D9.10
Marine subdomain white paper for sustainable data management

Work Package	WP9
Lead partner	IFREMER
Status	Final
Deliverable type	Report
Dissemination level	Public
Due date	30-04-2023
Submission date	07-06-2023

Deliverable abstract

The report provides the scope of the sustainability for the data management of the Marine subdomain, for the research infrastructures themselves and with respect to their contributions to ENVRI-FAIR and EOSC. Recommendations to consolidate the sustainability of the Marine subdomain contribution to EOSC beyond the ENVRI-FAIR project and the RI services are proposed.

ENVRI-FAIR (www.envri-fair.eu) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824068

DELIVERY SLIP

	Name	Partner Organization	Date
Main Author	Delphine Dobler	Euro-Argo ERIC	18-04-2023
Contributing Authors	Thierry Carval Peter Thijsse Justin Buck Katrina Exter Vincent Bernard Romain Cancouët Estérine Evrard Dominique Obaton Ivan Rodero Alex Vermeulen Steve Jones	Ifremer / Euro-Argo Maris / SeaDataNet NOC / SeaDataNet VLIZ / Lifewatch Ifremer / Euro-Argo Euro-Argo ERIC Euro-Argo ERIC Ifremer EMSO ICOS ICOS	
Reviewer(s)	Paolo Laj Yann-Hervé De Roeck	Université de Grenoble Alpes / ACTRIS Euro-Argo ERIC	29-05-2023 13-05-2023
Approver	Andreas Petzold	FZJ	07-06-2023

DELIVERY LOG

Issue	Date	Comment	Author
V 0.1	2023-01-17	Draft 1: First draft presented at the WP9 session during the 2023 ENVRI week	D.Dobler, T.Carval including inputs from Romain Cancouët and Estérine Evrard
V 0.2	2023-04-18	Draft 2: accounting for contributing authors inputs: transition from bullet points to full text. Addition of the Environmental section.	D.Dobler, J.Buck, K.Exter, P.Thijsse, including inputs from V. Bernard
V 0.3	2023-05-05	Draft 3: accounting for remarks after transition from bullet points to full text, adding recommendation section, conclusion, reference and glossary.	D.Dobler, K.Exter, P.Thijsse
V 0.4	2023-05-12	Draft 4: accounting for comments from D.Obaton, E.Evrard, R.Cancouët, J.Buck	
V 0.5	2023-05-15	Draft 5: accounting for comments from Yann-Hervé De Roeck	
V 0.6	2023-05-16	Draft 6: accounting for comments from Ivan Rodero and Alex Vermeulen	
V 0.7	2023-05-26	Draft 7: accounting for comments from Steve Jones	
V1.0	2023-05-30	Accounting for Reviewer's comments	

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the Project Manager at manager@envri-fair.eu.

GLOSSARY

A relevant project glossary is included in Appendix A. The latest version of the master list of the glossary is available at http://doi.org/10.5281/zenodo.4471374.

PROJECT SUMMARY

ENVRI-FAIR is the connection of the ESFRI Cluster of Environmental Research Infrastructures (ENVRI) to the European Open Science Cloud (EOSC). Participating research infrastructures (RI) of the environmental domain cover the subdomains Atmosphere, Marine, Solid Earth and Biodiversity / Ecosystems and thus the Earth system in its full complexity.

The overarching goal is that at the end of the proposed project, all participating RIs have built a set of FAIR data services which enhances the efficiency and productivity of researchers, supports innovation, enables data- and knowledge-based decisions and connects the ENVRI Cluster to the EOSC.

This goal is reached by: (1) well defined community policies and standards on all steps of the data life cycle, aligned with the wider European policies, as well as with international developments; (2) each participating RI will have sustainable, transparent and auditable data services, for each step of data life cycle, compliant to the FAIR principles. (3) the focus of the proposed work is put on the implementation of prototypes for testing pre-production services at each RI; the catalogue of prepared services is defined for each RI independently, depending on the maturity of the involved RIs; (4) the complete set of thematic data services and tools provided by the ENVRI cluster is exposed under the EOSC catalogue of services.

TABLE OF CONTENTS

D9.10 - Marine subdomain white paper for sustainable data management

1 Introduction

The ENVRI-FAIR project is promoting the Findability, Accessibility, Interoperability and Reusability of data (a.k.a. FAIRness of data) for datasets of the Environmental Research Infrastructures (ENVRIs). The RIs together increase the access to FAIR data, in which they converge in standards and technological solutions. One of the final goals and demonstrator is to allow an enhanced access and a synthetic visualisation to scientific, policymakers and end-users about the Essential Climate Variables (ECVs), and a common integration into the European Open Science Cloud (EOSC).

The ECVs were declared as "Essential" by the Global Climate Observing System (GCOS) sponsored by the World Meteorological Organization (WMO), the United Nations Educational, Scientific and Cultural Organization (UNESCO), the United Nations Environment Programme (UN Environment), and the International Council for Science (ISC) (https://public.wmo.int/en/programmes/global-climate-observing-system/essential-climate-variables). They are essential to monitor the health status and the impact of political decisions on the Earth system. The subset of ECVs that concern the marine subdomain are called the Essential Ocean Variables (EOVs). The data is collected by and distributed over different RIs, that can use, for the same EOV, different formats with respect to field names, encoding format, reference axes (across units, geographical, depth and time references, taxonomy references, etc.). These dataset specificities require a high-level of interoperability and smart mappings of each RI's datasets to achieve a simple access (via aggregated datasets) and a synthetic visualisation and monitoring (e.g. through data products) for these EOVs.

The ENVRI-FAIR project has increased the interoperability between the ENVRIs, not only among marine RIs but also at cross-subdomain level:

- by using more common technologies (common FAIR enabling resources or FERs);
- by cataloguing each ENVRI service (the catalogue of services), including APIs to access data in a machine-to-machine type of access;
- by developing an interface providing a centralised access to EOVs (metadata and data), which are spread among marine RIs, without the user (human or machine) needing to worry about the RI's underlying architecture (the interface is called the EOVbroker).

These developments were supported by adopting:

- web-exposed vocabulary servers (enhanced Accessibility and Interoperability);
- interoperable standards (enhanced Interoperability);
- ontologies' developments (ontologies are used in web semantics: they define the relationships among the metadata that are uploaded into "end-points". These end-points are web databases mirroring part of internal RIs databases, relevant for the final usages/services).

The pursued goal requests, as an essential prerequisite, a strong and sustained data management commitment in each RI. First, two questions should be answered: why being sustainable? What is in the scope of sustainability?

1.1 Why be sustainable?

The sustainability of the whole RI system is requested for several reasons. They are often in close relationships with each other. We need sustainability:

- for Marine RIs to provide strong data systems, with reliable metadata/data, that are maintained and that are interoperable (with the aim of being integrated within the EOSC framework);
- to support environmental and climate policies with sustainable decision-support tools (e.g. the environmental dashboard within the EOSC future project). Monitoring the climate and biodiversity changes and the impact of policy decisions on them requires long time series and, as such, sustainable upstream channels providing this information;
- to provide the data required to support the UN Sustainable Development Goals (SDG), in particular for goal 13 (*Take urgent action to combat climate change and its impacts*) and 14

(*Conserve and sustainably use the oceans, seas and marine resources for sustainable development*). Marine RIs provide input data related to topics such as Ocean Acidification, Ocean Heat Content, Sea Level Rise, Ocean Deoxygenation, … Several projects aims at providing synthesised information such as the annual Global Carbon Budget and global data synthesis efforts such as Surface Ocean CO_2 Atlas (SOCAT) and Global Ocean Data Analysis Project (GLODAP) for biogeochemical data;

- to support the Global Ocean Observing System (GOOS), the European Ocean Observing System (EOOS), their networks and evolution;
- to allow continuity in monitoring the EOVs (or more generally ECVs and their derived quantities, such as dissolved greenhouse gases, carbon pump, ocean deoxygenation, ocean acidification, etc.);
- to allow continuity in associated services, that build upon upstream channels to provide exploitable content in various fields, such as:
    - operational applications in the fields of marine safety (e.g. Search and Rescue),
    - weather and ocean forecasting (including ocean modelling, and hazards forecasting),
    - climate change and biodiversity monitoring;
- to reduce costs: re-inventing the wheel because of obsolescence, abandoned piece of code, or loss of human resource is first costly and second does not assure that the level previously reached will be reached again.

1.2 What topics are in scope for sustainability?

The sustainability of data management covers several aspects, which may not be strictly bound to "data management" *per se*. This can include:

- Financial funds and associated human resources;
- Data and metadata architecture and homogeneity through RIs, with reliable datasets (including e.g. quality level, uncertainty estimate, traceability of each measure …). This necessitates setting data management as a priority in RI strategic plans;
- Quality processes with certifications such as ISO9001, *RDA Core-Trust-Seal*, *IODE IOC national oceanographic data centre*. Quality processes cover various aspects such as those aiming at monitoring, adjusting and/or curating data and metadata content, those aiming at managing software or routines issues or requested evolutions (i.e., procedures for Quality Assurance and Quality Control – QA/QC), those aiming at monitoring the data systems flows;
- Securing data storage and upstream processes, through the back office computer architecture. This includes the assessment and enhancement of the robustness and performance of the system, and the risk assessment and management plan for contingency situations;
- Securing observations: Sustainability in the upstream inputs and evolution of the GOOS networks as community requirements evolve;
- Securing Interoperability of RIs (this topic is directly related to the EOSC goal);
- Sustainability with respect to environmental impacts of RIs (e.g., hardware, energy consumption and raw materials of Information Technology (IT) solutions and of *in-situ* observations means).

For each of these aspects, recommendations will be set up and summed up in the Recommendation section.

The next sections present the various aspects and/or topics that are related to data management sustainability. We decided to put as a first section the financial aspect, because this is one important aspect without which nothing else can be achieved. Then we present several inter-twined aspects throughout the next sections: the data, metadata and services design, the quality processes, the back office part, the input observational flow, and the interoperability of data/metadata/services. Lastly, we present the environmental impact that can and will somehow drive future developments regarding the whole system (from the observations, to the data storage, processing and the communalised services).

2 Financial sustainability and human resources

Financial long-term plans are essential for sustainability. These financial plans are most often built from a list of tasks and the associated human resources and/or purchase needed for each task. We can mainly differentiate two categories of tasks:

- those that are mandatory to maintain the existing services (at a very large sense: observations collection, qualification, processing, aggregation, distribution, analyses, communication, etc.) and the associated engagements;
- those that are dedicated to new or enhanced services (e.g. upstream research, prototypes developments, demonstrators developments, etc.).

Tasks that are mandatory to maintain the existing services should be under sustained funding lines, i.e. which are not restricted to the start and end dates of a project. It is noteworthy that some marine RIs also make a differentiation in the provenance of the sustained funding with respect to the type of data and/or services. Indeed, they consider that data and services meant to be deployed at the EU level (for an EU benefit or usage) should be/are financed by the European Union funding lines and that data services meant to be nationally deployed (e.g. data/cloud connectors) should be financed with national funding lines. The relevance of this differentiation in the provenance of the fund is specific to each RI.

Tasks that are dedicated to the new or enhanced services are often primarily funded on short-lived funded research projects (with a duration of typically 3 to 4 years). These evolutions can be differentiated in two kinds: the ones for which a one-time funding is sufficient, and the ones that will call for sustained funds if it leads to a new sustained service. Here, we want to emphasize that the term "enhanced services" can also include RIs core activities. Indeed, core activities must also evolve to match new user's needs, which are regularly evolving in relation with the increase in scientific knowledge and the changes in science and/or computer techniques. It is therefore a necessity for the RIs to provide new or enhanced services through a sustainable scheme. This will provide the insurance that a minimum "Research and Development" activities are always secured, to answer efficiently these new users' needs. The level of sustained funding dedicated to evolution-related activities will depend on each RI's priority. It is noteworthy that a way to secure funds is the pay-per-use approach; however, this is not the philosophy in marine RIs, and whether or not this approach should be considered deserves further discussions. Until now, data and services are distributed free of charge ('as open as possible') but the question of financial contribution from big and regular/operational consumers that deeply rely on them could be raised.

RIs should assess the ratio between the sustained versus short-lived funding lines that will allow them to be and/or stay in a financially healthy balance. Generally speaking, and echoing the precedent paragraph, there should remain a core of sustained activities strong enough to keep the human resource, their knowledge and the RI virtuous dynamics (through innovations, corporate culture, project follow-up, long-term strategy, opportunity for career evolutions, etc.). The computation of this ratio will depend on each RI's specificity. It can be/is also differentiated within the RI itself with respect to its members (which have various national funding policies and priorities). Indeed, some RIs are composed of a collection of institutes, which, in the end, provide the major part of the human resources and/or the "in-kind" services. These institutes should know (and report) which budget and funding lines (sustained/short-lived, EU/national) allow them to provide their human resources and/or in-kind services. It is noteworthy that this information is required by the European Strategy Forum on Research Infrastructures (ESFRI) monitoring and landscape analyses.

To compute the above-mentioned ratio, each RI should define/know as accurately as possible their budget and budget lines, the balance of their engagements at the European and/or national level compared to their sustained funding lines. RIs should also account for the budgets and funding lines of their members. This can be described in Long-Term Sustainability (LTS) plans at RI level.

Finally, it is important that the cost behind a variable (cost to collect and distribute an EOV for instance), the cost behind a service, and the cost behind a climate or biodiversity monitoring be known with sufficient accuracy, which is not an easy exercise. When new *in-situ* observational means are funded (most often on short-lived project lines), the associated data management costs through the typical lifespan of the observational means should not be forgotten. When new service developments are funded (most often on short-lived project lines), the associated operation and maintenance costs should be assessed and sustained funding searched and obtained, once the scientific pertinence and the policy primacy have been demonstrated.

We often encounter several caveats:

- the separation for each kind of task and associated needed human resources (and thus funds) is sometimes not well-enough described and/or known. It is noteworthy that this difficulty to separate tasks and associated needed budget can be caused by several factors: the complexity

of the RI, the complexity of the budget lines, the complexity to assess the "in-kind" value, the complexity for a single person to report accurately-enough the amount of time spent on each task, etc.;
- the sustained funds (i.e., not restricted to the start and end dates of a project) are sometimes not sufficient to cover the whole of the mandatory tasks/services;
- some of the foreseen evolutions (which can stem from a request at the EU level or from the community at large), that are initially launched through short-lived project lines and that call for sustained funds in a next phase can lack long-term visibility for these potential sustained funds. N.B. On the other hand, there also exist well-identified evolution strategies (for instance, the Copernicus Marine Services evolution strategy).

All these caveats could lead to degraded services because of:
- a lack of sustained funds for "mandatory" services;
- a loss of time in the research for short-lived project funds to sustain services that have already been agreed upon at a national, European or international level but that are not financially secured;
- a high turnover in the human resource, with part of the knowledge being lost and/or needing to be rebuilt.

Recommendation 1

Each RI should know as accurately as possible their budget, differentiating the sustained versus the short-lived funding lines, and detail them in correlation with the types of activities (mandatory or not). Each RI should assess the ratio between sustained and short-lived funding lines that will enable it to be/stay in a financially healthy balance and in a virtuous dynamic. In this assessment, RIs should consider the evolution of services as a necessity given that they have to respond to evolving users' needs (in relation with the increase in scientific knowledge and the changes in science and/or computer techniques).

Recommendation 2a

When a new *in-situ* observational means is funded (most often on short-lived project lines), the associated data management costs through the typical lifespan of the observational means should be clearly assessed and included in the definition of the project.

Recommendation 2b

When new service developments are funded (most often on short-lived project lines), the associated operation and maintenance costs should be assessed and sustained funding searched.

There have already been several advocacy papers including financial sustainability aspects and associated recommendations:
- Policy brief of the European Marine Board: "*Sustaining in situ Ocean Observations in the Age of the Digital Ocean*" [1]. In this policy brief, the Recommendation number 1 advocated for a change of funding model:
  "*Advocates the necessity for a change in the funding model*
  Recommendation 1: Recognize sustained in situ observations as a large-scale, essential, and enabling infrastructure generating global public-good data that create information and knowledge-based services, and advance its implementation with appropriate financing models to deliver systematic and long-term monitoring. A possible endpoint could be an international entity with a subscription-based or a binding Nationally-Defined Contributions model, with a backbone/core Ocean observing capability, overarching governance and institutional arrangements, and roles and functions for nations and the EU" (p.14).

- European Commission staff working document: "*Sustainable European Research Infrastructures: a call for actions*" [2], where one of the identified challenges is "*Establishing adequate framework conditions for effective governance and sustainable long-term funding at*

every stage in their life cycle" (p.9). They recall "*intangible assets tend to be underestimated*" (p27). They include computerised information in these intangible assets. "*A possible solution that could be explored is to see whether financial contributions to ESFRI projects and ERICs could be provided under national budget lines similarly as for international treaty-based organisations. This could provide the RI operators a sufficiently stable investment environment allowing them to concentrate on providing high quality services for their user communities instead of continuously looking for funding even for their basic operations.*" (p34), "*encourage synchronisation of national roadmaps and their alignment with the European RI roadmap*" (p.36).

- European Strategy Forum on Research Infrastructure (ESFRI) Long-Term Sustainability (LTS) working group: "*Long-Term sustainability of Research Infrastructure*" [3] "*National authorities should consider governance model which provide the right balance between long-term funding commitments (including operation costs and strategic developments) and regular evaluation of the RI performance*" (p57).

From these recommendations, the best funding model to follow is not obvious and may not have the same relevance depending on the constraints of each RI. Nonetheless, they all advocate for better coordination between national, European and international levels.

3 Data, Metadata and Services architecture

3.1 About the system FAIRness and its evolutions

For sustained data management, the FAIRness of the data, metadata and services should be continuously enhanced. Data and metadata must be well described, traceable (through globally unique and persistent identifiers), Findable, Accessible, Inter-operable, Re-usable. The continuous enhancement includes curation actions where needed, FAIR-enhanced functionalities where possible, accounting for evolution needs when requested. To provide concrete examples, enhancements could include accessible history of changes, use of cloud solutions (enhanced datasets accessibility), extended ontologies to match specific user needs, enhanced visualisation functions and ergonomics in the proposed services.

In this process of continual enhancement, the user feedbacks (or requirements) are one essential pillar. Indeed, the system as a whole must remain continually relevant for (and thus used by) both the human and machine audiences. They both should be accounted for when undertaking curation actions, when selecting FAIR-enhancing solutions or when implementing evolutions. In addition, the architecture needs to be flexible to allow accommodation of evolution needs with time.

To quantify the enhancements, the FAIRness assessment needs to go further. In the ENVRI-FAIR project, quantitative indicators were designed and assessed. Their evolution through the project timeline showed a better coverage of FAIR principles for all marine RIs and convergence toward the use of common technologies. This is a first great achievement. To move forward, as recommended in the ENVRI-FAIR Deliverable D9.9 [4], the quality (and not only the quantity) of the FAIR-enabling resources implementation should also be assessed. There remains to be defined what quality aspects can be included in this new assessment. This could include automatic "good-health" procedures for machine-2-machine, common set of non-regression procedures when services are upgraded, or human check procedures for human FAIR access to data.

Recommendation 3

The FAIRness assessment should be improved to include the assessment of the quality of implementation for the FAIR-enabling resources, and not only the quantity.

3.2 Metadata common set

To enhance the interoperability among RI assets, and if it is not already the case, it is recommended that each component of the observation network tends toward a harmonised subset of common metadata

description (referred to as “basic information” in the hereafter documents). For the marine subdomain, this basic information is detailed in the following document EuroSea D3.7 « Network harmonisation Recommendations » [5]. In this document, recommendations for each network are provided to have main metadata homogenised between the different networks and to fill in the gaps.

It is noteworthy that data often have many layers: the measurements themselves, the adjusted measurements when additional calibration is needed, the associated uncertainties, different types of quality flags, and so on. This adds to the metadata volume and requirements that need to be maintained. It also adds complexity to the data cross-integration. For instance, quality flags often have close but not exactly identical meanings for each data network. In this particular case, the building of correspondence tables allows the end-user to access the requested qualified data, in a cross-RI integrated way. This latter consideration was part of the ENVRI-FAIR Task Force 4 activity. Briefly, the outcome was to suggest the use of the IODE Quality Flags standard [6] and to map each RIs quality code to this standard**.**

Recommendation 4

To enhance the metadata integration, each RI should extend or enhance the subset of common metadata description, following the EuroSea D3.7 Deliverable recommendations, and if relevant (for quality code for instance), map their existing vocabularies to a common vocabulary.

3.3 Data and metadata volume increase

The volume of observational data is growing fast across the GOOS and data centres landscape. This increase often means an increase in the workload but most often, this is not reflected in the amount of human resource handling them. To circumvent this, the management of data and metadata need to be made more efficient. In particular, for metadata, this can be achieved:

- through procedures as automatic as possible (e.g., metadata retrieved from the observational device itself, instead of manual fill-in),
- through enhanced real-time quality control of the metadata (and data) collection (including the set-up of monitoring rules, refined acceptable ranges, automatic coherent checks),
- through the use of user-friendly Graphical User Interfaces (GUI) for the human access, and Application Programming Interfaces (API) for the machine access, and incorporating the necessary semantics.

The selected technologies should be tolerant to this volume increase of data and metadata. Their robustness with respect to the expected increase for the years to come should be systematically assessed.

Sustaining the data, metadata and associated services is very demanding and requests that data management be a priority in RI strategic plans, feeding into the data management practices at national level. There exist several frameworks providing guidelines for a sustained data and quality management, a robust IT infrastructure design, and more specifically, marine data centres requirements. In addition to these guidelines, there are specificities to meet more FAIRness within all these processes. This is detailed in the following section.

Recommendation 5

The selected technologies should be tolerant to the volume increase of data and metadata. Their robustness with respect to the expected increase for the years to come should be systematically assessed.

4 Quality processes and certifications

4.1 Certifications, accreditations

Quality processes and certifications promote good practices in IT architecture and software development, and risk management through dedicated plans. They incite to describe processes with an accurate-enough level, to define and operationalise system monitoring, to describe and operationalise upgrading processes and to define plausible contingency situations and their associated curative actions. As such, they are powerful tools to sustain data management. Their applicability will be specific to each RI. There exist several certifications, such as:

- the ISO9001 certification ([7], [8], [9]), first published in 1987, that set requirements for the quality management system;
- the ISO/IEC 27001 certification [10], standard for Information Security Management Systems (ISMS), that defines requirements an ISMS must meet;
- the Research Data Alliance (RDA) Core-Trust-Seal (CTS) certifications ([11], [12], [13], [14]), launched in 2017, which promote the development of sustained and robust infrastructures, and offers core level certification for Trustworthy Digital Repositories (TDR) holding data for long-term preservation;
- the International Oceanographic Data and Information Exchange (IODE) accreditation for National Oceanographic Data Centre (NODC) [15]. This accreditation is specific to the marine domain and provides a common set of requirements to ensure NODCs compliance with IODE standards (including regular monitoring and assessments of the quality of data and service, ability to provide secure long-term storage of and access to marine data). The IODE program was established in 1961.

In addition to quality management, sustainability and risk management plans, certifications should also focus on the actual state of the services with respect to their FAIRness and ergonomics (user-friendliness).

4.2 FAIR Quality processes in software developments

There is a growing need for code sharing among a community that is, in most marine RIs, geographically spread, in various institutions, using various associated IT architectures, having various preferred programming languages, and having different levels of financial means. This code sharing is also a necessity for transparency and reproducibility among and outside the community. To facilitate this code sharing, data management software should move toward open source and open science frameworks and/or cloud-native implementations, when possible and relevant. This will enable improvements through a greater cooperation in the development of these softwares, an increased shared knowledge, and a wider community of developers and data managers, either coming from the RI itself or from outside the RIs. All of these aspects will contribute to the long-term sustainability of tooling.

The open-source, open-science recommendations imply that the Quality Assurance/Quality Control (QA/QC) procedures should also follow this path. This can imply:

- management of bug reporting and evolution requests through an accessible framework (e.g., by the use of GitHub for ticket management, or any other easily available and shareable system).
- coding rules description, validation means, validation rules to allow/facilitate community review of changes and updates in software repositories. This code-sharing also implies that software developments are of Continuous Integration/Continuous Development (CI/CD) type.

Recommendation 6

To facilitate code sharing, data management software should move toward open source and open science frameworks and/or cloud-native implementations, when possible and relevant.

4.3 Securing data storage and processes: back office computer architecture

As already mentioned within the NODC accreditation, sustaining marine data management implies ability to provide secure long-term storage of and access to marine data.
This implies several requirements for the computer architecture management:

- Robustness assessment with respect to:
    - Redundancies for both the dataset and all the processes feeding and managing the dataset. These redundancies encompass various levels of restart procedures (including the needed infrastructure), also referred to as the failover system. The various levels of restart include cold restarts (several hours/days are needed), warm restarts (within minutes, almost imperceptible), and hot restarts (instantaneous). The implemented levels of restart depend on the purpose the data are used for (requirements from the users and downstream services). These procedures, including risk assessment and contingency plans, are often described within the ISO9001 certification framework;
    - Scalability to allow sustainability of the architecture for growing and evolving content. It is noteworthy that there exists off-the-shelf solutions for scalable architectures (e.g., open-source Kubernetes);
    - Cybersecurity aspects: these can be described within the ISO/IEC 27001 certification framework, within a cybersecurity plan.
- Performance assessment. The overall performance of RIs could be enhanced using emerging infrastructure types, namely cloud native solutions (such as those promoted through the EOSC programme). This will allow an enhanced access to data and services, processing time access, software and code online frameworks. The use of cloud-native solutions is quite a change and this should go with a thorough review of the risk management plan, where the use of an external solution (and submitted to unknown or uncontrollable contingencies situation, including failures, shut down or even hacking) should be backed up with an internal solution (even if the back-up service is degraded in terms of performances). When such an architecture change is foreseen, it is important that the cybersecurity plan be revised with respect to potential new threats.

It is noteworthy that datasets are stored at several levels:

- national Data Assembly Centre (DAC), Global Data Assembly Centre (GDAC), RI other storage means. These storages can be in national standard format, with operational data format conversion layers to international data exchange standards;
- cached/stored in EU aggregators like SeaDataNet, EMODnet, Eurobis, Copernicus Marine Service.

Data management is required at all levels and across them (lifecycle management). In particular, the dataset meant to be the reference should be clearly stated, and be uniquely identified with a persistent identifier (e.g., DOI). For the cached and/or duplicated storage access, the reference dataset should be clearly exposed with information about the provenance (location, persistent identifier, date and time and version number of the dataset). This traceability topic is further discussed in section 5 (sustain the interoperability and connections).

4.4 Securing the data flow: sustainability in the upstream chain

In the previous sections, we mentioned the data, metadata and services FAIRness, the quality procedures and subsequent certifications, the back-office robustness and performance. But at the origin, there are the *in-situ* observations (even if the RI does not collect them directly).

Various factors could lead to a reduced upstream chain, such as:

- single manufacturing point for part or totality of the observing system (electronic component, sensors, supporting vector);
- shortage in raw material, electronic device (used in *in-situ* observing systems, but this remark also applies to IT solutions);
- reduced deployment opportunity (war, disease, cost increase, etc.) and degraded maintenance of the instruments;
- issues in data transmission (satellite communication service, access to the instruments, etc.) or in data collection (e.g., failure of a sensor);

- a disruption in the connection from national data centres and RIs to EU aggregators.

A reduced upstream observations flow will impact RIs data and services. It is also noteworthy that many data systems are cross-calibrated. For instance, climate and oceanic models use *in-situ* data for instantiation, assimilation and validation, earth observation satellite data are cross-calibrated with *in-situ* data (most often, remote sensing are quite indirect measurements), autonomous Argo profiler needs accurate CTD cast observations for salinity calibration, and so on. Data without calibration and uncertainty estimates are of poor use, as they mean a low confidence level. Thus, these cross-calibrations are essential as they allow providing and reducing these uncertainties.

Thus, we cannot tackle data management sustainability without talking about the sustainability of the upstream observational collection.

What is included in the perimeter of the upstream chain will depend on each RI specificity. This can include:

- *in-situ* observations only,
- *in-situ* observation and upstream data centres for RIs that aggregate different kind of datasets,
- *in-situ* observation, upstream data centres, and national data centres that operate data layers and connections services to support the RI.

The upstream perimeter should be assessed for each RI. When relevant, an upstream risk-reduction plan should be defined and included in the risk management plan.

Recommendation 7

When relevant, a risk-reduction plan for the upstream channel(s) should be defined and included in the risk management plan of each RI.

5 Sustain the interoperability and connections

All the above-mentioned topics (costs assessment and funding, data management, quality management) apply to both upstream and downstream providers (downstream includes data collections and/or services such as those proposed by SeaDataNet, Global Telecommunication System (GTS), ENVRI-Hub, Environmental dashboard EOSC-future, Ocean Info Hub, etc.).

In particular, there is a gap to handle/close regarding the lack of traceability from downstream providers (e.g., aggregators). When RIs provide data to the downstream providers, the traceability can be incomplete or sometimes totally lost. This traceability information can encompass several pieces of information: the source RI, the reference storage location including the associated persistent identifier of the dataset, the version, date and time of the dataset. For a downstream product that performs statistics using data from several datasets, the relative contribution of each dataset could also be known. The lack of provider traceability may lead to incomplete statistics on the data usage at some point of the stream. This issue, in turn, has implications for the sustainability of RIs (e.g., data usage reporting and subsequent impact to governing bodies and funding agencies).

In addition, any adjustments made to data or metadata, associated method if any, quality code information, uncertainty information if available must be communicable to all points in the stream. Updates to datasets (in the form of new versions) must be communicated to all downstream partners. Similarly, issues with a dataset discovered downstream must be communicated upstream to the original source so the necessary corrections can be applied for all users at the earliest possible point in the chain, to reduce the possibility of specific corrections to data only being available from certain points in the chain. It is noteworthy that the lack of methods and version traceability may influence the user's confidence level and eventually affect the sustainability of the services.

It is noteworthy that the more interconnected the overall system will be, the more rigorous the quality management must be. Impacts of sub-system upgrades need to be assessed (and mastered) with respect to the various streams and services to end-users. This can become quite a difficult exercise with the fast growing number of services, of "machine2machine" end-users, and of the external and various technologies. On the other hand, impacts of subsystems outages or failures also need to be assessed. Mitigation actions shall be sought if relevant (e.g., use of local/cache replication of upstream datasets to circumvent outages). The relevance of the mitigation action will depend on the requirements from the service usage.

The development and maintenance of the integrated system at EOSC level has a cost that should be evaluated, on both sides of the interface (i.e., both on the data/service provider and on the data/service collector). Providers have a workload increase to account for collectors' requirements. Collectors must integrate various providers' upgrades, maintain the homogeneity and continuity in the proposed service. The impact of a lack of funds on these kinds of systems should be estimated, and priority set up when financial contingency arises. In preparatory phases, it is often difficult to estimate the financial cost for the maintenance of such systems, especially because it depends on many external input channels. However, it is important to try assessing the amount of human resources that will be needed for the maintenance.

Recommendation 8

Downstream providers (e.g. dataset aggregators, service providers, etc.) should report **traceability** information about the dataset(s) they aggregate/use. This includes source RI, the reference storage location including the associated persistent identifier, the version, date and time of the dataset, any quality code, uncertainty information, adjustments and associated adjustment method applied.. The RIs should provide means for this traceability (if not yet complete) and communicate to all points in the stream. The downstream users and/or providers should have means to report issues so that curation actions can be undertaken at the most upstream location.

Recommendation 9

The maintenance of the integrated system at an ENVRI, and at an EOSC level is driven by the pipeline of RI services. This has a cost that should be evaluated on both sides of the interface: the providers should assess the costs of being integrated, the integrator should assess the costs of integrating and sustaining such an integration. This should include the risk assessment and mitigation action plan for contingency situations at the integrated system level.

6 Environmental impact

All the "production line" from the raw material extractions to the sensor making, the sensor measurement, communication, data collection, storage, processing, web exposure has an environmental imprint. This environmental imprint can cover various aspects:

- Environmental imprint of the upstream *in-situ* data collection,
- Environmental imprint of the IT solutions,
- Environmental imprint of the human resources (office accommodation, transport, etc.).

As environmental Research Infrastructure, there is a growing demand to assess the carbon footprint of our activities. This is a great but important challenge, especially as one main aim of our activity is to protect the environment.

Some initiatives and action plans exist at local levels to estimate and reduce this environmental imprint of our activities (e.g., LABOS 1.5, MetOffice Journey to NetZero by 2030, NOC Net Zero Oceanographic Capability). If not yet the case, RIs should estimate their imprint and build action plans toward more environmentally friendly processes when possible and relevant.

Recommendation 10

Research infrastructures should also assess sustainability with respect to their environmental imprint in terms of IT solutions, of *in-situ* observatory technologies/strategies, of human resources accommodation and transports, etc. and undertake mitigation actions when relevant and possible.

7 Recommendations

N°	Description
R1	Each RI should know as accurately as possible their budget, differentiating the sustained versus the short-lived funding lines, and detail them in correlation with the types of activities (mandatory or not). Each RI should assess the ratio between sustained and short-lived funding lines that will enable it to be/stay in a financially healthy balance and in a virtuous dynamic. In this assessment, RIs should consider the evolution of services as a necessity given that they have to respond to evolving users' needs (in relation with the increase in scientific knowledge and the changes in science and/or computer techniques).
R2a	When a new in-situ observational means is funded (most often on short-lived project lines), the associated data management costs through the typical lifespan of the observational means should be clearly assessed and included in the definition of the project.
R2b	When new service developments are funded (most often on short-lived project lines), the associated operation and maintenance costs should be assessed and sustained funding searched.
R3	The FAIRness assessment should be improved to include the assessment of the quality of implementation for the FAIR-enabling resources, and not only the quantity.
R4	To enhance the metadata integration, each RI should extend or enhance the subset of common metadata description, following the EuroSea D3.7 Deliverable recommendations, and if relevant (for quality code for instance), map their existing vocabularies to a common vocabulary.
R5	The selected technologies should be tolerant to the volume increase of data and metadata. Their robustness with respect to the expected increase for the years to come should be systematically assessed.
R6	To facilitate code sharing, data management software should move toward open source and open science frameworks and/or cloud-native implementations, when possible and relevant.
R7	When relevant, a risk-reduction plan for the upstream channel(s) should be defined and included in the risk management plan of RIs.
R8	Downstream providers (e.g. dataset aggregators, service providers, etc.) should report **traceability** information about the dataset(s) they aggregate/use. This includes source RI, the reference storage location including the associated persistent identifier, the version, date and time of the dataset, any quality code, uncertainty information, adjustments and associated adjustment method applied. The RIs should provide means for this traceability (if not yet complete) and communicate to all points in the stream. The downstream users and/or providers should have means to report issues so that curation actions can be undertaken at the most upstream location.
R9	The maintenance of the integrated system at an ENVRI, and at an EOSC level is driven by the pipeline of RI services. This has a cost that should be evaluated on both sides of the interface: the providers should assess the costs of being integrated; the integrator should assess the costs of integrating and sustaining such an integration. This should include the risk assessment and mitigation action plan for contingency situations at the integrated system level.
R10	Research infrastructures should also assess sustainability with respect to their environmental imprint in terms of IT solutions, of in-situ observatory technologies/strategies, of human resources accommodation and transports, etc. and undertake mitigation actions when relevant and possible.

8 Conclusions

During the ENVRI-FAIR project, many steps forward have been made toward RIs cross-integration. The Marine RIs FAIRness has been greatly enhanced, making it possible to design and develop the ENVRI-Hub demonstrator and the EOV broker. The EOV broker effectively demonstrated that the various marine RIs dataset could be queried in a machine-to-machine way.
This effort must be continued to extend the RIs cross-integration and meet the operational requirements for an EOSC.

Throughout this deliverable, we showed that a sustainable data management had many interconnected branches. We provided 10 recommendations covering a wide range of aspects:

- financial (funding and costs assessments, recommendations n° 1, 2 and 9);
- enhanced FAIRness quality assessment (recommendation n° 3);
- extension of common vocabularies (recommendation n° 4);
- traceability (recommendation n°8);
- IT and code development methods (recommendations n° 5, 6);
- risk-reduction plan (recommendation n° 7);
- environmental imprint mitigation assessment (recommendation n° 10).

The recommendations 3, 4, 5, 6 and 8 are more directly related to the sustainability data management, focusing on the ENVRI, and beyond, the EOSC integration. The recommendations 1, 2, 7 and 9, even if indirectly related, are nonetheless very important as they impact or can impact the data management.

9 References

[1] European Marine Board (2021). *Sustaining in situ Ocean Observations in the Age of the Digital Ocean*. EMB Policy Brief No. 9, June 2021. ISSN: 0778-3590 ISBN: 9789464206081 DOI: https://doi.org/10.5281/zenodo.4836060

[2] European Commission, Directorate-General for Research and Innovation (2017). *Sustainable European research infrastructures: a call for action: Commission staff working document: long-term sustainability of research infrastructures*, Publications Office. DOI: https://data.europa.eu/doi/10.2777/76269

[3] European Strategy Forum on Research Infrastructures Long-Term Sustainability Working Group (2017). *Long-Term Sustainability of Research Infrastructures*. ESFRI Scripta Volume II.ISBN: 978-88-901562-8-1. URL: https://www.esfri.eu/sites/default/files/u4/ESFRI_SCRIPTA_VOL2_web.pdf

[4] Alviset Guillaume, Dobler Delphine (2022). *ENVRI-FAIR D9.9: Marine subdomain FAIRness assessment report*. DOI: https://doi.org/10.5281/zenodo.7505613

[5] Obaton Dominique, Pouliquen Sylvie, Antonio Novellino, Alessandra Giorgetti, Abed El Rahman Hassoun et al. (2022). *EuroSea D3.7: Network harmonization recommendations.* URL: https://eurosea.eu/download/eurosea_d3-7_network_harmonisation_recommendations

[6] Paris Intergovernmental Oceanographic Commission of UNESCO (2013). *Ocean Data Standards, Vol.3: Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data*. IOC Manuals and Guides, 54, Vol. 3. 12 pp. (English). URL: https://www.iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=10762

[7] International Organization for Standardization (2015). *ISO 9001:2015 Quality management systems: Requirements*. Reviewed and confirmed in 2021. URL:https://www.iso.org/standard/62085.html

[8] Ifremer ISO9001 qualification P8 processus (Collect and distribute marine data) online documentation (French). URL:https://w3z.ifremer.fr/qualite/Processus-ISO-9001/Approche-processus/P8-Recueillir-et-mettre-a-disposition-des-donnees-sur-le-milieu-marin

[9] Ifremer ISO9001 qualification P14 processus (IT design, development and management) online documentation (French). https://w3z.ifremer.fr/qualite/Processus-ISO-9001/Approche-processus/P14-Developper-et-administrer-les-services-informatiques

[10] International Organization for Standardization/International Electrotechnical Commission (2022). ISO/IEC 27001:2022 Information security management systems. URL:https://www.iso.org/standard/27001

[11] CoreTrustSeal Standards and Certification Board (2022). CoreTrustSeal Requirements 2023-2025 (V01.00). https://doi.org/10.5281/zenodo.7051012

[12] CoreTrustSeal Standards and Certification Board (2019). *Ifremer CoreTrustSeal assessment information* (English). URL:https://www.coretrustseal.org/wp-content/uploads/2019/11/IFREMER-SISMER.pdf

[13] Cotty Pierre (2019). *Ifremer-Sismer Core Trust Seal: document d'évaluation pour la certification* (French). DOI:https://doi.org/10.13155/76796

[14] L'Hours Hervé, Kleemola Mari, de Leeuw Lisa (2019). *CoreTrustSeal: From academic collaboration to sustainable services.* IASSIST Quarterly, 43(1), 1–17. DOI: https://doi.org/10.29173/iq936

[15] Intergovernmental Oceanographic Commission of UNESCO (2022). *Guide for Establishing an IODE National Oceanographic Data Centre, IODE Associate Data Unit or IODE Associate Information Unit (3rd revised edition)*. Paris, UNESCO, 26 pp. URL: https://unesdoc.unesco.org/ark:/48223/pf0000382882

10 Appendix A: Glossary

Acronym	Definition
API	Application Programming Interface
CF	Climate and Forecast (semantics for NetCDF)
CI/CD	Continuous Integration/Continuous Development
CTD	Conductivity - Temperature - Depth sensor
CTS	CoreTrustSeal
DAC	Data Assembly Centre
ECV	Essential Climate Variable(s)
EMSO	European Multidisciplinary Seafloor and water column Observatory
ENVRI	1) An environmental RI cluster FP7 project 2) Environment research infrastructures (in ESFRI level or upcoming) as a community
EOOS	European Ocean Observing System
EOSC	European Open Science Cloud
EOV	Essential Ocean Variable(s)
ERIC	European Research Infrastructure Consortium (legal entity type)
ESFRI	European Strategy Forum on Research Infrastructures
EU	European Union
FAIR	Findable Accessible Interoperable Reusable
FER	FAIR enabling resource(s)
GDAC	Global Data Assembly Centre
GCOS	Global Climate Observing System
GOOS	Global Ocean Observing System
GTS	Global Telecommunication System
GUI	Graphical User Interface
ICOS	Integrated Carbon Observation System
IEC	International Electrotechnical Commission
IOC	Intergovernmental Oceanographic Commission
IODE	International Oceanographic Data and Information Exchange
ISC	International Council for Science

ISMS	Information Security Management Systems
ISO	International Standardization Organization
IT	Information Technology
LTS	Long-Term Sustainability
NetCDF	Network Common Data Format
NODC	National Oceanographic Data Centre
NVS	NERC Vocabulary Serveur
QA/QC	Quality Assurance/Quality Control
RDA	Research Data Alliance
RI	Research Infrastructure
UN Environment	United Nations Environment Programme
UNESCO	United Nations Educational, Scientific and Cultural Organization
WMO	World Meteorological Organisation