Data Management Plan

Types of data

We will be collecting high-resolution plankton imagery, biological samples, and environmental data via the ISIIS, MOCNESS, the LISST-200X, and shipboard sensors. All of these data will be collected simultaneously, along with the ship's position and time, and will be recorded and stored in digital formats on 2 Terabyte (TB) hard drives on board the ship. ISIIS imagery data are taken at a very high scan rate and are during the deployments saved in the Audio Video Interleave (.avi file extension) format. The .avi format is used to stack the image frames into a single file to ease data transfer. Per cruise, roughly 48 TB of raw image data will be generated, leading to a total amount of 192 TB for the whole project (i.e., 4 cruises). All digital data will be backed up on board the ship onto duplicate 2TB hard drives.

Biological plankton samples will be stored in the HMSC Plankton Ecology Laboratory and Sutherland Lab until they are processed, and zooplankton and larval fishes sorted and identified. Sorted and identified larval fishes will be stored in vials in at HMSC until selected for otolith microstructure and/or gut contents analyses. Otolith microstructure data, elemental analysis and gut content data will be maintained in the HMSC Plankton Ecology lab (ichthyoplankton samples) and Sutherland Lab (gelatinous samples) in separate databases (with duplicate backup protocols as described above). All samples will be linked among databases via unique station and sample identification numbers. Once gut contents and otolith data are obtained, physical samples will be available for use by others via direct communication with PIs. Resultant data from gut contents, elemental analysis (C:N) and otolith microstructure will be accompanied by annotated metadata in laboratory notebooks, excel spreadsheets and word documents to maximize the usable lifespan of the data. Data originally recorded on physical paper datasheets or laboratory notebooks will be scanned weekly.

Post cruise, primary video data, along with all physical (i.e., sensor) data, will be uploaded to our local Center for Genome Research and Bioinformatics (CGRB) servers over a 100Gbps network-to-network up-link from where it will enter our data pipeline. Hard drives with backup data will be transported and stored in a separate location on campus, 50 miles inland. Once data is uploaded to the CGRB data storage, the pipeline can be run either locally or through cloud services. Depending on resource availability, we generally utilize XSEDE cloud services to increase our processing power and reduce time. Processing steps include segmentation of the video data into images that can be run through a Convolutional Neural Network (CNN). Using XSEDE allows our processing to take advantage of multiple CPU and GPU resources at the same time providing an easy to use resource that scales with our needs. We have found each 2TB of video data can be processed on average from start to finish in about 17 hours using the XSEDE services. If there are network issues accessing XSEDE the CGRB has local hardware resources available to process the data using less throughput.

Data and Metadata Standards

Together with OSU's CGRB, we developed an image database based on OmniSciDB, allowing for extremely fast multi-queries. This system captures metadata for all original and derivative images processed by the system, following Darwin Core standards. Metadata include

classification information and plankton lengths, geolocation as well as parameters such as conductivity, temperature, depth, dissolved oxygen, photosynthetically active radiation (PAR), and fluorometry. Where possible and applicable, these metadata will also be made compliant with existing metadata standards, such as CF-netCDF or the North American Profile of ISO 19115:2003 for geographic information, to make the data more easily consumable by the broader scientific community.

Policies for access and sharing and provisions for appropriate protection/privacy

The images derived from the live camera feed will be stored in a standard Unix file system format with a RDBMS used for file indexing, location, and user access and control (i.e., OmniSciDB). Presentation of the data will be provided through a standard web based CMS (Content Management System). We will also work with BCO-DMO both with respect to uploading the data and to coordinate access by the broader scientific community. The data will be made public within a 2-yr time period following initial collection. As these images do not contain any human information, there are no ethical or privacy concerns.

OSU provisions for re-use, re-distribution

We are planning for permanent sharing and reuse of these data. It is likely that many researchers within both the marine ecology and biological oceanography fields could utilize these data. Additionally, these data and our OmniSciDB based data-sharing system may have applications in other areas of marine science, education, and conservation management. We anticipate that this system for sharing and archiving will also serve for reuse/re-distribution. A web form will be made available to interested researchers. This form will be published with a persistent URL so that the form will be available for publication or other method of distribution as well as for future reference (Note: this is in addition to BCO-DMO held data).

Plans for archiving and Preservation of access

All ISIIS project data, including images, will be available through our OmniSciDB database, utilizing resources from NSF's XSEDE supercomputing cluster. The use of open-source OmniSciDB ensures that ongoing maintenance and support will not be tied to a proprietary system. Running on XSEDE's servers allows for the constant availability of the data to the public and means that database data are periodically backed-up by XSEDE itself. Original source data will be kept on hard drives in multiple locations. An index of metadata will be created and maintained. Specifically, for our training libraries, we will maintain a second location for public data access, the EcoTaxa platform. While our current Northern California Current (NCC) image library is already freely available on the EcoTaxa system, this project will expand on that library and complement it. This is part of our ongoing effort to build a *Global In Situ Plankton Image Library* consisting of data from other regions and oceans where we have deployed ISIIS, e.g, the Straits of Florida, the Gulf of Mexico, Mediterranean Sea. All of our oceanographic/ environmental data will be compiled into correct formats (per National Oceanographic Data Center, NODC) and sent to the NODC within 1 yr of the completion of the cruises. Archiving of zoo- and ichthyoplankton data will necessitate sorting and identification of samples, which will require more time. Once complete and quality-controlled, these data will also be archived in national databases (BCO-DMO) where the data will be freely available to the public. Archived biological samples will reside in both labs (OSU and UO)s until transfer to a permanent storage/curation facility such as the OSU Ichthyology Collection in Corvallis, Oregon.