PARSING THE BIOGEOCHEMISTRY OF MARINE CARBONIC ANHYDRASES

DATA POLICY COMPLIANCE

Identify any published data policies with which the project will comply, including the NSF OCE Data and Sample Policy as well as other policies that may be relevant if the project is part of a large coordinated research program (e.g. GEOTRACES).

The project investigators will comply with the data management and dissemination policies described in the NSF Award and Administration Guide (AAG, Chapter VI.D.4) and the NSF Division of Ocean Sciences Sample and Data Policy.

PRE-CRUISE PLANNING

If the proposed project involves a research cruise, describe the cruise plans. (Skip this section if it is not relevant to your proposal.) Consider the following questions:

1. **How will pre-cruise planning be coordinated? (e.g. email, teleconference, workshop)**
2. **What types of sampling instruments will be deployed on the cruise?**
3. **How will the cruise event log be recorded? (e.g. the Rolling Deck to Repository (R2R) event logger application, an Excel spreadsheet, or paper logs)**
4. **Will you prepare a cruise report?**

Pre-cruise planning will be conducted by conference calls and email. Sampling instruments will include the ships CTD system (hydrographic sensors and bottle data), Saito Lab Trace Metal Rosette (hydrographic sensors and bottle data), and McLane pumps. The actual sampling events will be recorded using the digital event log using the R2R event logger application. A final report will be prepared describing cruise activities and deposited to BCO-DMO and the foreign clearance nations.

DESCRIPTION OF DATA TYPES

Provide a description of the types of data to be produced during the project. Identify the types of data, samples, physical collections, software, derived models, curriculum materials, and other materials to be produced in the course of the project. Include a description of the location of collection, collection methods and instruments, expected dates or duration of collection. If you will be using existing datasets, state this and include how you will obtain them.

This project will produce observational and experimental datasets.
Experimental datasets

1. **CA activity and metalloproteomic experimental** datasets (ICP-MS and LC-MS and enzyme activity) will be produced.

Observational Datasets:

1. **CTD and Niskin bottle data:** CTD data collected using a SeaBird SBE CTD package; processing to be done using SeaBird's SeaSave software; data will include standard environmental measurements (such as pressure, temperature, salinity, fluorescence). We will provide processed field calibrated 2dbar CTD data. File types: Raw (.con, .hdr, .hex, .bl) and processed and .cnv, .asc, .btl) ASCII files. Repository: BCO-DMO

2. **Event log:** Cruise scientific sampling event log; will include event numbers, start/end dates, times & locations of instrument deployments. Will be recorded using the R2R event logger (if available) and on paper log sheets. File types: Excel file converted to .csv; scanned PDFs. Repository: BCO-DMO and Rolling Deck to Repository (R2R).

3. **Cruise underway data:** Routine underway data collected along the ship's track (including meteorological data, sea surface temperature, salinity, fluorescence, ADCP). Will be collected by the shipboard instrumentation. File types: .csv ASCII files. Repository: BCO-DMO and R2R.

4. **Macronutrient, dissolved metal, particulate metal, POC/PON, carbonic anhydrase activity measurements, and incubation experimental results** will be collected, compiled and submitted to BCO-DMO.

5. **Metaproteomic data:** Raw files collected in Thermo's proprietary raw format will be converted to nonproprietary format (either **mzIdentML**- or **PRIDE XML)** for submission. Processed datasets will be generated from our informatic pipeline as excel spreadsheets or CSV files and submitted to BCO-DMO (processed data) and the Ocean Protein Portal, while raw data will be submitted to ProteomeXchange/PRIDE.

DATA AND METADATA FORMATS AND STANDARDS

Identify the formats and standards to be used for data and metadata formatting and content. Where existing standards are absent or deemed inadequate, these formats and contents should be documented along with any proposed solutions or remedies. Consider the following questions:

1. **Which file formats will be used to store your data?**

2. **What type of contextual details (metadata) will you document and how?**

3. **Are there specific data or metadata standards that you will be adhering to?**

4. **Will you be using or creating a data dictionary, code list, or glossary?**

5. **What types of quality control will be used? How will data quality be assessed and flagged?**

Field observation data will be stored in flat ASCII files, which can be read easily by different software packages. Field data will include date, time, latitude, longitude, cast number, and depth, as appropriate. Quality flags will be assigned according to the ODS IODE Quality Flag scheme (IOC Manuals and Guides, 54, volume 3; http://www.iode.org/mg54_3). Metadata will be prepared in accordance with BCO-DMO conventions (i.e. using the BCO-DMO metadata forms) and will include detailed descriptions of collection and analysis procedures.
We will endeavor to use standardized parameter names from the onset of data collection to facilitate data sharing and ingestion into BCO-DMO based on the SeaDataNet Parameter Usage Vocabulary.

DATA STORAGE AND ACCESS DURING THE PROJECT

Describe how project data will be stored, accessed, and shared among project participants during the course of the project. Consider the following:

1. **How will data be shared among project participants during the data collection and analysis phases? (e.g. web page, shared network drive)**

2. **How/where will data be stored and backed-up?**

3. **If data volumes will be significant, what is the estimated total file size?**

The Investigators will store project data on laboratory computers equipped with automated backup systems. Field sensor data will be stored on portable hard drives and transfered to lab and cloud computers upon return to the laboratory. Mass spectrometry data is archived on WHOI on the Saito laboratory's 40TB server in the institution's climate controlled computer room as well as being deposited to PRIDE. Project data will also be shared using Google Drive cloud services.
All raw and processed CTD and discrete biogeochemical data will reside at BIOS on 2 mirrored Linux work stations supporting RAID1 data storage which in turn will be backed up daily and weekly archives, using a Crashplan near replication service.

MECHANISMS AND POLICIES FOR ACCESS, SHARING, RE-USE, AND RE-DISTRIBUTION

Describe mechanisms for data access and sharing, and describe any related policies and provisions for re-use, re-distribution, and the production of derivatives. Include provisions for appropriate protections of privacy, confidentiality, security, intellectual property, or other rights or requirements. Consider the following:

1. **When will data be made publicly available and how? Identify the data repositories you plan to use to make data available.**

2. **Are the data sensitive in nature (e.g. endangered species concerns, potential patentability)? If so, is public access inappropriate and how will access be provided? (e.g. formal consent agreements, restricted access)**

3. **Will any permission restrictions (such as an embargo period) need to be placed on the data? If so, what are the reasons and what is the duration of the embargo?**

4. **Who holds intellectual property rights to the data and how might this affect data access?**

5. **Who is likely to be interested in re-using the data? What are the foreseeable re-uses of the data?**

Immediately after completion of the research cruise, underway data and metadata will be submitted to the Rolling Deck to Repository (R2R) project. Laboratory and field results will be deposited to BCO-DMO upon completion of analyses and write-up at the time of manuscript submission. Raw mass spectrometry data will be submitted to ProteomeXchange upon submission of manuscripts (through EBI's PRIDE or San Diego's MASSIVE repository system). The Saito laboratory and BCO-DMO collaboratively created the Ocean Protein Portal prototype with the intention of making it lightweight enough so as to be sustainable long-term, and it is hosts processed ocean metaproteomic data from around the world. The data reuse policy for the OPP is modeled after the GEOTRACES data use policy, where if substantive use of data occurs, reaching out to the data generator is encouraged to foster collaboration, proper data use, and to encourage continued participation of data generators in the project.
Data produced by this project may be of interest to chemical and biological oceanographers, and climate scientists interested in the role of biogeochemistry in the global climate system. Protein data is also of interest to biochemistry (metals in biology) researchers who are interested in mechanistic aspects of protein biochemistry.
We will adhere to and promote the standards, policies, and provisions for data and metadata submission, access, re-use, distribution, and ownership as prescribed by the BCO-DMO Terms of Use (http://www.bco-dmo.org/terms-use).

PLANS FOR ARCHIVING

Describe the plans for long-term archiving of data, samples, and other research products, and for preservation of access to them. Consider the following:

1. **What is your long-term strategy for maintaining, curating, and archiving the data?**

2. **What archive(s) have you identified as a place to deposit data and other research products?**

R2R will ensure that the original underway measurements are archived permanently at NCEI and/or NGDC as appropriate. BCO-DMO will also ensure that project data are submitted to the appropriate national data archive. The PI will work with R2R and BCO-DMO to ensure data are archived appropriately and that proper and complete documentation are archived along with the data.
All processed data will be submitted to BCO-DMO. Sequence data will be submitted to NCBI and raw mass spectrometry data will be submitted to ProteomeXchange.

ROLES AND RESPONSIBILITIES

Describe the roles and responsibilities of all parties with respect to the management of the data. Consider the following:

1. **If there are multiple investigators involved, what are the data management responsibilities of each person**

2. **Who will be the lead or primary person responsible for ultimately ensuring compliance with the Data Management Plan?**

Each PI will be responsible for sharing his/her subset of data among the project participants in a timely fashion. Subhas will be responsible for the cruise event logs for the Bermuda - Woods Hole expedition and carbonic anhydrase activity measurements. Saito will be responsible for dissolved metal data and the mass spectrometry and proteomic datasets. Johnson will be responsible for the hydrographic and biogeochemical measurements associated with the BATS sampling program, and the biogeochemical measurements conducted on the Bermuda-Woods Hole Expedition. Subhas will be the primary PI responsible for ensuring compliance with the data management plan.