Data Management Plan

1. Types of data to be produced
Data produced by this project will include values collected in the field and processed in the laboratory to investigate the relationships between temperature, phosphorus, bacterial biomass composition and their physiological characteristics, in the laboratory and from ocean samples. Data will be collected and processed to assess bacterial biomass composition and physiological characteristics (e.g., growth rate) and will include (1) elemental content of bacterial biomass (2) measurements of cell size and othe properties (e.g., fluorescence) and 3) measurements of bacterial growth rates from culture work. The research data will primarily include observations from experiments in the laboratory, which will manipulate resources, and perform incubations, and multiple ocean sampling trips including a single large cruise. It will be critical to standardize treatment codes for the unique site community samples and reference each of the 40 isolates used in incubations to observed conditions and source samples from the field. The various aspects of the project information will be synthesized with the proposed work when appropriate and will inform how researchers approach data archival, processing, and/or analysis. The data management system for this study will accommodate relevant field measurements of environmental, biotic and chemical conditions, observations in the lab from manipulative experiments, and characteristics of preserved cultures resulting from experimental microbial communities. All environmental isolates will be archived and there location will be included as metadata that accompanies the data from each experiment in which they were used.

The data model used to manage these data will be designed in a hierarchy using an enterprise level relational database to perform data queries and organize data to facilitate statistical analysis, document and contain important contextual information from the field and generate metadata required for submission to online repositories (e.g. Dublin Core). The database schema used will be normalized and will include unique digital object identifiers where possible.

2. Facilities used for data processing, QAQC and archival
Our plan is to standardize content and identifiers, implement QAQC practices and track data and metadata from the field through lab processing and publication of results and data products. Quality control will begin with field and lab procedures that eliminate contamination and/or mislabeling. Collected water samples will be categorized and measured for physical and chemical attributes. Data characterizing biological and metabolite characteristics and responses to treatments will be produced through well-established laboratory techniques at EcoCORE at CSU and a workflow will be created within the local data management system at NREL (Natural Resource Ecology Lab) to prepare data for analysis and publication within ScienceBase and data catalog (https://www.sciencebase.gov/catalog/) and GreenGenes, the 16S rRNA Database (http://greengenes.secondgenome.com/). Data products, contextual metadata and important research results will also be made available to the public through the Natural Resource Ecology Lab website (www.nrel.colostate.edu/projects/),

3. Documentation of data and standards for metadata to enable data curation, data discovery, and interoperability
Our approach to data organization and documentation is to incorporate community standards unless an existing standard is not available, and uses OpenSource technologies. Data and metadata will be categorized and organized around the experimental framework upon which data collection will occur. Standards will be enacted to enable interoperability between different levels of data within the hierarchy of datasets, and with the online community repositories we use. Descriptions of the properties of bacterial communities from each ocean site and the description of the trait distribution of each isolate, will be made available to the broader scientific community as supplemental datasets or digital objects submitted with manuscripts when possible or through other public repositories, such as the USGS ScienceBase. ScienceBase is interoperable with other online repositories as it serves data and metadata through standards-compliant web services to enable other applications and web sites to use information resources

cataloged in ScienceBase. For all data generated, we will use Dublin Core XML schema to document metadata and facilitate data discovery in the future.

4. Policies for access, sharing and re-use

Tabular raw data or derived data products, with accompanying metadata will be available for download from both the NREL website and ScienceBase in a persistent, non-proprietary format, such as ASCII. Data products used in education and outreach activities will be made available through SUPER program and undergraduate courses for which PI Hall is the instructor (e.g. Microbial Ecology MIP/ESS 432). As data are being processed, they will be stored at the NREL on RAID servers and available upon request. CSU collaborators, and their staff and students have controlled access to the NREL system. The NREL system is reliable as it is maintained by dedicated NREL IT professionals, and preserved with off-site backup. All data and metadata produced by the project may be re-used and re-distributed as long as the data are properly cited with provided citations. The one exception is the content from scientific publications which may require permission from the publisher. We reserve the right for an appropriate embargo period for publication prior to submittal of data to public repositories. Our plan supports the production, quality assurance, organization, security, and preservation of data and metadata to allow for access to data products of value by scientific and/or educational communities. We will ensure our system provides people with disabilities access to information, data products and metadata by conforming to CSU accessibility standards.

5. Cyberinfrastrcuture and information technology team

The CI team for this project will be led by a dedicated information manager, who has over ten years experience in database design and management and has been a member of the LTER Network information management and Colorado State University research data committees. Our CI Team will be involved throughout the research process and stages of the data life cycle to ensure that data management best practices are employed and resulting data products are of the highest integrity. We will support dissemination of research findings by providing a data management service that accelerates that data flow pipeline from collection to publication. The information manager will manage and facilitate data processing for researchers and students and will train the next generation of scientists in approaches and tools in data management. Good data management practices create transparency and repeatability in scientific work and contribute to emerging NSF goals for accessible, well-documented, and useful data products. Data will be managed on NREL servers in relational databases until data are archived in ScienceBase or other appropriate data repositories. Data placed in local and domain specific repositories are backed up on redundant on-site web servers daily and on off-site web servers weekly.