DATA MANAGEMENT

PLAN

Data types

This study will produce and/or aggregate multiple types of data:

Objective A – Mesocosm experiments

- Algal mass, height
- Predatory and herbivorous fish identity and density
- Experimental conditions (i.e., water temperature; water quality; food ration mass; etc.)

Objective B – Artificial reef/halo experiment:

- Primary producer identity, percent cover; height (in-situ)
- Predatory and herbivorous fish/invertebrate identity and density
- Hyperspectral imagery and derived estimates of benthic primary producer coverage
- Environmental conditions (i.e., chlorophyll A; dissolved nutrients (N; P); solar irradiance; sea surface temperature; water velocity; sediment deposition
- Experimental treatment conditions (i.e., artificial reef size/location; water temperature; water quality; etc.)

Objective C – Global-scale synthesis:

- Satellite-derived (RGB) reef/halo location, size, nearest neighbor, etc.
- Predatory and herbivorous fish identity and relative density (Global Finprint database)
- Environmental conditions (i.e., chlorophyll A; solar irradiance; sea surface temperature; wind speed; etc.)

Data formats and standards

In general, all data will be collected, compiled and managed in accordance with the BCO-DMO *Data Management Guidelines Manual*'s best practice recommendations for collecting and sharing biogeochemical and ecological oceanographic data and metadata.

All data collected (other than imagery) will be stored in .csv format; aggregated data will be stored in its native format or converted to .csv. Imagery will be stored in its native format (e.g., geoTIFF).

Ecological metadata (date, time, location, experimental treatment, etc.) will be prepared in accordance with Ecological Metadata Language (EML) conventions. For each other type of data acquired, we will adhere to community-based data annotation standards.

Analytical code will be primarily written in R, with code for geospatial-specific analyses written in Python. All R analytical code will be written in accordance with the "Tidyverse" principles (Wickham and Grolemund 2017).

Data storage, access, re-use and distribution

While the project is ongoing, data and computational assets will be housed within the UHM ITS Data Center. The Data Center provides storage services including bulk storage of research data, digital media, and other digital assets. This storage is handled by an on-campus storage cloud, which provides advanced RAID technology to ensure the safety, security, and integrity of the stored data. GitHub will be used for version control throughout the project.

The lab will collectively retain preliminary data until publication or within 2 years of collection,

whichever comes first. After this point, all data (other than raw imagery), code, and model outputs will be made publicly accessible within a GitHub repository(ies). Where licenses permit, imagery will be made available via Zenodo. Storing data and code on GitHub a) ensures data security (i.e., protection against data loss due to computing failure, b) enables version control, and c) provides open access to data and code in line with best practices for reproducible science. Because GitHub maintains editable versions of data and code, data (including derived imagery data) and code used in every publication will also be permanently archived via a no-cost, GitHub-compatible Zenodo permalink. This permalink permanently maintains the unique version of a dataset/code that was used to generate figures in a particular publication, thereby ensuring permanently reproducible science. Data and all derived products will be discoverable through the BCO DMO web page. I will encourage collaborative re-use of these data once they are publicly available for future re-analyses, synthetic analyses, comparative analyses, meta-analyses, etc.

Intellectual property and data generated under this project will be administered in accordance with the UHM and NSF policies. Any software developed under this proposal will be made available under the GNU General Public License (GNU GPL) version 3.

Data archiving and preservation

In addition to publication-specific storage on Zenodo, long-term archival and preservation of all digital information produced will use either tape media or cloud-based archives at the UH ITS Data Center.

References

Wickham, H., and A. G. Grolemund. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. 1st ed. O'Reilly Media, Inc.