Royal Belgian institute for natural Sciences: R/V Belgica Data Management Plan template

Data summary

What types of data will the programme collect?

What sensors/instruments are used to collect it?

When, how and by whom are all the sensors read out?

How is the sensor data sent to shore?

What software is used to read out/analyze/transform the sensor data and generate raw data usable by a human? Ie. there might be a difference between the primary raw data files (voltages) vs processed files (that take into account configuration settings).

What version of the software is needed to read/analyze/transform the data and generate raw data usable by a human?

Describe the general steps used in the software to reach usable, human-readable files.

What are the file extensions of the primary output and of the human-readable files?

Example Answer:
.gizmo for the raw primary output, .txt for the human-readable output.

What further data processing (not data analysis) is performed on the human-readable files before they are archived? Quality control, data reorganisation,...

Data organisation

Please write down the file naming schema. Will you apply a different file naming schema from what the sensor software generates?

Example Answer:
Our file naming scheme is <year>_<campaign_id>_<device>.txt

Do you foresee a need for different versions of the data? E.g. for some analyses the data might need reorganisation from a common ancestor. Which versioning scheme do you have in mind?

Which finalized, global datasets will your research programme generate? Please list each of them. It is useful to already state their final names now. If they fit into an earlier data collection, state the name of the earlier data collection. Take into account the data license (it should be the same for all data points in a dataset!)

Data sharing and licenses

Which organisation is the data owner? Is co-ownership foreseen?

Are there agreements with third parties (contracts, Motions of Understanding,...) that have an effect on data ownership, sharing and reuse? What effects?

Which persons are the data authors? Mention them in the right order.

Which data is free for sharing and reuse? All or some? State which.

When is what data free for sharing and reuse? Possibilities: immediately, embargo x months after the cruise date, yearly releases,....

What is the license for each finalized, global dataset? Choose the following:

- CC0
- CC-BY
- CC-BY-SA

Guidance:

CC0: you release your data in the public domain and forego your right that a user of your dataset has to cite you or else face legal charges. Citation is considered scientific etiquette anyway and the norm, so CC0 is fine for scientific data. Strongly recommended for primary sensor data or datasets with large numbers of contributors.

CC-BY: users can do whatever they want with your data (even commercially), and are legally required to cite you. This is a formal citation process, not just scientific citation that we are used to. Recommended for works of art, shapefiles, or very derived datasets.

CC-BY-SA: same as CC-BY but a derived copy of the dataset must use the same license.

Metadata and documentation

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

An important element in dataset metadata is the lineage. Lineage is: "the different basic steps that have been performed to the

raw data in order to help the interpretation by others users (e.g. statistical analysis, geographical modification,...): global origin, previous file formats the data had, specific steps and transformations taken to clean, compile and present the data. Relations with other data sets." Please explain where you will note down all the transformation steps, so they are not lost for posterity. Where can this document be found?

Guidance:
Ideally, all documentation is stored forever on a github page, together with all software code used for analysis.

Please provide the url where the sensor (correct model and make) documentation (model fact sheet, user and installation manual) can be found. If the manufacturer has no url for this, please upload it somewhere.

Please provide the url towards the user manual for the software (correct version) needed to analyze/transform the raw data. If the manufacturer has no url for this, please upload it somewhere.

Please explain where and how you will store all calibration and maintenance sheets of each sensor of the sensor system, and how you will share them with others.

Which person will perform the sensor calibrations?

Do you use custom (external or self-built) software to transform the data from the raw to the short-term storage? If yes, please explain which and provide an url towards it (eg. closed source executable, GitLab, GitHub,...)

Do you use custom software to transform the data from the short-term storage to the long-term storage? If yes, please explain which and provide an url towards it (eg. closed source executable, GitLab, GitHub,...)

Immediate data storage

On which computer system are the primary data archived for short-term use (analyses, data sharing with colleagues,...)? If this computer is an internal RBINS server, please state the server name.

Guidance:
Possible data repositories are: Kept as files in a file directory, stored in a cloud-based system (eg. SharePoint, SeaFile,...), Ingested in a file repository (VLIZ MDA, BMDC DITS,...), kept in a scientific file server (eg. THREDDS), transformed into a database. In any case, state the name of the data repository and the vendor. Example: "CTD directory on D drive of MS Windows computer ODN-xyz; SeaGizmo database on Oracle 10.6; files saved in BMDC DITS; /home/seagizmo/files/data/CTD on zostera.rbins.be Ubuntu Linux"

What procedure is followed to get the data there?: automated procedure or manual upload after each campaign,...

Example Answer:
We use FileZilla to put the files on the ftp; we have a bash script that copies files over from a memory stick to the server.

How is the primary data backed up and how frequently?

Long-term data storage

Is the short-term storage system suited for long-term storage?

On which system is the data archived for long-term use (analyses, data sharing with colleagues,...)? If this computer is an internal RBINS server, please state the server name.

What are the reasons to store it on this long-term system? Possibilities: redundancy, easier to disseminate to internal and external partners,...

Data availability through requests and dissemination

Which public download links do you provide towards the datasets from within your own organisational platforms? Explain and give the url. Possibilities: a web page with simple file download links; a web page with search interface; a custom API/web service; a web service following some international standard,...

To which regional, federal or international external data portals is the data disseminated (GBIF, SeaDataNet, EMODnet,...)? Make a difference between actual dissemination and desired dissemination.

Do you also respond to bulk data requests if the data volumes would be too high for internal systems to cope with?

Security

How are the different systems (short term and long term) secured against unauthorized access, (accidental) data tampering and data loss?

How high do you judge the risk of data loss?

Responsibilities & contacts

Who (which teams) is responsible for adequate short-term data storage? Also report contact person if different from responsible.

Who (which teams) is responsible for adequate long-term data storage? Also report contact person if different from responsible.

Who (which teams) is responsible for adequate internal or external data requests? Also report contact person if different from responsible.

Who (which teams) is responsible for data dissemination? Also report contact person if different from responsible.

Who (which teams) is responsible for sensor metadata? Also report contact person if different from responsible.

Who (which teams) is responsible for dataset metadata? Also report contact person if different from responsible.