Data Management Plan

Data generation, storage, and formats

Data will be collected by personnel from the PIs' research groups Pedrosa Pamies and Ruff (MBL), Liu (UTMSI), Bochdansky (ODU) and Gallager (Coastal Ocean Vision) as described in the proposal. Data generated will include surface microscopy characterization and chemical composition of different polymers, before and after incubation in the deep sea, and microbial biofilm communities of the deployed polymers. The data generated during the microbial project parts will mostly be 16S rRNA amplicon and metagenomic/ -transcriptomic sequencing data, as well as microscopy data. Sequencing data will have the usual .fasta and .fastq formats and microscopy images will have .czi (Zeiss format) and .tif, once the images are converted. In addition, we will produce publicly available bioinformatics scripts using the software environment R available as .r or .RData.

Field notes including coordinates, date and times will be recorded in Rite-in-the-Rain notebooks, and these notebooks will be stored in the PIs' labs. These notes will be scanned and digitally stored and archived in the same manner as the field data. Laboratory analytical data will initially be recorded on the instruments used for those analyses or on associated laboratory computers as they are collected, and then copied and archived. The field and laboratory data will be transferred to desktop computer hard drives in the office for longer-term storage after this project.

This project will not involve the acquisition of either animal or human subject data.

Data and metadata standards

Field and laboratory data products will take the form of computer files, generally consisting of tab-or comma-delimited values with appropriate header information and accompanying metadata. Metadata will consist of, at a minimum, a plain text file noting the date, time, location, instrumentation, personnel, procedure of data collection, and important quality assurance information such as instrument precision, accuracy, calibration date, or other notes on data quality and uncertainty. Derived data, program code and results generated by analyses within programs such as MATLAB, Excel, and R will be stored in plain text formats to ease data portability. Quality control will be performed on the data. All datasets will consist of three files: 1) a README file with the description of the dataset, 2) a Metadata file describing all variables measured with their respective units, and 3) the specific dataset in .csv, or.tiff format. The .csv format will allow us to have easy access of the data from both commercial software such as Excel and scripting processing software such as R, MATLAB and Python. Sequencing data will be amended with comprehensive contextual data. For the contextual data we will use the latest standard for marker gene sequences MIMARKS (Minimal Information for Marker Gene Sequences) and metagenomic sequences MIMS (Minimal Information for a Metagenomic Sequence (Field *et al.*, 2008)) as recommended by the Genomic Standards Consortium (https://press3.mcs.anl.gov/gensc/). We will also archive high-quality metagenome-assembled genomes (MAGs; >90 % completeness, <5 % contamination) using the latest standard MIMAG (Minimal Information for a Metagenome-Assembled Genome (Bowers *et al.*, 2017)).

Backup and archiving

The office computers and hard drives mentioned above will be connected to a secure, daily, automatic backup program or service (e.g., Code42 CrashPlan backups every 15 min the files from individual computers, and Veeam backups the MBL servers every night, both administered by Marine Biological Laboratory IT staff). External hard drives containing duplicate data and results will also be maintained offline via manual backup.

Intended repositories

All physicochemical data derived from this project will be deposited in The Biological and Chemical Oceanography Data Management Office (BCO-DMO) at Woods Hole (bco-dmo.org) in comma-delimited text format. Code to reproduce results from our publications will be included as supplementary material in those papers linking to the datasets archived. This code will be available in Github (https://github.com), most likely in R or Python. Any data collected specifically as part of a student poster, presentation or thesis

will be made available after the student publishes it. No datasets will be covered by copyright or licensed and there will be no charge for accessing the data. We anticipate no ethical or privacy issues. No data with specific household information will be collected and no data collection will require Institutional Review Board approval.

16S rRNA gene amplicon raw reads will be processed (barcode and primer trimming) and archived in the NCBI Short Read Archive (www.ncbi.nlm.nih.gov/sra). Shotgun metagenomic data and metatranscriptomic data will be quality checked and high-quality fastq datasets will be archived at SRA as well. We will archive all sequencing data under a dedicated SRA Bioproject and add contextual data (i.e. metadata) to each dataset. Fluorescence micrographs and other microscopy imaging, as well as cell counts, biogeochemical and physicochemical parameters, anion/cation concentrations, metabolite concentrations, and all other contextual data will be archived in PANGAEA (www.pangaea.de). All data that were collected during the study will be available under one project, including links to the aforementioned sequencing datasets and mass spectra. The project (and even individual datasets) will get a DOI (Digital Object Identifier) and can be retrieved and cited independently.

Timeline for data release

Oceanographic context data will be released 12 months after fieldwork. It will be up to 36 months for data generated from chemical analysis, and 12 months for microbial data.

Policies for access and sharing

Data will be made freely accessible to the public according to NSF and BCO-DMO procedures and guidelines. As mentioned above none of the data will be proprietary or otherwise restricted. We will ensure that the data is publicly accessible at the above listed repositories either i) upon submission of the accompanying manuscript, or ii) twelve months after the project funding has commenced, whichever is first. After its release all data will be publicly available without restrictions and can be re-used by everybody by citing the respective publications or datasets. Preservation of the data is guaranteed by each of the abovementioned repositories.

We strongly believe in the FAIR principles for scientific data and stewardship (Wilkinson *et al.*, 2016), and thus ensure that our data is findable, accessible, interoperable, and reusable. All acquired and analyzed data will be available on publicly accessible data repositories. In addition, we will share the bioinformatics code that was used to analyze the data, so that the community can re-analyze our data, and re-use or expand our bioinformatics workflow for their purposes. An example for one of our workflows can be found here (https://sourceforge.net/projects/visuar/files/). Our workflows are very well annotated and commented, to ensure that even unexperienced users will be able to use it. We use these workflows as teaching material and believe that society can benefit from open access data and barrier-free sharing of knowledge.

Publishing

Results of the work will be shared principally through open-access peer-reviewed publications. When appropriate and not included in the main manuscript, data and results will be published as online supplements to the peer-reviewed publications. The process and results of the project will be shared with the public via the multiple methods of achieving broader impacts described in the proposal. Further details about and copies of data and methods will be made available upon request at the end of the project period, unless pending publication, in which case the details will be released concurrent with the publication.

References

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., et al. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**: 725–731.

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., et al. (2008) The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.* **26**: 541–547.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**: 160018.