Data Management Plan – Thrash, USC

This proposal will involve the collection and analysis of diverse data types and accompanying metadata, including physical samples (whole seawater, filtered biomass, and cultured microorganisms), environmental metadata, microscopy images, and genomic sequences and their accompanying computational analyses/outputs). The goal of this data management plan is to ensure that all data and products are archived, shared and accessible for the long term. Physical samples and data products will be managed and disseminated according to published recommendations for best practices; ***we will emphasize open access publication of all data products, manuscripts, and analyses, and we will additionally adhere to strict metadata standards to ensure maximum data reuse and accessibility.*** Our data management plan is compliant with the NSF Division of Ocean Sciences Sample and Data Policy (NSF 11-060). We will collaborate with appropriate data repositories, particularly the Biological and Chemical Oceanography Data Management Office (BCO-DMO; http://www.bco-dmo.org/), and all data from the cruises will be archived at the Marine Geoscience Data System (MGDS; http://www.marine-geo.org/) as well as at the Rolling Deck 2 Repository (R2R; http://www.rvdata.us/).

The data that will be generated as part of this project include:

1. **Environmental metadata** – Sample site locations and descriptions, including geographic coordinates and sample collection dates.
2. **Biogeochemical measurements** – Conductivity, temperature, and density (CTD) data, as well as fluorometry, dissolved oxygen, salinity, pH, and concentrations of major inorganic nutrient ions (nitrate, nitrite, phosphate, ammonium, silicate).
3. **Diagnostic DNA barcodes for species IDs** – 16S rRNA gene amplicons generated via Sanger sequencing and associated with individual axenic cultures.
4. **rRNA gene sequence amplicons, metagenomics, metatranscriptomics, and isolate genome sequence datasets** – Environmental PCR products and size-selected DNA/RNA sequences sequenced on the Illumina HiSeq/MiSeq platforms, including raw sequence reads, quality-processed sequences, collections of genome bins, and contigs resulting from assembly.
5. **Axenic cultures** – bacterial and archaeal isolates obtained from cultivation experiments.
6. **Microscopy images** – scanning and transmission electron microscopy imagery of axenic cultures collected as part of their characterization.
7. **Flow cytometric data** – output from both the Millipore Guava and BD Influx FACS during cell quantification and sorting.
8. **Bioinformatics scripts** – executable files and documentation (commands, parameters, software versions) used for processing and analysis of nucleic acid datasets.
9. **Outputs of statistical analyses and bioinformatics workflows** - intermediate files, measures of statistical significance, and summarized/clustered data (e.g. Operational Taxonomic Units generated from Illumina metabarcoding reads)

Data management: The USC High Performance Computing Center will support pre-publication data management for this project including: (1) effective and secure access to data files, (2) support for core methods of analysis and (3) secure data storage. ***Security:*** Data will be stored on secure servers at each institute, and all data will be backed up on site using both physical tape drives and mirrored storage.

Data Standards: We will follow metadata guides and references produced by the Marine Metadata Interoperability Project (http://www.marinemetadata.org), "Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS)" specifications for metadata (http://wiki.gensc.org/index.php?title=MIxS), and the principles set forth by Deines et al.

(2003) regarding the open reporting of analytical metadata from project inception. For biogeochemical measurements: 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 PI is responsible for the appropriate level of QA/QC that is consistent with best practices in his respective field. In some cases, there is no established protocol, and as such the PI 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. Adherence to these standards will ensure that water chemistry measurements and microbial physiological and taxonomic information are appropriately archived and accessible alongside the genomic datasets themselves.

Access and sharing: All data will be deposited in public archives and, in accordance with NSF policy, all data will be made publicly within two years of data collection. All data will be provided to BCO-DMO and linked to core cruise metadata in MGDS and R2R. Raw Illumina data will be archived within public international DNA sequence databases (GenBank, EMBL, and DDBJ) to ensure broad access to the research community. Bioinformatics scripts, intermediate files, and processed text file outputs will be stored on appropriate repositories for non-standard data types, such as GitHub (https://github.com) and Figshare (https://figshare.com).

Archiving and preservation: All data will be deposited into the BCO-DMO within two years of acquisition as described above where stable links to data housed in other repositories can be recorded. Long-term storage and access to DNA sequence data will be accomplished by submission to the routine, appropriate databases for the projects (e.g., GenBank, EMBL, DDBJ, Dryad, MG-RAST and the NCBI SRA). These archives are considered permanent and have no specified retention period. Electron microscopy imagery and flow cytometric data will be backed up in multiple locations including the USC HPC and Thrash Lab hard drives. Non- standard data types and formats such as bioinformatic scripts and computational pipeline outputs will be archived in open access formats on long-term repositories where it is possible to assign a permanent digital object identifier (DOI), such as Figshare and GitHub. Figshare automatically assigns DOIs when data is uploaded and published, and snapshots of GitHub repositories can be preserved as versioned "releases" and assigned a DOI using the Zenodo project at CERN (https://zenodo.org).

Microbial Cultures: All cultivated isolates will be cryopreserved and maintained in the Thrash Lab culture collection at USC for future experimentation and dissemination to other labs. The existing public list of available cultures will be updated at the Thrash Lab website (http://thethrashlab.com/culturecollection). Any formally described species will additionally be deposited in two international culture collections such as ATCC (https://www.atcc.org/) or DSMZ (https://www.dsmz.de/).

Educational Data: Educational data for the mCURE courses will be maintained in coordination with the USC Department of Biological Sciences. In addition to formal publication of the research data, protocols for specific CURE courses will be published in open access journals (e.g., the Journal for Microbiology and Biology Education- ASM Press), and survey data will be made publicly available on Figshare.