DATA MANAGEMENT PLAN

All data collected during the duration of this research will be archived in accordance with NSF policy (see AAG Chapter VI.D.4) on the dissemination and sharing of research results. Experimental data will be published in a timely fashion.

Three cruises are planned (1 each in years 1, 2 and 3) during which ocean biogeochemistry data will be collected. Pre-cruise planning will be done during PI meetings. Detailed plans for station locations, instrument deployment, water sampling strategy and water sample allocation will be written up as a science implementation plan for the cruise. The actual sampling events will be recorded on paper logs (scanned into PDF documents) and in a digital event log.

Soon after the completion of the cruise, the original underway data will be contributed by the vessel operator to the UNOLS central data repository at http://www.rvdata.us/catalog/ managed by the Rolling Deck to Repository (R2R) project. Also, R2R will ensure that the original underway measurements will be archived permanently at NODC and/or NGDC as appropriate for the data type. The biogeochemical measurements made by the science party will be managed by the Biological and Chemical Oceanography Data Management Office (BCO-DMO) and the data sets will be available online from the BCO-DMO data system (http://bco-dmo.org/data/). BCO-DMO will also archive all the data they manage at the appropriate national archive facility, such as NODC and NGDC. For some types of data generated from cruises (tag sequencing datasets, metagenomes and metaproteomes) the BCO-DMO may not be the best archive for the primary data, but we will deposit a dataset record with them to alert the community that these data exist, communicate the metadata, and point to the location where the data itself can be obtained.

16S rRNA tag sequencing will be conducted using protocols and analysis pipelines standardized by the Earth Microbiome Project (http://www.earthmicrobiome.org) and both sequence data and associated environmental metadata will be deposited there so they may to add value to future comparative analyses.

Genomic analyses will be conducted using pipelines currently in place in the Rocap lab. Our assembly, annotation and phylogenetic analyses use a combination of our own high memory server and compute cluster and Amazon Web Services (AWS) and Google Cloud. We use StarCluster cluster management software which allows for intensive and massively parallel compute operations and data storage. All sequence data from this project will be made publicly available in both GenBank and MG-RAST. Assembled metagenomic contigs will be submitted to GenBank as a Whole Genome Shotgun Sequence project. Individual reads will be submitted to the NCBI Sequence Read Archive. In addition, all raw reads will also be submitted and made publically available in MG-RAST (https://metagenomics.anl.gov), along with associated environmental metadata. Phylogenetic character matrices and trees will be formatted and deposited for inclusion within the Open Tree of Life (http://purl.org/opentree/data-sharing).

Proteomics analysis generates large volumes of data (~1 gb per run). Access to and manipulation of the data requires large storage and computational capacity. The MMRC facility that houses the mass specs has an open access archive of data and adequate computational space for manipulating data. We use the University of Washington's Hyak Cluster Computer Facility (http://escience.washington.edu/content/hyak-0). File format standards for these data sets and electronic dissemination and preservation plans are as follows: The raw data and instrument settings from each mass-spec run are saved in native file formats. For proteomics analysis, the files will be converted to the open community standard mzXML file format. These data will be accessible upon request through the collaboration file system. Metadata and tabular results will be made available through the collaboration file system in a timely fashion and deposited with federally funded clearing houses for electronic dissemination.

For archival data storage for mass spectra we will use, Lolo, a file-based storage service for research computing customers at UW. It includes two file systems, Archive and Collaboration. The archive service is intended as a repository for data that you may rarely access but that you want to ensure is safe and available over the long term. Users write files to disk as they would with any other network file system. Within a day all files are automatically transferred from disk to tape in a primary campus datacenter. A second tape copy is created in another campus datacenter within an additional day. Recall of files is automatic and accomplished by simply opening and reading the file. The MMRC currently provides users with up to 8 TB of archival data storage.

Output of model simulations from this project will be archived for data sharing, on servers maintained by the Deutsch. Model output will be made available to the broader scientific community and the general public at the time of publication of the scientific results. The archived data will be in the Network Common Data Format (netcdf) and placed on a raid-5 storage array to ensure maximum stability over time. Effort will be made to sustain access to the data indefinitely, subject to availability of external funds to maintain the necessary computer hardware. Associated model code will also be made available to researchers upon request.

We retain rights to “first publication” of our data. This data management plan was written after consulting NSF document 04-004 “Division of Ocean Sciences Data and Sample Policy”.