**Data Management Plan**

***Data Products:*** During the proposed field trips to Moorea in 2016 and 2017 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 nucleic acid sequence data describing planktonic and coral-associated microbial communities in the context of environmental surveys and laboratory experiments. Additional measurements of coral physiology (oxygen metabolism, photosynthetic efficiency, and color), dissolved organic matter chemistry (neutral and amino sugar concentrations, fluorescence matrices) and environmental parameters (nutrient and dissolved organic concentrations, pH, temperature, etc.) will be collected in conjunction with the experiments and surveys. 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 Rohwer. 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 Carlson's existing Project areas: MCR-LTER and Coral DOM. Data sets and associated metadata will be made available in Excel-compatible spreadsheets. 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. 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.

***Short term data storage and organization*** **-** Data will be collected and added to lab databases that will be mirrored on laptop computers, the UH-C-MORE and SDSU-Rohwer local in-house network servers, and in the UH-managed Google Drive computing cloud, 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, doc, et) to facilitate future data usage.

***Data archiving*** **–** All 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 under Rohwer and Nelson respective accounts in the MG-RAST for public access and

searchability. All phylogenetic marker gene amplicon data will be made available in an open access area of the Nelson Lab website in conjunction with phylogenies used for classification as part of a new effort to establish consistent marker gene classifications across studies (Nelson is currently involved in collaborative efforts to develop standards for marker gene sequence open access repository deposition).

***Metadata Formatting and Archiving:*** Metadata on survey and 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). Ancillary data products from other programs, including oceanographic data, reef benthic data, and meteorological data, will also be important in interpretation and will be linked through open-access data repositories. Carlson and Nelson are 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. 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/).

***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 amongst other details.

***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. To ensure accuracy and data tracking, the project will have a specific data use policy including:

- User requests require current and valid contact information that will be used by the PI for tracking and documenting data usage.
- Users are required to cite the project publications and acknowledge the NSF as the original funding source.
- PIs and project participants will conduct quality control on the primary data and ensure accuracy of the primary data to the best of their abilities.
- The PIs and project participants will not release any private or confidential information to the public, and in-house databases will be password protected.
- The PIs and project participants will retain intellectual property rights, except where explicitly required to be released for publication and documentation.