DATA MANAGEMENT, ANALYSES AND INTEGRATION

Buchan-Steen-Stubbins-Spencer

Data for the current project will be generated at four institutions, UTK, SkIO, WHRC, and at the Max Planck Marine Geochemistry Group in Oldenburg, Germany where Stubbins and the SkIO student will run FT-ICR MS analyses. All institutes will back up data immediately to their central servers. Data will be managed among the research partners within a framework designed to integrate among measurements, disseminate results, and enable hypothesis testing. Our overall goal is to optimize the availability and utility of data to all members of the research team, facilitating communication and ultimately the publication of results. All PIs will plan the overall sampling campaign. Individual responsibilities are outlined in the proposal and budget justification. Detailed plans for sample locations, biodegradation timetables, water sampling strategy, and water sample allocation will be written up as a science implementation plan. The actual sampling events and experiments will be recorded on paper logs (scanned into PDF documents), samples split and prepared for the required analyses.

Data analyses will involve comparisons within and between measurements and experimental data sets. Two all-participant meetings, at the start of the project and in year 2, will ensure coordination between sites. Data analysis will be carried out in the framework of "reproducible research" (e.g. Mesirov 2010 Science 327: 415-416); a subject on which PI Steen has developed and taught a graduate course during Spring 2013. While the individual labs that generate the various data streams will be responsible for maintaining records of data quality (standard curves, measures of analytical error, etc.), the collated data will also be screened for anomalies. Where possible, re-analyses of archived samples will be completed to check anomalous values. Possible outliers included in the final, submitted data set will be flagged to alert subsequent data users.

Raw data will be processed to the extent possible using script-based software environments, including the R Statistical Package and MATLAB (Some molecular data will likely need to be processed using non-script-based software packages). During year 3, the all-project database (including all chemical and physical analyses) will be sent with relevant metadata to the ***Biological and Chemical Oceanography Data Management*** (BCO-DMO) repository. With respect to distribution to the scientific community, sequence data will be deposited in the ***National Center for Biotechnology Information*** **(NCBI)** as well as within the *Community Cyberninfrastructure for Advanced Microbial Ecology Research & Analysis* (CAMERA) or the *Metagenomics Analysis Server* (MG-RAST) which will allow us to couple molecular and site specific data for use by other researchers. Taxonomic composition will be included in all documentation (*i.e.,* often as supplemental data sets to papers). After year 3 of the study, we will make residual nucleic acids publicly available upon request through a web/email interface. Scripts for data analysis will be published as supplemental information in peer-reviewed reports of our work, so that outside workers will have all the tools necessary to audit our data analysis procedures and recreate our findings. Prior to submission to any of the aforementioned repositories, all data will be collated and managed at UTK and SkIO. Both institutions will provide a backup to the data storage systems.

For DOM datasets, data will be managed through a collaboration led by Stubbins and Dr. Fatland, Program Manager, Microsoft Research Connections (see Letter of Support). This Biogeochemistry Data System (BGC-DS) is being built around cloud technology (Microsoft Azure) enabling remote data upload, standardization and processing. The BGC-DS beta system is scheduled to be tested in November 2013, before the project start date. The BGC-DS is being built to link directly to the BCO-DMO and DataONE NSF cyberinfrastructure. The team at Microsoft involved in the BGC-DS has worked to lower the barrier to publishing data to the DataONE archive and produced "DataUp" (http://dataup.cdlib.org), a prototype for DataONE connectivity developed in collaboration between Microsoft Research Connections, DataONE, the Moore Foundation and others. For the current project we will use a next

generation product similar to DataUp to manage and link our DOM data system to DataONE and BCO-DMO. During the course of the study we will continue to develop the BGC-DS with Microsoft, including planned expansion to allow the DOM BGC-DS cyberinfrastructure to link with genetic databases to allow the more efficient merging of these datastreams before they are submitted to advanced statistical and informatics approaches. Developing the BGC-DS is motivated by a need to collate, standardize, and integrate biogeochemical data from ecosystems and research groups around the world to make the most of this hard won information.

Date Use, Privacy and Sharing Policies.

Any data and software generated will be open access and made available for educational, research and other non-profit purposes. Software developed with Microsoft Research will be open and licensed through a Microsoft Research Licensing Agreement. Other code and software will be licensed through Microsoft Research or other NSF-approved licensing agreements. Distribution will be via participant websites, and the BCO-DMO, NCBI, CAMERA, MG-RAST, BGC-DS and DataONE archives as appropriate where products will have associated metadata and unique DOIs. Hyperlinks to products, as well as how to cite them, will be noted on project websites, in publications, and in metadata. Intended and foreseeable users are microbial ecologists, organic geochemists and aquatic scientists.