Data Management Plan for NSF-OCE 1829835

Casey Dunn (PI), Steve Haddock (co-PI), Anela Choy (co-PI)

Biological samples

The biological samples collected for this project will include:

- Formalin preserved morphological vouchers of siphonophores and representative taxa from their prey fields. These will be used to verify specimen identity.
- Mounted slides of siphonophore tentacles for morphological analysis.
- Frozen siphonophore tissue, gut contents, and potential prey taxa for genetic analysis.
- Frozen siphonophore tissue and and tissue from potential prey for stable isotope analysis.

These biological samples, except when consumed in the process of analysis for data acquisition, will be deposited in the Peabody Museum of Natural History, where PI-Dunn is the curator of Invertebrate Zoology. Formalin specimens will be deposited in the wet collections. Slides will be deposited in the dry collection. Frozen tissue will be deposited in the Cryo Collection (http://peabody.yale.edu/collections/cryo-facility). All specimen data will be fully searchable via the Peabody collections web interface (http://collections.peabody.yale.edu/search/). Specimens will be cross-referenced to other data (including sequence, distribution, environmental, and isotope data) via their collection numbers. Specimens will be available to other researchers according to established museum lending practices. We will provide BCO-DMO with metadata that points to where the physical samples are deposited.

Data archival and dissemination

Below is a list of the data that will be generated by this project, along with the public archive it will be deposited in. Data will be deposited and publicly available when it is published or within two years of collection (whichever is sooner).

- Stable isotope data generated by this project, along with relevant environmental metadata, will be stored in the publicly-accessible Biological & Chemical Oceanography Data Management Office (BCO-DMO, https://www.bco-dmo.org).
- All molecular sequence data will be deposited at the National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov). Sequence data will be submitted along with metadata that connects them to other deposited data, including museum vouchers, location, and environmental data. Links to data deposited in GenBank and NCBI will be provided to BCO-DMO.

- Species distribution data generated by this project will be deposited in the Ocean Biogeographic Information System (OBIS, http://www.iobis.org). Links to data deposited in OBIS will be provided to BCO-DMO.
- Siphonophore tentacle images, videos, and measurements will be deposited in the Dryad Digital Repository (https://datadryad.org). Links to the data deposited in Dryad will be provided to BCO-DMO.

Software and analysis code

All software, training materials, and documents (including user manuals and manuscripts) will be developed, managed, and released in git repositories at GitHub (eg, https://github.com/caseywdunn). These repositories will be made publicly available at or before the time of publication. We will make extensive use of GitHub's tools for tracking user-reported bugs, feature requests, general project management, and web site hosting. The Haddock, Dunn, and Choy labs already use git repositories extensively.

In addition to fully adopting and integrating git for this project, we will adopt the best practices that are now widespread in industry for using these tools, including code reviews by project participants. For example, feedback will be provided through the integrated issue trackers. All code development in git will follow a test-driven development paradigm, facilitated by continuous integration tools including Travis-CI. These tools test all code each time it is modified, catching bugs early and allowing for clear measurement of progress toward formally defined project goals that are specified as tests before any other code is written.

Each manuscript published with support from this project will have an associated git repository with all code needed to rerun analyses, the data (or links to the data in public archives), the text of the manuscript, metadata, and any other files associated with the manuscript. This is already common practice in our labs. Data-driven manuscripts will be developed as executable documents that include embedded analysis code using established platforms such as knitr and Jupyter. This will facilitate reproducibility and transparency of our work, and make it easier for others to extend and repurpose our analyses. Code will be released under open-source under the GNU Public License v3. Links to all git repositories associated with this project will be provided to BCO-DMO.