iFADO Good practice:
Oceanographic underway data management

Good data management is key to knowledge discovery and innovation and should be carried out using a set of community-agreed guiding principles and practices to easily discover, access, appropriately integrate and re-use, and adequately cite, vast quantities of information. This is the vision proposed by Wilkinson *et al.* (2016) who formulated of a set of foundational principles that all research objects should be Findable, Accessible, Interoperable and Reusable (FAIR) both for machines and for people. These principles have been applied, for example, in the data cataloguing system developed at and in use at the Marine Institute, Ireland (Leadbetter *et al.*, 2020).

Collecting data is usually an exciting task however, the gathering, organizing and quality assessment process it's not so appealing (Figure 1), and data sometimes ends in some hidden folder, poorly documented, with the risk of getting lost. Therefore, the best practice is to prepare the data to be shared, writing careful documentation about the data collection process. This requires a great deal of effort and resources, demanding a practical and easy way to do it. This guide intents to explore and explain the best path to follow since data collection until their submission to an appropriate data cataloguing and sharing service.

This guide is focused on data collected using underway sensors as Thermo-salinometers (TSG) during the oceanographic surveys, but the same general principles are applicable to all environmental data. Usually there are some questions that arise after data collection and we will try to answer them throughout this text, using a simple and informal layout.

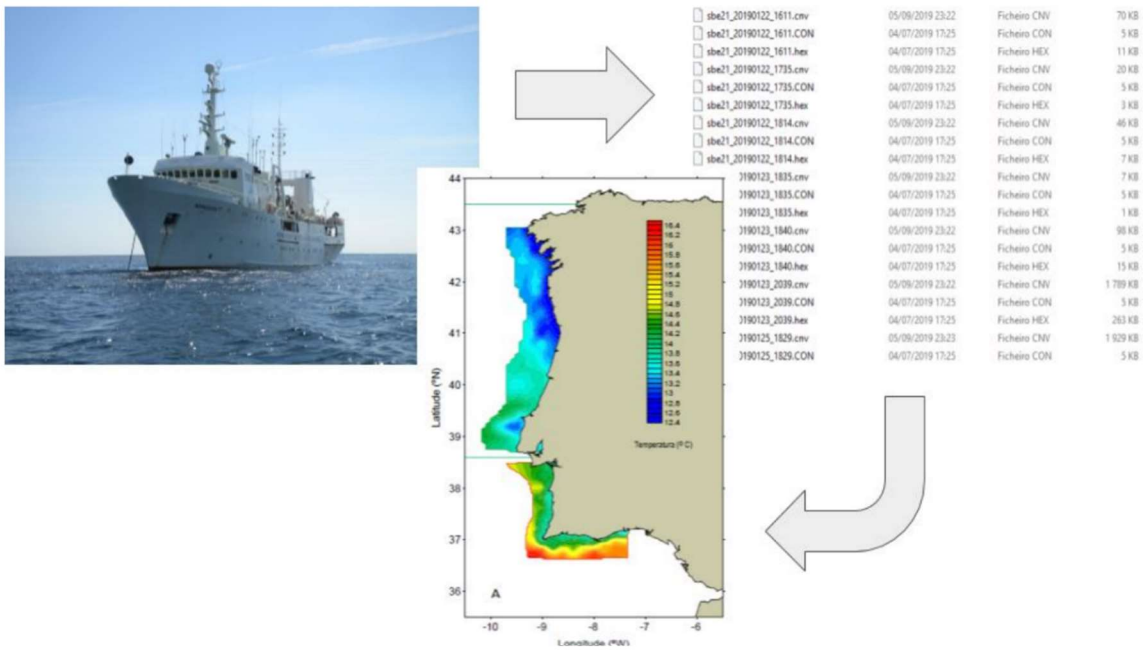

Figure 1 : Data collection, assessment and analysis

How to make raw data look clean and easy to understand?

This is the first step and is essential, because it should be understandable for every user, it will depend on the equipment outputs (formats, file display, etc), but the compilation of the raw data should be composed of the collected variables (ex: temperature) accompanied by time and position always, otherwise the data is useless. The folder management should be intuitive and the way to compile every information collected during the survey should be easy and fast and must keep data that is not analysed yet.

Should I use the sampling time-step? Does it make the final output too big?

Generally, the underway sensors collect data with a high frequency (0.25Hz-1Hz) and are on during all the mission, this makes the amount of data quite big. The choice must be reasonable with the study subject, area and vessel speed. It is common to have a final data output with time-steps including values from 1min to 5min (0.1nmi - 0.7nmi), these values are considered enough to detect differences in temperature and salinity at the surface.

How should I do the data quality assessment?

Again, it will lean on the data collecting procedure, each data retriever will have an individual procedure, equipment and reliance on the job done. So the analyst should first perform it is own quality assessment, just establishing when was the equipment measuring flowing sea-water or not and then use a standard quality control procedure, for example, the TSG-QC from the GO-SUD project, to establish the flags explained in Table 1.

Should I use Quality Control flags?

Yes, this is critical to assess the data usage. It's preferable to elimination of the suspicious measurements, letting the data user to decide if, although the data is not "top quality", it may be usable for his purposes. To apply quality flags the need of a calibrated equipment to compare with another strict measurement (ex: water samples, other TSG) is mandatory. A consistency between the two measurements is the main factor to apply a QC value 1 (Good data, cf. Table 1). Typically, a threshold on the difference between the two measurements is applied. For example, set a QC = 1 if the difference is below 0.1 for temperature and 0.05 for salinity. Besides equipment maintenance, other cases can be subject of suspicion during the data analysis and assessment, such as flow on the conductivity cell lower than required or air bubbles, we qualify these cases with QC value 3. Identified cases during the survey are qualified as QC value 4, such as electronic failures, large debris, seacock closed or insufficient flow in the conductivity cell.

Our quality flags (Table 1) were created based on our equipment and our expected problems, this is quite variable from user to user, however following this nomenclature, we highly recommend at least the flags described by bad data and harbour Quality Control (QC) definition. This is common to every TSG used during an oceanographic survey. This table was inspired by GO-SUD QF.

Table 1 : Definition of quality flags and how are they applied

QC Value	QC Definition	Cases when it is applied
0	No QS was performed	Not enough information to do QC
1	Good data	TSG consistent with other measurement (ex: water samples, other calibrated TSG)
2	Probably good data	TSG doesn't have comparing measurement but was calibrated before the survey
3	Probably bad data	TSG may or may not be collecting good data, not enough information available (ex: TSG not calibrated before survey and no comparing measurement) Flow on the conductivity cell lower than required or numerous air bubble reduces significantly the salinity
4	Bad data	Seacock closed insufficient flow in the conductivity cell shell or large debris inside the conductivity electronic failure
5	Value changed	Manually changed value after expert judgment
6	Harbour	The ship has entered a bay or harbour
7	Not used	Free space for another annotation
8	Interpolated value	Applies to position only
9	Missing value	Typically, NaN or -9999999

How can I write my metadata and catalog the final dataset??

To meet Wilkinson *et al.* (2016) FAIR principles of data management, Leadbetter *et al.* (2020) proposed that a dataset should be described by rich metadata in a searchable resource and the dataset should be assigned a clearly labelled persistent, unique identifier (DOI).

In cases of recurrent datasets (ex: yearly campaigns, fixed buoys, etc.) or organisations that share a big volume data, is usual that they have their organisation details associated to some repository. Later in this guide we will discuss some platforms for data submission, for now we will take SeaDataNet example of how they connect data with metadata. There are a lot of "types" of metadata, if we can call it like that, but metadata varies from organisation name to instrument calibration and SeaDataNet separates it really well and provides a bunch of metadata services. It has his own repository of organisations (EDMO) and projects (EDMERP), this repositories contain a lot of information and make it easy not only to submit data as for data users (example). These repositories are associated with the data submitted in the portals: CDI, EDIOS and EDMED. SeaDataNet also provides a portal where the cruise metadata (CSR) can be shared, however is not associated with the submitted dataset. In Figure 2 we can see how all these repositories are connected.

What is the ideal format to save the data?

This is a very addressed matter, however the answer it is again dependent on the user demands. There are two common formats, NetCDF (Network Common Data Form, file extension .ncdf or .nc) and CSV (Comma Separated Values, file extension .csv). With a NetCDF file all the information is in one file, including the metadata, it's usually a smaller file and, once the users are equipped with the appropriate tools, is easy to manipulate. The main obstacle is that it takes some time to create the first file but, after that, the process will be faster for further surveys. The CSV format it's common fast view table, it's easy to create and manipulate in every programming tools and provides a way to easily look at the data. This type of format on the other hand can turn in a massive file and the metadata must be written in a specific secondary file. Your decision can be helped if you want to submit your data in an online platform as we are going to discuss.

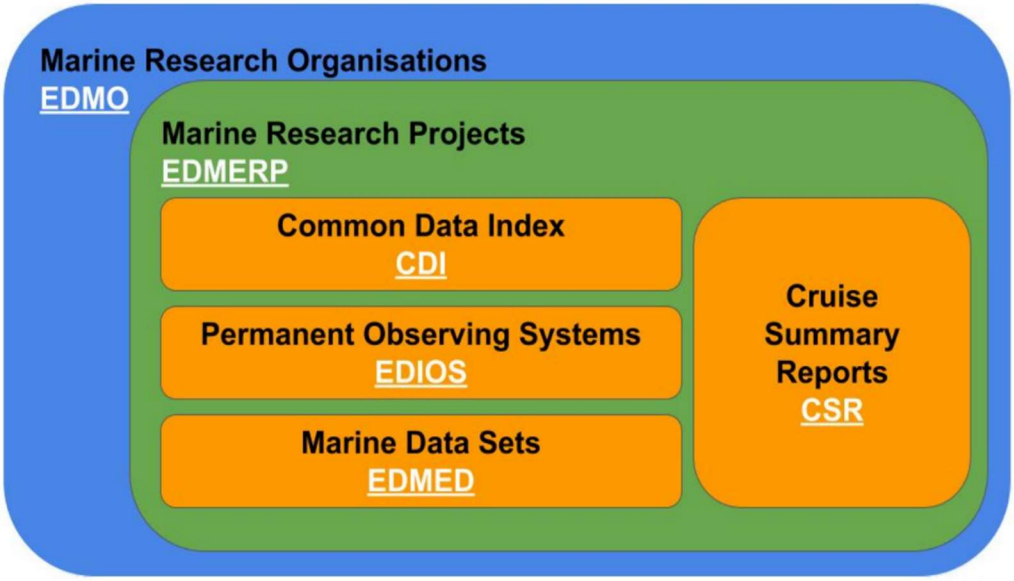

Figure 2 : SeaDataNet Metadata and data portals and connections

How can I share the final data?

There are many places where you can share/save your data since your pen drive to your institution server, nowadays oceanographic data sharing world is presented with a lot of solutions. In this guide we got into the main four European data publisher portals and did an analysis on some main topics that not only concern the data supplier but also the data user. We can see that report in the Table 2.

Table 2 : Data Publisher Portals

	Data Publisher Portals			
	PANGAEA.	SEANOE	SeaDataNet	EMODnet
Platform size	Big - Multidisciplinar - Earth and Environmental Sciences	Small - Oceanography dedicated	Big - Oceanography dedicated	Big - Multidisciplinar - Earth and Environmental Sciences
Data submission	Intuitive and fast	Intuitive and fast	Uses SEANOE as data submitter	Time consuming
Information needed	Dataset name, Authors, keywords, description, geolocation	Dataset name, Authors, Affiliations, abstract, keywords, geolocation	Uses SEANOE as data submitter	Dataset name, Authors, Country, Organization, geolocation
Persistent identifier	provides a DOI automatically	provides a DOI automatically	Uses SEANOE as data submitter	optional DOI
Submission formats	ncdf, csv	ncdf, csv	Uses SEANOE as data submitter	ncdf, csv
Platform possible duplications	ICSU WDS, WIS, GEOSS, OpenAIRE, GBIF, OBIS,GFBio, DataONE	EMODnet, SeaDataNet, EurOBIS	EMODnet and EurOBIS	SeaDataNet, EurOBIS, EGDI, ICES, COGEA
Data latency	Long-term	Long-term	Long-term	Long-term
Platform characteristics	Created to search for a specific dataset (ex: per equipment)	Created to search for a specific dataset (ex: per equipment)	Created to look for specific type of data in a specific region, it also has submitted data and survey reports	Created to look for specific type of data in a specific region, it also has submitted data
Data download	Easy and most of it free Depends on dataset restrictions imposed by the author	Easy and free Depends on dataset restrictions imposed by the author	Easy and free Provide aggregated products	Easy and free Provide aggregated products
Hosting institutions	AWI and MARUM	IFREMER	Governmental Agencies	Governmental Agencies
Years active	since 1995	since 2010	since 2006	since 2017

Conclusions

To sum up the need of answering the questions analysed before, it's important to understand the chronological sequence to reach the final product (Figure 3). This order is not mandatory because each dataset has its own characteristics and sometimes demands other additional tasks.

References

Leadbetter, A., Meaney, W., Tray, E. et al. A modular approach to cataloguing marine science data. Earth Sci Inform (2020). https://doi.org/10.1007/s12145-020-00445-w

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

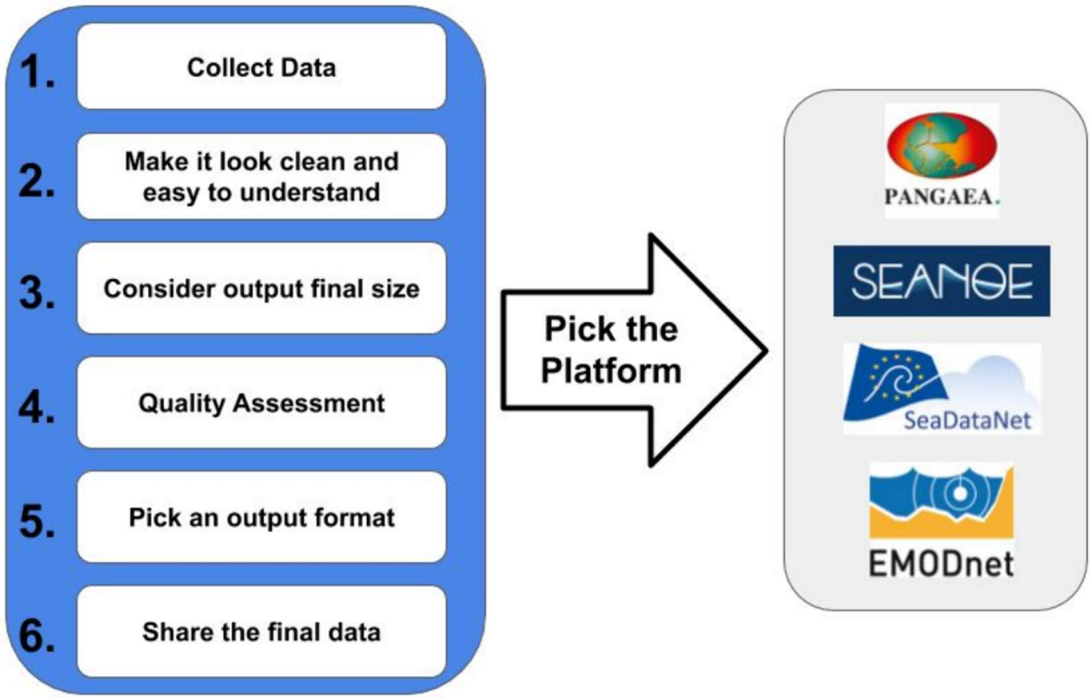

Figure 3 : Chronological sequence from collecting to sharing