

Data Management Plan

The project PI's support sharing scientific data broadly in ways that provide maximum benefits and value to all end-users. We will include multiple activities to provide unrestricted access for other researchers and the public as soon as final results are available. All field data collected under this program will be made available as per NSF guidelines within 2 years of collection via published manuscripts, publicly available NSF final reports, the BCO-DMO data management office at WHOI (<http://bco-dmo.org>), the Antarctic Master Directory database (<http://gcmd.gsfc.nasa.gov/>) and archiving with NODC (<http://www.nodc.noaa.gov>). In the past, hydrographic data reports from our major Ross Sea expeditions (ROAVERRS and CORSACS) have also been published and made publicly available by the Stanford group led by Dunbar.

Expected Types of Data and Samples:

Expected data types include both water column and on-deck incubation data. Water column profiles will include: (1) raw and post-processed hydrographic CTD data and metadata, (2) macronutrient and trace metal profiles, (3) HPLC pigment data, (4) POC/PON data, (5) Dissolved inorganic carbon data, (6) Volatile organic gases, (7) raw and processed proteomic data, (8) Photosynthetic parameters including productivity, photosynthetic efficiency and PE light curves, (9) Bacterial abundances, (10) DNA and RNA sequence data, and (10) expedition metadata (cruise track, dates, analytical protocols etc). Many of these same measurements will be collected from our on-deck incubation experiments and will be reported separately. Many of the above measurements will also be completed for sediment trap samples and made publicly available.

Data and Metadata Standards:

Formats utilized for the on-line publication of data will include Excel, Tiff, Matlab data structures, and HTML formats. For some of the datasets, the generated data will consist of raw mass spectra generated by the mass spectrometers that will be held and made available as semi-processed text files that can be used across multiple platforms and software environments. Proteomics processed datafiles will be in spreadsheet format, and submitted to BCO-DMO and raw data is in proprietary RAW formats and open mzXML formats.

Policies for access and sharing including provisions for appropriate protection of privacy:

We will post the data on several websites, including the Stanford Digital Repository. The Stanford Digital Repository (SDR), a service offered by the Stanford University Libraries, provides digital preservation, hosting, and access services that enable researchers to preserve, manage and share research data in a secure environment for long-term citation, access and reuse. The data will be updated several times a year as various analyses are completed. The data will also be publicly posted and advertised at journal sites whenever their publications include online data access. Any data included as part of a peer-reviewed publication is formatted per the journal's editorial policy and made available as supplementary material where applicable. We retain the right to exclusive use of project data for two years following the cruise end date. However, we expect that some data sets will immediately be made publicly available. For other data sets once the 2 year period has ended or if data derivatives are published before then we will release the accompanying data. The data is not copyrighted and no licenses pertain to it.

Policies and provisions for re-use, re-distribution, and the production of derivatives:

We will not impose any restrictions on the use of the data or require permission to use the data once it has been released. In fact, we anticipate that our data set will be useful for the Italian LTER Program at Terra Nova Bay and will encourage the transfer of the data to an appropriate Italian database. Any unpublished data submitted to a publically-accessible data facility would be automatically available for re-use, redistribution and the production of derivatives.

Plans for archiving data, samples, and other research products, and for preservation of access:

Data, samples and other research products will be archived at each of the co-PI's home institutions. The long term strategy will include regular monthly back-up of the data on University mainframe servers. In addition, we will explore the possibility of delivering processed data in a compact form (as Excel spreadsheets and text files) as part of the 'Supplemental Data' for any journal articles, as we have done previously for proteomics datasets (e.g. Mackey et al., 2015; Bertrand et al., 2013; Saito et al., 2011). This approach would utilize existing, long-lived publishers' archives as one way of satisfying the requirement for public access to the data. In general, the supplemental data associated with a journal article is 'free' for download.

Currently, there is no environmental repository for raw proteomics mass spectrometry, and the available recently launched biomedical repositories do not have the capability for incorporating and searching for key environmental metadata (location and sampling dates). As a result, we will store large raw data files at WHOI on a secure and backed-up server in the computing facility, and these files will be made available upon request. We have recently proposed the development of an prototype Ocean Protein Portal within the NSF EarthCube program (in collaboration with BCO-DMO) to formalize the storage, sharing, and interrogation of ocean protein data by a broad scientific community. If both proposals are funded, the proteomics datasets collected as part of this study will be incorporated into the Ocean Protein Portal.