Data Management Plan

1) Data Description, Collection Standards, and Retention Policy

The data products for this project will derive from: i) laboratory experimentation. Primary forms of data will be raw analytical data from flow cytometry. A smaller dataset will consist of the interpreted data that stems from these analyses and experiments and that is transformed into manuscripts and presentations.

The data will be collected following the rigorous standards applied in our laboratories, which meet the requirements of global oceanographic data collection protocols (e.g., World Ocean Circulation Experiment data quality). Data resulting from this project will be retained and preserved by Bigelow Laboratory for a minimum of at least 10 years after the end of the project, according to institutional policy.

2) Data Storage and Preservation

Bigelow Laboratory is positioned to leverage its High Performance Computer Cluster to handle a diverse range of scientific data processing needs. For the purposes of this study, the data warehouse component of the cluster consists of over 200TB of high performance, highly available storage being served up via a multitude of open protocols, allowing for a great deal of flexibility in the environment to support all operating systems seamlessly. The networking component of the cluster consists of the NUMALink 6 interconnect to connect all compute nodes at speeds of 56Gbps, while client side networking is served via 10GbE. This allows flow cytometry and imaging cytometry instrument data processing units to be connected directly to the storage facility and to be routinely, automatically backed-up. Bigelow Laboratory currently has a 100Mbps connection to the Internet to facilitate fast transfers of large datasets, with the ability to scale to 1Gbps on both the public Internet as well as Internet2.

Bigelow Laboratory has a robust, multi-tiered data backup and retention strategy in place. All data in the organization is stored on a clustered, enterprise SAN/NAS platform that utilizes RAID DP. The platform utilizes a very large battery backed cache, allowing for guaranteed protection of data should a power interruption take place while data is actively being written to disk. Storage volumes on the platform utilize compression and data duplication for space savings, and implement snapshots for instantaneous retrieval of deleted files. Snapshots are performed on an hourly basis (for 72 hours), daily basis (for 30 days), and weekly basis (for 52 weeks). Storage volumes are also synchronized nightly to a secondary storage tier, should the primary tier experience a catastrophic failure. Lastly, volumes that are deemed to be ready for permanent archive are encrypted and stored on Amazon Glacier.

Off campus secure access to the individual SRS directories on this system is permitted; this approach will allow Archer, Poulton, Posman and Lubelczyk to store **all** data files on the Bigelow server, and not on the drives of their personal computers, thus taking full advantage of the institutional storage and backup capacities.

3) Data Formats and Dissemination Methods

Microbial composition and rate data will take the form of spreadsheets and graphs. We will work with staff at the Biological and Chemical Oceanography Data Management Office (BCO-DMO) http://bco-dmo.org/data/ to insure that all data are managed and archived at the National Oceanographic Data Center http://www.nodc.noaa.gov/.

Laboratory notebooks containing: raw data generated by this project, detailed procedures and methodological approaches, deviations from protocols, specific equipment, and chemical reagents utilized for this project, will be cataloged by the PI. Digital copies of project-associated

data, including spreadsheets, databases and data summaries, will also be kept to promote data security and prevent physical loss via water or fire damage. Notebooks and digital copies thereof will serve as permanent records of the project and will be available upon request after results are published, following allowable guidelines on intellectual property and dissemination established by the Federal OMB guidelines.

Teaching and public lecture material resulting from the proposed research, primarily in the form of digital presentations and large-format research posters, will be made available on the websites of the PI for educational use after data have been published or any IP has been protected. Archer and Poulton expect to incorporate results from the proposed research into public lectures, teaching assignments and website content to achieve the widest possible dissemination of the data.

4) Access to Data and Data Sharing Practices and Policies.

All aspects of our research will be well documented in both physical (lab notebooks) and digital formats. A running log will be kept by the PIs for all samples collected and analyses performed. Data resulting from the proposed work will be shared openly after publication or at the end of the project following the NSF Policies on Dissemination and Sharing of Research Results: Award Administration Guide, Section VI.D.4. and Polar Programs Guidelines and Award Conditions for Scientific Data. The PI will provide updates to the NSF program manager in NSF annual reports and to the general public.

Metadata files for all data sets described will be submitted in the form of Directory Interchange Format (DIF) entries to BCO-DMO, when our final report is submitted to NSF.

5) Roles and Responsibilities for Data Management and Retention

The PI will take responsibility for data management associated with this project to ensure that analytical tasks are being performed for timely publication of the results. Data will be shared between researchers via the central institutional servers (primary and backup) as soon as they become available. This practice of open-access prior to publication of data will ensure that data are retained if key personnel depart the project at any time. Bigelow Laboratory is committed to the goal of responsible data management, security, and retention and a considerable amount of time and effort have been expended over the past year to improve internal data management policies and practices in response to the updated NSF Policies and Procedure Guide.