DATA MANAGEMENT PLAN

DATA POLICY COMPLIANCE

Our project will include several forms of data, including (1) quantitative data on microbial agents over time-course experiments, (2) metabolomics analysis, and (3) molecular data. The project investigators will comply with the data management and dissemination policies described in the NSF NSF Grant Proposal Guide (Section II.C.2.j.) and the NSF Division of Ocean Sciences Sample and Data Policy.

DESCRIPTION OF DATA TYPES

Direct and plate counts: Virus, protist, *Halobacteriovorax,* and prey bacteria counts will be taken by the classical plate count culture methods and, in the case of viruses and protists, also by the direct microscopic count method.

Metabolomic: Mass spectrometer data will be generated using a liquid chromatography system (Eksigent Technologies) coupled via a heated ESI (H-ESI) source to a hybrid linear ion trap FT-ICR mass spectrometer. For small metabolites (<150 Da in mass), the mass measurement will be based on the time-of-flight mass analysis. Lists of known and unknown metabolites will be generated based on SUMMIT MS/NMR approach.

Molecular data: Molecular data from liquid samples in the microcosms will consist primarily of 16S rRNA gene amplicon sequences and Quantitative Polymerase Chain Reaction (qPCR) data.

Alongside the experimental data, code, manual, additional outputs from the research study will be in the form of papers in peer-reviewed journals and other technical publications, and oral or poster presentations at regional and national conferences. In addition, several laboratory notebooks will be produced.

DATA AND METADATA FORMATS AND STANDARDS

The majority of experimental data will be obtained from analytical instruments and collected in computer files using instrument-specific electronic formats (e.g., .xls, .pdf, .jpg). Some data (e.g., pH) will be read from the instrument output and manually entered into Excel spreadsheets. All notes describing experimental conditions, procedures, and methodologies will be entered in laboratory notebooks in sufficient detail for others to reproduce the materials and experiments. All experimental data, except imaging data and 16S rRNA gene sequencing data, will then be converted into assembled Excel spreadsheets for plotting and tabular evaluations. Data contextual details include experimental run numbers, dates, and environmental conditions such as temperature, salinity, and nutrient concentration. 16S rRNA gene sequence data will be collected and curated according to the Genomic Standards Consortium Minimum Information about a (Meta) Genome Sequence (MIGS/MIMS) standards. These standards outline a uniform format for the minimum information required to accurately describe 16S rRNA gene data, including metadata, with the goal of facilitating inter-study comparisons and transparency.

All FT-ICR MS data will be stored as. DAT files. All time-domain data will be Hanning apodized, zero-filled, and fast Fourier transformed to yield magnitude-mode mass spectra. Other data will include the mass to charge ratios, LC retention times, and peak areas for all metabolites. 2D 13C–1H HSQC, 2D 1H–1H TOCSY, 2D 13C–1H HSQC-TOCSY, and 3D 13C–1H HSQC-TOCSY NMR spectra will be collected and converted to MATLAB format.

Viral and *Halobacteriovorax* counts will be taken by the classical plate count culture method using the double agar overlay method with LB or other appropriate media for numbers of viral PFU and polypeptone 20 medium for *Halobacteriovorax* PFU. In the case of viruses, direct microscopic counts of viral-like particles will be done using an epifluorescence Axioskop Trinocular Microscope (Zeiss,Thornwood, NY). qPCR data will be acquired using the commercially available CFX ManagerTM software.

DATA STORAGE AND ACCESS DURING THE PROJECT

Initially, all data will be archived on computers in the respective labs of the PIs, and backed up on remote servers and/or external hard drives. We will submit annotated data files to BCO-DMO for environmentally-focused grazing experiments. We will also make data available upon request. All appropriate genetic files will be deposited into GenBank. Links to this public sequence data will be made available through BCO-DMO. Metabolomic data will be generated on facility computer systems and simultaneously mirrored onto a backup server, then stored independently on the global NHMFL server (mirrored in a separate geographic location and stored on microfilm tapes indefinitely), copied to personal laptops, and copied onto DVD or flash drives. Once processed and interpreted, all data will be simultaneously stored in at least three different storage locations. The ICR facility stores data in an NHMFL-defined format, and software available free-of-charge to the scientific community to read and analyze data on laboratory websites and software storage areas. The PIs have a history of reporting data in figures as well as in supplementary tables for ease of use by other researchers and will continue to publish data in this format.

MECHANISMS AND POLICIES FOR ACCESS, SHARING, RE-USE, AND RE-DISTRIBUTION

Data sets produced by the science party will be made available through the BCO-DMO data system within two-years from the date of collection. There will be no permission restrictions for these data. The data may be of interest to researchers in chemical and biological oceanography. Data will be freely available for commercial and non-commercial reuse after publications in peer-reviewed journals with associated funding.

PLANS FOR ARCHIVING

Data will be submitted to public databases (BCO-DMO, NCBI and EMBL) and published, both in print and online as journal articles or supplementary material. All data collected at NHMFL will be archived by coPI Chen on the NHMFL server, which is archived every ~10 years. In addition, all data files will be retained at least five years after the end of the project, as well as stored at the local home institutes. All data will be stored electronically on the NHMFL server and archived on CDS or other similarly permanent media. Metabolomics data and metadata collected will be submitted to publically accessible databases including Metabolomics Data Repository managed by Data Repository and Coordination Center (DRCC).

ROLES AND RESPONSIBILITIES

Each PI will be responsible for sharing his/her subset of data among the project participants in a timely fashion. The Lead PI, Henry N. Williams, will coordinate the overall data management and sharing process and will submit the project data, including GenBank accession numbers, and metadata to the BCO-DMO who will be responsible for forwarding these data and metadata to the appropriate national archive.