Data Management Plan

All data generated by this project will be made available to the scientific community in a timely and efficient manner via the World Wide Web.

Types of data. This project will generate samples of phytoplankton cultures and seawater for extraction of DNA and RNA and the analysis of DOC and dissolved constituents. Extracted nucleic acids and DOC concentrates will be stored in freezers at UW for two years after the project ends or until the final manuscripts are published. I do not anticipate creating any long-term curated physical collections. This project will generate chemical data of organic matter and next-generation sequence data for prokaryotic and eukaryotic genes and transcripts.

Internal Data Management. I will use a SQLShare database (developed by the eScience Institute at UW) for merging and analyzing the disparate data types in order to facilitate downstream bioinformatics and statistical analyses. The database schema will allow archiving of chemical data in tables that contain information on the sample, compounds detected, and concentrations. Also through the database schema, sequence data will be archived in tables to contain bioinformatic taxonomic and functional annotations for each read (closest taxon, closest functional gene, COG assignment, KEGG assignment, sFam assignment, BLAST quality scores, percent identity). Chemical and biological data tables will then be queried through SQL to create custom datasets more appropriate for input into R and network analysis software.

Data and Metadata Standards and Release. Chemical data, environmental context data, and metadata will be made available through the BCO-DMO website.

Environmental Data – Physio-chemical measurements, microbial rate measurements, and metadata taken at each sampling time will be submitted to the BCO-DMO website for posting to the project site. Tabular, downloadable data files will be made available in text and Excel-compatible file formats.

Chemical Data – Mass spectrometry data files will be archived at the BCO-DMO project site in Excel compatible spreadsheets. For MS data, concentrations and individual organic compounds will be recorded, generating spreadsheets of, at most, 50 rows by 10 columns for each sample, within the size range appropriate for Excel-compatible format.

Meta-genomic and -transcriptomic Data – Meta-omic libraries will be subjected to three QC steps: removal of low-quality reads, removal of rRNA sequences (in the case of metatranscriptomes), and joining/trimming of the paired-end Illumina reads. Following QC, sequence libraries will be deposited in FASTA format in public repositories at NCBI and EMBO. Previously, sequence data would be deposited along with metadata in CAMERA; however, CAMERA is no longer be accepting public data. Unfortunately, NCBI and EMBO do not accept accompanying metadata, although they can include direction to the BCO-DMO project site. Thus, all sequence data, chemical data, and environmental metadata for each sample will be directly accessible at the BCO-DMO website. Sequence data and metadata will conform to the Genomic Standards Consortium specifications for the “Minimum Information about a Metagenome Sequence” (MIMS). Data will be made available at the time of the first publication that uses the dataset or within two years of collection.

Policies for Access and Sharing. This project will conform to NSF standards for data access and sharing. Metadata will be available immediately after data are archived, and data files (e.g. chemical and sequence data) will be openly and freely available on the web within two years from data of collection. RNA, DNA, and DOC concentrates will be stored for at least two years. Raw data will be archived for at least two years following the end date of the project. Any sulfonate standards made in-house will be provided directly to individual researchers upon request.