Data Management Plan

The PIs will conform to NSF policy on the dissemination and sharing of all research results. Proposed work will generate a diverse set of multifaceted physical, biological, chemical data that will be managed by the respective PIs and Senior Personnel. Details on the types of data and materials that will be generated in the course of this 5-year project, as well as our policies for access and sharing, distribution and archiving are detailed below.

We are proposing a multi-pronged approach to understanding the role of that virus infection of ballasted phytoplankton plays in mediating export. We will use a combination of lab-based, mechanistic studies using model coccolithophore and diatom host-virus systems and observational studies from two proposed field campaigns (California Current and the Northeast Atlantic)

Laboratory-based experiments will generate the following data types: cell abundance, diagnostic staining of cell physiology measured by flow cytometry; virus abundance (measured by flow cytometry, most probable number and/or plaque assays), photosynthetic rates measured by fast repetition rate fluorometry (FRR); transparent exopolymer particle (TEP) and Coomassie staining particle (CSP) concentration via microscopy and UV-Vis spectrophotometry; visualization of particle aggregation and characterization of particle size spectrum via FlowCam analysis, VMP, and LISST; particle characteristics and speeds from PIV data; flow velocities and echo intensity from Aquadopp data. Raw data will be collected in their respective format and subsequently exported into relevant analysis programs (e.g. Excel, R, or Matlab).

Observational, field-based data will consist of physical, chemical, biological, and biogeochemical parameters that will be linked to the *in situ* community composition, physiological cell state, and viral infection dynamics. Data from water column sampling will consist of: hydrographic and oceanographic data from CTD hydrographic casts; temperature, salinity mixed layer and euphotic zone depth; dissolved nutrients (NH_4^+, NO_x, PO_4^{3-}, silicate); particulate matter (POC, PON, PIC, BSi); pigments (chlorophyll a, 19'-HEX, HPLC accessory pigments); Algal and virus abundances via flow cytometry using autofluorescence and SYBR Green/Gold staining; Physiological characteristics (live/dead, ROS, nitric oxide) via diagnostic fluorescent stains (SYTOX-Green, DCF-DA, DAF-FM) and flow cytometry; cell characteristics (size, calcification, free coccoliths) via flow cytometry (side scatter, forward scatter, perpendicular and parallel FSC); photophysiology measurements via FRR; TEP and CSP measurements (both via bulk extract and FlowCAM); Fluorogenic enzyme activity assays; Mathematical model simulations and output data; grazing- and virus-based mortality rates from dilution experiments (using bulk chlorophyll and flow cytometry). Raw data from the VMP-250, LISST-200, and images from the Plankton and Niskin bottle camera will be processed for quality control using Matlab and tagged with metadata including instrument type and serial number, geospatial coordinates, calibration information, and QA/QC steps. From these data streams, the turbulent dissipation rate, vertical eddy diffusivity, the vertical flux of nutrients, the particle size distribution, and particle fall velocity will be determined.

Lipid data in the form of raw HPLC/MS output (mzXML files) will be analyzed at WHOI using a custom-designed lipidomic workflow (LOBSTAHS; Collins et al. 2016) that can be used with environmental or experimental data from a variety of systems and is freely available at https://github.com/vanmooylipidomics/LOBSTAHS. Data will be filed within an in-house preprocessing database. After quality control, data are posted to the Van Mooy website hosted on WHOI servers, as is the current practice for existing projects. There is no accepted repository of lipidomic data, but if one should emerge over the course of this award, we will use it. These data are highly specialized, and we expect that primarily the PIs and their lab group members will use these data. Nonetheless, all data that passes quality control will be open to the public.

All data will be submitted and disseminated through the Biological and Chemical Oceanography Data Management Office (BCO-DMO) in accordance with the NSF OCE policies on data archiving and dissemination.

All software and analysis code will be shared publicly on GitHub. All data sets will be deposited at dataDryad (https://datadryad.org/stash). All manuscripts arising from primary work in Prakash lab will be submitted to BioRxiv at the time they are also submitted to a journal for peer review. All model and analysis software as well as relevant outputs necessary to reproduce the results from the NCAR

team will be made public upon completion of the project or publication (whichever comes first) via GitHub/Zenodo (zenodo.org) and the NCAR DASH repository.. Data will be uploaded in netcdf format and include metadata. The data will be shared immediately via shared access to the computer system for project collaborators. The NCAR lead has experience making model output and analysis scripts publicly available with doi via GitHub, zenodo, and figshare.com. He also has been trained in the use of DASH, which is available to all NCAR employees.

Data Access and Sharing. Raw and processed data will be maintained in the most appropriate and relevant format (i.e. laboratory notebooks, excel spreadsheets, standard flow cytometry .fcs files, FlowCam raw image files). Experimental data and observations will be published within 2 years of collection and made publically available through PI's Bidle and Thamatrakoln's custom-designed, laboratory information management system (LIMS). Through a partnership with Big Rose Web Design LLC (http://www.bigrosestudio.com), Bidle and Thamatrakoln have built an integrated, query-based database management system that allows for data entry, secure data storage, visualization, and analysis (Fig. 1). The goal of the LIMS is integration of diverse datasets in an easily-accessible format that allows for rigorous and robust interpretation and identification of potential linkages between experimental findings that may otherwise go unnoticed. Through our secure lab website, authorized users are able to add/edit/query/analyze/export data, which are stored in a MySQL database. MySQL data tables are created using standards based data schemas, ensuring that experimental data can be shared with other laboratories and exported to public databases in uniform formats. The database is housed internally within DMCS at Rutgers on secure-servers maintained by the Information Technology group and backed up daily. A project within the LIMS specific to the work proposed here will be created and all relevant personnel at all institutions will be granted authorized, secure access and receive the relevant training on the use of the system. This will greatly facilitate ease of data sharing between our groups and organize our data in a standardized format. Throughout the duration of the project, Big Rose will provide support and maintenance of the database, as well as any new customizations that may be deemed necessary during the course of the research.

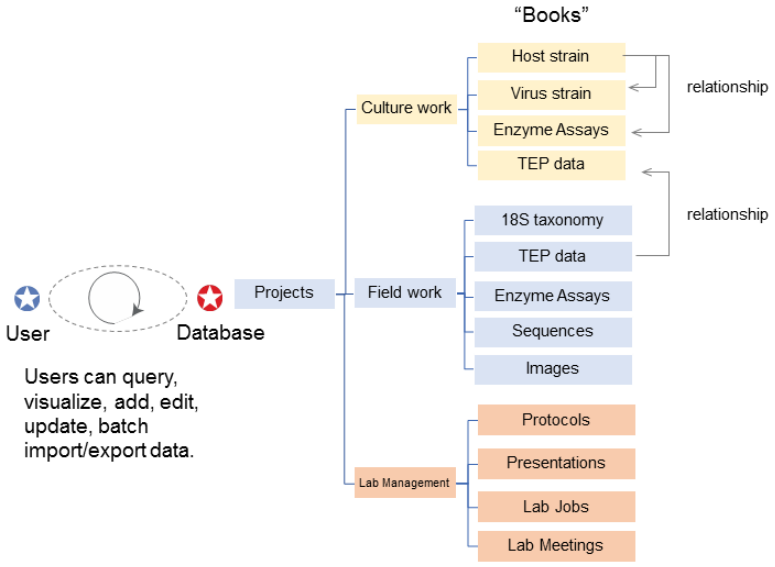

Figure 1. A schematic illustrating the organization and connectivity of the *'laboratory information management system'.* Users can query, visualize, add, edit, and batch import/export any part of the system and can generate customized data reports/files with composite data from different experiment and field campaigns across multiple tables to identify relationships. Data are contained within thematic 'books' which are further contained within a project-specific library. Example data types are shown for conceptual illustration.