

DATA MANAGEMENT PLAN

1. Types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project.

Field samples will include frozen filters containing cells for pigment, nucleic acid, and mass spectrometry analysis. Water samples will also be collected for nutrient quantification, extraction of dissolved organic matter, and uptake rates. Upon analysis, these samples will primarily produce nucleic acid (DNA and RNA) sequences, mass spectrometric data, and images from the NanoSIMS and microarray experiments. Other data will include meta-data for field collections. These data will be archived and made available through the Biological and Chemical Oceanography Data Management Office BCO-DMO (<http://www.bco-dmo.org/>). Measurements may include conductivity and salinity, temperature, DO, TIC, nitrate, nitrite, phosphorus, silicate, pigments, POC, TOC, and bacterial abundance. Some, but not all corresponding data will be collected during shipboard sampling trips as needed for the project.

2. Standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies).

We will follow the best management practices for metadata and data outlined by the Biological and Chemical Oceanography Data Management Office (BCO-DMO), which are available on line (http://www.bco-dmo.org/files/bcodmo/BCO-DMO_Guidelines.pdf). All data will be collected and archived using most commonly accepted data formats. Mass spectrometry data will be preserved as “.raw” data files and also converted to XML compatible file formats such as mzDATA and mzXML. Data will therefore be accessible through spread sheet software and can be opened via software programs such as Xcalibur and See-MS.

All DNA and RNA data will be stored as read files containing quality information in FASTQ format, as well as the raw read data generated by the sequencing machine. Environmental metadata for field collections will be collected and summarized in excel spread sheets and where possible converted to data formats compatible with Ocean Data View (odv.awi.de). Where possible, tab-delineated text files containing data and ratios will be produced and archived with datasets, in case future software updates no longer support readout of original data file formats.

3. Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements.

All data produced as part of this grant is intended for publication in peer reviewed journals and/or will be made available to the public without reservations, restrictions, or limitations. Sequence data will be archived through publically available databases such as NCBI's GenBank (ncbi.nlm.nih.gov), EMBL-EBI (www.ebi.ac.uk/metagenomics/), and/or MG-RAST (metagenomics.anl.gov/). Relevant publications will cite accession numbers. Mass spectrometry data will be shared through data repositories such as DRYAD (www.datadryad.org) or EarthCube (www.earthcube.org).

4. Policies and provisions for re-use, re-distribution, and the production of derivatives.

Publications will cite requisite information, references, and accession numbers to allow public access to data. Dataset that are not compatible with public databases and journal formats will be made available through the PI's web sites, though we anticipate that all relevant data can be deposited to public databases. Oceanographic data from field samples will be made available through BCO-DMO.

5. Plans for archiving data, samples, and other research products, and for preservation of access to them.

All data will be retained for at least three years beyond the award period, as required by NSF guidelines. For short- and long-term storage of raw data, images, and data files, as well as their derivative products and downstream analysis and work, the data will be backed up in duplicate on LTO-6 tape drives (\$100 per 2.5 TB per tape cartridge) through the Oklahoma OU's PetaStore long term data storage facility (See Facilities and Equipment). We will purchase cartridge as needed out of the general supplies budget of

the grant. We anticipate that several TB of data may result from the proposed work. We will preserve data as long as public databases and the Oklahoma PetaStore allow us to do so. Additionally, we have a 12 TB RAID 5 of network attached storage (NAS) in the Wawrik lab, where data will be housed for additional back up and rapid ftp-based data access.