Data Management Plan

The PIs of this proposal understand that research has become both increasingly data-intensive and highly collaborative. They firmly believe in the open-access principle and will provide all generated data to the community in a presentable and usable form, ideally prior to publication. We will adhere to the Division of Ocean Sciences Data and Sample Policy and will archive data and/or links to data repositories within the 2-year timeframe as stipulated. All data types will either be submitted directly through the Rolling Deck to Repository program, directly to BCO-DMO, or linked from the existing database to BCO-DMO.

A. Data Generation Activities: This proposal will generate data ranging from oceanographic data to community distribution and abundance data to large amounts of genetic sequence data. This proposed research program will generate the following types of data:

(1) Video and still photographic data from dives will be generated and stored in raw and processed formats according to the specifications of the submergence assets.
(2) Standard oceanographic (temperature, salinity, depth) data plus dissolved oxygen, as well as pH, alkalinity, and calculations of aragonite saturation state. These will all be georeferenced typically in .csv or .txt formats.
(3) Porewater geochemical, isotopic, and cell count data from sediment cores and carbonates will be organized in .txt and .csv formats.
(4) Georeferenced species abundance data in .csv and .txt formats. Scripps also maintains a database of species distribution data in a .kml format, and we will contribute to that effort.
(5) DNA sequence data, including but not limited to the following file types: .fasta, .sff, .fna, .fsa. Gene trees and species trees in newick and nexus formats.

B. Research Cruise Data: Cordes, as the lead P.I. and Chief Scientist will be primarily responsible for producing and distributing metadata. All cruise participants will be involved in generating and capturing data on board the cruise. The Cordes lab maintains two small, portable 32-TB RAID servers that will serve as a central repository for all of the data generated during the cruise. All data will also be stored on the local machines from which it is generated, including all video data on on-board servers and hard drives. Metadata will be generated continuously by each of the cruise participants. Cordes will assign one of the participants from his lab to collect and QA/QC metadata from each of the participants after every dive. This will include the dive number, site, time of sampling, latitude, longitude, depth, sample type, tentative identification, primary co-P.I., and subsample distribution. All data will be available to all cruise participants, and it is the responsibility of the participants to supply the required drive space to house any data that they require. Following the cruise, Cordes and the PIs are committed to providing unlimited access to data for collaborators and the community, and will submit a cruise report and all cruise metadata to BCO-DMO following the cruise.

C. Faunal Specimen management: Samples will be sorted into morphospecies and photographed, with vouchers kept for long-term vouchering and subsamples taken for the molecular work (95% ethanol and/or freezing). Unique numbers are assigned to each specimen or specimen lot and this is tracked though all subsequent processing, including DNA sequencing. Voucher specimens will be accessioned into the Benthic Invertebrate Collection (BIC) at Scripps Institute of Oceanography (SIO). Shipboard data management will be conducted using Lightroom, Filemaker Pro and Excel.

D. Cataloging: Specimen images and specimens from the Costa Rica cruises will be vouchered for long-term archiving into the Benthic Invertebrate Collection at Scripps Institution of Oceanography (SIO-BIC) (http://collections.ucsd.edu/bi/). Samples will be processed by the SIO-BIC manager, Dr Harim Cha, who is employed full-time by SIO to manage this Collection. Each specimen lot will minimally include locality, coordinates, depth, collecting date, collector, collecting method and fixative. Any habitat notes or remarks will also be included. A unique catalog number is issued for a specimen lot. Cataloged lots are moved to the taxonomically arranged compactor system (room temperature), -20˚C or -80°C freezers. As with all specimens held by SIO-BIC, the new accessioned collection will be open to access by the research community and available for loan for morphological and DNA/RNA-based studies within two years, at no fee. The loanee is expected to pack the specimen/s appropriately for return shipping and to the cover the costs of the return of the loan, also using a courier.

E. Local Databasing and online database: All cataloging information is entered into a MS Access-based relational database. This database automatically issues one unique number (Catalog_ID) for each lot entry and it links to a Collection ID that represents a collection event. Each catalog entry requires a taxon name and spelling of each taxonomic name is checked with WoRMs (http://www.marinespecies.org/). For collection data, locality information such as Latitude and Longitude are cross-checked with Google Earth software to ensure the locality name and coordinates are matching. Newly entered catalog data are uploaded monthly to the server maintained by UCSD along with an image of a specimen (preserved or live, if available). Low-resolution (72dpi) images are uploaded with higher resolution images available on request. Once uploaded to the server, catalog information is available through a web-based collection database for the Benthic Invertebrate Collection (http://collections.ucsd.edu/bi/search/).

F. Disseminating data to the Ocean Biogeographic Information System (OBIS) and iDigBio To reach a wider audience, all specimen data generated from this project, and SIO-BIC in general, will be disseminated to OBIS-USA (http://www.usgs.gov/obis-usa/) and iDigBio. For iDigBio, each specimen lot will be assigned an institutional catalog number and a GUID, according to iDigBio guidelines (https://www.idigbio.org/content/guid-guide-data-providers-0). The associated data and images will be provided to iDigBio portal through the Integrated Publishing Toolkit (IPT).

G. Molecular sequence data: All DNA data generated by direct (Sanger) sequencing will be deposited on GenBank with appropriate referencing to the voucher specimens held at SIO-BIC. Raw reads generated from the Illumina sequencing will be deposited at the GenBank Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/Traces/sra/) and also we will establish a Data Dryad repository for our data http://datadryad.org/.

H. Protocols: Any new protocols developed during this program will be deposited at the Protocols.io database (https://www.protocols.io/)

I. Data Quality: Data QA/QC will follow specific plans for each type of data generated. Careful attention will be paid to sample custody and metadata during the cruise. Because this is such a large, complex, inter-disciplinary project, it is simply impossible within the space provided to detail all of the QA/QC procedures for genetic, oceanographic, chemical, ecological, and other data sets and analyses.