DATA MANAGEMENT PLAN (created using DMP Tool)

Creators: Joan M. Bernhard, David Brankovits, Elizabeth Borda
Affiliation: Woods Hole Oceanographic Institution (WHOI), Texas A&M University-San Antonio (TAMU-SA)
Template: National Science Foundation (NSF)

DATA POLICY COMPLIANCE

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

DESCRIPTION OF DATA TYPES

The project will produce several observational and experimental datasets which are listed below. Samples will be collected from caves along the coast of the Yucatan Peninsula and Cozumel. Five sampling campaigns will be completed, two during each of Yrs. 1 & 2 and one during Yr 3. Sampling trips will last 10-16 days which will allow for sampling and processing at several sites.

Observational datasets:

Water profiling and sampling: Physicochemical parameters and organic matter (OM) availability will be characterized through a combination of sensor-based (e.g., YSI EXO2) methods and discrete sampling.

a) Spatial and temporal profiling of physico-chemical parameters: Hydrologic Data will be collected using sensor-based measurements. The EXO2 multiparameter sonde deployed in the water column of the caves and sinkholes along depth profiles will collect fundamental water properties (e.g., depth, temperature, salinity, pH, dissolved oxygen, and fDOM). Data loggers will collect information (e.g., water level, water temperature, salinity, dissolved oxygen, rainfall, and atm. pressure) in a time-series style. File types: Excel. Repository: BCO-DMO.

b) Discrete water sampling: Water samples will be collected from various depths within the water column using manually operated 60 mL syringes (in replicates) during sampling trips 1-4 and cm-scale water samplers along vertical depth profiles (10 depths) during sampling trips 2-4 at each sampling site. We estimate a total of 150 samples over the course of the project. Samples will be sub-sampled for and preserved separately for different laboratory analyses. Additional water samples will be collected to quantify the particulate fraction. From each sample, we will collect a variety of chemical information through laboratory analyses (i.e., concentration and stable carbon isotopic values, for select samples, of CH_4, DIC, DOC, POC and concentration of chloride, dissolved N and P, sulfate and sulfide) to generate large amounts of Geochemical Data. File types: Excel. Repository: BCO-DMO.

Experimental datasets:

1. Sediment sampling: Sediments will be collected to determine benthic eukaryote (nanobiota, foraminifera, metazoan meiofauna) distributions, diversity (via barcoding), gene expression (RNAseq), cell ultrastructure (foraminifera only), and sedimentary geochemistry as follows:

a) Foraminiferal distribution and viability: Triplicate quantitative cores (ethanol-fixed) will be incubated in CellHunt Orange CMTMR. A total of approximately 180 (2.5-cm diameter) cores will be analyzed. Multivariate statistics will be used for analyses of benthic community data in combination with environmental characteristics (salinity, temperature, distance from cave opening, methane, ammonium, sulfide, DOC, POC, DIC, sulfide, O_2, NO_3^-, SO_4^{2-}). All statistical analyses will be completed in the software RStudio (version 4.0.2, RStudio, 2020) and OmicsBox Bioinformatic workflows. Unconstrained Q-mode cluster analysis will be performed using Bray-Curtis dissimilarities of square-root transformed data to identify ecologically meaningful assemblages. A distance-based RDA (db-RDA) will be completed to examine how environmental variables explain species variations. Specimen micrographs will be hosted on BCO-DMO. File types: Excel files of abundance data, images (.jpg, .TIF files). Repository: WHOAS, BCO-DMO.

b) Sanger and NGS sequencing: Sediment/water samples (e.g., frozen, preserved in ethanol, DNA/RNA Shield) will serve as the source for DNA and RNA extraction, PCR products and sequencing libraries and stored in Co-PI Borda lab -80°C freezer in San Antonio, TX. Sample data (collection/voucher information; storage location, barcoding sequence data) will be made available on NCBI and BOLD. Sediment and select specimen reference vouchers will be deposited at the UNAM National Collection. File types: .ab1 files, .fastq file, .fasta files. Repository: NCBI SRA, NCBI TSA, BOLD.

c) Cellular Ultrastructure: Individual foraminifera from the most common species that appear laden with cytoplasm will be isolated from the glutaraldehyde/paraformaldehyde-fixed sediment samples and analyzed with Transmission Electron Microscopy (TEM). At most, we will analyze 50 specimens. TEM micrographs will be hosted on BCO-DMO. File types: images (.TIF, .jpg files). Repository: BCO-DMO.

d) Foraminiferal nitrate content and denitrification rates: Nitrate concentration as well as the $\delta^{15}N$ and $\delta^{18}O$ of nitrate will be measured in ~200 foraminifera. Also, pools of conspecific foraminifera (n~10-20) will be incubated to determine nitrate reduction and production rate of N2 (denitrification). File types: spreadsheets (.xls, .csv). Repository: BCO-DMO.

e) Bulk organic carbon and nitrogen: Collected sediment samples (approx. 15 samples per trip) and select faunal samples (approx. 10 specimens per trip) will be frozen as soon as possible after collection (-20C) and shipped on dry ice to WHOI, where they will be freeze-dried, and ground for analysis that will generate the following data: Total Organic Carbon and Total Nitrogen as well as stable isotopic values ($\delta^{13}C$ and $\delta^{15}N$). A total of approximately 125 sediment/faunal samples will be analyzed combined. File types: Excel. Repository: BCO-DMO.

2. Environmental Response Experiment: Physiological experiments carried out on foraminifera collected in the cave system and grown under controlled environmental conditions; dataset will include data on the experimental treatments and counts of individuals at the end of the experiment. Experiments will be conducted in the lab. File types: Excel file(s). Repository: BCO-DMO.

DATA AND METADATA FORMATS AND STANDARDS

Geochemical data will be maintained in XLSX Excel files, stored in plain CSV text files and the associated metadata in read.me files. Sanger sequencing reads will be saved as .ab1 files and consensus sequences will be saved as .fasta files. High-throughput sequencing reads will be saved as .fastq files. Assemblies will be saved as .fasta files. Raw data will be uploaded to the NCBI SRA database. Consensus sequences for 16S and COI barcodes will be submitted to the NCBI Nucleotide database as well as the Barcode of Life Database (BOLD). Assembled meta-transcriptomes will be uploaded to NCBI TSA. Analytical procedures will follow standard, published protocols for QA/QC and be compared to certified standards. Hydrologic data, including vertical profiles and time-series, will be downloaded from the instruments after deployment and maintained/stored the same way as the geochemical data. All sensors will be calibrated according to manufacturer specifications. All data will be geo-referenced. There are no standards for the various foraminiferal analyses. The functionality of CHO will be clear because fluorescent organisms will be present in all samples. There are also no standards for TEM and SEM. Finally, outreach assessment data will be archived in digital format and databased along with metadata such as number of views, content and themes of the video, and other data potentially valuable to future scientific outreach research. These databases will be available to all team members as soon as QA/QC measures are taken.

DATA STORAGE AND ACCESS DURING THE PROJECT

The investigators will store all project data (e.g., spreadsheets, images, text files, scanned field/lab logs) on office computers which are backed up daily by the WHOI Information Services (IS). The Principal Investigator (PI) will also establish a shared account (Cloud-based storage) for data storage and sharing among project members. Personal computers will be backed up daily to an onsite external hard drive (e.g., WD 20TB My Book Duo Desktop RAID External Hard Drive).

MECHANISMS AND POLICIES FOR ACCESS, SHARING, RE-USE, AND RE-DISTRIBUTION
The investigators of the project will hold the intellectual property rights to the data. However, we will not place any restrictions on data access, use and re-use, except that we be attributed when data are used for future studies. To ensure public access, we will be depositing data in publicly accessible online databases (e.g., NCBI, BCO-DMO, and USGS Science Base) as well as making our scripts and code available (e.g., GitHub). New code and custom scripts will be licensed under CC-BY-SA (attribution and sharealike required). The majority of the data we will collect do not present any privacy issues and can be made publicly available without restrictions. We plan on publishing all of our results in peer-reviewed open-access journals to ensure that there are no restrictions on access to our data or published results. Any survey data collected in association with the outreach assessments (pending IRB approval) would be kept anonymous and secure following standard protocols for human subject research.

PLANS FOR ARCHIVING
BCO-DMO will ensure that all project data is submitted to the corresponding national data archive. In addition, the PI will work with BCO-DMO to ensure that data is archived appropriately, including any associated documentation. Genetic sequencing data will be stored in a publicly accessible, permanent database (e.g., NCBI).

ROLES AND RESPONSIBILITIES
Each PI will be responsible for sharing his/her subset of data to all research members in a timely fashion. D. Brankovits and S. Rohret will be responsible for collecting all sediment and water column samples. J.M. Bernhard will oversee graduate student S. Rohret in all foraminiferal sampling processing and analyses (e.g., viability, morphological documentation (SEM), ultrastructure). E. Borda will be responsible for overseeing the training of S. Rohret, MS student and TAMUSA undergraduates on data process for meta-barcoding and meta-transcriptomics. S. Rohret and D. Brankovits will be responsible for all discrete sampling chemical analyses (e.g., methane, sulfide, DOC, POC, TOC). Co-PI E. Borda will coordinate data management and sharing among investigators as well as submission of project data to NCBI (GenBank accession numbers) and BCO-DMO (metadata), who will then transfer the data to the appropriate national archive.