Consiglio Nazionale delle Ricerche

R/V Gaia Blu Data Management Plan

Version 1.0 - October 2024

Table of contents

1 INTRODUCTION

This data management plan has been developed to support the research activities conducted aboard the oceanographic R/V Gaia Blu, a key scientific resource provided by the National Research Council (CNR). Equipped with the most advanced bathymetric and oceanographic study tools, the R/V Gaia Blu is designed for multidisciplinary ocean and ecosystem research. The primary goal of the missions aboard the R/V Gaia Blu is to collect diverse and detailed marine environment data, including oceanographic, geological, geomorphological, biological, and microbiological information. This data is crucial for understanding and monitoring marine dynamics, underwater geological processes, marine biodiversity, and the impacts of environmental changes. The data management plan outlines procedures for collecting, storing and sharing data, ensuring efficiency, reliability, and accessibility for the global scientific community. In accordance with FAIR principles, the data will be made findable, accessible, interoperable, and reusable, using open standards and well-defined metadata. Additionally, data sharing through public repositories is actively encouraged to support research, education, and environmental policy development.

Data management involves several key tasks: properly describing data and metadata (including future contacts and processes), ensuring long-term storage and preservation, documenting dissemination and reuse options through licenses and data sharing, and managing procedures for sensitive data. These tasks must be documented so all partners understand and act according to their responsibilities. This DMP will outline these practices. As Gaia Blu's practices evolve, influenced by the number and objectives of cruises and other deliverables, this DMP will be a living document.

A cruise-specific DMP complements this DMP (see Sec. 4) to carefully document the data specifically collected during the campaign and the management practices.

2 GAIA BLU RESEARCH DATA, SUPPORTING INFRASTRUCTURES, AND DATA FLOW IN BRIEF

There are three major classes of research data collected during a R/V Gaia Blu campaign: (i) data originating from R/V Gaia Blu scientific equipment, (ii) physical samples collected during the campaign, and (iii) data originating from cruise-specific equipment onboarded for the campaign. Suitable metadata must accompany all these research data (see Sec. xx).

Data and metadata originating from R/V Gaia Blu scientific instruments are mainly stored on board the vessel by relying on the onboard data center and transferred, as soon as possible, to a data center guaranteeing long-term storage and availability. The ITINERIS data center, i.e. the Data Center developed by the ITINERIS project[1], will acquire and host these data and the accompanying metadata.

Metadata characterising physical samples are mainly stored on board the vessel by relying on the onboard data center and transferred, as soon as possible, to the ITINERIS data center for long-term storage and availability.

Data and metadata stemming from cruise-specific equipments will be managed according to the cruise-specific data management plan (see Sec. 4). However, this cruise-specific data management plan should be developed thus to guarantee that data and metadata will be managed according to the overall R/V Gaia Blu data policy and to FAIR data management (see Sec. 8).

[1] https://itineris.cnr.it/

2.1 Data Flow from the R/V to Data Center(s)

Details concerning the procedures and strategies implemented by the R/V Gaia Blu on board data center aiming at guaranteeing that data and metadata are acquired, securely stored and properly managed (e.g. backup protocols) when on board are described in a specific document (to be produced).

These data and metadata are transferred from the onboard storage system to LTO tapes.LTO tapes have been selected as Data Transport media due to their large capacity (an LTO8 cartridge offers up 30Tb of data compressed), durability and ease of transport, being light (about 200gr) and robust (e.g. less prone to mechanical disruption)

Before the campaign, sufficient LTO tapes are prepared, labeled and logged in an inventory. The PI is responsible for data transferring to the LTOs. In particular, the PI must verify that data integrity is guaranteed. The tapes are then stored in a secure and climate-controlled environment on the vessel. Inventory of all LTO tapes is maintained.

When the ship arrives at the port, the LTO tapes are prepared for transportation to the main ITINERIS data center. Tapes must be securely packed and labelled to prevent damage or loss during transit. The label should be a progressive number, the cruise name, and the date (CN_YYYY_MM_DD_NN).

- CN is the Cruise Name
- YYYY_MM_DD is the date of the Cruise
- NN is a progressive number

On arrival at the ITINERIS data center, LTO tapes are logged into the main inventory and inspected for any signs of physical damage. Data is then transferred from the tapes to the main storage system, with integrity checks to ensure data completeness. The data is archived in a designated long-term storage and regular backup procedures are implemented to ensure preservation.

In the ITINERIS data center, regular back up procedures are implemented to ensure preservation. The tapes are then stored in a secure and climate-controlled environment and inventory of all LTO tapes is maintained.

3 CRUISE DATA AND METADATA

Every R/V Gaia Blu cruise campaign is characterized by a rich collection of information ranging from research data to administrative data. Part of this information is cruise-agnostic, meaning that its management is not affected by cruise-specific decisions, while there is a part that inevitably depends on the peculiarities of the cruise. Independently from that, this R/V Gaia Blu Data Management Plan describes the shared practices and approaches aiming at guaranteeing comprehensive and homogeneous handling of information stemming from Gaia Blu activities.

3.1 Data summary

The following main categories of data are acquired and managed during R/V Gaia Blu cruises:

- *Data acquired or produced on board*: includes (i) data acquired through equipment available on the R/V Gaia Blu; (ii) data acquired through the processing of physical samples conducted on board with equipment available on the R/V Gaia Blu; (iii) data acquired or physical samples processed using equipment owned by the researcher/research group embarked during the

campaign; (iv) geographic data related to the execution of the campaign. For example: sampling stations, transects, routes.

- *Data related to onboard acquisitions*: includes (i) data resulting from the processing of acquisitions conducted on board; (ii) data resulting from the onshore analysis/processing of physical samples collected during the campaign.

Because of the variety of these data, it is impossible to find a single solution that fits all of them. Moreover, these categories of data will be detailed and transformed into more specific datasets or collections whose management will be carefully detailed in future versions of the DMP.

3.2 Research Dataset Metadata

Research datasets stemming from R/V Gaia Blu activities are charactarised by the classical typologies of metadata including:

- Descriptive Metadata conveys information about the content of the dataset and helps to identify and discover it. They include information like identifiers, title, authors/creators, description, keywords, and dates (e.g. data collection, creation, or publication);
- Structural Metadata conveys information about how the dataset is organized and its internal structure. They include file formats, file size, data models, versioning;
- Administrative Metadata conveys dataset management information, including access, preservation, and rights details.
- Technical Metadata conveys technical aspects of the dataset, such as formats, software and tools required to deal with the data, and storage.
- Statistical/Analytical Metadata conveys information necessary to understand, analyze, and interpret the data including units of measurement and sampling methods.
- Provenance Metadata conveys information about the origins, history, and changes made to the dataset over time.
- Geospatial Metadata conveys geographic information like Coordinates, Projection, Spatial Resolution and extent.
- Reference Metadata convey information about standards and context under which the dataset was created or collected.

Depending on the typology of the research dataset and the solutions exploited to store and publish them, specific metadata standards and formats might be exploited, e.g. ISO 19115 is the most prominent international standard for describing geospatial data. Metadata are a key component of the FAIR data management (see Sec. 8).

3.3 Physical samples metadata

If the collection of physical samples (water, sediments, biological samples, etc.) is planned during the cruise, it is the responsibility of the PI to ensure that samples are properly documented and the relative metadata be published in the Gaia Blu Cruise Inventory (see Sec. 8).

For each sample the following information should be captured by the metadata characterizing it: Sample identifiers; Sampling station identifier; Geographical coordinates; Sample type; Sample container; Sample quantity; Sample quantity unit; Description of the sample processing; Where the samples will be

stored, for how long and who is responsible for storage. Moreover, it must refer to the vessel, the specific cruise and the corresponding project. The publication of these metadata is supported by the Gaia Blu Cruise Inventory.

Sample names must remain consistent throughout the cruise and reflect the sequential order of their collection.

Samples collected during v Cruise activities are expected to be transmitted to established archiving repositories: the archive curator's location and name and contact information should be documented in sample metadata.

4 CRUISE-SPECIFIC DATA MANAGEMENT PLAN

Applicants must develop and implement a cruise-specific data management plan, i.e. a data management plan that complements the data management practices promoted by this document with cruise-specific information concerning data management. In particular, cruise-specific data management plans must identify the datasets the campaign is targeting as well as any data management practice that diverges from what is established by the Gaia Blu Data Management Plan (this document).

Two versions of the cruise-specific DMP are expected: a *preliminary version*, submitted as a PDF during the application, and an *actual version*, to be released after the cruise is granted but before it starts. The actual version of the cruise-specific DMP is a public and living document, i.e. it can be modified by the PIs (whenever convenient and/or necessary) in consultation with the Gaia Blu Data Management Group.

Cruise-specific DMPs must be developed by filling in a specific template[2] and must be compliant with the Gaia Blu Data Policy[3]. They complement the Gaia Blu DMP by explicitly referring to it. Whenever a cruise-specific DMP introduces a data management practice derogating from the Gaia Blu DMP it should also justify the deviation.

Cruise-specific DMPs, both preliminary and actual versions, must contain:

1. Data types gathered during the cruise
2. Data purposes
3. Expected data volumes in GB
4. Brief data processing description
5. Data release plan in open data
6. How metadata of the datasets will be documented
7. Data storage description (i.e. filenames, versioning, folder structures)
8. Information about releasing licenses and data security

The cruise-specific Data Management Plan must precisely outline the data processing timelines and release procedures, detailing any embargoes or release restrictions that differ from the Gaia Blu Data Policy.

[2] Gaia Blu Cruise-specific DMP template

[3] Gaia Blu Data Policy

Researchers or research groups partaking in a campaign will be granted an embargo period of **24 months** during which they will have exclusive use of the data collected unless they decide to release the data. After the 24-month embargo period, the data will be shared with the global scientific community under open science conditions, making it accessible to any interested researchers, institutions, and the public. During the embargo period, data sharing will be limited to the researchers involved in the project and their collaborators, as specified in the initial research agreements.

Each data management activity will be the responsibility of the individual researchers or research groups involved in the project. Specifically, the PI will write the cruise-specific preliminary DMP that must be completed when the cruise is approved and extended when new information is available. The PI will also oversee the data collection on the cruise.

5 STORAGE AND BACKUP

The data management strategy for the oceanographic cruises aboard the R/V Gaia Blu includes comprehensive provisions for data storage, backup, and recovery to ensure the integrity and accessibility of all collected data.

The onboard data center can handle the storage of ship instrumentation data, including the substantial datasets generated by activities such as multibeam and ROV operations. While the onboard data center is generally sufficient for storing the data collected during these cruises, the specific storage needs will be assessed based on the activities planned for each mission. If it is determined that additional storage capacity is required, the costs for these services will be incorporated into the project budget.

To safeguard against data loss, robust backup systems are in place. Data will be backed up using the onboard data center's backup mechanisms. Shared folders accessible to all scientific personnel on board will have backup strategies implemented to ensure data integrity. Furthermore, researchers will have access to virtual machines and applications as a service (such as GIS applications, RStudio, Jupyter Notebook, etc.) for specific data processing and computational tasks to be performed on board. These virtual environments will also be included in the backup protocols.

The PI, in collaboration with the technical support team on the R/V Gaia Blu, will oversee the backup and recovery processes. They will ensure that all data is regularly backed up and that recovery procedures are well-documented and tested. In the event of an incident leading to data loss or corruption, the recovery protocols of the onboard data center will be activated. The technical support team will execute these protocols to restore the data from the most recent backups.

Detailed documentation and training will be provided to all scientific personnel to familiarize them with the backup and recovery procedures. This preparation will ensure that everyone on board is aware of the steps to take in the event of data-related issues and can assist in the recovery process if necessary.

In addition to Gaia Blu on board services, a Virtual Research Environment is available on D4Science premises and accessible via a plain web browser.[4] Any content stored in the VRE and any use of the services offered by this VRE is regulated by specific terms of use.[5]

[4] Gaia Blu Lab Virtual Research Environment https://services.d4science.org/web/gaiablulab

[5] Gaia Blu Lab Terms of Use https://services.d4science.org/terms-of-use

6 SECURITY

To effectively manage access and security for data collected aboard the R/V Gaia Blu, a comprehensive strategy will be implemented, considering the division of responsibilities between the CNR and individual scientists or the PI.

Risks to Data Security and Their Management. Data security risks include unauthorized access, data breaches, and accidental data loss. To mitigate these risks, the onboard data center employs advanced security measures such as encryption, secure user authentication, and regular security audits. These measures ensure that data stored within the system is protected against potential threats. Additionally, all personnel will be trained in data security protocols to minimize the risk of human error.

Control of Access. Access to the data will be tightly controlled using role-based access controls (RBAC). Only authorized personnel, such as the PI and designated research team members, will have access to specific datasets. Each user will have a unique login and permissions will be assigned based on their role and the nature of their work. This system will help ensure that sensitive data is only accessible to those who need it for their scientific activities.

Secure Access for Collaborators. To facilitate secure access for collaborators, virtual private networks (VPNs) and secure file transfer protocols (SFTP) will be used. These methods will allow external collaborators to access the necessary data without compromising security. Additionally, data access can be monitored and logged to track any unauthorized attempts and ensure compliance with data management policies.

Safe Transfer of Field Data. When creating or collecting data in the field, it is crucial to ensure its safe transfer to the main secured systems onboard. Data will be encrypted during transfer and securely transmitted using protected communication channels. Portable storage devices used for data transfer will also be encrypted and handled according to strict protocols to prevent unauthorized access or data loss. Once the data is transferred to the onboard data center, it will be integrated into the primary storage system and backed up immediately.

7 SELECTION AND PRESERVATION

The primary focus for safeguarding will be on the raw and/or original data, as these are the most critical for future research and validation.

Data that must be retained or destroyed for contractual, legal, or regulatory purposes will be managed in strict compliance with relevant regulations and agreements. This includes ensuring that sensitive data, which might be subject to privacy laws or specific contractual obligations, is handled appropriately. Any data required to be destroyed will be securely deleted to prevent unauthorized recovery.

Decisions regarding the retention of other data will be guided by their potential long-term value for scientific research and their relevance to ongoing or future projects. Criteria for this decision will include the data's uniqueness, its potential for reuse in future studies, and its relevance to the scientific community. Raw and original data will be prioritized for preservation due to their foundational role in research integrity and reproducibility.

The foreseeable research uses for the data include further analysis and comparison in subsequent studies, validation of research findings, and contribution to longitudinal studies that track changes over time. By retaining and sharing this data, we facilitate new discoveries and support a collaborative scientific environment where data can be reanalyzed with new methodologies or combined with other datasets.

Data will be retained and preserved for the long term, with specific timeframes depending on the data and regulatory requirements. Raw and original data will be preserved for at least **10 years or longer** if deemed necessary for ongoing research and potential future use. This extended preservation period ensures that valuable scientific data remains accessible for verification, reuse, and integration into broader research efforts.

8 FAIR DATA

FAIR is an acronym for "Findable, Accessible, Interoperable, and Reusable". It is a set of guiding principles that were proposed to improve the management and stewardship of scientific data.[6] These principles are fundamental and do not prescribe any specific technology, standard, or implementation method. Gaia Blu Research Data must be FAIR. In the remaining part of this section it is reported how the Gaia Blu data management practices are organized to implement the FAIR principles.

8.1 Making data findable, including provisions for metadata

Gaia Blu data are discoverable thanks to a specific data catalogue developed via the Gaia Blu Lab virtual research environment. This general-purpose catalogue is configured to host the metadata characterising the typologies of resources to be published, including a ***cruise*** resource to be filled for every campaign. Thus the catalogue is expected to realize the Gaia Blu ***cruise inventory***. Every cruise item is a collector of all cruise-related items discussed in Sec. 4, e.g. the cruise summary report, the CDI, the cruise-specific DMP, technical reports stemming from the cruise, etc. The items linked by every cruise item can be published by the same Gaia Blu general-purpose catalogue as well as by other recognised catalogues and repositories. Cruise resource metadata are inspired by the Cruise Summary Report.

In addition to the general-purpose catalogue, the virtual research environment offers a spatial data catalogue for publishing spatially referenced resources. Resources published into the spatial data catalogue are also automatically harvested and made discoverable via the general-purpose catalogue, thus making them seamlessly discoverable with the rest of the resources stemming from Gaia Blu activities.

All the resources published in the Gaia Blu general-purpose catalogue and the spatial data catalogue are equipped with a *persistent identifier* (an handle). In the near future it is planned to provide Gaia Blu resources with DOIs.

8.2 Making data openly accessible

Gaia Blu datasets will be collected by the Gaia Blu cruise inventory and made openly available by default unless diversely defined in cruise-specific data management plans.

The primary repository of raw datasets is the Gaia Blu data center (see Sec. 2). In addition to the storage of raw data, and depending on the peculiarities of the datasets, the Gaia Blu data center will make available specific repositories. For spatial data, the Gaia Blu data center will offer a standard-based spatial data infrastructure comprising a catalogue service (based on GeoNetwork), several GeoServer repositories and a THREDDS Data Server. This array of services can grow by either making available

[6] 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

additional ones as well as considering as endorsed repositories any repository planned to be used by cruise-specific data management plans.

Standards for accessing Gaia Blu data will be advertised by the metadata characterizing Gaia Blu cruise inventory; namely, every cruise resource will contain a link and enough metadata for accessing the datasets collected by the specific cruise. In addition to that, every dataset will be published on its own with enough metadata, a link to access the data and a link to the cruise resource representing the cruise the data originate from.

8.3 Making data interoperable

The use of the Cruise Summary Report metadata format facilitates data exchange with initiatives collecting this specific typology of data, e.g. the SeaDataNet, the Eurofleet.

The use of spatial data related standards, namely OGC W*S protocols, promoted by the usage of the services offered by the Gaia Blu data center facilitates both data discovery and access.

The content of the entire Gaia Blu cruise inventory is offered by general purposes standards for catalogues, namely DCAT and OAI-PMH, in addition to a REST API, thus facilitating the discovery and access to the contents published by the Gaia Blu Catalogue.

8.4 Increase data re-use (through clarifying licenses)

Cruise summary data will be accompanied by a suitable license promoting its re-use. However, the license of the overall cruise will not impact on the license of the single datasets stemming from it. These licenses are carefully identified by the cruise-specific data management plan.

According to the Gaia Blu Data Policy and what is defined in Sec. 4 concerning embargo, Gaia Blu data will have an embargo period of 24 months maximum. However, the PI is encouraged to act thus to keep this embargo period as short as necessary. Embargo periods, if any, might vary per dataset, i.e. every dataset stemming from a cruise might have its release plan.

9 GLOSSARY

Data Management Plan: a formal document that outlines how data will be handled both during a research project and after the project is completed. It details the strategies for data collection, storage, protection, sharing, and preservation. The purpose of a DMP is to ensure that data are well-managed, accessible, and reusable in the future, facilitating transparency, reproducibility, and long-term preservation of research outcomes.

DMP: see Data Management Plan;

Embargo: a temporary restriction placed on the availability of specific data or information, during which access is limited or prohibited to certain individuals or groups. Once the embargo period ends, the data becomes accessible according to predefined terms and conditions.

Research Vessel: a Research Infrastructure primarily consisting of a ship specifically designed and equipped to conduct scientific studies at sea. Research vessels are used for a variety of purposes, including the collection of oceanographic, biological, geological, and environmental data. They are equipped with advanced instrumentation and onboard laboratories to perform analyses and experiments directly at sea. Research vessels can operate in diverse environmental conditions and are often utilized by research institutions, universities, and government agencies to explore and better understand marine ecosystems and oceanic processes.

R/V: see Research Vessel

10 REFERENCES

https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

https://www.nature.com/articles/sdata201618?error=cookies_not_supported&code=7f02a178-f361-4c2c-9992-f395b4ca303f

https://www.eurofleets.eu/download/Deliverables/D4.4%20Data%20Management%20Plan%20Template%20for%20Funded%20Cruise_Final%20.pdf

https://aquarius-ri.eu/wp-content/uploads/2024/09/D6.2-Data-Management-Plan-V1-MARIS.pdf