National Research Council (CNR) - Institute of Marine Science (ISMAR)

Data Management Plan (DMP)

February 2024

V2.0

HISTORY OF CHANGES			
Version	**Date**	**Changes**	**Authors**
1.0	24/11/2023	First draft	Mauro Caccavale, Lorenzo Corgnati, Valentina Grande
2.0	01/02/2024	Second draft	Mauro Caccavale, Lorenzo Corgnati, Valentina Grande, Consiglio di Istituto ISMAR

Venezia
Tesa 104 - Arsenale,
Castello 2737/F
30122 - Venezia, IT
+39 041 2407911
protocollo.ismar@pec.cnr.it
www.ismar.cnr.it

Bologna
Area della Ricerca
di Bologna –
Via P. Gobetti 101
40129 - Bologna, IT
+39 051 639 8891

Lerici
Forte Santa Teresa,
Pozzuolo di Lerici
19032 - La Spezia, IT
+39 0187 1788900

Napoli
Calata Porta Di Massa
Porto di Napoli snc
80133 - Napoli, IT
+39 081 5423802

Roma
Area della Ricerca
di Roma 2 - Tor Vergata
Via del Fosso del Cavaliere 100
00133 - Roma, IT
+39 06 45488634

Trieste
Area Science Park
Basovizza - Edificio Q2
Strada Statale 14, km 163.5
34149 - Trieste, IT
+39 040 3756872

List of definitions

Best practice: methodology that has repeatedly produced superior results relative to other methodologies with the same objective; to be fully elevated to a best practice, a promising method will have been adopted and employed by multiple organizations.

Data: individual objects or records of any nature (physical or digital), at any level of processing and organization (as per ISMAR Data Policy). In this document we deliberately use the word data very broadly, to comprise data (stricto sensu) and the ecosystem of digital things that relate to data, including metadata, software and algorithms, as well as physical samples and analogue artefacts (and the digital representations and metadata relating to these things).

Data formats: definition of the structure of data within a database or file system that gives the information its meaning. Structured data is usually defined by rows and columns, where columns represent different fields and each field has a defined type, such as integers, floating point numbers, characters, and Boolean. Rows then represent individual records that fill in each column with its corresponding value. Unstructured data includes audio or video objects with a format that can be recognized and played back by software capable of decoding the data from that object. According to FAIR principles, data formats should be self-describing, portable, scalable, appendable, sharable and archivable.

Data Management Office: an organization unit or entity such as a group, physical department, or virtual department within the organization, responsible for facilitating and coordinating data management and supervising efforts and designing, developing, and updating a strategic plan, to manage data and compliance to the related national and the global policies and standards. The main tasks of the Data Management Office are to develop the data strategy, to design the data governance and to write and keep updated the Data Management Plan, to design and coordinate the data management infrastructure, to ensure data quality and usability by continuously assessing the current data landscape (i.e. assess data sources, storage, and processing systems, evaluate data quality, data governance, and security practices, analyze the organization's data capabilities, and support data-driven decision-making).

Data Management Systems (DMS): systems built on data management platforms (a software platform used for collecting and managing data), including a range of components and processes that work together to extract value from data. These can include database management systems, data, data integration tools, analytics, and more.

Data Producer: an individual, or individuals organized into teams, who conduct the observations and/or experiments and generate the data and metadata and, directly or after processing, organize them, according to policies defined by the owner, into records such that they can be shared and used by third parties.

Data product: object derived from any analysis of the raw data, that facilitates an end goal through the use of data.

Data Steward: personnel specialized in responsible and ethical data planning and management (stewardship), he/she plays a crucial role in promoting open science. He/she acts as an intermediary, providing technical, disciplinary, and legal support in data stewardship. The data steward works closely with data producers to ensure that research data are managed responsibly and with respect for privacy and intellectual property rights. He or she can help data producers in identifying best practices for data management, ensure quality, interoperability, accessibility, sharing, and compliance of data with the data policy.

Data value chain: is a series of steps and information flows needed to generate value and useful insights from data

Dataflow: the movement of data through a system comprised of software, hardware or a combination of both.

Digital Persistent Identifier (PID): a long-lasting reference to a digital resource.

FAIR: the FAIR principles act as a guide to data publishers and stewards to assist them in evaluating whether their particular implementation choices are rendering their digital research artefacts Findable, Accessible, Interoperable and Reusable.

Metadata: "data about data". This means that metadata are descriptions or information about an object, an entity, a variable or a datapoint, and the primary purpose of metadata is to describe or annotate the data it accompanies in a machine-readable way.

Naming convention: a convention (generally agreed scheme) for naming things. Its intents may include allowing usuful information to be deduced from name based on regularities, showing relationships, and ensuring that each name is unique for same scope.

Ontologies: a representation of knowledge, generally of a particular domain, written with a standardized, structured syntax that describes the relationship between concepts, also called resources, that serve to characterize the domain[1].

Thematic Hub: an organization unit or entity such as a group, physical department, or virtual department within the organization, that acts as a knowledge centre on a specific research field. It brings together expertise and knowledge on a specific thematic framework and supports (i) the Data Management Office in developing the data strategy and the solutions for implementing the data governance on its specific research field and (ii) the Data Stewards in identifying best practices, scientific procedures (e.g. Quality Assurance of the measurement collection, Quality Checks procedures for thematic data) and operational solutions (e.g. metadata schemas and data formats to be adopted) for guaranteeing the application of the requirements of the Data Management Plan.

Thesauri: similar to ontologies in that they can describe hierarchical and associative relationships between terms. However, they are generally used to facilitate indexing and retrieval of written and recorded items[2].

Vocabularies: a set of terms (words, codes, etc.) that is used by a specific community. They provide a way to make sure that members of the community all agree on the meaning of a term or code[3]. An example is the BODC parameter dictionary[4].

[1] https://marinemetadata.org/guides/vocabs/ont/definition
[2] https://marinemetadata.org/guides/vocabs/voctypes/voccat/thesaurus
[3] https://marinemetadata.org/conventions
[4] https://www.bodc.ac.uk/data/codes_and_formats/parameter_codes/

1. Introduction

A ***Data Management Plan (DMP)*** documents how data will be collected, organized, documented, stored, quality assured, protected, shared and archived along its life cycle (*Figure 1*). The DMP provides descriptive details of the data, the processes, the decisions, the roles and responsibilities and includes a long-term data sharing and preservation plan to ensure data are publicly accessible and reusable, according to FAIR (Findable, Accessible, Interoperable, Reusable) principles[5].

"Data Management Plans (DMPs) are a key element of good data management."

Since data flows and data structures changes during time, related to observational and scientific activities and needs, a DMP is a living document, i.e. its nature is to evolve by anticipating changes in data framework, thus being periodically and continuously reviewed and integrated. This DMP is based on the Horizon 2020 FAIR Data Management Plan template[6] and provides the framework for all data acquired and managed within ISMAR, from acquisition and processing to dissemination, i.e. along the whole data life cycle, in order to maximize the data value chain.

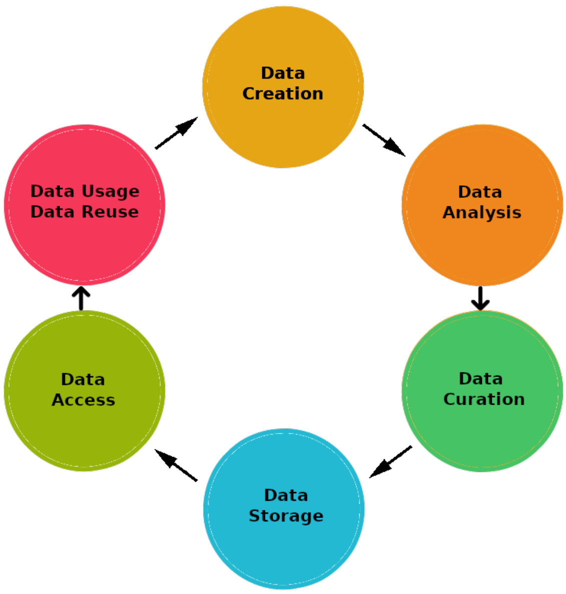

Figure 1 – Schema of the different steps of the data value chain.

[5] https://www.go-fair.org/fair-principles/
[6] European Commission Horizon 2020 FAIR Data Management Plan Template, https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Venezia
Tesa 104 - Arsenale,
Castello 2737/F
30122 - Venezia, IT
+39 041 2407911
protocollo.ismar@pec.cnr.it
www.ismar.cnr.it

Bologna
Area della Ricerca
di Bologna –
Via P. Gobetti 101
40129 - Bologna, IT
+39 051 639 8891

Lerici
Forte Santa Teresa,
Pozzuolo di Lerici
19032 - La Spezia, IT
+39 0187 1788900

Napoli
Calata Porta Di Massa
Porto Di Napoli snc
80133 - Napoli, IT
+39 081 5423802

Roma
Area della Ricerca
di Roma 2 - Tor Vergata
Via del Fosso del Cavaliere 100
00133 - Roma, IT
+39 06 45488634

Trieste
Area Science Park
Basovizza - Edificio Q2
Strada Statale 14, km 163.5
34149 - Trieste, IT
+39 040 3756872

The **data value chain** is a series of steps and information flows needed to generate value and useful insights from data (Curry, 2016[7]).

In addition to these steps, it is important to make well thought out decisions on the types of data to collect and to have good sampling design, which influences the end-use and value of data. The data value chain is the centre of the future knowledge economy, bringing opportunities of digital developments to traditional sectors as part of the Big Data Value Public Private Partnership[8] (BDV PPP) (Zillner et al., 2017[9], European Commission 2013[10]). The same value chain applies for exploiting the value from data within marine science, leveraging large volumes of complex and heterogeneous data originating from multiple data sources. As shown in *Figure 1*, the key steps in the data value chain are:

- **Data Creation:** generation of the data, whether collection of samples and measurements or production via processing algorithms.
- **Data Analysis:** elaborating raw data for making them available for decision-making, other researches or any other activity. It involves exploring, standardizing and transforming data info data products. This includes the use of machine learning and data mining approaches.
- **Data Curation:** the management of data over its lifecycle to ensure it meets quality requirements. It is performed by data curators that are responsible for ensuring trustworthiness and FAIRness of data.
- **Data Storage:** the persistent and scalable management of data to enable rapid access and end-user applications.
- **Data Access:** the ability to retrieve and utilize data stored in a database or other storage system. It encompasses a range of activities, including data retrieval, management, analysis, and protection.
- **Data Usage / Data Reuse:** the activities and applications that use data, e.g. increased automation for data analysis or decision-making based on data products. Reuse is the usage of the data by third parties who may have partially different objectives than the "acquirers" (leading to greater exploitation of the acquired data) and/or may better develop research based on such data. Data reuse can result in the creation of new data.

[7] Curry, E. (2016). The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. In J. Cavanillas, E. Curry, & W. Wahlster (Eds.), New Horizons for a Data-Driven Economy. https://doi.org/10.1007/978-3-319-21569-3.
[8] https://www.bdva.eu/PPP
[9] Zillner, S., Curry, E., Metzger, A., & Auer, S. (2017). European Big Data Value Strategic Research and Innovation Agenda. Retrieved from http://www.bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf
[10] European Commission, Elements of a data value chain strategy, https://digital-strategy.ec.europa.eu/en/library/elements-data-value-chain-strategy

The main parts of the data life cycle are defined as follows in the UNESCO "Guidelines for a Data Management Plan"[11]:

- **Acquisition (collection):** this encompasses the activities by which the data are collected in-situ or remotely. This often includes processes to filter noise, convert from engineering to physical units and even some data smoothing processing. Typically these are actions that do the initial data processing in preparation for post-acquisition processing. Data acquisition bears to data creation.
- **Assembly:** this includes the activities of combining into a single data set, all of the data collected during the acquisition phase. The data may be from a single instrument, or multiple, depending on the nature of the acquisition activities. Appropriate metadata are combined with the data at this stage. Sensor calibration, whether done prior to acquisition or as part of data assembly, should be clearly described in the plan. Data assembly is part of the data curation process, relies on data analysis and enables data access and data reuse.
- **Processing:** this includes all of the data and metadata processing needed to prepare the data for the archives. Typically, it includes quality checks, duplicate resolution, perhaps unit conversions, format changes, etc. Some of these may have occurred in earlier stages and appropriate metadata describing this should be brought together at this stage. Data processing is part of the data curation process, relies on data analysis and enables data access and data reuse.
- **Archiving:** this includes insertion of the data into archives that will provide for later dissemination. Actions can include the typical create, modify, replace, delete actions in an archive. Ensuring long-term storage and preservation is also part of this activity. Data archiving is the core of data storage and enables data access and data reuse.
- **Dissemination:** this activity includes requests for data and the responses to those requests. It includes identification of appropriate repositories and data catalogue systems. Capabilities such as on-line and off-line data access, reformatting and download are included in this aspect of data management planning. Data dissemination enables data access and data reuse.

[11] Guidelines for a Research Data Management Plan. Paris. Intergovernmental Oceanographic Commission of UNESCO,16pp.2016. (IOC Manuals and Guides, 73) (English.) (IOC/2016/MG/73)

Venezia
Tesa 104 - Arsenale,
Castello 2737/F
30122 - Venezia, IT
+39 041 2407911
protocollo.ismar@pec.cnr.it
www.ismar.cnr.it

Bologna
Area della Ricerca
di Bologna –
Via P. Gobetti 101
40129 - Bologna, IT
+39 051 639 8891

Lerici
Forte Santa Teresa,
Pozzuolo di Lerici
19032 - La Spezia, IT
+39 0187 1788900

Napoli
Calata Porta Di Massa
Porto Di Napoli snc
80133 - Napoli, IT
+39 081 5423802

Roma
Area della Ricerca
di Roma 2 - Tor Vergata
Via del Fosso del Cavaliere 100
00133 - Roma, IT
+39 06 45488634

Trieste
Area Science Park
Basovizza - Edificio Q2
Strada Statale 14, km 163.5
34149 - Trieste, IT
+39 040 3756872

The European Marine Board Future Science Brief 6[12] states that marine science is rapidly entering the digital age and the big data framework, and stresses the importance of increasing the value of the wealth of marine big data by making it openly shared, interoperable and integrated into complex transdisciplinary analyses.

"This is only achievable by strictly adopting FAIR principles and by operating the data value chain via well structured data management systems."

ISMAR is in the process of building a **Data Management System (DMS)** based on a Marine Data Archive and a Marine Data Discovery platform for managing data entities and for providing services for the delivery of high quality environmental data, supporting excellence in marine research to better answer societal and policy needs. The ISMAR DMS will take care of data generated by the main data and research infrasctructures, and data that is not yet structured for following a proper and FAIR value chain.

It has to be noted that a well structured DMS adopting the FAIR principles immediately becomes an invaluable resource for processing and collaboration services, e.g Virtual Research Environments (VREs), i.e. services giving access to solutions and "facilities as services" for researchers and users, creating products for EU marine core services and users. This DMP conceptually describes the elements and services of the DMS and refers to it while detailing the data life cycle steps and operations.

The content of this document includes the data types that are managed by ISMAR, the standards that are applied, the actions for data processing, archiving and long-term maintenance, the access policies and the quality standards. Furthermore, the ISMAR data FAIRness is extensively described in a dedicated section together with a number of actions for data archiving and long-term maintenance as well as for the interoperability and the re-use of the collected data. The DMP defines also the roles and responsabilities of the people involved in the data management flow. Finally, brief information about the necessary allocation of the resources and the rules on the data security is also provided in the document.

[12] Guidi, L., Fernandez Guerra, A., Canchaya, C., Curry, E., Foglini, F., Irisson, J.-O., Malde, K., Marshall, C. T., Obst, M., Ribeiro, R. P., Tjiputra, J., Bakker, D. C. E. (2020) Big Data in Marine Science. Alexander, B., Heymans, J. J., Muñiz Piniella, A., Kellett, P., Coopman, J. [Eds.] Future Science Brief 6 of the European Marine Board, Ostend, Belgium. ISSN: 2593-5232. ISBN: 9789492043931. DOI: 10.5281/zenodo.3755793

2. Data summary

2.1. Purpose of the data collection/generation

ISMAR conducts basic and applied research in remote sensing, oceanography, biology and ecology, marine geology, and anthropogenic impacts in Mediterranean, oceanic and polar regions aiming at contributing to the study of ocean processes and climate variability and to the development of systems/services for the observation, protection and sustainable management of the marine and coastal environment. The purpose of ISMAR data management is to provide a continuous and valuable marine data flow to stakeholders, integrating physical, biogeochemical, biological, and ecological information.

Furthermore, ***ISMAR should comply with the common principles that collected observation data sets should be shared in order to become available for uptake and redistribution by the European marine infrastructures***, such as:

- EuroGOOS ROOS's and Copernicus Marine Service for operational oceanography exchange of Near Real Time (NRT) data;
- SeaDataNet for delayed mode and exchange of validated data sets;
- EMODnet, whereby data will provide extra input for EMODnet thematic data products;
- Global Biodiversity Information Facility (GBIF) for dataset on biodiversity;
- ESFRI Research Infrastructure (RI)[13].

Moreover, ISMAR data sets should also be accessible through the Data Management System in order to promote the results of activities and progress in establishing a more streamlined data flow.

2.2. Origin of the data and generated data types and formats

The ISMAR Data Policy[14] defines ***data as individual objects or records of any nature (physical or digital), at any level of processing and organization***. The term "data" is thus used in its most general sense and refers both to the raw data acquired from the sensor, or the physical sample of any nature, and to a product derived from any analysis on the data at any level of processing whether automatic or manual. In this document we refer to the Beijing Declaration on Research Data[15] definition, where the term data comprises data (stricto sensu) and the ecosystem of digital things that

[13] https://roadmap2021.esfri.eu/projects-and-landmarks/view-the-table/
[14] https://www.ismar.cnr.it/rapporti-tecnici/n-23-ismar-data-policy/
[15] https://zenodo.org/records/3552330

relate to data, including metadata, software and algorithms, as well as physical samples and analogue artefacts (and the digital representations and metadata relating to these things).

The ISMAR data policy document classifies data as follows:

- **Raw data**: also known as source data or atomic data, is information that has not been processed to be displayed in any presentable form. The raw form may seem very unrecognizable and almost meaningless without processing, but it may also be in a form that some people can interpret, depending on the situation (e.g. raw standard parameter environmental data with ancillary data, sequence data, biological specimen, water samples, core drilling samples, etc.);
- **Processed data:** data that have undergone processing, understood as the process of collecting raw data and transforming them into usable information. Examples are: calibrated and quality-controlled environmental data, data from non-standard chemical tracers, (e.g., rare earths, isotope ratios, organic pollutants, etc.), data produced by in situ or laboratory experiments, data resulting from routine numerical simulations;
- **Products**: intellectual outputs like theoretical and physical models, data on biological communities and populations, data interpreted and processed both automatically and manually, synthetic graphic representations, thematic maps, source code, software, algorithms, etc.

This DMP adopts these definitions and treats all data classes, as they become available along data life cycle, as summarized in *Figure 2*.

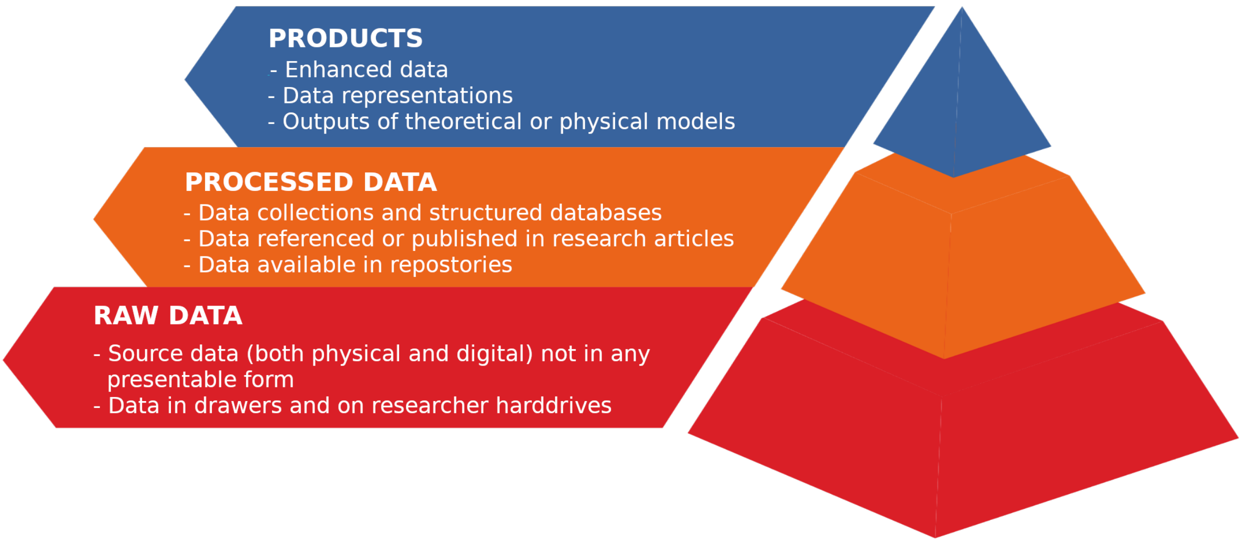

Figure 2 – General data classification.

Furthermore, data can be classified based on the temporal nature of its generation as:

- **Near Real Time (NRT) data**: data generated only with the delay required for electronic communication and automatic data processing. This implies that there are no significant delays, i.e. less than 24 hours. NRT time are automatically collected, streamlined and processed without human intervention.
- **Delayed Mode (DM) data**: data that have been submitted to a data management infrastructure after collection. It is implicitly assumed that this kind of data has undergone rigorous quality evaluations and calibration checks. If this is not the case, then data should be submitted as NRT data. Most of the time, DM datasets can be considered as reference datasets, especially when applied quality control and calibration procedures are openly acknowledged and described.

ISMAR manages a diverse and non-homogeneous data framework which will merge the outputs from the all the scientific communities. A general description of the nature of ISMAR data portfolio can be found hereafter:

- **Physical Samples**: physical objects collected via sampling.
- **Digital Data**: records that are stored on a digital medium, like a computer or a mobile device. They are created, accessed, and maintained in digital form.

Considering the data acquisition methods within the ISMAR framework, data origin can be summarized in the following four categories:

- **Remote sensing data:** data about the Earth's surface or atmosphere collected without actually being in contact with it. Most common remote sensing data are satellite data and High Frequency Radar data. Remotely sensed data have a variety of spectral, spatial and temporal resolutions and different processing levels.
- **In situ data:** data measured or collected at specific known locations where the instruments are present. In contrast to remote sensing data, the difficulty of in-situ data is based on the diversity of parameters: platforms, technique of sampling, processing of data, sampling frequency.
- **Laboratory data:** data collected from experimental studies involving to take measurements which can be manipulated by human intervention. In a laboratory experiment the researcher manipulates the variables under consideration and tries to determine how the manipulation influences the other variables. Laboratory data needs to observe good practice guidelines and the EU legislation/regulations and describe the way an experiment was carried out to enable a specific form of conclusion to be drawn, to provide some level of evidence regarding the way the experiment was conducted, the appropriate protocol or practice used.

- **Modeling outputs:** computational data generated by models, intended as methods for reproducing the reality, schematizing its major features and simplifying the main driving processes with an acceptable loss in representativeness. Models are used to study processes that cannot be directly measured. All model outputs can be reproduced, given information about the model type, setup and input data, and computing time availability.

The data description for ISMAR data will be under continuous development in order to guarantee a standardised data characterisation with better precision and consistent quality during the lifespan of the data. Data are expected to become available in a variety of formats. However, they will have to comply with the relevant international standards before being released through the European Data Infrastructures and via the ISMAR data mangement system.

Many datasets collected within the ISMAR framework are already handled by existing data management systems either operated or joined by the institute research groups and by international partners. The ISMAR Data Management System connects these individual systems to a unified virtual data management system.

2.3. Data distribution towards the main European data infrastructures

During the last decades, a series of standards for data and metadata formats together with exchange protocols have been suggested by international organisations, initiatives and projects like JCOMM, RDA (Research Data Alliance), EuroGOOS, EMODnet, SeaDataNet, Copernicus Marine Service, OGC, GBIF, which have been adapted by the research marine community with data management guidelines that should be followed by ISMAR for publishing of the data and their efficient upload to the major European data portals. Only by implementing these practices, it can be guaranteed that good quality data compliant to FAIR principles will be readily available.

ISMAR should make optimal use of existing infrastructures, standards, services, softwares, best practices and initiatives, to compose its data flow through the use of elements of this data management plan.

Many datasets generated by ISMAR are already integrated into the main European marine data portals (Copernicus Marine Service, SeaDataNet NODCs, EMODnet, etc.), that are built upon the state-of-the-art data management principles and ensure the re-use of the data. The ISMAR DMS itself will implement operations that will facilitate the re-use of the data, based on FAIR principles.

2.4. Expected size of the data

The total amount of data managed by ISMAR is yet not known. The expected volume of the data that will be collected by ISMAR cannot be estimated yet. The total data volume could be several terabytes, but the ISMAR Data Management

System will be set up to efficiently manage such an amount of data. Work is underway for identifying the details of core datasets, in collaboration with the development of the ISMAR Data Management System.

2.5. Data utility

The challenge for ISMAR is to ***ensure that data collected and generated by the institute activities will be accessible and useable by a wider community*** which includes the international ocean science community and other stakeholders in this field.

By applying the actions described in this DMP, ***ISMAR will provide high quality, FAIR scientific data in physical, chemical and biological oceanography and marine geology*** in Mediterranean, oceanic and polar regions aiming at contributing to the study of ocean processes and climate variability and to the development of systems/services for the observation, protection and sustainable management of the marine and coastal environment.

As stated in the institute mission[16], ISMAR data and products will be high valuable resource for understanding (i) the evolution of oceans and their continental margins, studying submarine volcanoes, faults and slides and their potential impacts onshore; (ii) the influence of climate change on oceanic circulation, acidification, bio-geochemical cycles and marine productivity; (iii) submarine habitats and ecology, and the increasing pollution of coastal and deep-sea environments; (iv) the evolution of fish stocks with a view to keeping commercial fishing within sustainable limits and improving mariculture and aquaculture practices; (v) natural and anthropogenic factors impacting economically and socially on coastal systems from pre-history to the industrial epoch.

[16] https://www.ismar.cnr.it/en/institute/

3. Data Management System (DMS)

ISMAR data will be managed through a Data Management System (DMS) consisting in an integrated set of components for collecting, storing, and processing data and for providing information, knowledge, and digital products.

Data Management Systems (DMS) are built on data management platforms (a software platform used for collecting and managing data) and include a range of components and processes that work together to extract value from data. These can include database management systems, data, data integration tools, analytics, and more.

"The DMS will provide several user services to classify, store, organize, document, share and publish, find and reuse the data in a FAIR perpesctive. "

The main platforms for users will be:

- **Marine data archive,** an online but secured data system to manage data files and file versions for a specific context (project, report, analysis, monitoring campaign); to be used as a personal or institutional archive or back-up system, and as a repository for scientific products.
- **Marine data discovery platform**, a service in charge of managing numerous marine data, databases, and information systems which ISMAR is responsible for implementing. The discovery platform is an intelligent system designed to help people explore what's interesting, by giving access to the information (metadata) needed for detailing the available data, and unlock the data value.

The Marine Data Archive (MDA) stores replicated copies of the digital data, and manages the data accesses through authentication and authorizations mechanisms. The Marine Data Discovery Platform (MDDP) hosts the metadata of digital data and physical samples. This platform is a metadata catalogue allowing to integrate, search, find, access, share, preserve, and harvest metadata. The catalogue will allow to manage internal and external Digital Persistent Identifiers (PID), and use/access policies. The MDDP will describe also the physical samples managed in the ISMAR repositories that could be digitized and archived in the MDA. The link between the metadata form and the physical sample will be implemented using a barcoding (*Figure 3*).

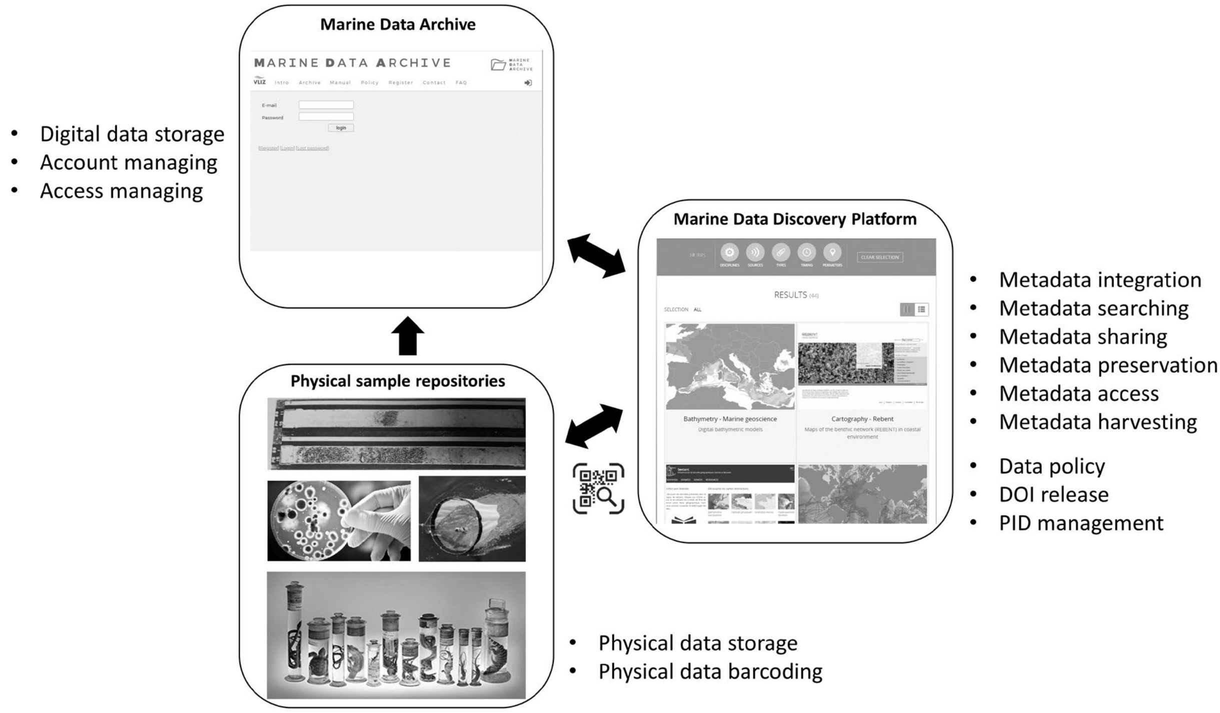

Figure 3. Main components of the ISMAR Data Management System (DMS).

The main objective of the system is to provide access to services about metadata, data, software, algorithms, physical samples, following FAIR principles, and to collect data from data producers. The main services provided by the DMS are described in the following subparagraphs.

3.1. Metadata service

This service allows the data producer to describe the physical or digital resource using a metadata form containing all the information related to the resource (see paragraph *"4.1. Making data findable, including provisions for metadata"*). The metadata service integrates all ingested metadata for feeding the Data Discovery Interface.

Metadata is "data about data". This means that metadata are descriptions or information about an object, an entity, a variable or a datapoint, and the primary purpose of metadata is to describe or annotate the data it accompanies in a machine-readable way.

"Data without associated metadata are simply unusable."

Metadata are the ID card of the resource, they provide essential information for its identification and its comprehension. Metadata also make the document more easily retrievable through search by other users. Generally speaking, metadata compile all the necessary information required to use the data effectively. There are 3 levels of metadata:

- The first level allows data to appear in the inventory of available data;
- The second level informs potential users of the data where to find and how to obtain them (listing conventions of use where relevant, licences);
- The third level allows users to estimate the quantity and adequacy of the data before accessing them (objectives, resolution, quality, limits of use, etc.).

3.2. Quality Check service

The use of data acquisition and processing technologies produces considerable quantities of data. It is extremely important to establish good data management practices in order to describe how the data are acquired and processed. Today this notion of quality is indispensable for ensuring data interoperability and traceability, especially as the notions of quality and level of validation are not "universal" and can vary according to the type of data exploitation.

Beyond the aforementioned mechanisms designed to ensure that data management systems function correctly and are accessible to users, emphasis is also placed on the quality of banked data, in close association with the scientific teams coordinating observations at sea, and in accordance with the major international programmes (ARGO, IODE/IOC, etc.).

The aim of this quality control is to enhance the reliability of data delivered to users, but also to allow the exchange of consistent and comparable data at European and international levels. In compliance with the international standards defined by UNESCO's IOC, ICES and the European Commission (MAST programme) for oceanographic data archiving, the ISMAR Data Management System will perform Quality Checks (QC) on data and metadata. These QCs are divided into three main groups:

- QC 0: Checking of the format;
- QC 1: Checking of the data/metadata location - Navigation check;
- QC 2: Checking of the data.

The first (QC0) and second (QC1) point will be managed by the Data steward (see paragraph *"5. Role and Responsability"*), while the **data producer has the responsibility of the data quality (QC2), that can be guaranteed using best practices and standards.**

A ***best practice*** is a methodology that has repeatedly produced superior results relative to other methodologies with the same objective; to be fully elevated to a best practice, a promising method will have been adopted and employed by multiple organizations.

The Ocean Best Practice System[17] (OBPS) supports the entire ocean community in sharing methods, developing best practices and capacity development in their use. The Ocean Best Practices System under the auspices of the IOC supports the end-to-end best practices value chain. OBPS is a secure, permanent global repository of ocean research, operations, data/information management and applications methodologies (also known as "Best Practices"). The OBPS invites the ocean community to submit their own methodologies to share globally with their colleagues.

The QC service of the ISMAR Data Management System will allow the data producer to define and configure the quality control procedures to be applied on data, and will integrate the related information into the metadata of the dataset.

3.3. Organization service

A well-organized research project will save time, improve reproducibility, and make it easier to share. Deciding how to organize data depends on the specific characteristics of the research activities. A well-organized project includes:

- a well-thought-out folder structure where files can be easily located;
- sustainable file formats that will last in time and are independent of specific software
- an appropriate file naming convention that will make file names comprehensible
- detailed information on the data collection and processing procedures
- Documentation (e.g. README file or similar solutions) that describes the organization system.

"Organizing your data well will enable efficiency and reproducibility of your work."

The ISMAR Data Management System will rely on a well organized data and metadata structure and will provide automatic and seamless features for allowing data producers to organize their data and metadata in flexible and effective ways for maximizing the discovery and access possibilities.

3.4. Classification service

The classification of data depends on how much protection the data requires and it impacts the choice of storage medium, security protocols, and limits the ability to share or archive data.

[17] https://search.oceanbestpractices.org/

By following the state of the art, the ISMAR Data Management System will provide four solutions in term of data classification in according with the ISMAR Data policy:

- **Unclassified**, available for general disclosure. Information that may or should be available to the general public, with no special access restrictions;
- **Restricted**, not for general disclosure. Information which is not open for everyone. There are no laws or regulations saying that the information should be open;
- **Confidential**, available for someone who can be entrusted with information. Information which ISMAR is obliged to protect by law, agreements and other regulations;
- **Secret,** kept or meant to be kept private, unknown, or hidden from all but a select group of people. This category encompasses the same type of information as "Confidential", but where special circumstances makes it necessary to protect the information even more. Demands on protection and safety are to be written down in agreements or other written documentation;
- **Top secret**, of the highest secrecy.

The ISMAR DMP will provide tools for supporting the users in selecting the desired classification level and in managing their data and metadata accordingly.

3.5. Storing/banking service

Banking means storing the data/metadata in files and in databases. Depending on the type, quantity and volume of data and metadata, it may be stored using "simple" files or using databases. The ISMAR Data Management System will provide to data producers a secure space to store data and metadata.

3.6. Citation service

The citation service of a Data Management System gives appropriate attribution to research for improving data discoverability, describing their usefulness and fitness, providing citable contributions to the scholarly record, and supporting long-term reusability. The citation service is based on the usage of identifiers.

An ***identifier*** is a label which gives a unique name to an entity: a person, place, or thing. Unlike URLs, which may break, a persistent identifier reliably points to a digital entity. A ***Digital Persistent Identifier (PID)*** is a long-lasting reference to a digital resource. An ORCID is an example of a persistent identifier for a person. A DOI (digital object identifiers) is a persistent identifier for things or entities such as journal articles, books, and datasets.

*A **Digital Object Identifier or DOI** is a digital identifier of an object, any object — physical, digital, or abstract.* A DOI is a unique number made up of a prefix and a suffix separated by a forward slash. Designed to be used by humans as well as machines, DOIs identify objects persistently. They allow things to be uniquely identified and accessed [18].

"DOIs solve a common problem: keeping track of things".

Crossref and DataCite are the main organizations assigning DOIs for these purposes in scholarly communication. Since 2023, CNR is member of the **DataCite**[19] consortium offering the possibility to attribute a DOI to research outputs and resources, from data and preprints to images and samples.

"Like a citation of a scientific article, it provides visibility for a dataset and its author."

CNR is also member of **mEDRA**[20] that is a DOI Registration Agency for commercial and open access publications.
The **benefits** deriving from the use of DOI are many. Among these, in particular, there are the following:

- persistence: the DOI will continue to function correctly even if the location of the material is changed, they help to keep track of things;
- cooperation with other data, coming from other sources;
- extensibility: the DOI name can be extended, adding new features and new services through the administration of DOI Name Groups;
- platform independence, designed to be used by machine;
- dynamic updates: dynamic metadata, application, and service updates.

The usage of PIDs has increased in the last years and improved the FAIRness of various research related objects (e.g., data, software, researchers, instruments, and research organisations). Wider adoption of PIDs for physical aspects of research can improve the findability and accessibility of these resources, which will allow for data to be put into more detailed context. **By using PIDs all the information about a physical sample could be more easily available, allowing for persistent links to other sources of relevant information** (*Figure 4*). Collection and curation of physical samples is a time-consuming endeavour and deserves recognition in the form of attribution, which could be enabled by making physical resources citable. If digitized, they may become a source of digital data, including but not limited to use of

[18] https://www.doi.org/the-identifier/what-is-a-doi/
[19] https://datacite.org/
[20] https://www.medra.org/en/index.htm

persistent identifiers for samples to support unambiguous citation and linking of information in distributed data systems and with publications, metadata standards for documenting samples and collections and for landing pages, access policies, and best practices for sample and collection catalog, including a broad range of issues from interoperability to persistence (see paragraph “4. FAIRness”).

Table 1 provides an overview of the most commonly used persistent identifiers for physical research components.

Persistent identifier	**Full name**	**Object**	**Webpage**	**Starting year**
IGSN	International Geo Sample Number	Physical samples	https://www.igsn.org/	2007
ORCID	Open Researcher and Contributor IDentifier	Researchers	https://orcid.org/	2009
RRID	Research Resource Identifiers	Resources (antibodies, model organisms and software projects)	https://www.rrids.org/	2014
FundRef	Open Funder Registry	Funder	https://gitlab.com/crossref/open_funder_registry	2016
RAiD	Research Activity Identifier	Research activities	https://www.raid.org.au/	2017
ROR	Research Organization Registry	Research organisations	https://ror.org/	2019
PIDINST	Instruments	Persistent Identification of Instruments	https://www.rd-alliance.org/groups/persistent-identification-instruments-wg	
LSID	Life Science Identifiers	Information resources on the life sciences domain	https://www.lsid.info/	2007
DOI	Digital Object Identifier	Any object — physical, digital, or abstract.	https://www.doi.org/	2010 approved, 2012 published, 2022 revised

Table 1. Overview of persistent identifiers for physical research components, such as samples, resources, instruments, funding and research institutes and researchers[21].

Benefits of linking:

- are easily discoverable through unique identification standards and metadata catalogues and reusable for other analysis;

[21] Plomp, E., 2020. Going Digital: Persistent Identifiers for Research Samples, Resources and Instruments. Data Science Journal, 19(1), p.46.DOI: https://doi.org/10.5334/dsj-2020-046

- are related to other entities, such as sampling methods, persons, instruments, and to the different sub-entities generated from the same samples;
- can be attributed to people and cited (citable PID);
- are semantically enriched (with thesauri and controlled vocabulary).

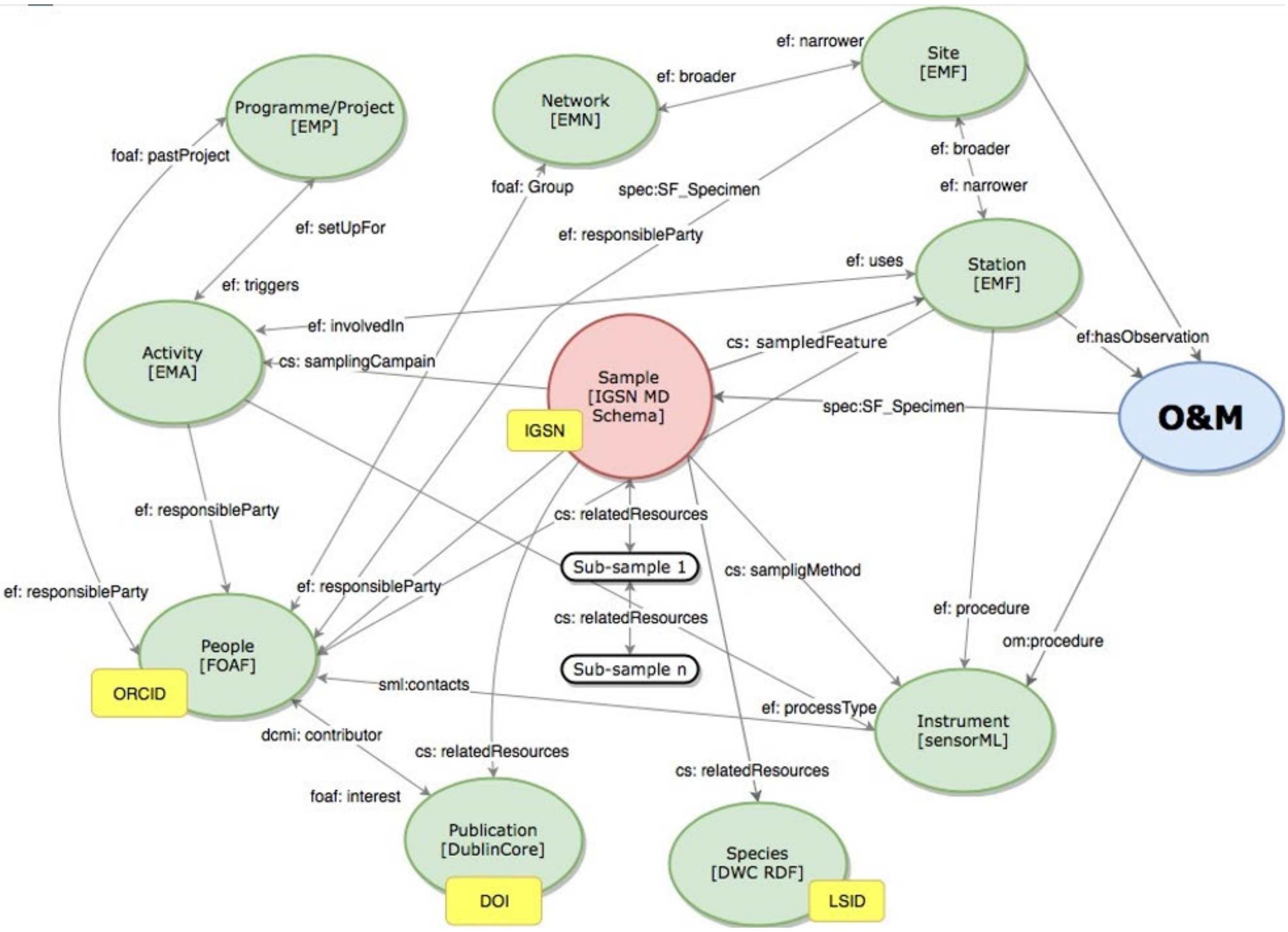

Figure 4[22]. Example of geological sample (e.g. core) described using the IGSN metadata schema and identified by the IGSN code. The physical sample is related to Sub-samples (specimen collected along the core) through the field "relatedResource", the species associated to the sample are described with Darwin Core metadata and identified with the persistent identifier LSID, the publications are described with Dublin core and identified with a DOI, the related peaple are described with the ontology FOAF and identified with the ORCID, ect.

[22] https://doi.org/10.5281/zenodo.1478535

3.7. Sharing service

There are many different reasons highlighting why it is beneficial to share and archive data, apart from complying with publishers' and funders' requirements. **Benefits** of sharing data:

- makes research more robust and reliable by helping prevent data manipulation and fraud
- Increases the citation rates of the project outputs;
- helps avoiding data duplication, that is different research groups collecting the same data;
- increases scientific transparency (by making research more open) and reproducibility (by allowing other researchers to reuse your data and reproduce your results);
- gives more visibility to research;
- opens up for new uses of scientific data and facilitates collaborations with other researchers around the world.

This means that archiving and sharing data contributes to both making science more open and improving research quality and visibility.

"Data sharing, archiving and publishing facilitates data re-use and contributes to more transparent and reproducible research. Here you will find information on which data to archive and publish, and how to do it."

The ISMAR Data Management System will provide the state-of-the-art protocols for sharing the ingested data and for maximising their impact in the scientific community.

3.8. Finding and reusing service

Data's life cycle does not end on publishing or archiving – it can be re-discovered and re-used. When finding and reusing data, scientists are working with secondary data, as opposed to primary data that are directly collected.
Benefits of reusing data:

- Collecting new data takes a lot of resources, both for data acquisition and storage;
- Different sets of primary data can be redundant and different researchers might collect the same types of data unknowingly;
- Increasing amount of data poses high demands for limited storage space;
- Reusing data promotes research transparency and reproducibility;
- Increase data quality, the more others' data are used, the more likely inconsistencies or errors in datasets can be spotted.

3.9. Data access tracking service

The Marine Data Discovery Platform will provide access following FAIR principles and will adopt specific services to evaluate the amount/kind of data search and downloaded to guarantee, always, the optimisation of the system.

4. FAIRness

The **FAIR principle**[23] act as a guide to data publishers and stewards to assist them in evaluating whether their particular implementation choices are rendering their digital research artefacts Findable, Accessible, Interoperable and Reusable (*Figure 5*).

The increasing availability of online resources means that data need to be created with longevity in mind. Providing other researchers with access to data facilitates knowledge discovery and improves research transparency.

In this context, during the Lorentz Workshop "Jointly Designing a Data FAIRport" (2014)[24], participants formulated the FAIR data vision to optimise data sharing and reuse by humans and machines, which resulted in the publication of The FAIR Guiding Principles for scientific data management and stewardship, published in Scientific Data[25].

The FAIR principles describe how research outputs should be organised so they can be more easily accessed, understood, exchanged and reused. Major funding bodies, including the European Commission, promote FAIR data to maximise the integrity and impact of research investment[26].

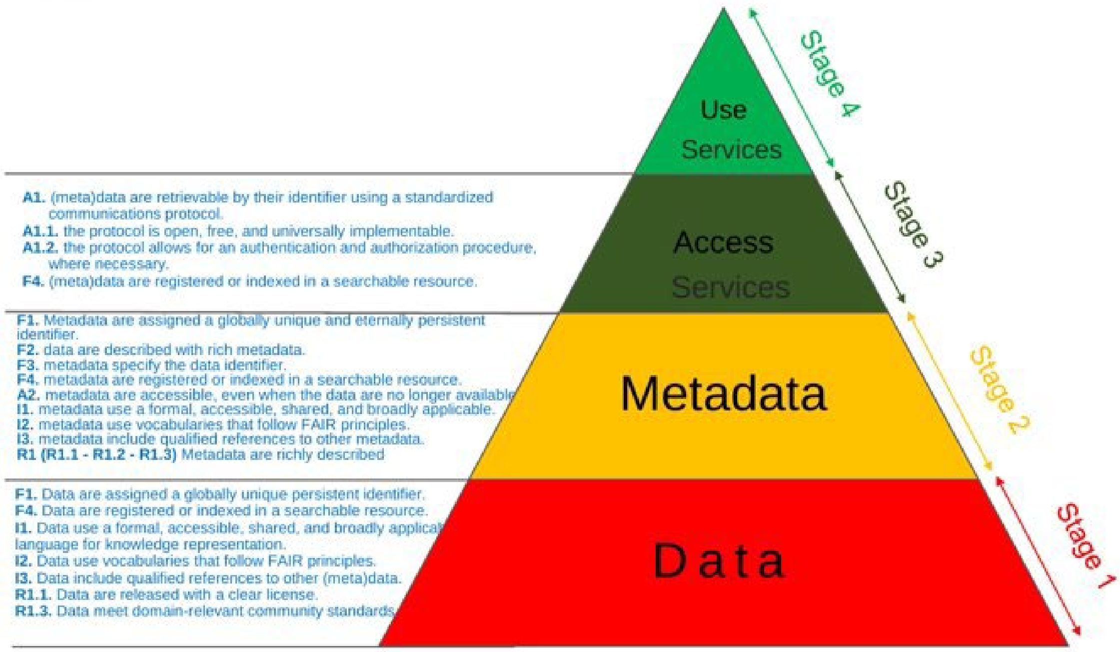

Figure 5. Process to achieve FAIRness. The pyramid shows the hierarchy of the different stages of information content and service capability of the data entities. Each stage is complemented on the left with the actions to be undertaken

[23] https://force11.org/info/the-fair-data-principles/

[24] https://www.datafairport.org/, https://www.lorentzcenter.nl/jointly-designing-a-data-fairport.html

[25] Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18

[26] https://www.openaire.eu/how-to-make-your-data-fair

Venezia
Tesa 104 - Arsenale,
Castello 2737/F
30122 - Venezia, IT
+39 041 2407911
protocollo.ismar@pec.cnr.it
www.ismar.cnr.it

Bologna
Area della Ricerca
di Bologna –
Via P. Gobetti 101
40129 - Bologna, IT
+39 051 639 8891

Lerici
Forte Santa Teresa,
Pozzuolo di Lerici
19032 - La Spezia, IT
+39 0187 1788900

Napoli
Calata Porta Di Massa
Porto Di Napoli snc
80133 - Napoli, IT
+39 081 5423802

Roma
Area della Ricerca
di Roma 2 - Tor Vergata
Via del Fosso del Cavaliere 100
00133 - Roma, IT
+39 06 45488634

Trieste
Area Science Park
Basovizza - Edificio Q2
Strada Statale 14, km 163.5
34149 - Trieste, IT
+39 040 3756872

for implementing that specific stage. Each action is labeled with a letter representing the FAIR principles: F=findability, A=accessibility, I=interoperability, R=reusability.

The ISMAR Data management system will follow the Guidelines on FAIR Data Management in HORIZON 2020 (version 3.0, 26 July 2016)[27]. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation solution. This version of the ISMAR DMP is not intended as a strict technical implementation of the FAIR principles, it is rather inspired by FAIR as a general concept.

ISMAR will apply the FAIR Principles not only to data (stricto sensu), but also to software, algorithms, tools, workflows the led to that data, metadata, as well as physical samples and analogue artefacts (see paragraph "2.2. Origin of the data and generated data types and formats").

The European Code of Conduct for Research Integrity[28] defines as research result, but is not limited to, publications, data, metadata, protocols, code, software, images, artefacts, and other research materials and methods. The Beijing declaration[29] refers to physical samples and artefacts as data, which can include samples such as biological specimens, minerals, soil, sediments, rocks, water, air, art, maps and physical texts, archaeological and synthetic materials, and tissues from humans and animals. The application of the FAIR principles to physical samples, as well as reagents, instrument, software, workflow, etc, can address several challenges that these disciplines are currently experiencing, such as find information that have been used in previuos researches (the detailed sample or resource descriptions are often not publicly available as they are rarely listed in the published literature), formalize a naming convention, track samples or resources across studies. There thus remains a need to well document the information generated using physical samples, that shoud be persistent available, findable, verifiable, and reusable. This practice is more efficient than new samples and field campaigns (not always possible) in terms of costs, time, and resources. Making information and data generated from physical samples more widely available in a standardised manner will facilitate collaboration across different research groups and disciplines, as it will be easier to identify which analyses have already been performed and see where the gaps in knowledge persist. There is an urgent need for better integrating these physical objects into the digital research data ecosystem, both in a global and in an interdisciplinary context to support search, retrieval, analysis, reuse, preservation and scientific reproducibility.

[27] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[28] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/european-code-of-conduct-for-research-integrity_horizon_en.pdf
[29] https://zenodo.org/records/3552330

4.1. Making data findable, including provisions for metadata

ISMAR data will be discoverable with metadata, identifiable and locatable by means of a standard **identification mechanism**. For data and documents DOIs will be used, URLs for metadata. Physical samples will be described with a metadata form; the URL generated by the catalogue will be used to obtain a barcode (a method of representing data in a visual, machine-readable form) to be located on the sample. Data managed through international, European, national or regional data infrastructures will mantain the PID provided by the external system.

To organize and make data findable avoiding mistakes or duplication, following a **naming convention** is a best practice. Suggested principles for developing a naming convention are as follows:

- be consistent;
- give to each file a brief and descriptive name, which may include a combination of elements such as date, location, and researcher's surname;
- put the elements in a particular order to facilitate sorting;
- use underscores or hyphens instead of spaces to separate the elements;
- use letters and numbers instead of special characters like ~ ! @ # $ % ^ & * ( ) ` ; < > ? , [ ] { } ' ");
- standardize the date for chronological sorting, i.e., YYYYMMDD or YYYY-MM-DD;
- include a number at the end to indicate the version (if applicable), e.g.:
    - Original document: Bisondata_1.0
    - Original document with minor revisions: Bisondata_1.1
    - Document with substantial revisions: Bisondata_2.0
    - Document the naming convention
    - Ensure that all members on the research team understand and use the convention

A ***naming convention*** is a convention (generally agreed scheme) for naming things. Their intents may include allowing usuful information to be deduced from name based on regularities, showing relationships, and ensuring that each name is unique for same scope.

Versioning (create a new version of) should be taken into consideration in the process of developing a naming convention. A simple method consists in put the version at the end of the file name, so that, files can be grouped by their name and sorted by version number (i*mage1_v1.jpg, image1_v2.jpg, image2_v1.jpg*). But this method can lead to strange, unhelpful results when sorting is performed (*image1_v1.jpg, image1_v10.jpg, image1_v2.jpg*). A good practice that helps avoiding these problems is to use an international standard date to designate version numbers

(YYYYMMDD or YYYY-MM-DD). Using this order will help avoiding confusion when collaborating with researchers or systems that use different data formats (such as DD-MM-YYYY or MM-DD-YYYY). Also, versions can be sorted in chronological order (*image1_20151021, image1_20151214, image1_20160123*). If files are created and/or edited collaboratively by a team, incorporating the team members' initials into the file naming conventions can be considered. This way, it will be clearer who has created which version of the file:

dataset1_20160402_KES

dataset1_20160501_WTC

dataset1_20160814_GSC

To make data easily findable, ISMAR will create **metadata** using a template with basic information. The metadata form will include the fifteen "core" metadata elements developed by the Dublin Core Metadata Initiative (DCMI) - a global initiative to formally standardize metadata terms - extented with specific elements from the International Standard Organization (ISO) 19115. These fields cover also the mandatory properties of the DataCite metadata schema.

ID	Label	Description	Domain
1	Identifier	An unambiguous reference to the resource within a given context.	String or number conforming to a formal identification system. Examples: Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL), Digital Object Identifier (DOI), International Standard Book Number (ISBN).
2	**Locator**	Location (address) for on-line access using a Uniform Resource Locator address or similar addressing scheme. Location for physical sample.	String or number conforming to a formal identification system. Examples: Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL).
3	Title	The name given to the resource. Typically, a Title will be a name by which the resource is formally known.	Free text.
4	Subject and Keywords	The topic of the content of the resource. Typically, a Subject will be expressed as keywords or key phrases or classification codes that describe the topic of the resource. Search keywords are words used for tagging resources, thus making searches more effective.	Controlled vocabulary or formal classification scheme.
5	Description	An account of the content of the resource. Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content.	Free text.
6	Type	The nature or genre of the content of the resource. Type includes terms describing general categories, functions, genres, or aggregation levels for content.	Controlled vocabulary (for example, the DCMIType vocabulary). Example: collection, dataset, image, physical object, service, software.
7	Format	The physical or digital manifestation of the resource.	Example: pdf, shapefile, netCDF, rock, biological specimen.
8	**Lineage**	Statement on process history and/or overall quality of the spatial data set. Where appropriate it may include a statement whether the data set has been validated or quality assured, whether it is the official version (if multiple versions exist), and whether it has legal validity. The process history may be described by information on the source data used and the main transformation steps that took place in creating the current resource.	Free text.

9	Date	A date associated with an event in the life cycle of the resource. Typically, Date will be associated with the creation or availability of the resource.	ISO 8601 [Date and Time Formats, W3C Note, http://www.w3.org/TR/NOTE-datetime] Example: YYYY-MM-DD.
10	Date type	Event used for reference date The default value is “creation”.	CI_DateTypeCode <CodeList>. Example: creation, publication, revision, deprecated
11	Language	A language of the intellectual content of the resource.	RFC 3066 [RFC 3066, http://www.ietf.org/rfc/ rfc3066.txt] which, in conjunction with ISO 639 [ISO 639, http://www.oasis-open.org/cover/iso639a.html]), defines two- and three-letter primary language tags with optional subtags. Example: "en" or "eng" for English.
12	Source	A Reference to a resource from which the present resource is derived. The present resource may be derived from the Source resource in whole or part.	String or number conforming to a formal identification system.
13	Relation	A reference to a related resource.	String or number conforming to a formal identification system.
14	Coverage	The extent or scope of the content of the resource. Coverage will typically include: - spatial location (a place name or geographic co-ordinates), - temporal period (a period label, date, or date range) or - jurisdiction (such as a named administrative entity). Where appropriate, named places or time periods should be used in preference to numeric identifiers such as sets of co-ordinates, bounding box or date ranges.	Controlled vocabulary (for example, the Thesaurus of Geographic Names [Getty Thesaurus of Geographic Names: http://www. getty.edu/research/tools/vocabulary/tgn/]) .
15	Creator	An entity primarily responsible for making the content of the resource. Examples of a Creator include a person, an organization, or a service. Typically the name of the Creator should be used to indicate the entity. The following properties are expected: • organisation name (free text) • individual name (free text) • position name (free text) • electronic mail address (free text) • onlineResource (String or number conforming to a formal identification system, for example the ORCID) • role (CI_RoleCode <CodeList>). The default value is “creator”.	
16	Publisher	The entity responsible for making the resource available. Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity. The following properties are expected: • organisation name (free text) • individual name (free text) • position name (free text) • electronic mail address (free text) • online resource (String or number conforming to a formal identification system, for example the ORCID) • role (CI_RoleCode <CodeList>). The default value is “publisher”.	
17	Contributor	An entity responsible for making contributions to the content of the resource. Examples of a Contributor include a person, an organization or a service. Typically, the name of a Contributor should be used to indicate the entity. The following properties are expected: • organisation name (free text) • individual name (free text) • position name (free text) • electronic mail address (free text)	

		• online resource (String or number conforming to a formal identification system, for example the ORCID) • role (CI_RoleCode <CodeList>). The default value is "contributor".	
18	**Point of contact**	This is the contact who can best answer questions about the data. The following properties are expected: • organisation name (free text) • individual name (free text) • position name (free text) • electronic mail address (free text) • online resource (String or number conforming to a formal identification system, for example the ORCID) • role (CI_RoleCode <CodeList>). The default value is "pointOfContact".	
19	**Responsible party**	Identification of, and means of communication with, person(s) and organization(s) associated with the resource(s). The following properties are expected: • organisation name (free text) • individual name (free text) • position name (free text) • electronic mail address (free text) • online resource (String or number conforming to a formal identification system, for example the ORCID) • role (CI_RoleCode <CodeList>). The default value is "responsibleParty".	
20	**Classification**	Name of the handling restrictions on the resource. The default value is "unclassified".	MD_ClassificationCode <CodeList>. Example: unclassified, restricted.
21	**Legal constraints**	Restrictions and legal prerequisites for accessing and using the resource or metadata. The default value is "none".	Free text.
22	**Rights**	Information about rights held in and over the resource. Typically a Rights element will contain a rights management statement for the resource, or reference a service providing such information. If the rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource. The default value is "license".	MD_RestrictionCode <CodeList>. Example: copyright, patent, license.
23	**License**	Formal permission to do something.	Controlled vocabulary or formal classification scheme (for example Creative Commons 4.0[30]). Example: CC BY, CC BY-NC-SA. See the ISMAR Data policy.
24	**Other constraints**	Other restrictions and legal prerequisites for accessing and using the resource or metadata. The default value is "Acceptance of the ISMAR Data policy available at the link:(https://www.ismar.cnr.it/rapporti-tecnici/n-23-ismar-data-policy/"	Free text.
25	**Citation**	Standardized resource reference.	Free text.
26	**Metadata file identifier**	It is a Universally Unique Identifier (UUID) for the metadata and not for the resource. This term is automatically filled by the system.	String or number conforming to a formal identification system. Example: 619a4b95-1a82-4006-be6a-7dbe3c9b33c5.
27	**Metadata point of contact**	Party responsible for the metadata information. The following properties are expected: • organisation name (free text) • individual name (free text) • position name (free text) • electronic mail address (free text) • online resource (String or number conforming to a formal identification system, for example the ORCID)	

[30] https://creativecommons.org/licenses/

		• role (CI_RoleCode <CodeList>). The default value is "pointOfContact". This term is automatically filled by the system.	
28	Metadata date	Date that the metadata was created. This term is automatically filled by the system.	ISO 8601 [Date and Time Formats, W3C Note, http://www.w3.org/TR/NOTE-datetime]. Example: YYYY-MM-DD.
29	Metadata language	Language used for documenting metadata. This term is automatically filled by the system.	RFC 3066 [RFC 3066, http://www.ietf.org/rfc/ rfc3066.txt] which, in conjunction with ISO 639 [ISO 639, http://www.oasis-open.org/cover/iso639a.html]), defines two- and three-letter primary language tags with optional subtags. Example: "en" or "eng" for English.
30	Metadata standard name	The name of the metadata standard used to document the resource. This term is automatically filled by the system.	Free text.
31	Metadata standard version	The version of the metadata standard used to document the resource. This term is automatically filled by the system.	Free text.

Table 2. ISMAR metadata profile. In grey, the fields defined by the Dublin Core schema[31], and in blue the fields suggested by the International Standard Organization (ISO) 19115[32,33]. The fields underlined answer to the DataCite metadata schema v4.4 mandatory properties[34].

The idea is providing a simple template applicable to all ISMAR data typologies and formats, allowing not only to describe but also to summarize all the information related to the resource, such as internal and external persistent identifiers (e.g. DOI, EMODnet code, IGSN ID), related people (ORCID) and publications (DOI). Besides these core elements, different disciplines will have different standards for creating and structuring metadata. It is recommended to familiarize with the existing discipline-specific metadata standards before creating the metadata records. The Research Data Alliance (RDA) provides a Metadata Standards Directory[35] that can be searched for discipline-specific standards and associated tools.

4.2. Making data openly accessible

ISMAR data policy[36] defines the policy for access data and information made available by the institution, with the aim of facilitating its dissemination, access and use within the institute and externally. The document defines some

[31] https://www.dublincore.org/specifications/dublin-core/usageguide/elements/
[32] https://inspire.ec.europa.eu/documents/Metadata/MD_IR_and_ISO_20131029.pdf
[33] https://www.seadatanet.org/content/download/4543/file/CDI-profile-V13.0.0.pdf
[34] https://support.datacite.org/docs/datacite-metadata-schema-v44-mandatory-properties
[35] https://rdamsc.bath.ac.uk/subject-index
[36] https://www.ismar.cnr.it/rapporti-tecnici/n-23-ismar-data-policy/

guidelines for sharing data and information to strike a balance between the rights of those who produce them and the need for free, open, and unrestricted sharing and exchange.

The **availability** of the ISMAR data will be decided by the data producer according to the ISMAR data policy. If certain datasets cannot be shared or need to be shared under restriction, the data producer will choose the most appropriate license and will describe the choice in the metadata. The ISMAR data policy addresses also the data produced in multi-beneficiary project or in collaboration with internal and external collegues.

The data **storing** will be performed using one or more repositories connected each others and accessible through a catalogue via web. ISMAR will use an open-source Data Management System for powering a data portal, a web application holding data from different sources, organized under subsets or categories to make it simple for the users of the site to find the data they are looking for. Users will be able to access the data portal easily, no matter what device they are using (desktop, tablet, or smartphone).

The **methodology** to access the data, the software tools, the documentation about the software and its code will be the focus of the next version of the ISMAR DMP, that will define all technical products to implement the ISMAR Marine Data Archive (data repository) and the ISMAR Marine Data Discovery Platform (metadata catalogue). The access to the repository will be managed by the Data Management Office, according to the data policy assigned to the data by the data producer. The metadata catalogue will be free available online to general public, indeed, metadata are always accessible according to the ISMAR Data Policy.

The **conditions of access** are described in the metadata form, where a machine-readable license will be expressed using specific fields, such as “Classification”, “Legal contraints”, “Rights”, “Other constraints”, and “License”. The ISMAR Data policy will guide the data producer in the choice of the most appropriate license.

The main objective of the Marine Data Discovery Platform is to provide a free access to all metadata according to the ISMAR Data policy. For this purpose, the initial access to the Marine Data Discovery Platform will be without any login or user account. The data access will be guarantee using different user levels, with specific privileges. The final user will have to accept the ISMAR Data Policy together with the license indicated by the data producer to download and reuse the data. The access to the Marine Data Archive will guarantee to privileged user (i.e. Data stewards, ICT office, Data Management Office) to upload new data, verify old ones and limit the possibility of data modification and corruption.

The registration info will be treated following the GDPR[37] and used to evaluate the infrastructure impact over the community.

4.3. Making data interoperable

Data will be as much as possible interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available open software applications, and in particular facilitating re-combinations with different datasets from different origins).

To make data interoperable and allow inter-disciplinary interoperability, ISMAR will foster the use of data and metadata **vocabularies**, **thesauri**, **ontologies**, standards and methodologies. In case it is unavoidable to use or generate specific ontologies or vocabularies, ISMAR will provide mappings to more commonly used ontologies.

ISMAR will foster the usage of standardised **formats** for the representation of the data, for example OGC geospatial services allowing requests for geographical features across the web using platform-independent calls, or UniData NetCDF (Network Common Data Format), a set of software libraries and machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data, that is also a community standard for sharing scientific data.

Metadata are readable and thus interoperable for machines without any requirements thanks to the usage of standards. Records are exposed as XML (Extensible Markup Language) on the Internet (over HTTP) by a CSW (Catalogue Service for the Web). Meta(data) can be connected to other online resources through links to people (ORCIDs), scientific papers (DOIs), projects (webpages, DOI), or other meta(data) catalogues. The field “Relation” in the metadata form allows the data producer to define role and relationship with other resources described in the ISMAR catalogue or in other external catalogues.

4.4. Increase data re-use (through clarifying licenses)

To permit the widest possible re-use, data will be licensed in the metadata form using predefinite fields and vocabularies and in accordance with the ISMAR Data policy. The latter answers to questions about the accessibilty and reusability of the ISMAR data, such as: how the data will be licensed, when the data will be available for re-use, if an

[37] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1

embargo is sought to give time to publish, if the data will be used by third parties, how long it is intended that the data remains re-usable.

Concerning the quantity and the quality of metadata provided in order to enhance data reusability, metadata will include information about the provenance of the data, such as origin, history and workflow in the metadata elements “Format” and “Lineage”. To ensure a cross-community understanding and the interoperability with other platforms, metadata records will be filled in English, with the possibility to add other languages indicated in the element “Metadata language”. The standards used for metadata will be indicated in elements “Metadata standard name” and “Metadata standard version”, that are compliant with community standards and machine-understandable. The metadata records are exposed as XML with the HTTP protocol and the CSW allows other infrastructures to automatically harvest them.

To foster the re-use, the data format will be described in the fields “Format” and “Locator” of the metadata form. ISMAR will encourage the usage of standards compliant with machine-understandable community standards, such as the OGC geospatial services for the spatial representation of the data in other infrastructures or desktop environments.

Data generated within ISMAR will be integrated in the major existing data integrators infrastructures, promoting, and facilitating their re-use and redistribution. As mentioned above, the data will be distributed with a minimum set of metadata making possible the traceability of their originator and the feedback on their use.

5. Roles and responsibilities

The organization chart responsible for implementing the actions described by the ISMAR Data Management Plan is built of four entities: the **Data Management Office**, the **Thematic Hub**, the **Data Steward** and the **Data Producer** (*Figure 6*). The coordinated actions of these actors allows for having all data generated by the institute organized, stored, and properly described and formatted for being compliant with FAIR principles.

It is meant that, if not differently stated by the data policy, the data owner is the ISMAR institute.

The **Data Management Office** is an organization unit or entity such as a group, physical department, or virtual department within the organization, responsible for facilitating and coordinating data management and supervising efforts and designing, developing, and updating a strategic plan, to manage data and compliance to the related national and the global policies and standards. The main tasks of the Data Management Office are to develop the data strategy, to design the data governance and to write and keep updated the Data Management Plan, to design and coordinate the data management infrastructure, to ensure data quality and usability by continuously assessing the current data landscape (i.e. assess data sources, storage, and processing systems, evaluate data quality, data governance, and security practices, analyze the organization's data capabilities, and support data-driven decision-making).

The **Thematic Hub** is an organization unit or entity such as a group, physical department, or virtual department within the organization, that acts as a knowledge centre on a specific research field. It brings together expertise and knowledge on a specific thematic framework and supports (i) the Data Management Office in developing the data strategy and the solutions for implementing the dat governance on its specific research field and (ii) the Data Stewards in identifying best practices, scientific procedures (e.g. Quality Assurance of the measurement collection, Quality Checks procedures for thematic data) and operational solutions (e.g. metadata schemas and data formats to be adopted) for guaranteeing the application of the requirements of the Data Management Plan. ***A Thematic Hub per each thematic area covered by ISMAR activities will be established: Remote Sensing Thematic Hub, Oceanography Thematic Hub, Biology and Ecology Thematic Hub, Marine Geology Thematic Hub and Anthropogenic Impact Thematic Hub.***

The **Data Steward** is specialized in responsible and ethical data planning and management (stewardship) and plays a crucial role in promoting open science. He/she acts as an intermediary, providing technical, disciplinary, and legal support in data stewardship. The data steward works closely with data producers to ensure that research data are managed responsibly and with respect for privacy and intellectual property rights. He or she can help data producers in identifying best practices for data management, ensure quality, interoperability, accessibility, sharing, and compliance of data with the data policy. ***Each ISMAR section will have a Data Steward*** who gives, under the counseling of the interested Thematic Hubs, operational support to the data producers of his or her section in implementing all the practical steps for generating FAIR data and for ingesting them into the ISMAR Data Centre.

The **Data Producer** is an individual, or individuals organized into teams, who conduct the observations and/or experiments and generate the data and metadata and, directly or after processing, organize them, according to policies defined by the owner, into records such that they can be shared and used by third parties. Thus, the data producers are possibly all ISMAR researchers and employees generating data.

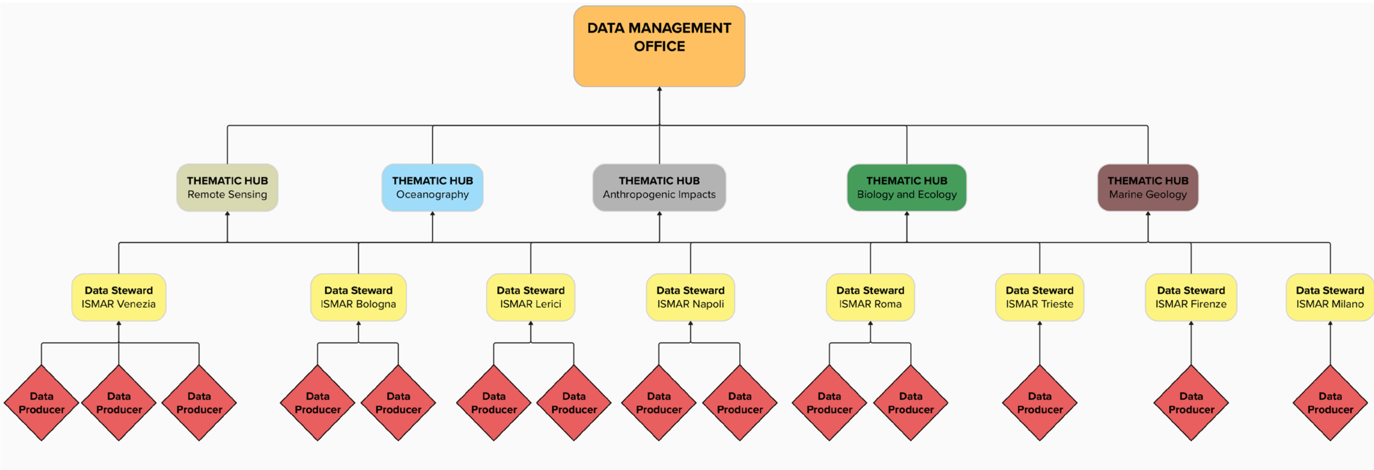

Figure 6 – Organization chart of ISMAR Data Management.

6. Data flow

This paragraph defines the flow of data through the ISMAR Data Management System. A **Dataflow** is the movement of data through a system comprised of software, hardware or a combination of both. The systematic sequence of tasks can be summarized in the following phases (*Figure 7*):

Phase 1: the Data Producer generates and organizes data and metadata following the best practices described in the ISMAR DMP with the support of the Data Steward of the relative ISMAR section and the Thematic Hub of the related research field;

Phase 2: the Data Steward and the Thematic Hub validate data and metadata (steps QC0 and QC1, see paragraph *"3.2. Quality check service"*);

Phase 3: the Data Steward prepares data and metadata to be ingested in the ISMAR DMS, he/she releases data in the Marine Data Archive (MDA) and metadata in the Marine Data Discovery Platform (MDDP), when possibile data and/or metadata will be also deployed in external data and research infrastructures on the basis of Thematic Hub recommendations.

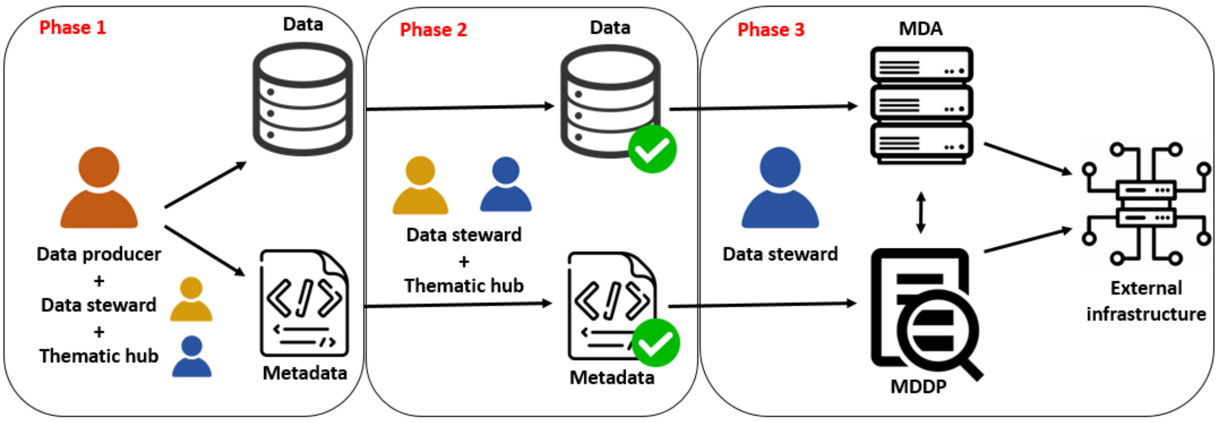

Figure 7. Data flow diagram that syntetizes the phases (or steps) that the actors involved should execute to successfully deliver data and metadata.

The data/metadata produced by the ISMAR will be accessible through the Marine Data Archive (MDA) and the Marine Data Discovery Platform (MDDP). Through the ISMAR Data Management System (DMS), it will be possible to discover available datasets, download data and access to the services of the marine domain (e.g.: data/metadata conservation or organization).

The access to the ISMAR interfaces, the amount of data searches and data downloads will be managed by specific services to guarantee, always, the optimization of the data/metadata provided and their impact over the community and the final users.

The data and metadata will be updated as soon as recorded (Near Real Time – NRT), at regular intervals (e.g.: hours or daily) and/or according to the specific needs/requirements of the Data producer in order to always guarantee the maximum efficiency of the system for data/metadata identification and retrieval. Where possible, data will be shared immediately after acquisition and/or production. For the other cases, see the moratorium periods described in the ISMAR Data policy.

7. Allocation of resources

The ISMAR Data Management System represents the main structure of the marine institute and provides access to the data/metadata collected. The infrastructure, endowed with considerable computing capacity and storage space, guarantees discovery and access to all ISMAR data based on the FAIR principles in coordination with the international repositories (European RIs and marine data portals) and provides primary data storage (or a backup copy of the data of the partners who request it). Specific services can also be provided at the request of the ISMAR section and/or specific projects according to future needs.

For the provision of these services and the management of a network of geographically distributed sections, the Data Management System is structured so that all its components (in terms of redundant hardware and software) can guarantee the efficiency, stability and operational continuity of the services and security of the stored data/metadata. The computing power, the data/metadata storage capacity and the overall performance of the system guarantee the possibility of adopting different approaches of "data flows" within the domain and with the outside each based on specific interconnections and data exchange.

The connection with the international data repositories will be realized via the state-of-the-art sharing protocols and, where possible, using specific Application programming interface to automatize the data workflow.

The infrastructure provides all information and communication technology (ICT) services for all users and the public. The system services should be adapted and will evolve according to the requirements of the research activities and specific project needs. However, some basic services can be identified:

- Storage and access to data/metadata for all users.
- Providing the possibility of a security copy of the data.
- Providing the possibility of a long-term off-line copy for data.
- Provide access and interchange of data/metadata with other institutions and with other infrastructures (IODE, COI, SeaDataNET, ESFRI Research Infrastructures, etc)
- Other services that will be highlighted in the future from ISMAR or scientific community.

The main components of the Data Management System are the repository, the data and metadata catalogue, the metadata registry, the service catalogue and the portal.

ISMAR data and metadata represent a very voluminous and heterogeneous set due to their naturally diverse research areas and standards. This creates difficulties in adopting a geographically widespread and federated approach at section level for the management, storage and dissemination of the data produced by the institute.

This difficulty is also often encountered at the level of the project or research group, thus making necessary and desirable to define shared strategies at institute level to guarantee the correct data workflow and availability over time and secure storage.

The harmonization of data/metadata and the axis procedures applied by the different local levels will contribute to the rapid establishment of the ISMAR Data Management System as a distributed and cooperating system to ensure the continuity of data/metadata and services.

Regarding the data formats used, there is no uniformity among the different research areas. The prevailing trend is to provide the same data in different formats to encourages a shared and non-exclusive approach.

Particular attention is paid to the possibility/need to guarantee different data output formats only on demand. This solution makes it possible to diversify the output format for the end user without necessarily duplicating the stored data.

In the current situation, it is not possible to have a precise and detailed overview of every aspect related to the resources needed to store the data produce. Neither it is possible to have a detailed overview of the costs of running such an infrastructure or to make the data FAIR.

At this stage it is not possible to assess the **costs** of making the data/metadata available and compliant with the FAIR principles. An assessment can be made once the long-term data/metadata retention policies and the precise flow of data/metadata have been established.

However, the distributed and federated structure that is being set up has been designed in such a way to guarantee high efficiency in terms of data/metadata security, efficiency in its dissemination and interoperability, and the necessary computing and storage resources for current and future long-term data.

As mentioned before, the present and future total amount of data is not known. However, the data centre infrastructure is calibrated to easily manage large amount of data and grove with it. By way of example only, it should be noted that, as a major value, a single observation node can produce more than 100 TB of data per year.

The ISMAR Data archive will be distributed (both in terms of hardware and data/metadata copies), and federated through secure interconnections for information exchange and data/metadata retrieval.

8. Data security

One of the priorities in ISMAR is the data safeguard. The value of the scientific data produced by the institute is considered invaluable to the understanding of the marine system and, if lost, they are no longer recoverable or reproducible. Extreme care is being taken to ensure that the data is always protected from any errors or malicious actions that could compromise its integrity.

A backup copy of the data stored at the Data Archive will replicate at an off-site location regularly, ensuring availability regardless of possible risks and a full recovery of any potential lost data.

As this is a distributed facility, efforts are made to establish secure connections between the sections and the end-users of the data, as between the institute and other institutions and/or with other infrastructures (IODE, COI, SeaDataNET, etc.) connected with the infrastructure, to ensure the security of what is transferred and always used. Certified public and virtual private encrypted connections are favoured, as point-to-point (P2P) connections cannot be guaranteed.

Currently, the quality of data and associated metadata check is a responsibility of the data producer and their contact persons.

The security requirement of the data/metadata storage is demanded to the Data Management Office in coordination with the ICT office and their contact persons, who undertake actions to ensure the security of the data and its restoration in the event of a disaster.

The suggested strategy is to adopt the so-called 3-2-1 rule: there should be at least three copies of data, two different backup formats, and one backup stored off-site.

Although the 3-2-1 backup approach was developed about 20 years ago for the effective and long-term preservation of photo archives, the basic concepts are still the benchmark for data protection today.

In short, the 3-2-1 backup approach provides:

Three copies of the data: Copies of all critical data (RAW) must be made at regular intervals. There should be at least three copies: original data and two backups.

Two types of storage: copies should be kept on two different types of storage: different machines to prevent the breakage of a single machine from affecting all copies of the data. The types of storage devices can be either in-house or cloud based.

An external location: at least one copy of the data must be stored on a storage medium located in a geographically different location from the other copies. This copy may be a replica of the primary storage site or an off-line (tape library type) copy of the data.

The 3-2-1 approach is a first step in defining a standardised procedure for secure data storage within the marine domain. A discussion is currently underway to define the technical details and some improved strategies to the approach to adapt it to an infrastructure with multiple geographically distributed nodes.

The actions taken and being defined ensure, as far as possible, that there are no single points of failure for the data.

All nodes of the distributed and federated system will be requested to prove that 3-2-1 approach for data secure and storage is adopted.

9. Ethical aspects

ISMAR ethical conduct aims at ensuring responsible data management by making research results available to stakeholders and encouraging societal participation.

Whenever the ISMAR Data Management Office project will implement surveys, questionnaires, or collect personal data for any reason (e.g., attendance to organized events), European GDPR law will be used as reference and the user will be informed about the use of personal data.

Also the registration phase in the ISMAR Data Management System will follow the European GDPR.

In general, ISMAR will not transfer personal data (e.g., email addresses) to other entities and the only use will be setting up a distribution list to inform users about project progress. Users will be always able to change their consent and ask to be removed from the distribution channel.

10. Other issues

Nothing to report.

11. Further support in developing the DMP

Nothing to report.

Venezia
Tesa 104 - Arsenale,
Castello 2737/F
30122 - Venezia, IT
+39 041 2407911
protocollo.ismar@pec.cnr.it
www.ismar.cnr.it

Bologna
Area della Ricerca
di Bologna –
Via P. Gobetti 101
40129 - Bologna, IT
+39 051 639 8891

Lerici
Forte Santa Teresa,
Pozzuolo di Lerici
19032 - La Spezia, IT
+39 0187 1788900

Napoli
Calata Porta Di Massa
Porto Di Napoli snc
80133 - Napoli, IT
+39 081 5423802

Roma
Area della Ricerca
di Roma 2 - Tor Vergata
Via del Fosso del Cavaliere 100
00133 - Roma, IT
+39 06 45488634

Trieste
Area Science Park
Basovizza - Edificio Q2
Strada Statale 14, km 163.5
34149 - Trieste, IT
+39 040 3756872