Data Management Plan

Introduction: Data management will be coordinated by the Principal Investigators, and will be carried out by all project participants. The plan encompasses four areas: use policies, data preservation and archival, standards, and sample preservation and archival. We will leverage existing systems wherever possible; we have begun working with the Moorea Coral Reef Long Term Ecological Research (MCR LTER) project, which has a sophisticated data management system and personnel experienced with standards and data curation. The LTER network has discussed offering a training course that will be available to our participants (M. Gastil-Buhl, pers. comm.). Data storage services during the course of the project will be provided by computing systems at UCLA, the University of Miami, and the University of the Virgin Islands. However, long-term preservation of data products will use BCO-DMO.

Our project includes diverse types of data, some of which have specific data management requirements to contribute to specific metadata repositories such as sample tracking to GADZOOX and genomic data to Genbank. We will also maintain searchable access to our data thru CoRIS (Coral Reef Information System). Data include ecological monitoring data, data from experiments, physical/chemical data, and genetic data. Ecological monitoring of up to 40 years duration for 10 reef sites includes data on coral diversity and species recovery, coral symbiont abundance and diversity, recruitment, reproductive activities, algal cover, sea urchin population densities and bioerosion, corallivore impacts, coral tissue regeneration after partial mortality, fish guild abundance, and trophic interactions. Photoquadrats will be added during this RAPID funding cycle. Data from experiments are usually shorter term (weeks to years), but of the same overall nature as monitoring data. Physical/chemical data include current meters, *in situ* temp records, CTD casts, and nutrient concentrations. Physical samples that must be archived and made available include preserved tissue for DNA analyses and preserved cryptic organisms.

Use policies: Data collected under the project will be made available to the public with as few restrictions as possible. We (and users of our data) will abide by the LTER Network Data Access Policy (2005). Under these policies, we plan for the publication of most data with metadata after primary publication of results, or at most two years after the completion of the study. Certain datasets may warrant proprietary restrictions, such as locations of endangered species, and in these cases it is our policy to publish metadata with instructions for requesting data delivery. We are also committed to protecting the privacy and accuracy of users' confidential information. Our data and document storage will have 3 tiers of access: public access of data packages; internally shared pre-release or controlled-access data packages; and areas accessible only by the data owner. This enables us to secure new data against loss early in the process before it is ready for review and internal sharing, yet ultimately all data will move through to the public access tier. Since we have adopted the policies of the MCR LTER project, the following are conditions for use of our data: 1) The user of data agrees to contact the data owner (*i.e.,* the project investigator responsible for data) prior to publishing, and where appropriate, users whose projects are integrally dependent on our data are encouraged to consider collaboration and/or co-authorship with the data owner; 2) the user agrees to cite our project in all publications that use our data by including the following statement in the Acknowledgments using this statement: "Data were provided by the [PROJECT NAME] Project funded by the US National Science Foundation ([AWARD NUMBER])"; 3) the user agrees to send the full citation of any publication using our data to our project email address; 4) users are prohibited from selling or redistributing any data provided by our project; 5) extensive efforts are made to ensure that online data (from website) are accurate and up to date, but we will not take responsibility for any errors that may exist; and 6) any violation of the terms of this agreement will result in immediate forfeiture of the data and loss of access privileges to other project data sets.

Plans for archiving and preserving data: Metadata and data will be submitted to existing catalogs (see below for genetic data, photoquadrats). As does the MCR LTER, we propose to use the Knowledge Network for Biocomplexity (KNB), a national network to facilitate ecological and environmental research on biocomplexity that develops software for use in managing data. This management system is especially designed for data from a highly-distributed set of field stations, laboratories, research sites, and individual researchers. A common framework is provided by use of the Ecological Metadata Language (for describing ecological data) and the Metacat metadata server (a flexible database based on XML and built

for storing a wide variety of metadata documents). This repository provides a reliable place to store metadata and data (replicated on multiple machines), and has an interface for information entry and retrieval. The metadata includes all of the information needed to understand what the described data are and how to use them: a descriptive dataset title; an abstract; the temporal, spatial, and taxonomic coverage of the data; the data collection methods; distribution information; and contact information. Each information provider decides who has access to information (the public, or just specified users), and whether or not to upload the original data together with the metadata.

Our contributions are planned to be generally in concordance with the practices of MCR LTER, the LTER network, and their information management partners. This will assure long-term preservation and access to our data.

Standards and formats to be used for metadata and data: We plan to publish datasets using an established XML specification (Ecological Metadata Language, EML) that is adaptable for a variety of data types and is widely used in the ecological and environmental sciences, including the LTER Network. We plan to leverage software in use by MCR LTER for creating and managing datasets, as well as other tools that might become available. We will make use of guidelines for data and metadata developed by the LTER Network that are designed to promote long-term usability. These include high-precision georeferencing, the use of standard keywords, units and measurement descriptions, and the inclusion of methods and/or protocols with all datasets.

Data will be archived with metadata, unless there are proprietary restrictions (as described above). As much as possible, data will be archived in ASCII format, which is the most flexible and readable over the long term. We will archive data in tabular formats that have been proven successful when sharing data among the project collaborators. It is expected that certain data products, such as those used in GIS applications, may be more easily preserved in their native (binary) formats. To the best of our ability, data in binary formats will be kept up to date and readable by current software. Descriptions of metadata containing photoquadrats can be uploaded to KNB, with a link to a url with the photographs as is standard for the MCR LTER. (http://mcr.lternet.edu/data/db/census/coralImage.php).

Preservation and archival of physical samples: Organisms will be preserved in the field (70% ethanol), transported to the University of Miami, and voucher specimens placed at the Invertebrate Museum at the University of Miami. Curated by Prof. Nancy Voss, the collections will be made available to any qualified biologists.

Genetic samples and data: Coral samples are collected as skeletal fragments (most samples ~1cm^2) with associated tissue. We typically digest the tissue *in situ* using an SDS lysate buffer and Proteinase K. Small aliquots of this lysate are then removed for DNA extraction on an as-needed basis. We retain archived lysates (consisting of SDS-lysed tissue and a small piece of skeleton) that are stable long-term at room temperature. These archives will be made available 2 years after the end of the project, or following publication of the data in the peer-reviewed literature, whichever is sooner.

Because corals are listed under Appendix II of the Convention on International Trade in Endangered Species (CITES) we are obliged to ensure that all samples comply with CITES regulations when shipped internationally. RSMAS is registered as a scientific institution with the CITES secretariat (Code: US 138-A) which greatly facilitates the exchange of samples between other scientific institutions worldwide.

Metadata for all samples [consisting of Sample ID, Coral species, Site name, Latitude and Longitude, Depth at time of collection, Date of collection and time of collection (important for calculation of MLWS depth) and Coral condition] will be made available online within two months of collection on GADZOOX (**G**lobal **A**ssessment of **D**iversity in **ZOOX**anthellae). This will advertise the availability of samples to potential collaborators within a rapid timeframe. GADZOOX will be maintained by the Baker Lab and hosted at www.rsmas.miami.edu/groups/reef-futures. Data will be archived in ASCII format.

Novel DNA sequences obtained by this project will be deposited in Genbank www.ncbi.nlm.nih.gov/genbank/. Accession numbers will be reported in publications as well as in the metadata maintained on GADZOOX.