the project members' compliance and will work with NSF to resolve any issues should they arise.

Within one year of collection, all 16S and 18S rRNA sequences will be made publicly available in the NCBI's GenBank non-redundant nucleotide database (which serves as the target for their BLAST search services and is a composite of both GenBank and the European Molecular Biology Laboratory (EMBL) database; http://www.ncbi.nlm.nih.gov). 16S and 18S rRNA sequences will also be deposited in the QIIME repository (http://qiita.microbio.me) and the SILVA next-generation sequencing (https://www.arb-silva.de/ngs) databases. Metagenomic and transcriptomic data will be publicly available in the MG-RAST repository (https://metagenomics.anl.gov). In addition, sequence data will be stored indefinitely on Girguis laboratory computers, as well as on the Harvard Research Computing cluster, Odyssey, which offers over 20 Petabytes of raw storage for its users and includes both daily checkpoints and off-site backups of stored data. Raw and calibrated geochemical data generated by the *in situ* mass spectrometer or other measurements obtained in the field will be stored indefinitely and made available on Harvard's Odyssey computing cluster, the Biological and Chemical Oceanography Data Management Office (BCO-DMO; www.bco-dmo.org), and VENTDB (www.earthchem.org/ventdb) within one year of acquisition. Biophysical data collected during experiments will be made available, after publication, to other researchers upon request. This data will be stored indefinitely on Girguis and El-Naggar laboratory computers, as well as on the Odyssey cluster at Harvard (space has been delegated for investigators to allow 20+ years archival of data) and the USC Digital Repository (USCDR), which provides fee-based consulting and services to help USC researchers meet NSF requirements. Services include digitization, cataloging, preservation, archiving, and long-term online access.

Additionally, all biological samples collected will be sub-sampled and photo-documented for delivery to the Ocean Genome Legacy (OGL, www.northeastern.edu/cos/marinescience/ogl), a facility funded to archive biological material for any and all investigators who require the materials (free of

charge). All contextual data (e.g., sample collection site, weights, etc.) will be provided to the OGL and the BCO-DMO.

Finally, in the interest of fostering "open design" among the scientific community, relevant sampling technologies or methodologies that are developed during the course of this project will be made publicly available through our websites and societies dedicated to microbiology and bioelectrochemistry (e.g., the International Society for Microbial Electrochemistry and Technology http://www.ismet.ugent.be).

Post-Project Data Management: Upon completion of this project or as dictated by NSF guidelines, whichever comes first, all data will be deposited into the aforementioned repositories, as well as placed on common, easy to use web-based servers for rapid dissemination (e.g., Google Drive™ or Dropbox™). 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 third parties generate data products integral to the successful completion of the proposed project, we will ensure that these too are made publicly available along the same timeline.

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.

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.