Ref. Ares(2024)1153019 - 15/02/2024

Project acronym: **GROOM II**

Project title: Gliders for Research, Ocean Observation & Management: Infrastructure and Innovation

Grant agreement no. 951842

D6.2

Data management road map for the GROOM RI

Due delivery date: M34

Actual delivery date: February 2024

Organization leading for this deliverable:

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS)

Dissemination level		
PU	Public	**x**
CO	Confidential, only for members of the consortium	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951842

Author(s) – in alphabetical order	
AMU (CNRS)	Anthony Bosse
NOC	Justin Buck
NOC	Emma Gardner
GEOMAR	Johannes Karstensen
PMM	Anouchka Le Roux
NOC	Alvaro Lorenzo Lopez
WMO	Victor Turpin

Deliverable number	6.2
Deliverable responsible	CNRS
Work Package	WP6 - Technical benefits of a Glider European Research Infrastructure

Document revision history			
Version	**Date**	**Modifications introduced**	
		Change reason	**Modified by**
V0	05/2023	outline	Victor Turpin, Justin Buck, Alvaro Lorenzo Lopez
V0.1	10/2023	Initial draft	Victor Turpin, Justin Buck, Alvaro Lorenzo Lopez, Emma Gardner, Johannes Karstensen, Anouchka Le Roux, Anthony Bosse
V0.2	11/2023	Review	Pierre Testor
V0.3	12/2023	Final draft	Victor Turpin, Justin Buck
V0.4	02/2024	Final review	Pierre Testor
final	02/2024	Final document	Yves Ponçon and Clara Flack

Deliverable Executive Summary
The GROOM data management roadmap presented in this document is the outcome of an extensive review that meticulously considered the requirements of various stakeholders in the field. Encompassing data infrastructure, tools and services, network management, and skills and training, our comprehensive approach lays the foundation for effective data management within the Marine Autonomous Systems considered by GROOM RI. Notably, this roadmap refrains from specifying resources—both human and financial—deliberately deferring this critical aspect to the subsequent phase: the implementation plan. This roadmap provides a forward-looking perspective, delineating short, mid, and long-term visions spanning from 2 to 10 years. By doing so, it serves as a strategic guideline, identifying priorities and charting the evolution that the GROOM RI data management community should embrace when formulating its implementation plan. The careful consideration of these temporal horizons ensures a robust and adaptable framework, positioning the GROOM RI for sustained success in managing its data landscape. In order to effectively navigate this data management roadmap, the GROOM RI must establish as soon as possible a dedicated team, drawing upon the collective expertise of all GROOM RI members. This team should be characterized by strong leadership and a clearly defined mandate to ensure seamless implementation. While the precise resources committed by GROOM RI and its members for this purpose have not been quantified in this study, the forthcoming implementation plan must meticulously assess these aspects, acknowledging the potential hidden costs associated with data management. Notably, the comprehensive scope of the roadmap underscores the imperative for a collaborative effort, underscored by the active support of the GROOM RI community. Only through a concerted and shared commitment can we guarantee the sustained success and trajectory of our research infrastructure.

Table of Contents

Table of Tables

Table of Figures

1. Introduction

The first attempt to organise glider data management in Europe was made during the GROOM FP7 project, in 2011 with the support of the COST Action ES0904 "Everyone's Glider Observatory (EGO)". This effort has led the European glider community to a very well-established network in the landscape of the European Ocean Observing System represented by a growing EuroGOOS glider task team. This deliverable builds on this substantial track record on data management with efforts initiated in the original GROOM project supported by the EGO community. These activities lead to the definition of the first GOOS glider format (EGO)[1], the first versions of processing software, initial versions of near real time quality control, and the DAC-GDAC structures. Efforts toward the goals of developing glider data workflows to the European and international data aggregators were made by a range of actors, including Coriolis, US IOOS, SOCIB and IMOS. In 2017 the OceanGliders programme was ratified as a GOOS network. This reignited the momentum for international harmonization of processes and tools in the ocean glider community across the major glider operators. In 2021, the GROOM II and the EuroSea projects supported, together with the EuroGOOS Glider Task team, an OceanGliders workshop to introduce ocean science concepts into the OceanGliders community with the creation of the OceanGliders GitHub repository[2] that is used to manage and govern harmonization of activities like sensor handling, data quality control, data formats and provenances. Thus, the evolution of glider data management is rapid and pertinent, and from the perspective of a European glider infrastructure (GROOM RI), it is now time to think further and define what could be the organization of the European glider data management for the coming decade. At the very start and likewise the heart of the glider data ecosystem (as it is true for all ocean observations) are the vehicle operators that enable real-time data transmission, are knowledgeable on many of the data provenances, and being involved in the process to create the delayed mode quality-controlled data that ultimately should find its way to the data aggregators. In many cases there are also one or more national data centers, often linked to the operator's institutional data centers, that will be the first choice for submission of data. In fact, many data aggregators are in dialogue with these nations' data assembly centers to get access to the operators' data to save time and avoid the need for having a dialogue with each single operator.

This report is an attempt to draw a road map to achieve a unified vision of data management for gliders in Europe.

The objectives of this document are:

- Review the current data landscape in Europe and beyond.

[1] https://repository.oceanbestpractices.org/handle/11329/1253
[2] https://github.com/OceanGlidersCommunity

- Identify the GROOM RI data stakeholder community.
- Collate requirements from the GROOM RI data stakeholder community.
- Define the GROOM RI data product.
- Define a roadmap to connect data creators at national level with European and global data users.

To achieve this ambitious program, we gathered the data management community during the duration of this task, involving them in the decision-making process and collecting their feedback with regard to glider data management. We organized:

- Two weeks of virtual EuroGOOS/GROOM II/EuroSea data workshop in Summer 2022.
    - Scoped the broad landscape.
    - Identified stakeholders.
    - Identified gaps in capabilities and community priorities.
- A virtual review of stakeholders after the workshop.
- A virtual formulation of a GROOM data product canvas to organize the outputs into a data product for a GROOM RI.

This document is a direct output of consulting the community, turning the data product canvas into a roadmap with 1-2 years, five years, and ten years ambitions for a GROOM RI.

We provide the classic element of a complete road map. First, we propose a GROOM RI data vision, we identify the landscape and the milestones to achieve that goal, reporting the expected results and the challenges associated with the proposed vision. Finally, we propose a road map to guide the future GROOM RI towards a data management organization aligned with the evolution of the glider technology and the requirements of all the GROOM RI stakeholders.

2. The vision

The GROOM RI data management system is a technical and management framework to serve data and metadata to the multiple interests of a large landscape of stakeholders. It is a coordinated, robust, harmonized, sustained, and evolving infrastructure able to tackle the anticipated growing use of marine autonomous systems in Europe and the expansion of its observing capacities. This represents an exciting vision for the GROOM RI that will favor the enrollment of new academic, public and private partners.

The capacity of Marine Research Infrastructures to engage in the ESFRI road map is assessed through many aspects (see ESFRI public guide[3]), one of them being the services for the scientific community and the technical maturity of the RI. The data management road map

[3] https://www.esfri.eu/sites/default/files/ESFRI_Roadmap2021_Public_Guide_Public.pdf

described in this document aims to raise the GROOM RI data management system to this level of requirements of the ESFRI standards.

As will be demonstrated in this report, the GROOM Research Infrastructure is nested in a wide and complex landscape of stakeholders made of ocean observers, data and research infrastructures, national entities, political consortiums, public and private partners. This landscape represents both a resource and a threat for the GROOM RI. On one hand, technical and human resources exist, national and international infrastructures and coordination bodies have been in place for many years, and the GROOM community is highly visible in the European MRI landscape. On the other hand, multiple stakeholders come along with multiple expectations that can lead to some sort of disorganization if not considered and taken into account in the early design of the infrastructure. That is one of the hurdles the GROOM RI data management system must overcome.

The GROOM RI data management system, along with the data management roadmap is encouraged by the general enthusiasm of the partners towards delivering the initial version of the system. The data management system acknowledges the diversity of the stakeholders' expectations and promotes the best organization to serve the multiple interests identified in the community workshops.

The GROOM RI data management system will be built on what exists, what already works, and what is required by members and partners. It will clarify the role of each element of the data management chain, enable better coordination and development across partners, increase sustainability and support the expansion of glider observations in academic, public and private sectors.

The implementation of the GROOM RI data management system described in this document requires cooperation amongst partners to engage in the road map and support the development of the missing elements. Beyond what already exists and is operational, major work needs to be done to implement the road map, sometimes with an important impact on the existing procedures. It is imperative that core partners of the GROOM RI align politically and technically with this strategy on data management.

What we are proposing is a revision of the previous data management roadmap initiated ten years ago, incorporating recent advances in technology and emerging observation requirements, ensuring the data management roadmap is fit for the GROOM RI while being inclusive of the different national strategies regarding data management. The GROOM RI data management system we promote benefits from a decade of learning process and development toward efficiency, with higher consideration of the diversity of stakeholders toward a shared vision of the GROOM RI.

3. The landscape

The GROOM RI data management landscape is a complex panel of actors interacting with the marine autonomous data management system. Each of them has a different level of connection with the data management system that is impacting the design of the future GROOM RI data management system.

To draw the landscape, we have been requesting the European glider community to provide their view on the stakeholders involved in the data management. This exercise lead to the following mapping (see figure 1).

This figure, made of 3 concentric circles (core team, involved, informed) describes the relationship between the GROOM data management system and the stakeholders. The circles are split into three areas, top-down stakeholders, bottom-up stakeholders, third parties.

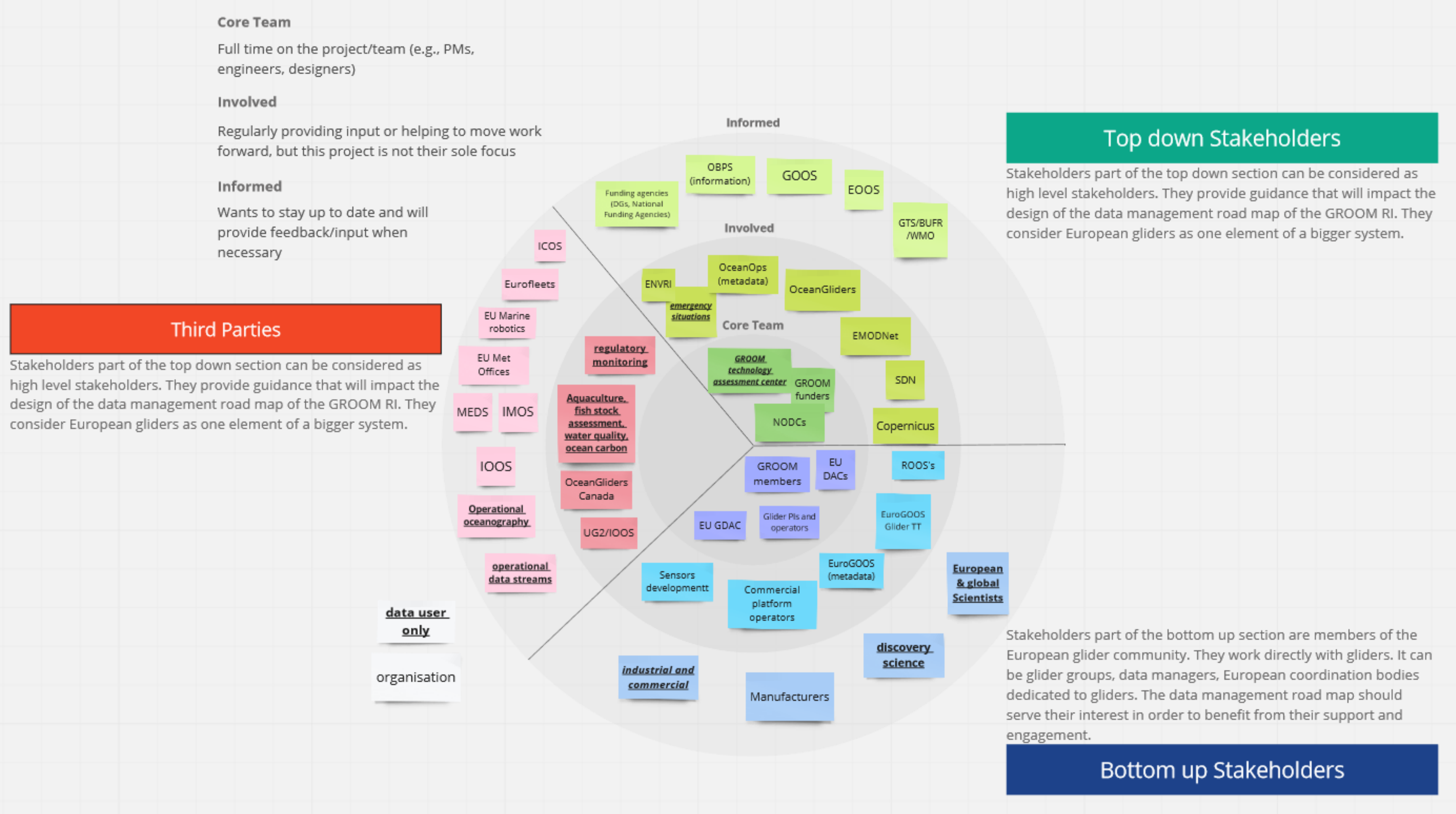

Figure 1 - Relationship between the GROOM data management system and the stakeholders.

Core Team: Stakeholders part of the "core team" circle directly impact the data management roadmap. Their requirement for data management must be answered by the road map. They will be part of the GROOM RI data management team and will contribute to implementing the road map.

Involved: Stakeholders part of the "involved" circle will have an indirect impact on the GROOM RI data management scheme. The GROOM RI data management roadmap should facilitate the implementation of their requirements but is not the main driver of the road map. They provide regular and valuable input, helping work progress, but the GROOM RI is not their sole focus.

Informed: Stakeholders part of the "informed circles" could influence the design of the data management roadmap through requested feedback and long-term policies. They could be impacted by the data management of the GROOM RI and must remain up to date with the activity of the GROOM RI data management team.

Bottom up: Stakeholders in the bottom-up section are members of the European glider community. They work directly with gliders. It can be glider groups, data managers, or European coordination bodies dedicated to gliders. The data management roadmap should serve their interest to benefit from their support and engagement.

Top down: Stakeholders part of the top-down section can be considered as high-level stakeholders. They provide guidance that will impact the design of the data management roadmap of the GROOM RI. They consider European gliders as one element of a bigger system.

Third parties: Stakeholders part of the third-party section will interact with the GROOM RI but will not influence the design of the data management roadmap. However, they have an interest in the outcome of the roadmap. Some of them provided input to the definition of the data management roadmap.

This figure demonstrates the complexity of the landscape shaping the GROOM RI data management roadmap. By mapping this landscape, we are bringing together multiple considerations about glider data management, providing a complete overview of the expectation from an exhaustive list of stakeholders. This unique exercise gives us the capacity to draft a data management strategy answering as best as possible the many requirements brought by all the stakeholders, avoiding the risk of narrowing the question to a limited number of partners. The design of the GROOM RI data management system will be driven by the very diverse requirements of stakeholders with different expectations of the RI. Those requirements and expectations are exhaustively described in this document.

4. The waypoints

This section is divided into three subsections to describe the drivers of the GROOM RI data management roadmap. The first subsection sets out several examples, non-exhaustive but illustrative, of the kind of challenges the GROOM RI data management system is facing. The second subsection stresses the multiple expectations of the data management system of the GROOM RI. The third section emphasizes the risks of not implementing the GROOM RI data management system. Mitigating those risks is a steering force of the GROOM RI data management roadmap.

It is worth noting that the glider community, through dedicated meetings[4][5] and workshops (see annexe 1) have been deeply consulted to identify the requirements described below.

4.1. The challenges

These examples or use cases address the data management challenges the GROOM community is currently facing.

4.1.1. A growing community in the private and public sphere

The recent survey about European glider capacity made during the GROOM II project drew up a list of 18 European countries (20 including Turkey and Israel) equipped with marine autonomous systems funded by public money. This represents about two hundred instruments currently operated in European seas by public operators to benefit academic research, monitoring and operational services. As shown by D5.1 : Glider Services for Public and Private Needs, glider, and other MAS, represent key elements for the future of Ocean Observing, for both public and private sector, and with the rapidly evolving technology, MAS importance will only grow on the coming years.

Adding to this number of units, the private sector is also very active with marine autonomous systems in industrial fields like offshore wind farms, aquaculture and fisheries and oil and gas. These actors represent a huge potential of activity for the GROOM RI through research and development that no organisation could support individually.

Also, the public sector is expressing a growing interest in the technology of the marine autonomous vehicle to support marine ecosystem monitoring obligations. The Marine Strategy Framework Directive, the common fishery policy and the Marine Spatial Planning Framework Directive represent an important opportunity for public authorities to capitalize on the usage of marine autonomous vehicle technology.

4 https://github.com/OceanGlidersCommunity/meeting_notes/blob/main/2022/European%20Glider%20Data%20Management%20Workshop%20Agenda%20Week%201.md

5 https://github.com/OceanGlidersCommunity/meeting_notes/blob/main/2022/European%20Glider%20Data%20Management%20Workshop%20Agenda%20Week%202.md

To put the number of existing units and their potential into perspective, the number of individual gliders with data reaching the European global data assembly centre over the last 24 months is less than 60.
This gap demonstrates the difficulty the glider community in Europe is having today in engaging new players in the marine autonomous data management system. And this will become even more difficult for non-academic operators willing to benefit from the GROOM RI data management system.

The road map carefully considers this situation and aims to build an easier access system for any marine autonomous system operator.

4.1.2. Increasing diversity of sensors

The current glider data management system was developed a fifth of a century ago following the development of the Argo and OceanSITES programs. It is well suited to support the data flow for temperature, salinity and pressure. It is also pretty efficient for the standard set of BGC variables (Oxygen, Chlorophyll-a, Backscatter, Nitrate, Irradiance, PH). D4.3 "*GROOM RI contribution to statutory monitoring frameworks*" highlights the rising need for monitoring of the ocean and the role that GROOM RI, and more largely, MAS, will have to meet these challenges.

Marine autonomous vehicles can carry a greater diversity of sensors onboard than Argo floats. In the last five years, many developments of sensors (PCO2, pH, turbulence, camera, acoustic sensors, ADCP, hydrocarbons, …) have occurred in the oceanographic world along with integration on autonomous vehicles (BioGlider, ZooGlider, BRIDGES, …). This trend will not just stop but accelerate in the coming years.

The current data management system needs to be updated for the new variables that can be much more complex to be properly qualified and often too heavy to be archived and delivered to the users through the existing procedures. The case of oxygen, which is not a "new" variable, is symptomatic. Less than 10% of oxygen datasets available today on the global data assembly center can be used for scientific research and operational services due to a lack of or imprecision in metadata. This example demonstrates the difficulties in managing sensors out of the standard CTD set. Without rethinking the data management architecture for new variables, these numbers would become even lower for new sensors. This case study highlights the crucial importance of a fit-for-purpose data system to enable the full realisation of OceanGliders and GROOM RI data as the observation capabilities evolve.

The future data management system design along the road map will diversify the data pathways to enable formatting, qualification and delivery of the growing diversity of data delivered by marine autonomous systems.

4.1.3. Reinforcing the role of national data infrastructure

This use case will demonstrate how the current data management system is bypassing the national data infrastructures and consequently lowering the system's sustainability.

The Argo approach deeply inspired the current glider data management system, designed during the GROOM FP7 project (2011-2014). It bases most of the effort on data management on the concept of data assembly centre (DAC hereafter). In the case of Argo, only two DACs exist in Europe, Coriolis in France and BODC in the UK. These two DACs support the data management of the 11 national European Argo programs. The rationale for this approach in 2011 was motivated by the idea of the development of national MAS infrastructures similar to national Argo programs and the success of the Argo data management system.
The higher capabilities of the MAS and certainly the higher investment cost set the conditions for both approaches, the national infrastructure and the institution investments in the technology. This situation led to an organic growth of institution-owning gliders in Europe. In addition, MAS can run tens of missions during the lifetime of an Argo float. This highly impacts the real-time data management effort for the DACs to process the data.
The proliferation of actors increasing the number of missions to process pushes the Argo-like DAC, GDAC approach to its limits. It is time to organise the workload differently.

To support and sustain this growing activity on data management, the option on the table is to rely more and more on National Ocean Data Centers (NODC hereafter). Many countries have an established NODC, with national scientists already committed to pushing the data to their national facility, as the financial resources for data management exist through this route. However developing an entire MAS processing chain takes work, it requires investment, time and human resources. The MAS data management system should focus on developing tools to support the NODCs in the implementation of the GROOM data management system.
However, we are not advocating to scrap the DAC/GDAC concept. The countries with existing NODC will become DACs for the GROOM RI, while countries with no existing NODC will rely on key data assembly centers to process the MAS data according to the GROOM RI requirements. The GDAC concept will be key to centralising MAS data sets, but it will be adapted to the different data types (EOVs, acoustic, videos, etc.) or applications (operational services, physical oceanography, ecosystem research, etc.).

This approach should contribute to increasing the sustainability of the GROOM RI data management system by relying on existing national infrastructure and providing the capacity to the GROOM RI to fulfil its engagement to users concerning data management.

4.1.4. Data quality

4.1.4.1. Real time data

Gliders can transmit data in near real time, any time they surface and manage to connect with satellites. Iridium coverage gives the capacity to receive the data at the DAC level within an hour and process it within 3h in the best case (about 60% of Argo data is made available at GDAC within 3h).

While for temperature and salinity data the real time data quality is acceptable for operational application, the quality of the other standard sensor carried by a glider (Fluorometer, backscatter, dissolved oxygen sensor…) is not good enough to be used in real time, mainly because the information transmitted by the glider is "intermediate parameters". Therefore, managing the quality of those data in real time require expertise from the operator and the data assembly center. Operator should configure the sensor properly and provide the associated relevant metadata, while data assembly center should compute the data base on its configuration.

Tackling this challenge will require clear guidance and common best practices within the GROOM RI members to deliver good data in real time.

4.1.4.2. Delayed mode data

Achieving delayed mode quality for gliders requires, in addition to the sensor information (model, calibration coefficient, configuration), the collection of the technical parameters needed to correct the effect of the flight of the glider.

Similarly to real time data, the challenge of delayed mode data lies in clear requirements and good documentation. This is certainly a challenge GROOM RI, in line with the international community, should tackle.

4.1.5. Emergency access

Accessing MAS to respond in emergency situations (oil spills, natural disasters…) is a key aspiration for the GROOM RI (see D5.1). Some specific situations deserve the capacity to deploy instruments at sea rapidly. To build this capacity, the GROOM RI must also build the capability to qualify the data and make it timely available automatically.

In such a situation, the robustness of the data management is central. As soon as the system is deployed, the data must fly to the GDAC timely and seamlessly, to support rapid decision making, to meet unforeseen requirements in glider piloting, effective adaptive sampling, etc.

4.2. The Community requirements

4.2.1. Top down requirements

OceanGliders, the glider program of the Global Ocean Observing System (GOOS) is the main international actor for the GROOM RI through the OceanGliders data management task team. The task team is delivering the following glider data management requirements that align with the highest international standards (see introduction).

- Delivering free and unrestricted ocean data, in real-time, through the GTS (Global Telecommunication System) for met-ocean application.
- Delivering free and unrestricted ocean data, in real-time and delayed time, at the highest quality to a GDAC (Global Data Assembly Center) for scientific research and archiving.

At the European scale, the European Ocean Observing System (EOOS) and its implementing component EuroGOOS must be considered as the cross-institutional/cross-national representation of nations interests and the principal institutional player for GROOM RI. The EuroGOOS Data management, exchange and quality working group (EuroGOOS DATAMEQ) is the data management component of EuroGOOS. EuroGOOS DATAMEQ is completing GOOS/OceanGliders requirements with the following:

- Ensure data distribution through the European ocean data infrastructures: Copernicus Marine Service (CMEMS[6]), SeaDataNet[7], and EMODnet[8].
- Enhance the overall quality of Essential Ocean Variables (EOVs[9]) through harmonized quality control procedures for real-time and delayed mode data.
- Ensure Interoperability of the EOOS ocean dataset through the management of metadata and controlled vocabularies.

The GROOM RI data management roadmap must comply with GOOS/OceanGliders and EOOS requirements on data management to provide the best quality ocean data for operational oceanography, as well as for ocean and climate scientists, efficiently.

Complying with GOOS/OceanGliders and EOOS requirements on data management will ensure that the high-quality ocean data acquired by the GROOM RI will efficiently serve operational services and ocean and climate research, two key services for the GROOM RI.

4.2.2. Bottom-up requirements

The GROOM RI must also answer the glider community needs. The community is made of Principal Investigators (i.e. scientist responsible for a glider mission), glider operators and data managers. To succeed, and to engage more and more members, the GROOM RI Data Management Road Map must consider the specific needs of each link of the data acquisition chain.

4.2.2.1. Glider operator

Operators are key providers of the mandatory information coming along with the datasets (WMO identifier, sensors attached to the glider, calibration coefficients, etc.). Metadata is

[6] http://eurogoos.eu/cmems-in-situ-tac/
[7] http://www.seadatanet.org/
[8] http://eurogoos.eu/emodnet/
[9] https://www.goosocean.org/index.php?option=com_content&view=article&id=283&Itemid=441

fundamental to achieving FAIR principles and fulfilling the top-down requirements cited above. Facing the surge of operations at sea, maintenance of gliders and sensors, many operators have difficulties fulfilling the metadata requirements of GROOM RI data management. This situation can lead to low community engagement, despite initial interest.

Recommendation: **The GROOM RI Data Management Road Map must acknowledge this situation and find solutions to simplify data management procedures to increase glider operators' engagement in the GROOM RI data management procedure.**

4.2.2.2. Principal Investigator

Principal Investigators have the responsibility to provide access to data of known quality, either directly or via their respective data dissemination plans, to the GDAC. Data quality is assessed as outlined in the respective GROOM RI endorsed Best Practices for data handling and, most importantly, the underlying procedures for estimating the uncertainty of the observations. PIs are responsible for ensuring data is best processed and delivered to users, following GROOM RI or respective global network (e.g. OceanGliders) best practices and protocols. Data access could be granted by data transmission to the GDAC, serving of data on originator APIs (e.g. ERDDAP), or providing a respective handle/link (e.g. DOI) for the data. This refers to the real-time and delayed mode quality control of the data set produced.

After more than a decade of sustained ocean observations with gliders in Europe, it is evident that the current situation remains unsatisfactory. With the evolution of joint European glider operations, it can be stated that most of the real-time glider datasets reach the appropriate repositories, also following documented RT quality control. For the delayed mode quality controlled (DMQC) data, many nations use data dissemination pathways that are defined by national or European data archiving obligations. It is unknown how many DMQC data sets are publicly available but it can be stated that (G)DACs such as *Pangaea.de* or *Seanoe.org* host many of such data sets in real time, and just a few glider datasets are quality controlled in delayed time. As for other networks and infrastructures (e.g. ICOS[10]) a data hub via a dedicated GROOM RI portal could be set up and provide access to distributed datasets via DOI links. It can be envisioned that a GROOM RI or global networks such as OceanGliders would be maintaining such a server eventually linked to European data aggregators (e.g. Copernicus CMEMS or EMODnet). Such an approach would enable GROOM RI to effectively manage both mature EOVs (salinity temperature, oxygen, nutrients, etc.) where data management pipelines are well established and immature EOVs (such as marine noise or eDNA) where data pipelines are nascent with community consensus still forming. The combination of documentation in data quality and data access will allow overcoming the current state of not knowing the versioning of data sets and provide access via a virtual central

[10] https://www.icos-cp.eu/data-services/about-data-portal

hub, enabling national contributions to be visible at the national level and fulfilling national data requirements but also serving the European and global needs.

We must acknowledge that principal investigators do not have the capacity and the guidelines yet to do the delayed mode quality control and to share appropriately the best version of the datasets they produce. Considering this situation, GROOM RI data management roadmap must lead to a situation where delayed mode quality-controlled data management is achievable.

- GROOM RI should publish best practices for delayed mode quality control to support and facilitate the work of PIs.
- GROOM RI should develop the tools, open to the whole glider community to cover the full data lifecycle from real time data transmission to data assembly center, to the delivery of the best quality data set (delayed mode data) to users.

4.2.2.3. Data assembly centers (DAC)

The role of data assembly centers is to convert raw data into a common format, quality control the data in real-time, curate metadata and distribute the best reliable dataset to the “outside” world (including GDAC). A key outcome for a GROOM RI would be for the data workflow to be made most efficient and thus ease adoption by the community.

Many of the current National Ocean Data Centers (NODC) supporting the role of glider DAC in Europe are missing guidelines to fully take on this responsibility. Other data centers are lacking resources to complete the different tasks of a glider DAC.

Recommendation: The GROOM RI Data Management Road Map must clarify the tasks falling onto the DAC’s responsibility, supervise the relationship with DACs, facilitate its implementation, provide access to resources (guidelines, tools, best practices, shared human resources) and consider the monitoring of the performance of the GROOM RI data flow.

4.2.3. Requirements from third parties

Third parties encompass the private sectors (large and small companies), NGOs and the public sector. Those users of the infrastructure can be interested in both environmental data and technical data (if we take the case of companies developing new sensors, and tools for glider operation or services). In the past, only ocean related public institutes were taken into account to define the data management rules. The integration of this new player in the design of the road map led to some key questions about how the GROOM-RI should consider the data it produces.

First: What is the role of the GROOM RI with regard to third parties’ data provision?

The role of GROOM RI is not only to acquire raw environmental and technical data for third parties but also to deliver high-quality datasets easily exploitable by the third parties, considering its level of data management know-how. Taking advantage of the knowledge available in the Research Infrastructure is part of the responsibility of GROOM RI. The case of third parties owning and operating gliders without enough resources to properly manage the data should also be considered. GROOM RI could be subcontracted by the third party for this aspect.

Second: Is that data public?

This is certainly a question that the GROOM RI executive should answer by the end of the project. At the time of writing the question has not been solved yet. However, it is likely the funder will have the last word on it but that would strongly encourage to make it public and available, at least in real time for the basic variables, for the sake of Ocean forecasting.

If yes, then the same data management procedure should be applied to the datasets. Specific formatting, when stipulated in the contract with third parties, can be applied to the dataset.

If not, then the contract between third parties and GROOM RI should stipulate data management needs and related costs.

4.3. Risk in not implementing the road map

One of the drivers for the design and implementation of the GROOM RI data management road map is to avoid the risk related to the status quo. After almost two decades of regular glider observations in Europe, extensively funded by National or European research projects, the need for a robust, harmonized, sustained and planned data management system for gliders seems obvious.

“Maturity” of an ocean observing network is a mandatory requirement in the case of evolving towards a Research Infrastructure (RI) (amongst other criteria). An established, sustainable and mature data management infrastructure is often used to assess maturity. Within the European glider community, there is a strong consensus to strengthen the data management procedures in order to reach the increasing level of expectations of the funding agencies. Without effective coordination on data management, the risk of not being evaluated as mature enough to become a RI represents a significant threat to the GROOM RI community. In addition, adopting a clear and long-term data management strategy will encourage engagement in the research infrastructure.

5. How to get there? - The GROOM RI Data Management Road Map

5.1. Reminder of the vision

"*The GROOM RI data management system is a technical and management framework to serve data and metadata to the multiple interests of a large landscape of stakeholders. It is a coordinated, robust, harmonized, sustained, and evolving infrastructure able to deal with the anticipated growing use of marine autonomous systems in Europe and the expansion of its observing capacities*."

We will break the vision down into six categories taking in account three different timescales required for their achievement (1-2 years, 5 years and 10 years goals):

- Data Infrastructure
- Tools and Services
- Network management
- Network evolution
- Skills and training
- The GROOM RI user community

The roadmap will align services to the GROOM RI use cases and should be used in combination with other GROOM RI component roadmaps.

5.2. Data infrastructure

Data infrastructure encompasses each link of the data processing chain, from operators to OceanOPS, DAC and GDAC. It also includes the technologies used to support the data flow.

- 1-2 years
    - Proposed structure for future development, including cloud-native services in the European Science Cloud (EOSC) and open-source community development.
    - Scope and agree evolution of the DAC and GDAC infrastructure/ architecture updates between national DACs and international stakeholders, including OceanGliders.
    - Define the scope of GROOM RI data portal to allow federated access to data across GROOM RI nodes.

- 5 years
    - Develop operational examples of enhanced data flows of DAC and GDAC components managed with the open-source community frameworks to be used as templates to evaluate enhancement and as templates to evolve the data system globally.

- Operational exemplars of DAC and GDAC deployed in EOSC (and other cloud providers) to be used as templates to evaluate enhancement an as templates to evolve the data system globally.
- New and enhanced tools and services aligned with international (IOC and GOOS, WMO, OBPS) policies with roadmaps for updates as the policy landscape evolves.
- Operational GROOM RI data portal to allow federated access to data across GROOM RI nodes.

- 10 years
    - Open-source solutions for the management of DACs and GDACs, developed and managed by OceanGliders community, evolving with technology, enabling DAC in a boxed modular assembly of new data system nodes (DAC/GDAC). GROOM RI will adopt the open-source solution to manage data.
    - Robust, clear, and seamless data flows to all the GROOM RI data users (EMODNet/CMEMS/research community/Government/etc).

5.3. Tools and services

Tools and services encompassed the format, the software, the data access and the visualization of data. It also takes into account the tools to measure the performance of the GROOM RI data flow.

- 1-2 years
    - Scoping of FAIR data, data policies, and UN Decade data integration within GROOM RI.
    - GROOM community meeting the OceanGliders 1.0 format.
    - Scoping of data visualization requirements, consultation and assessment including the development of a roadmap for enhanced community services.
    - Scope evolution of enhanced GROOM RI services for operational oceanography (including WIS 2.0[11] in the con tools enable evolution to BUFR by DACs).
    - RTQC consensus on standard tests in OceanGliders.
    - DMQC - Priorisation of EOVs for DMQC.
    - Scoping of QC development for GROOM RI stakeholder and road map for enhanced quality control services.
    - Integration GROOM RI data methodologies into the OBPS frameworks

- 5 years

[11] https://community.wmo.int/en/activity-areas/wis/wis2-implementation

  - FAIR data, alignment of data policy with IOC data policy (or equivalent), including integration of GROOM RI data into the UN decade services (ODIS architecture[12]).
  - Operational exemplars of enhanced data visualization and user interface for highest priority GROOM RI users.
  - Enhanced GROOM RI serviced for WIS 2.0 implemented across the GROOM RI community.
  - RTQC - standard RTQC tools applied by GROOM DACs.
  - DMQC - 1 or 2 EOVs.
  - Cloud native DMQC tools for 1-2 EOVs.
  - Alignment and publication of GROOM RI data methodologies with the OBPS frameworks.

- 10 years
  - Complete data visualization and user interface services to meet the full breadth of GROOM RI stakeholders with a roadmap for continuous enhancement.
  - RTQC - Operational BGC RTQC and assimilation by operational centers.
  - DMQC - 90% of observed EOVs allowing for new EOVs to come through.
  - A sustainable, efficient route for the addition of new EOVs into the data system.
  - Cloud native DMQC tools for 90% EOVs.
  - Indicators to operationally measure performance of data flow.

5.4. Network management

This section considers the management with the other actors involved in the data system of the different existing elements of the GROOM RI network.

- 1-2 years
  - Agreement on the scope of MAS falling under GROOM RI and set up rules for periodic review process.
  - Scope tools to enable harmonized metadata management and sharing within GROOM RI.
  - Scope planning tools for EOOS observations across marine RIs and the role of GROOM RI in these.

- 5 years
  - GROOM RI sensor and platform metadata integrated into the EOOS.

[12] https://book.oceaninfohub.org/thematics/variables/index.html

  - Complete vocabulary collections for OceanGliders under community governance on a recognised vocabulary service.
  - Tools to enable harmonized metadata management and sharing within GROOM RI.
  - Planning tools for EOOS observations across marine RIs and GROOMs role in these.

- 10 years
  - GROM RI planning and network management integrated into the EOOS.
  - Globally recognised processes to entrain new sensors and platforms with GROOM RI having leadership.

5.5. Network evolution

The new sensors, the new glider model and the expansion to other Marine Autonomous Systems is covered by this section.

- 1-2 years
  - Scope the network (e.g OceanGliders) working groups on emerging sensors and platforms and make recommendations for governing bodies (e.g. EuroGOOS/OceanGliders/etc) for their formation.
  - Scope processes to entrain new sensors and platforms into GROOM RI.

- 5 years
  - Process to entrain new sensors and capabilities into GROOM RI as part of wider OG activity.

- 10 years
  - Globally recognised processes to entrain new sensors and platforms with GROOM RI having leadership.

5.6. Skills and training

The GROOM RI data management road map is also considering the delivery of services to increase knowledge and capacity within its members.

- 1-2 years
  - Data skills audit across GROOM RI members encompassing open science frameworks, quality control, best practices, staff evolution & succession, etc.

- 5 years
    - Training courses on the gaps in data skills leveraging existing capability such as Glider School[13].

- 10 years
    - An established training network and activities within GROOM RI to ensure a sustainable data skills base.

5.7. The GROOM RI data user community

- 1-2 years
    - Define the GROOM RI data user community and establish coordination groups and stakeholder engagement forums.

- 5 years
    - Establish coordination groups and committees to govern GROOM RI data management with regular stakeholder engagement.

- 10 years
    - A sustainable GROOM RI data user community with governance and sustainable leadership.

[13] https://gliderschool.eu/

6. Conclusion and summary of recommendations

The "*Gliders for Research, Ocean Observation and Management - Infrastructure and Innovation*" (GROOM II) project aims to scope a research infrastructure (RI) for Marine Autonomous Systems (MAS) in Europe. With the vision:

"*Be the European Research Infrastructure harnessing the advantages of Marine Autonomous Systems (MAS) to provide high-quality ocean observation data and services for the benefit of society, enabling scientific excellence and moving towards net-zero activities*."

And the mission:

"*This European Research Infrastructure integrates national infrastructures for Marine Autonomous Systems (MAS) to provide access to platforms and services to the broadest range of scientific and industrial users, as well as other ocean observing RIs. It maintains a unique centralized provision of cyber-infrastructure, data and knowledge for the optimized use of MAS to study climate and marine environments, and to support operational services and the blue economy*."

The evolution of a stakeholder-focused data system, software and infrastructure is a crucial element of a future MAS RI to meet the needs of the growing number and diversity of platforms.

The GROOM II project has developed a roadmap to scope and shape the data ecosystem to be included in any potential MAS RI. The roadmap involved extensive stakeholder engagement, including dedicated online community workshops in the summer of 2022, followed by project-level workshops to distill the results into a roadmap.

The roadmap outlined recommendations for the next 1-2 years, five years and ten years, with key themes in the recommendations including the move toward open source community-governed software environments, the need to address key gaps in capability such as quality control, the need for tools and service to be readily adopted by new nodes as the MAS networks grow, and that training of the research community so ensure a sustainable future.

The roadmap also has broader implications for the wider Global Ocean Observing System (GOOS) with the need for interaction between networks on aspects such as common infrastructure (e.g. for data assembly centers to enable rapid addition of new nodes) and quality control where common sensor types are used across GOOS networks with the essential alignment of data quality to enable data across networks to be readily utilized.

The GROOM II data roadmap is currently being integrated into the wider GROOM II RI scoping. This paper aims to present the results of the data roadmap to the EuroGOOS community to solicit feedback on the results that will be incorporated into the GROOM design of a future European MAS RI.

The following table summarizes the GROOM RI data management road map as detailed below.

Table 1 - GROOM RI data management road map summary

Roadmap — GROOM II — **CHECK OUR DETAILED ROADMAP HERE**

Research Infrastructure data management	1-2 years	5 years	10 years
Infrastructure	• Proposed structure for future developement (cloud-native services & open-source community development) • International agreement • Define the data portal's scope	• Operational exemplars of DAC & GDAC managed whith open-source community & deployed in EOSC • Tools and services aligned with international policies • Operational data portal	• Open source solutions for DAC & GDAC management (OceanGliders community) • Unambiguous and seamless data flows
Tools and services	• Scoping of FAIR data • Meeting with OceanGliders 1.0 format • Data visualisation requirements • WIS 2.0 - BUFR implementation • RTQC consensus on standard tests • Integration data methodologies into OBS framework	• FAIR data alignment with IOC data policy • Data visualisation and user interface for priority • WIS 2.0 - BUFR implemented • RTQC - standard tools • DMQC - 1-2 EOVs + cloud native tools • Alignment and publication of data methodologies whit OBS framework	• Data visualisation and user interface to meet the diverse range of users • RTQC- Operational BGC RTQC • DMQC - 90% of observed EOVs allowing for new ones to come through + cloud native tools • Sustainable & efficient route for new EOVs
Network management	• Agreement of the scope of MAS • Scope tools to harmonize metadata management & planning tools for EOOS observations	• Sensor and platform metadata integrated into the EOOS • Complete vocabulary collections • Agreement of the scope of MAS • Tools to harmonize metada management & planning tools for observations accross marine RI	• Planning and network management integrated in the EOOS • Globally recognised processes to entrain new sensors and platforms
Network evolution	• New networking groups on emerging sensors and platforms • Scope processes to entrain new sensors and capabilities	• Process to entrain new sensors and capabilities as part of wider OG activity	• Globally recognised processes to entrain new sensors and platforms
Skills and training	• Data skills audit	• Training courses on the gaps in data skills	• Training network and activities
The GROOM RI user community	• Define the user community	• Establish coordination groups and committees	• Sustainable data user community

To engage in this road map, the GROOM RI must organize around a data management team, involving all GROOM RI members, with strong leadership and clear mandate to implement it. The resources dedicated by GROOM RI and its members to this task haven't been assessed in this study, but the next step, the implementation plan, should consider this very accurately and not underestimate the hidden cost of data management. What is obvious with regard to the number of items covered by the road map is that only, a mutualized effort supported by the GROOM RI community can maintain the research infrastructure on track.

List of recommendations:

Recommendation: **The GROOM RI Data Management Road Map must acknowledge the glider operators' difficulties in fulfilling the metadata requirements of GROOM RI data management and find solutions to simplify data management procedures to increase their engagement in the GROOM RI data management procedure.**

Recommendation: The GROOM RI Data Management Road Map must clarify the tasks falling onto the DAC's responsibility, supervise the relationship with DACs, facilitate its implementation, provide access to resources (guidelines, tools, best practices, shared human resources) and consider the monitoring of the performance of the GROOM RI data flow.

Recommendation: To engage in the data management road map, the GROOM RI must organize around a data management team, involving all GROOM RI members, with strong leadership and clear mandate to implement it.

7. References

[1] EGO https://repository.oceanbestpractices.org/handle/11329/1253
[2] OceanGliders GitHub repository https://github.com/OceanGlidersCommunity
[3] ESFRI public guide :
https://www.esfri.eu/sites/default/files/ESFRI_Roadmap2021_Public_Guide_Public.pdf
[4-5] European Glider Data Management Workshops Meeting notes:
https://github.com/OceanGlidersCommunity/meeting_notes/blob/main/2022/European%20Glider%20Data%20Management%20Workshop%20Agenda%20Week%201.md
https://github.com/OceanGlidersCommunity/meeting_notes/blob/main/2022/European%20Glider%20Data%20Management%20Workshop%20Agenda%20Week%202.md
[6] CMEMS: http://eurogoos.eu/cmems-in-situ-tac/
[7] SeaDataNet: http://www.seadatanet.org/
[8] EMODnet: http://eurogoos.eu/emodnet/
[9] Essential Ocean Variables :
https://www.goosocean.org/index.php?option=com_content&view=article&id=283&Itemid=441
[10] ICOS: https://www.icos-cp.eu/data-services/about-data-portal
[11] WIS 2.0: https://community.wmo.int/en/activity-areas/wis/wis2-implementation
[12] ODIS architecture: https://book.oceaninfohub.org/thematics/variables/index.html
[13] Glider School: https://gliderschool.eu/

8. Annexe 1 - Community workshop on requirements

Table 2 - GROOM data services product features canvas