Planet Microbe: Toward the integration of oceanographic ‘omics, environmental and physiochemical data layers

Ponsero Alise Jany[1], Blumberg Kai[1], Bomhoff Matt[1], Wood-Charlson Elisha M.[2], and Hurwitz Bonnie[1]

[1]University of Arizona
[2]Lawrence Berkeley National Laboratory

November 16, 2022

Abstract

Oceanographic research cruises produce abundant data, using a wide range of methods and equipment; very often through large collaborative efforts. These research endeavors span a broad array of disciplines and are critical to investigating the interplay between biological, geological, and chemical processes in the ocean systems over space and time. The advent of genomic sequencing technologies allows for the analysis of gene expression in a variety of environmental settings, to measure the distribution and significance of metabolites and lipids in organisms and the environment. Despite scientists' best efforts to carefully curate and share their data with collaborators to advance individual studies and publications, no systematic, unifying framework currently exists to integrate ‘omics data with physical, geochemical, and biological datasets commonly used by the broader geoscience community. As a result, the moment each sample leaves the ship is often the last time each data component appears together in a unified collection. Typically, ‘omics datasets are submitted to nucleotide sequence repositories, whereas contextual environmental data are submitted and stored in specialized data-repositories, or only made available within published papers. This makes it difficult to fully reconnect in-situ data, therefore limiting their reuse in other studies. The development of resources to facilitate the aggregation, publication and reuse of biological datasets along with their physicochemical information is critical for studying marine microbes and the biogeochemical processes in the ocean that they drive. We present Planet Microbe, a cyberinfrastructure resource enabling data discovery and open data sharing for historical and on-going oceanographic sequencing efforts. Several historical oceanographic ‘omics datasets (Hawaii Ocean Time-series (HOT), Bermuda Atlantic Time-series (BATS), Global Ocean Sampling Expedition (GOS)) have been integrated into Planet Microbe along with new oceanic large-scale datasets as the Tara Expeditions and Ocean Sampling Day (OSD). In Planet Microbe, these 'omics data have been reintegrated with their in-situ environmental contextual data, including biological and physicochemical measurements, and information about sampling events, and sampling stations. Finally, cruise tracks, protocols and instrumentation are also linked to these datasets to provide the user with a comprehensive view of the metadata. Additionally, Planet Microbe integrates computational tools using National Science Foundation (NSF) funded Cyberinfrastructure (CyVerse) and provides users with free access to large-scale computing power to analyze and explore these datasets.

www.planetmicrobe.org

Keywords: FAIR principles, marine oceanography, ‘omics datasets

THE UNIVERSITY OF ARIZONA

PlanetMicrobe

HurwitzLAB

Toward the integration of oceanographic ‘omics, environmental and physiochemical data layers

A.J. Ponsero, K. Blumberg, M. Bomhoff, K. Youens-Clark, and B.L. Hurwitz
Department of Biosystems Engineering, University of Arizona

Context

Like a pirate's “buried treasure”, oceanographic data are incredibly valuable allowing us to deepen our knowledge of ocean systems at a higher spatiotemporal resolution. Unfortunately, such heterogeneous data are often separated upon collection and are rarely fully re-united. Despite careful curation, data are sometimes shared only with specific collaborators preventing complete dataset reassembly. Additionally, data are often used only to a address specific questions, limiting their reuse.

Currently, no unifying systematic framework exists for the integration of newer ‘omics data with traditional physical, geological, geochemical, and biological data sets commonly generated by the broader oceanographic community.

‘omics data → Public repositories (NCBI, EBI…)

Geochemical measures → Public repositories (BCO-DMO, Pangea…)

Study-specific measurements → Published papers

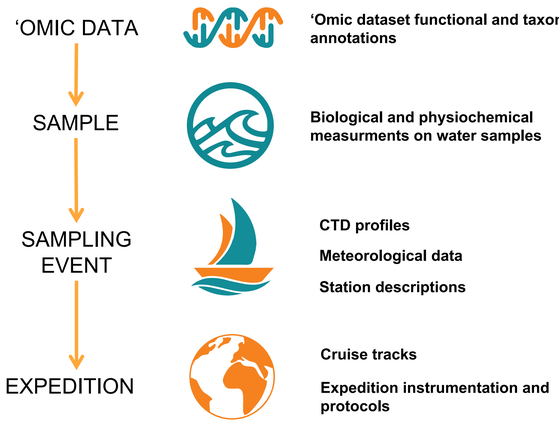
Here, we present here a new web-based cyberinfrastructure platform, **Planet Microbe**, tailored for the reintegration of marine ‘omics datasets with their environmental context.

Our aims

- Reintegration of ‘omics data with their environmental context
- Standardization of semantics for increased data interoperability
- Providing community driven analysis and visualizations tools

Reintegrating ‘omics data in their environmental context

Planet Microbe aims to re-unite taxonomic and functional microbial information from historic ‘omics datasets with their environmental context.



Environmental context can be provided by biological and physio-chemical measurements performed directly on a water sample. At a larger scale, environmental context can be provided by metadata from the sampling event and the sampling station.
Finally, cruise protocols and instrumentation can provide users with a better understanding of the metadata.

Historic datasets

We aim to integrate major historical oceanographic datasets for which abundant environmental metadata were collected.

BATS · TARA OCEANS · OSD Ocean Sampling Day

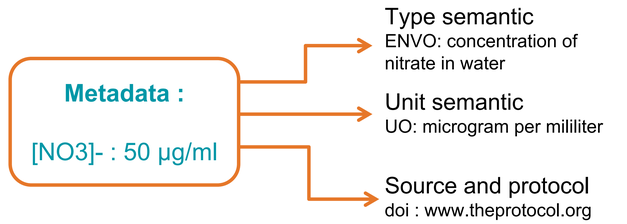
The project includes time-series such as HOT and BATS, as well as large-scale projects such as Tara oceans expeditions and OSD.

Metadata stored in public repositories like BCO-DMO, Rolling Deck repository, CCHDO, Pangea, … or published in research papers will be reintegrated with their respective publicly available ‘omics datasets.

Semantic standardization for data interoperability

Planet Microbe makes use of and extends standardized semantics provided by the Open Biological and Biomedical Ontology (OBO) Foundry and Library.

In order to ensure for well-represented, interoperable metadata **Planet Microbe** combines semantics from a variety of OBO Foundry ontologies including the Environment Ontology (ENVO) to represent metadata types, as well as the Ontology of Units (UO) to represent units.

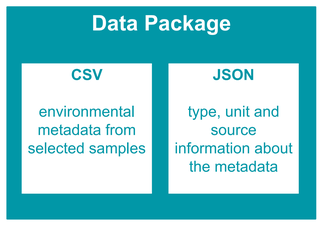


Additionally, a link to the metadata's source protocol is also provided, allowing users to access a detailed description of the instrumentation and protocol used to perform measurements and circumvent measurement bias and error.

Data containers for dataset interoperability

Planet Microbe makes use of frictionless data packages to make datasets interoperable, as well as shareable between systems.

A Data Package is a simple **container** format used to describe and package a dataset. The frictionless data format serves as a convenient wrapper by which to manage and share interoperable data.

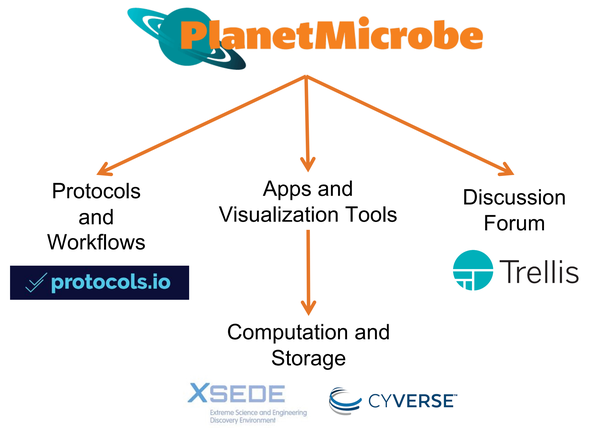


Data Packages can be used to package any tabular data by providing important additional descriptive information. Here, the Data Package allows for the association of each column in the csv metadata file to its OBO type and unit semantics, as well as source description.

Community-driven tools and visualizations

Planet Microbe is not an isolated data platform but operates in a larger cyberinfrastructure system which provide users with the capability to analyze and visualize datasets.

Links to protocol.io and Trellis provides the community with means to share and discuss protocols and workflows.



Planet Microbe provides users with a set of tools and apps for metagenomic analysis and visualization. Compute time and storage is available freely for **Planet Microbe** users through CyVerse and XSEDE resources.

Tool containers for reproducibility

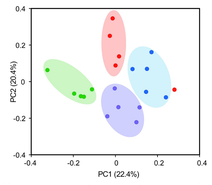
Apps and tools deployed in Planet microbe rely on community-developed docker/singularity containers.

Containers ensure the reproducibility of tools and analytical results in **Planet Microbe** by preserving the code, configurations, and dependencies.

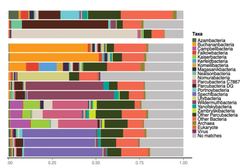
Future directions

GEOTRACES — An International Study of the Marine Biogeochemical Cycles of Trace Elements and their Isotopes · SCOR International Science Council Scientific Committee on Oceanic Research

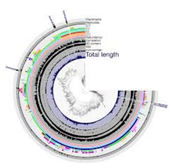
Adding new datasets (Geotraces, Baltic sea DB…)



Adding new tools to analyze datasets



Functional and taxonomic annotation and search function



Adding derived datasets (MAGs, gene/protein catalogues)

This work is supported by the NSF EarthCube Planet Microbe Building Blocks award #1639588 to Dr. Bonnie Hurwitz, and does not necessarily reflect the views of the NSF.