Establishing “Best practice” data workflows in marine research at GEOMAR, Kiel

Hela Mehrtens, *GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel (Germany)*, hmehrtens@geomar.de
Pina Springer, *GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel (Germany)*, pspringer@geomar.de
Claas Faber, *GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel (Germany)*, cfaber@geomar.de
Lisa Paglialonga, *GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel (Germany)*, lpaglialonga@geomar.de
Carsten Schirnick, *GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel (Germany)*, cschirnick@geomar.de

Background

The joint GEOMAR Data Management Team is a cooperation of GEOMAR Helmholtz Centre for Ocean Research Kiel and several large-scale research projects including collaboration with other marine research institutions. The coalition has established a permanent data management team and infrastructure in Kiel for marine research supporting the entire lifecycle of research data from description through storage and sharing for collaboration to publication. The infrastructure is continuously improved by extensions developed in close cooperation with scientists and data centres. In order to improve the findability and reusability of data, we are on the way to establish and describe reliable and reusable data workflows.

The role of data management in marine science

Marine research is often based on observations made at sea. Research cruises are therefore the starting point of the data workflow. Information on the acquired data (metadata) have to be preserved and collected from the vessels, deployed devices navigation and control systems. The information about research actions on a cruise is used to feed our institutional data information system (OSIS) and forwarded to a central world data centre for publication (PANGAEA) (Fig. 1). This allows to document the existence of data while data analysis is still ongoing. Given a data policy which sets timelines for periods between data sampling and data publication, scientists are supported in tracking their data flow. They are reminded to share their data in reusable formats within their working group and to start the data publication process in parallel or well before publication of their scientific results.

Who is needed to establish a data workflow

A close collaboration between scientists and data manager is necessary to establish a reliable workflow from cruise planning to published and reusable data. Each scientific community has to agree on a set of metadata relevant for their field of study and to discuss them with the data manager to make these metadata visible and searchable in the data repositories. The question of how to describe parameter, methods, quality levels, versions on the one hand and authorship, titles of datasets and references in a consistent way have to be answered along existing datasets and with different groups of interest.

Tools and services to support the data workflow

- data management plan (DMP) tool
- metadata tool

- application of persistent identifiers

- versioning systems

Data management plans describe the expected data and how and when they will be handled, stored and made available. The idea behind the latest tools is not only to assist scientists in writing the plan but to keep it up to date and keep track on the data deliverables and their success, thus leading to a comprehensive data management record.

Metadata tools can serve both as the instrument to collect the necessary metadata and also as an information system, allowing the search and dissemination of metadata and data. It is important to experience that only the metadata description efforts allow a successful search and retrieval later on.

Persistent identifiers (PID) assigned to data and metadata in their different levels (raw data, calibrated data, data products) help to make the data workflow reproducible, especially if the data handling can be documented by adding calibration routines or analysis scripts. Not only data and code but also samples (IGSN), platforms, devices and their sensors, or scientists will be made unambiguously identifiable via PIDs which allows to link back to enhanced descriptions and relations when necessary. Last but not least it allows the citation in papers to foster scientific credit.

Versioning systems play an important role in the data lifecycle. They are used to link data with scripts and documentation, allow structured sharing and collaboration among researchers and provide a convenient way for backup and restoring. One example of such a workflow is a Git project for processing AUV bathymetry data: The data is processed and at the same time documented in a Jupyter Notebook. Data and notebooks are versioned locally while on the ship. After the cruise, the repository is synchronized with the GEOMAR GitLab server, thereby automatically transfering data, scripts and documentation while preserving the processing history.

Results

Workflows for marine data publication are now fairly established at GEOMAR with respect to physical oceanography, sediment cores and underway bathymetry. The workflows for ocean modeling and (mesocosm) experiments are available but still need improvement. One major challenge is to find committed researchers or technical staff with a longterm perspective and standing in the community to ensure the continuity of the data workflow.

We also observe an ongoing need to train both scientists and data managers on existing and new systems and to explore new ways to include data output of new instrumentations. One of the essential data management experiences is the need for on-site personal contact and training of researchers during the entire data lifecycle.

Figure 1 : Data Workflow