

DATA MANAGEMENT PLAN

Expected work

We will participate in cruises in the northern Benguela Upwelling System OMZ to collect samples. At the outset, the PIs will discuss the science implementation plan that details timelines, procedures, and sampling strategies that are presented in the project narrative. There will be three cruises total, with one in austral summer 2021/2022 two back to back cruises in austral fall April-May 2023. Pre-cruise planning will be coordinated between the lead institution (Mason at Florida State University), the collaborating institution (Casciotti at Stanford University) and Hanselmann from the Swiss Federal Institute of Technology (ETH) Zurich (see letter of support). Mason and Casciotti will speak regularly by phone and through email with Hanselmann. All field notebooks will be digitized to PDF format.

The proposed work will generate chemical and molecular data from the water column. Data will be in the form of water column chemistry, dissolved gas concentrations, isotope measurements, natural abundance stable isotope data and molecular data. Molecular data will consist of 16S rRNA gene amplicon and shotgun metagenome and metatranscriptome sequences. Metatranscriptome data will be in the form of cDNA sequences.

Data formats and dissemination

All chemical data will be published in manuscripts resulting from this work. The PIs have a history of reporting data in tables for ease of use by other researchers, and will continue to publish all data in this format (see for example earlier work describing data from the DWH oil spill in the Gulf of Mexico in Mason et al 2012 and 2014). Metagenomic, metatranscriptomic and 16S rRNA gene sequence data will be collected and curated according to the Genomic Standards Consortium (http://gensc.org/gc_wiki/index.php/Main_Page) Minimum Information about a (Meta)Genome Sequence (MIGS/MIMS) standards (http://gensc.org/gc_wiki/index.php/MIGS/MIMS). These outline a standardized format for the minimum information required to accurately describe 16S rRNA gene and metagenomic/metatranscriptomic data, including metadata, with the goal of facilitating inter-study comparisons and transparency. Metagenome and metatranscriptome sequence data will be annotated, stored and made available to the public via the Department of Energy's Argonne National Laboratory metagenomics analysis server, MG-RAST. MG-RAST hosts 188,358 metagenomes (and metatranscriptomes). Metagenome, metatranscriptome and 16S rRNA gene sequences will be deposited at the Sequence Read Archive at the National center for Biotechnology Information (NCBI) in compliance with MIGS/MIMS. Contigs assembled from metagenomic sequence data will be made available through the Joint Genome Institute's IMG server, which allows users to carry out numerous bioinformatics analyses of the data. Additional microbial data will be published in supplementary tables as was done previously in e.g. Gillies et al (2015). We will also provide NCBI accession numbers for 16S rRNA gene, metagenome and metatranscriptome sequences in publications. All sequence data will also be made available to the public on the Mason Laboratory server: <http://mason.eoas.fsu.edu> (see for example the DWH oil spill data that is currently available on this server). The associated metadata and environmental data will be available on BCO-DMO. PI Mason will maintain responsibility for management, retention and submission of data to these resources throughout the duration of the project.

All laboratory/field notebooks and datasheets will be scanned and stored electronically, along with other information relevant to the collection, processing, and analysis of the samples. Metadata and measured data will be stored in electronic spreadsheets (e.g., Microsoft Excel, Access, or similar program). Sensor data will be downloaded and stored on PI computers.

Data stewardship, preservation, and sharing.

Inventories and data information will be kept on the computers of the PIs, who have procedures in place to regularly back up their data. This data will be shared between the PIs using Dropbox, which is backed up daily. Data will be maintained using database software, such as Microsoft Excel. As discussed above sequence datasets will be made publically available through NCBI, MG-RAST and PI Mason's server. All chemical and biological data generated through this project will be managed by the Biological and Chemical Oceanography Data Management Office (BCO-DMO); metadata, chemical and isotopic data will be made available online from the BCO-DMO system (<http://bco-dmo.org/data/>). BCO-DMO archives data at appropriate national facilities, such as National Oceanographic Data Center (NODC). All data will be publically available within 2 y of collection.

Some of our analytical practices are destructive, but for the samples for which it is appropriate, we will store aliquots of samples and make them available to other researchers upon request, after publication of data derived from these samples.

Period of data retention

Data generated will be published in peer-reviewed journals and sequence data will also be available directly from the PI via the Mason Lab server, through MG-RAST (data will be made available within 6 months after depositing, regardless of publication status) and through NCBI upon publication. Manuscripts will be prepared promptly, and we will make an effort to select open access journals where appropriate. When available, relevant datasets will be published in the form of online appendixes and made available through the website of the relevant journal.

Roles and Responsibilities

PI Mason is responsible for ensuring compliance with the Data Management Plan.