Data Management Plan – Collaborative Research: Diel dynamics of dissolved organic matter production and remineralization as a driver of coral reef nutrient recycling

Data Products: During the proposed field trips to Moorea in 2020 and 2021 we will be collecting daily samples from field surveys, field deployed experiments and laboratory microcosm experiments. The core data produced in this project will be (1) oceanographic data, (2) mass spectrometry data, and (3) nucleic acid sequence data. Metadata on survey and experimental designs will be crucial to interpreting the data, and our management strategy ensures consistent formatting and archiving of linked data products. Metadata, biogeochemical data, and sequence similarities will be incorporated into a relational database facilitated by Lead PI Kelly. This will enhance accessibility of the data to all PIs providing a searchable platform to promote interpretation and connectivity of all datasets.

Biogeochemical data: All field and experimental data biogeochemical data will be submitted to the Biological and Chemical Oceanography Data Management Office (BCO-DMO) for archiving and public dissemination. The data will be contributed to BCO-DMO within 2 years of their production to comply with NSF OCE data dissemination and archiving policy. The data will be accompanied by all relevant activity logs (mesocosm set-up, perturbations etc.). The PIs are already in contact with Cynthia Chandler, the BCO-DMO data manager, regarding required metadata and metadata standards optimum format for data submission, as they have been submitting similar data collected during prior proposals; data collections will be deposited within BCO-DMO linked with Nelson and Kelly's existing Project areas: MCR-LTER and Coral DOM2. Data sets and associated metadata will be made available in standard delimited text files (spreadsheets/flatfiles). Where appropriate, metadata will be submitted on the metadata forms developed by BCO-DMO. Metadata will include variable names, derived units, experimental set ups, analysis methods, descriptions of synthesis or calibration procedures where appropriate, data location, season, and quality control information. Nelson is currently associated with the Moorea Coral Reef Long Term Ecological Research program, and as such use the data and metadata repository available through the Knowledge Network for Biocomplexity (KNB, http://knb.ecoinformatics.org/index.jsp), an international online data repository. Variable names, keywords and metadata standards will follow guidelines available from the Marine Metadata Interoperability Project (marinemetadata.org) and the vocabulary and OA data management best practices being developed by the LTER network. Adherence to these standards will allow metadata to be shared and to be searchable between different databases. Ancillary data products from other programs, including oceanographic data, reef habitat data, and meteorological data, will also be important in interpretation and will be linked through open-access data repositories as appropriate through the LTER network.

MS data: All MS data, irrespective of size and number of files, will be uploaded, managed, analyzed and shared with the community through the MassIVE (Mass spectrometry Interactive Virtual Environment) interface of the Global Natural Products Social (GNPS) Molecular Networking platform (http://gnps.ucsd.edu; Wang et al. 2016). This interface is a community resource developed by the NIH-supported UCSD Center for Computation Mass Spectrometry, to promote, global, free exchange of MS data. We will make all spectra (MS1 and MS/MS), in open file formats (raw and mzXML), publicly available for analysis by the entire scientific community as soon as they are available, prior to publication. This is the policy for all projects within Prof. Dorrestein's Collaborative Mass Spectrometry Innovation Center. Computed feature tables (derived data) from MzMine linked to MS/MS spectra will be also be deposited there as .mgf files. Tables containing networking results, GNPS library annotations, elemental composition and hierarchical classifications from SIRIUS and Canopus as well as ancillary data on sample collection and other biogeochemical properties of the samples will also be stored there. Furthermore, any code and scripts developed in this grant will be made available through Github https://github.com/ or equivalent algorithm and code repository.

Nucleic acid sequence data: Sequence products will be permanently raw-archived with appropriate metadata in the NCBI Sequence Read Archive (SRA). Annotated metagenomic, metatranscriptomic and 16S data will be archived publicly under Kelly and Nelson respective accounts in the MG-RAST server.

***Short term data storage and organization*:** Data will be collected and added to lab databases that will be mirrored on laptop computers, in-house network servers, and in managed Google Drive computing clouds, with weekly back-up on hard drives. As data are generated, they will be collated and stored on shared internal databases and will be accessible by all persons involved in the project. Data will be archived in the original data format and also in more common, non-proprietary formats (e.g., tiff, csv, xml, et) to facilitate future data usage.

Metadata Formatting and Archiving: Metadata on experimental designs will be crucial to interpreting the data, and our management strategy ensures consistent formatting and archiving of associated data products. Metadata forms in Ecological Metadata Language (EML) are utilized to create and organize the overall project database for local use, and upon publication for increased ease of data dissemination (see "Publication" section below). Metagenomic, marker gene amplicon, and transcriptomic metadata will be carefully formatted according to the Genomic Standards Consortium recommendations (MIMS, MIMARKS, and the draft recommendations of MINSEQE, respectively). Long-term archival and curation of genomic metadata will occur through KNB and in the NSF-funded Dryad digital & open source archived data repository (http://datadryad.org/) as per MCR-LTER standard approaches.

Metadata will follow BCO-DMO guidelines, including variable names, derived units, experimental set ups, analysis methods, descriptions of synthesis or calibration procedures where appropriate, data location, season, and quality control information. Variable names, keywords and metadata standards will follow guidelines available from the Marine Metadata Interoperability Project (marinemetadata.org) and the vocabulary and open access data management best practices.

Methodological documentation - Documentation for this project will include the formation of written methodologies for sample collection and processing, made openly available on our respective lab websites as formal Standard Operating Procedures (SOPs), and published as peer-reviewed methodologies where applicable. Quality control will be conducted at each stage of the data acquisition, processing and analyses, including the development of metadata forms detailing the outline of the project, instrumentation used, format of data, QA/QC standards and controls, and funding source.

***Policies for Data sharing and Public Access*:** All of the data generated in this study will be made publicly available upon publication in a peer-reviewed journal or within 2 years of the completion of the project. In addition to publication in peer-reviewed journals, with all relevant attempts to provide original datasets in open access publication repositories, the data generated from this project will be made publicly available online wherever possible. The project will have a specific data use policy including requiring active contact information, citation, funding source acknowledgement, quality control and intellectual property rights agreement required of all data users.