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

This proposal seeks funding for a field program to quantify nitrogen pools and biogeochemical rates in the field, as well as associated changes in microbial community composition and transcription in both the field and in experiments. The types of data to be collected include concentration and biogeochemical rate data, and 16S rRNA gene and metatranscriptomic sequence data from field and experimental samples using water from Saanich Inlet, British Columbia. Each sample will have a unique identifier and associated metadata (e.g., station location/experiment, depth, date, sampling personnel, etc). Field samples will include frozen water samples, frozen filters, and poisoned dissolved gas samples. All measurements/data generated, including meta-data and hydrographic data, will be archived and made available through the Biological and Chemical Oceanography Data Management Office (BCO-DMO). Lead PI Grundle will be responsible for submitting all field data to BCO-DMO.

Rapid dissemination of sequence data (from co-PI Stewart) will be a priority in this project. Our proposed sequencing involves multiple runs on Illumina HiSeq and MiSeq instruments, as well as long-read Sanger sequence data. Rapid dissemination of these data to the broader community will be a priority in our project - all sequence data will be made publicly accessible within one year of generation. We will fully abide to the Minimum Information about a Genome Sequence and Metagenomic Sequence standards that have been recently established by the scientific community (MIGS and MIMS, respectively). Following automatic quality assurance filtering on the Illumina system, demultiplexed raw sequencing data with combined quality scores (FASTQ format) will be archived and stored on servers at Georgia Tech and for public access in the Sequence Read Archive (SRA) at the NCBI. SRA data will be assigned a single BioProject identifier with linked metadata. Our submissions will be annotated with detailed descriptions of the sampled environment or experimental treatments (project description, lat/long, date, habitat type, hydrography, chemical measurements, etc), including brief summaries of any associated physiochemical variables. Additionally, .pdf copies of all protocols used in the generation of sequence data (if not prohibited by manufacturer copyright restrictions) will be linked to the data submissions, either directly or via instructions for accessing copies on the PIs website or via BCO-DMO. Sanger-based sequences will be submitted to NCBI's GenBank and annotated with associated domain characterizations and metadata as above. All analysis results will be stored using internal data formats within the utilized software packages and common formats such as TIFF, ASCII, and Excel. Whenever possible, we will use portable data formats (such as ASCII files) to facilitate data exchange and subsequent analysis.

2. Standards to be used for data and metadata format and content

We will follow the best management practices for metadata and data outlined by BCO-DMO, which are available at (http://www. bco-dmo.org/files/bcodmo/BCO-DMO_Guidelines.pdf). All data will be collected and archived using most commonly accepted data formats. As all data are generated they will be loaded into Excel spreadsheets, and where possible converted to data formats compatible with Ocean Data View (odv.awi.de).

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 are intended for publication in peer reviewed journals

and/or will be made available to the public without reservations, restrictions, or limitations.

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. Any 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 lab repositories of the data generator. In the Grundle lab, all data are backed up in near real time using Time Machine® and all data generated by these labs are stored in an online data repository in a different building. This is institutional long-term storage. All data collected by the Stewart lab is backed up daily. The anticipated molecular sequence data occupies a relatively small amount of space relative to our computing capabilities and storage resources; thus, long-term preservation will be easily accomplished by keeping several copies of the data on local computers at Montana State University.