DATA MANAGEMENT PLAN

1. Expected Data Types and Formats

a. Physical and Environmental Data: We will collect data from both biological samples (epifaunal biomass and community composition, epiphyte biomass, lesion prevalence) and temperature records from dataloggers deployed at each site. Invertebrate samples will be sorted, counted and identified to the lowest taxonomic unit possible, and samples will be archived at the Smithsonian. Eelgrass leaf samples for lesion scoring, qPCR quantification of pathogens, and microbial samples will be collected and preserved.

b. Sequence Data: As outlined in the proposal we will sequence 16S rRNA gene segments for bacterial community description from eelgrass leaves at each site using an Illumina HiSeq 4000 platform at the DNA Technologies Core at UC Davis. We will adapt existing bioinformatics pipelines for the analysis of 16S rRNA data.

c. Drone data: Drone imagery from each study site will be saved as orthomosaic tiff files; point, line and polygon GIS data. All primary data for the drone work will be collected through the use of DJI Phantom 4 pro drones.

d. Disease data: The Eelisa app for measuring disease lesions will generate segmentation masks from expert, student, citizen annotations, and from machine inference; metadata about the source and parameters of each segmentation; and summary statistics for analysis based on these segmentations. Source code for the user interface and the machine learning classifier will be replicated in a remote Git repository for version control and backup. The Cornell Institute for Computational Sustainability operates a dedicated HPC cluster called Atlas with 56 computer nodes, a total of 672 cores, and 3.2TB RAM, a head node, a 30TB file server, and a 40Gbps QDR Infiniband network. Atlas will be used to develop Eelisa and other machine learning algorithms.

2. Data and Metadata standards

a. Biological and environmental data: Each of the various data types collected will include a consistent set of metadata to be agreed and finalized at the first PI meeting in year 1. Minimally, each dataset will include a title, description, location of source population (latitude & longitude via GPS), date, time (UTC), PI name and contact information, and co-PIs. Each data set will also include, as appropriate, standards used for measurements, instrumentation used, analytical methods used, data processing information, sampling procedures, and access restrictions. Biological data will follow Darwin Core conventions, and habitat metadata will follow CMECS classification. Data will be entered into a standardized digital format to be agreed upon by the PIs and in consultation with BCO-DMO staff. These will likely be Excel spreadsheets convertabvle into machine-readable CSV files for submission to BCO-DMO, Smithsonian MarineGEO, OBIS, and other repositories for final archiving, as we have done with similar data in the past (see Project Description: Results from Previous NSF Support). A standard file-naming procedure incorporating date as a file extension will be used.

b. EELISA app: Segmentation masks will be generated using standard image formats such as PNG, and summary statistics and metadata for each segmentation will be generated in CSV tabular format. Eelisa code will be written primarily in Javascript and Python using well-established libraries and frameworks. README files will document the source and usage of resulting code and data collections

c. Drone data: Drone imagery and related GIS data will be included and stored in ArcGIS geodatabase format in an open, publicly accessible mapping portal hosted by ArcGIS Online with consultation from the UCF team. All data will also include detailed metadata. Federal Geographic Data Committee (FGDC) standards will be used for all geospatial data, since these are the “gold standard” of disciplinary and GIS professional metadata creation and geospatial data cataloging. Example metadata will include: creation date, author information, data purpose/use, flight specifications, data collection method, and data sharing restrictions.

3. Roles and responsibilities

Each partner will collect and maintain copies of their site's data at their own site and send identical copies of all data to the lead PI (Duffy) as soon as possible for QA/QC and assurance of standardized formatting across sites. Co-PI Hawthorne will be responsible for the maintenance, processing and provision of drone data. The Cornell team will be responsible for Eelisa data analysis and qPCR. Co-PI Stachowicz will be responsible for processing, analyzing and posting of all microbial sequence data. We will use a restricted-access dropbox site to facilitate sharing of data and other communications among the PIs and other partners. Primary responsibility for maintenance of the database and website will reside with the Smithsonian team, with aid from the Smithsonian MarineGEO Data Manager.

4. Dissemination

Within two years of collection, all data and metadata will be made available online through the Biological and Chemical Oceanography Data Management Office (BCO-DMO). Results will be shared via publication in academic journals and presentations at both academic and public forums. Voucher specimens of species studied will be photographed, preserved, and maintained the Smithsonian Institution with putatively novel species or samples of special interest deposited into the permanent collections of the National Museum of Natural History. Raw sequence data will be deposited in NCBI's short read archive and processed data will be available through digital repositories such as DRYAD or BCO DMO. Analytical pipelines will be shared on github. Physical collections stored at the Smithsonian will be open access. Drone results and data will be posted to the Citizen Science GIS website, StoryMaps, and social media platforms. The Citizen Science GIS team will create an online, open source geospatial data portal, so that community partners, the general public, project personnel and interested academics/practitioners will be able to access project data, maps and resources.

5. Data sharing

Environmental and biological data will be made fully publicly available in online repositories within two years of collection through BCO-DMO. There will be no restrictions on the release of drone imagery. The team anticipates that the data will be widely distributed throughout multiple disciplines, including marine biology, geography, geology, biology, sociology, and other related disciplines. Given the community-focused nature of the drone imagery collection, it is also expected that multiple community organizations, national/district/local governments, and community residents will be interested in the data.

6. Archiving

All electronic data, including sequencing data, images, and metadata, will be deposited in public repositories noted above, but will also be backed up, with careful version control, on servers and personal computers at each partner institution. Where possible, processed data and metadata will also be stored in a shared cloud service (google drive or dropbox). Physical laboratory notebooks will be retained in the labs for a minimum of 5 years and scanned electronic copies will be retained for a minimum of 10 years. All spreadsheets will be organized through established naming conventions and a standardized system structure. The Citizen Science GIS team at UCF maintains a GIS Data Repository in which the majority of data will be archived for future use. All data will be backed up on an additional password-protected server. The data will be maintained for a minimum of 10 years beyond completion of the grant.

The PI and Co-PI's have made a good faith effort to provide all information required in the 'Data Management Plan', in compliance with new policies outlined by the National Science Foundation.