Fine scales shaping nitrogen fIxation in the GUlf stream (FIGURE)

A Data Management Plan created using DMP Roadmap for Eurofleets+

Creators: Mar Benavides, First Name Surname, First Name Surname, First Name Surname, First Name Surname

Affiliation: Your Research Institute

Template: LEFE, ANR

ORCID iD: 0000-0001-9502-108X

Grant number: several

Project abstract:

The biological fixation of dinitrogen (N2) by marine microbes called ‘diazotrophs’ sustains ~50% of primary production in the ocean, boosting CO2 absorption and mitigating climate change. Our knowledge of diazotroph diversity and activity (diazotrophy) derives from studies conducted at very distant spatiotemporal scales: i) discrete and short duration measurements in small seawater volumes isolated from the environment, and ii) spatial extrapolations and global models of diazotrophy projected over decades to centuries. The knowledge gap between these spatiotemporal scales impedes constraining nitrogen inputs and thus quantify and predict the ocean’s potential to withdraw CO2. This gap lies at the fine scales: dynamic seawater structures <200 Km wide and <2 months lifetime. The poor spatiotemporal resolution of oceanographic in situ sampling is incapable of resolving fine scales. FIGURE will bridge this gap by implementing on-cruise high-resolution diazotrophy measurements >10-50 times faster than those available today, focusing on the Gulf Stream. Fine scales will be characterized by underway sensors of current speed, temperature and salinity, vertical nutrient fluxes and satellite altimetry data. The community composition will be examined by molecular biology methods. Diazotroph activity will be measured using high sensitivity trace gas analysis. Physical and biological data will be correlated to elucidate the effect of fine scales on diazotrophy and to assess their impact on nitrogen inputs to the ocean. The achievements of FIGURE will imply a break-through advance in oceanography and stimulate applications in biotechnology and environmental science, providing new tools, approaches and knowledge for climate change adaptation and mitigation.

Last modified: 14-07-2022

Fine scales shaping nitrogen fIxation in the GUlf stream (FIGURE) - Phase 2: Full DMP

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Data will be collected to answer questions about the hydrographical influence on populations of nitrogen-fixing microorganisms. The focus of the data collection is the upper 200 m of the water-column. The purpose is to generate unprecedented spatiotemporal resolution data of diazotroph abundance, identity and metabolism in the ocean.

The cruise proposed here provides a unique sampling opportunity for 5 research projects granted to the chief scientist and two of the collaborators.

The results will be publicly disseminated as well as published in peer-reviewed scientific journals and scientific meetings.

What types and formats of data will the project generate/collect

The project will generate:

"Physics data", including current velocity and direction (ADCP), hydrography profiles (CTD), underway hydrography (thermosalinograph) and weather data.

Raw CTD data will include .hex file with raw cast data, a .bl file with time and scan for each fired bottle, a .hdr file with basic configuration info, a .xmlcon file with detailed sensor info, and additionally a .asc file with SBE 35 thermometer measurements for each fired bottle.

Processed CTD data will be provided as NetCDF files.

Raw SADCP data is produced by the UHDAS program.

Processed SADCP data will be provided as NetCDF files.

"Biogeochemical data": including discrete nitrogen fixation rates (IRMS), diazotroph abundance (qPCR), and amplicon sequencing.

IRMS and qPCR data are provided as .cnv files. Amplicon sequence data are provided as .fastq files.

How is the original data gathered and stored on board and how do you transfer it to shore?

Physics data

All physical data will be obtained onboard as raw data, which will be copied to the ship’s server and to an external hard drive as backup. Preliminary data analyses will be done on board, including processing CTD data with the SBE SeaData Processing and processing the SADCP data with the appropriate software (UHDAS data collection software; https://currents.soest.hawaii.edu/docs/adcp_doc/index.html). Likewise the thermosalinograph, position and meteorological dataset will be explored daily to ensure a proper data acquisition, and saved into the ship’s server and an external hard drive for backup. SADCP data will only be saved.

All these physics datasets will be copied into portable hard drives at the end of the cruise and taken to shore where they will be uploaded to the scientist's lab servers as a backup. Once on shore, the data will be further processed and prepared for database submission.

Biogeochemical data

Biogeochemical data are obtained from physical samples (either seawater or plankton biomass) that need to be transferred from the demobilisation port to each of the partner's labs in cold packages. Specific analytical techniques are then applied and the results thus obtained in each lab (see next section).

What processing on the raw data do you plan? Please differentiate between data quality assurance (handling of outliers, missing and suspect values, null observations) and data harmonisation (code, label and QC flag explanations, consistent use of headers, data formatting, conforming to standards).

Physics data

Raw Data Processing (Common to all Physics Data)

1) Data Quality Control includes identification and flagging of outliers, missing/null and suspect values.

2) Delivery of data at original sampling rates and bin-averaged over 1-m intervals in order to eliminate the very high frequency variability

3) Computation of derived variables

CTD

1) The raw data will be acquired with the SeaBird Data Acquisition software processed with the SeaBird DataProcessing toolbox.

2) Any further spikes will be taken apart.

3) Initial and final samples will be compared to seek for any time-deteriorating trend on the sensors.

Thermosalinometer and weather station

1)Data from the thermosalinograph will be calibrated with the uppermost pressure recorded with the SBE 911+CTD of the ship
2) If available GCPS data from the ship navigation instruments will be compare to that of the thermal and weather station
3) We will obtain true wind by removing the ship’s heading and speed from the manometer.
Acoustic Doppler Current Profiler (SADCP)
1) Use of the UHDAS program https://currents.soest.hawaii.edu/docs/adcp_doc/index.html
Data Harmonisation (Common to all Physics Data)
1) Definition of Headers and a standard code to flag values according to Data Quality Control
2) Data Formatting: use of header, standard variables/units and flagging of values according to 3) Data Quality Control

Biogeochemical data
IRMS
IRMS data will be calibrated with IAEA standards, linearity tests and filter blanks performed with every batch of samples as detailed in Bonnet et al. 2018.
Bonnet, *et al.*, In-depth characterization of diazotroph activity across the western tropical South Pacific hotspot of N2 fixation (OUTPACE cruise). *Biogeosciences* **15**, 4215–4232 (2018).
qPCR
Diazotroph abundance will be estimated by quantitative PCR targeting main diazotroph groups. Samples will be analyzed in triplicates with no-template controls and freshly generated standard curves. Inhibition tests will be performed on a random subset of samples before proceeding with analyses. Raw data will be QCd for outliers before submission.
Sequence data
Amplicon sequencing. Equimolar concentrations of amplicons will be pooled and sequenced (Miseq-Illumina). The R package Phyloseq will be used for alpha & beta diversity analysis & to create co-occurrence networks among sequenced taxa and *nifH* phylotypes.

When do you plan to perform these processing steps?

Physics data
Will be ready for submission to databases within 3 months after the cruise.
Biogeochemical data
qPCR data will be available and submitted to EMODnet within 12 months after the cruise ends.
Raw sequences from amplicon sequencing will be uploaded within 24 months of sequencing and released upon publication (2 year embargo requested). Sequences will be submitted to NCBI and thus publicly available.

What is the expected size (Megabyte to Terabyte range) of the data?

Physics data
SADCP <1Tb
CTD <1Tb
Thermosalinograph <1Tb
Weather <1Tb

Biogeochemical data
IRMS <1Tb
qPCR ~1Mb
Sequencing data ~2Tb

Who will be the principal users of the data? Users are both active (those that clean up or analyse the data) and passive (those that read or assess the data).

All the embarking scientists will use the data for scientific publications. PhD students and postdocs in their labs will also use the data.

2. 1. Making data findable, including provisions for metadata

What naming conventions for your data files will you follow?

All data files will be named with a code including:

- Project name/acronym.
- Data type
- Date file created/generated (in YYYY-MM-DD format)

Example: figure_thermosal_20220712

Can you list some search keywords? The purpose of keywords is to optimize the findability of the datasets.

Gulf Stream, nitrogen fixation, temperature, salinity, current velocity

Do you foresee a need for different versions of the data? Both for your own internal use and when publishing the data? E.g. for some analyses the data might need reorganisation from a common ancestor. Which versioning scheme do you have in mind?

Both raw and processed data will be stored as cnv files according to the requisites of EuroFleets+ for DMPs. Furthermore, physics datasets will also be exported in ODV and netcdf to facilitate use in common oceanographic data analysis software. Users may convert these to other formats for their personal/scientific use.

The Eurofleets+ data repository will allow you to create the metadata when uploading the data. For a list of metadata elements, please refer to the data management guidelines. How will you document all this information before submission, especially lineage information (i.e. processing and QC steps)?

In accompanying readme text files.

2.2. Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions or embargo), explain why, clearly separating legal and contractual reasons from voluntary restrictions. Do this for each type of dataset you will create.

Physics data
CTD, thermosalinograph and weather data will be readily available after the cruise. ADCP will need further processing after the cruise, being made available in databases within 3 months after the cruise.

Biogeochemical data
The scientists involved in this cruise request all the biogeochemical data generated to be embargoed until publication. This is a voluntary restriction and does not imply any legal or contractual constraints.

Do you plan to make the data and metadata available on another repository than the EuroFleets/SeaDataNet data repository, for instance an institutional, national or general data repository?

Physics data
CTD bottle and underway data will also be posted:
-on GO-SHIP https://www.go-ship.org/DataDirect.html
-on SEANOE https://www.seanoe.org/ , this provides datasets with a DOI number that can be cited by downstream users in their publications.

Biogeochemical data
IRMS and qPCR data along with metadata (temperature, salinity, nutrients) will be posted on the MAREDAT database, which currently hosts a global diazotroph database.(https://doi.pangaea.de/10.1594/PANGAEA.817715). This database is expected to be updated in 2022.
Sequence data will be available in NCBI within 24 months.

What methods or software tools are needed to access the data?

Raw cnv files can be accessed with any text or spreadsheet processor. Code used to clean up both physics and biogeochemical data will be posted along with data on the EMODnet Ingestion portal.

Is documentation about the software needed to access the data included?

Only open source software will be used (R, Python).

Is it possible to include the relevant software (e.g. in open source code)?

Yes, we commit to open a GitHub with all relevant code.

Where will the documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

GitHub.

2.3. Making data interoperable

The data management guidelines document the specific meta-information surrounding the data that is needed. Please specify how you plan to capture and store the specified individual meta-information elements.

All onboard operations will be documented in detail in the cruise report to be later incorporated into the databases generated.
Data will be stored in each researcher's drives and institution networks and further transmitted to databases as required.

Notwithstanding the work the reference data centres will perform, do you plan to already make use of standardized definitions (stored in vocabularies) to store the above meta-information?

Indeed.

Do you estimate that you will use uncommon terminology or will generate novel (e.g. new sampling techniques and devices) or project specific scientific terminology? If yes, how will you communicate this information in the above meta-information and make sure they are seen as novel?

The data generated already exists from other cruises and projects, so we will use the same documentation procedures.

2.4. Increase data re-use (through clarifying licences)

How will the data be licensed to permit the widest (by any party) re-use possible?

CC0 license applies.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Physics data will be readily available open access after the cruise. Biogeochemical data will be publicly available after scientific publication in peer-reviewed journals.
Amplicon sequencing data will need a longer embargo (24 months). The data from these samples will not exist until they're processed and sequenced in the lab, an estimated 24 months after the cruise. We are happy to upload these data and the comprehensive associated metadata immediately, but would like to maintain the embargo on dissemination until we have analysed them and submitted a draft publication. This will take longer than usual since analysis is a large time investment and this labour isn't funded - we will process them with similar data from funded projects. We are aware of a few cases where these data were released prior to publication and were then analysed and published by someone who found them online, and we want to prevent this type of problem.

3. Allocation of resources

Who will be responsible for data management in your project?

The PI.

4. Data security

Before the data is transferred to the Eurofleets+ data repository, what provisions are in place for data security (including backups, secure storage and transfer)?

Physics data are obtained onboard as raw data and further processed on land (transfer in hard drive and uploaded to the scientist's lab servers as a backup).

All other biogeochemical data are obtained from physical samples (either seawater or plankton biomass) that need to be transferred from the demobilisation port to each of the partner's labs in cold packages. Specific analytical techniques are then applied and the results thus obtained by each lab (see section 1).

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing?

The waters where the cruise will take place include Bermuda (UK) and USA exclusive economic zones (EEZs). The UK is a signatory of the Nagoya protocol, but the USA is not. After discussion with EuroFleets+ and the Captain of the R/V Atlantic Explorer, we indicate that the vessel is US-flagged but treated as a local vessel in Bermuda. As such, the R/V Atlantic Explorer does not need a clearance to operate in either UK or US waters. The objectives of the project are exclusively scientific.

6. Other issues

On top of the infrastructure and procedures that Eurofleets+ provides, which national/sectorial/ departmental procedures for data management are you following?

N/A.