D4.2
Policy landscape in the ENVRI domain

Work Package	4
Lead partner	University of Helsinki
Status	Final
Deliverable type	Report
Dissemination level	Public
Due date	31 March 2020
Submission date	26 July 2020

Deliverable abstract

This document provides a snapshot of the ongoing landscape analysis of key policies of the ENVRI-FAIR environmental research infrastructures. This document outlines the role of the Policy Working Group in the context of the project, the strategy for the policy work, the initial understanding of the policies and the intended policy framework. The main part of the document describes the (ongoing) analysis of the policy landscape, and the initial conclusions derived from this first phase of work.

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	Ari Asmi	UHEL	01 July 2020
Contributing Authors	H. Glaves	UKRI BGS	
Reviewer(s)	A. Petzold	FZJ	09 July 2020
Approver	A. Petzold	FZJ	26 July 2020

DELIVERY LOG

Issue	Date	Comment	Author
V 0.1	01 July 2020	First complete version	AA
V 0.2	14 July 2020	Revised version based on the review	AA

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.3465753.

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

D4.2 Policy landscape in ENVRI domain

1 Introduction and purpose

Wider adoption and implementation of the FAIR principles[1] has been instrumental in cementing open science practices throughout the ENVRI research infrastructure community and beyond. However, these principles are extremely general by design and, despite the commentaries provided by the FORCE11 group[2] and other associated initiatives (e.g. GO-FAIR[3]), much of the detail and definitions are left open to interpretation as part of the implementation layer. This makes alignment with these principles relatively straightforward, but creation of an integrated data system is actually more difficult. However, due to the usefulness of the FAIR principles, they have been widely adopted. For the purposes of this deliverable, the most important application of FAIR is integration of the ENVRI RIs with the European Open Science Cloud (EOSC), which has determined these principles as some of the most important requirements for being part of the EOSC (data) services.

The requirements of the FAIR principles can be approached from a technical perspective through identification of the solutions required for their implementation. In the context of ENVRI-FAIR, this analysis has been undertaken by WP5 and documented in the recently published deliverable D5.1 on Requirement Analysis, Technology Review and Gap Analysis of Environmental RIs[4]. However, this technical perspective can sometimes be incomplete because the changes in organisational structures, stakeholder requirements, technical management and (perhaps most importantly in this context) external technical developments can lead to any such landscape analysis being incomplete in terms of the overall goals and objectives of the individual research infrastructures (RIs). A comprehensive understanding of existing relevant policies, rules, and strategies is key, and in many cases the existing practices might not have been properly considered from the perspective of the organisational goals and needs. That is to say that practices might have been organically created at some level of the organisation that, while fitting a specific need, might not map well on to the long-term objectives of the RI as a whole.

Another aspect of this deliverable is the concept of *organisational debt*[5] in the current environmental ESFRI RIs with regards to FAIR data services. This term means that some aspects of an RI’s operations have not been completely decided or clearly defined during the RI development phase. Sometimes these kinds of organisational debts are a direct result of the non-alignment of organisational development, due to outside influences, or a lack of decision making bodies. Like in any debt, the longer

[1] https://doi.org/10.1038/sdata.2016.18

[2] https://www.force11.org/fairprinciples

[3] https://www.go-fair.org/fair-principles/

[4] https://zenodo.org/record/3884998

[5] This term is analogous to ”documentation debt” or “programming debt” used in software development, where suboptimal or poorly commented code is allowed to be a part of the production software due to time or resource constraints. Usually such practice will lead to increased overall costs (i.e. “interest”), but can be useful or even critical for success.

key aspects remain undecided, the “interest” – in this case the effort to actually codify already existing possibly divergent practices – increases and ultimately requires more work. Absorbing this type of debt is a natural result of any rapid organisational development, but it also needs to be understood and mapped, and the associated risks should be readily known in the RI management.

ESFRI RIs in the environmental sciences are highly heterogeneous, and have a wide range of organisational, technical and operational readiness levels, which is often evident form the existing policies that have been put in place. The Policy Working Group (PWG) is one of the key aspects of ENVRI-FAIR. The PWG aims to develop a common understanding of existing RI policies that are necessary for their integration into the EOSC, facilitating further interoperability, and developing an environment for environmental scientists compliant with the FAIR principles. The first step for the ENVRI-FAIR PWG is to analyse the current policy needs of the environmental RIs (at least as far as they are already known), and determine the current level of implementation. This work will be used to create a suggested policy framework for the entire community of ENVRIs, enabling the participating RIs to analyse their own policy needs and objectively consider any differences from this framework.

This kind of analysis is commonly called a *landscape analysis* and it is the subject of this deliverable. However, for the landscape analysis to be useful[6], it must be well targeted (which kinds of organisations are involved), well scoped (what kind of information is collected), and feasible (how much information is collected). All of these aspects are considered in the next section.

2 Methods

2.1 The policy working group (PWG)

Although this deliverable is focused on the overall landscape analysis, the composition and mission of the PWG is also relevant for this activity. (The composition and terms of reference for the PWG have been documented in deliverable D4.1.) The composition of the PWG was originally intended to include selected directors of the more mature RIs in the ENVRI cluster across the four environmental subdomains (marine/water, solid earth, atmosphere, ecosystems).

However, it was subsequently decided that the composition of the PWG should be modified to include representatives from the selected RIs that were more familiar with the policy aspects of the individual research infrastructures.

2.1.1 Strategy for policy development

One of the key aspects discussed in the PWG was the strategy for policy development, which included the following:

1) Clarifying the role of the PWG and WP4 in the context of ENVRI-FAIR, BEERi and participating RIs. Who is the policy framework intended for, and

[6] This section is based on the experiences from H2020 RISCAPE project, which analyzed the landscape of the international (i.e. non-EU) research infrastructures globally (manuscript in preparation, but some aspects available in the final report: https://zenodo.org/record/3539254)

how it is to be developed within ENVRI-FAIR? What is the level of interoperability needed or desired?

2) Analysis of the policies needed for integration of the participating ENVRI RIs or harmonization between the ENVRI RIs in the context of EOSC. This work includes analysis of the FAIR principles from a policy perspective, and discussion within the consortium regarding other important policies necessary for the required level of interoperability.
3) Landscape analysis of the policies defined in 2), based on interviewing the RI representatives, and creating a consistent dataset of existing (or developing) policies in the RIs involving the key aspects of FAIRness. The analysis should be made as transparent, useful, and consistent as possible, but still include the necessary information and reflect the existing diversity of RIs.
4) Based on this analysis (3), the PWG will then consider the overall policy framework and individual policy options that could be refined in consultation with RI representatives, BEERI and other stakeholders. These consultations are then directly included in the policy framework.
5) Based on these consultations, and with support from the legal experts participating in WP4, the PWG will draft the overall policy framework that provide suggested set(s) of policies that include optional disciplinary, service, and legal boundary conditions.
6) The overall framework will then be published and shared with stakeholders as part of the outputs from the ENVRI-FAIR project (i.e. this framework has no actual approval mechanisms outside of the project itself). However, general adoption is then supported for the period of the ENVRI-FAIR project.

This deliverable is focused on step 3) above, with some outline information regarding the first two steps, and gives initial guidance on the subsequent steps.

2.2 Definitions and targets

As mentioned in the introduction, a landscape analysis first needs to clearly define the goal of the analysis, the kinds of information collected, identify the target organisations, and, based on these, select the methodology and analysis methods.

The PWG defined the following analysis framework:

1) Goal of the PWG is to understand the minimal set of key policies (see section 2.3) needed for FAIR data services in the context of ENVRI-FAIR, and to support RI interoperability with the European Open Science Cloud (EOSC). These policies need to be either harmonized, or at least documented in a harmonized way, as part of a common policy framework. To achieve this objective, WP4 is conducting a landscape analysis across the participating RIs to determine the current status and immediate plans for adopting and/or implementing the identified key policies in a comparable and useful format.
2) The information collected needs to be relevant for creating common policies across the ENVRI-FAIR RIs for delivery of interoperable FAIR data services to the EOSC.
3) Target organisations are the ENVRI-FAIR RIs or the representing partners where there is currently no single legal entity.
4) The chosen methodology was based on conducting virtual interviews using a series of predefined questions that were supplied to the interviewee before the

meeting. This format was selected to allow a wider ranging and in-depth discussion, which would potentially provide more information and new perspectives. It was also decided to use this approach to improve the response rate, which is often low using other forms of online questionnaire. The format of the interviews followed that of the previous RISCAPE project where, for consistency, the same interviewer would conduct all of the interviews. The interviewer discussed the meaning of both the question and the answer with interviewee whilst completing a template with the responses.

The overall interview process was as follows:

a. The interviewee was contacted by email to explain the overall purpose of the interview, and to share the information package (see chapter 5 Appendix)
b. Meeting was organised with RI representatives via a virtual platform e.g. Zoom. In some cases, this was done several times for an individual RI in order to capture input from different personnel where a single meeting was not possible.
c. During the interview, the RI personnel were first informed about the purpose of the interview and the process involved. Their consent to record and report their responses was then confirmed.
d. The interviewer shared a Limesurvey page showing the questions, and each question was discussed in detail with the interviewee(s). Each answer was agreed mutually before finalising the survey page.
e. After the interview, the survey link and the answers were also shared with the interviewees to give them an opportunity to make corrections.

2.2.1 What is a policy?

> *1b) a **definite** course or method of action selected from among alternatives and in light of given conditions to **guide and determine present and future decisions***
>
> *Merriam Webster Dictionary*
>
> *The term is used in many different ways, varying from institution to institution, organisation to organisation and sometimes within institutions and organisations as well. It can be hard to pin down, but there are some central features common to all good policy:*
>
> - *it states **matters of principle***
> - *it is focused on **action**, stating what **is to be done** and **by whom***
> - *it is an **authoritative statement**, made by a person or body with power to do so.*
>
> *Above all, good policy is a tool which makes administration easier, and allows people to get on with the organisation's core business more efficiently and effectively.*
>
> *Australian policy handbook*
>
> *A policy is a deliberate **system of principles** to **guide decisions** and achieve rational outcomes. A policy is a **statement of intent**, and is **implemented as a procedure or protocol**. Policies are generally adopted by a governance body within an organization. Policies can assist in both subjective and objective decision making. Policies to assist in subjective decision making usually assist senior management with decisions that must be based on the relative merits of a number of factors, and as a result are often hard to test objectively. In contrast policies to assist in objective decision making are usually operational in nature and can be objectively tested, e.g. password policy*
>
> *Wikipedia on "Policy", retrieved 10.7.2019*

We consider organizational rules, best practices, and operating methods of the research infrastructure to be *policies* in this document. They are formal rules of operations, and typically are not explicitly technology dependent, but instead meant for the staff and contributors (humans) to implement inside their organization (RI in this case). They are the local standards for people working within the organization. However, it is important to note that the interaction between technical standards and policies is complex, and in many cases they do not intersect in a way that leads to specific technical requirements, or the existing technological limitations directly influence policies.

Policies can exist in a written policy document, but sometimes they are only unwritten practices or de-facto operational procedures within an organisation. However, some policies may be completely lacking, or there can be many conflicting versions of existing practices, and even different interpretations of written policies.

2.2.2 General expected aspects of ENVRI RI policies

There are some general principles for these policies that can be deduced from the needs of the ENVRI RIs and general good governance practices:

- **Authorized,** meaning that the originators of the policy have the necessary mandate to make such a policy. This also includes the requirement to have clear and documented authorship of the policy and the process how it came to exist. Note that the authority does not need to be an actual organizational subordination (subordinate *Relationship*), but can be derived from conditional or contractual obligations. E.g. in a data policy, the use of data can also add requirements for the data user, which they need to separately agree (contract), or that to be "labelled" by an RI as a qualified data source would require fulfilling a set of policies (condition for acceptance).

- **Targeted**, i.e. the policy must define specific people within or outside of the RI (*Positions*), who are the targets of the policy. They are bound by the rules defined in the policy and are required to comply with the policy requirements. These targets should be defined as *Positions*, and not as specific positions of individuals, otherwise each change in the RI organization would require change in each of the policies associated to that *Position*. However, the RI must then have a way to map all of the *Positions* required in the policies to specific positions in a separate document (or policy).
- **Scope**, which means that the policy must have clearly defined mandate, *Actions* and *Relationships* which it controls, *Positions* it affects, and a specified time limit it is active, or it needs renewal (even if "until further change"). The boundaries must be defined in the relationship to the other policies and actors.
- **Defined**, meaning that all terms and notations are clearly defined in the policy or are properly referenced in openly available documentation, and are understandable to a "non-expert reader" (with additional clarifications if needed).
- **Strategic,** meaning that the policies suggested are in line with the RI and other relevant organizational strategies, and are not in conflict with the organizations' vision, mission and strategic plans. These strategies must be reflected in the policies in relevant way. This is crucial for many ENVRI policies, otherwise their potential for implementation will be limited.
- **Consistent,** with itself, other RI policies, and the policies and regulations of other relevant authorities (national, EU, stakeholders such as Copernicus, etc.). This consistency should be properly referenced.
- **Available,** meaning that policies and documents they refer to are findable and openly available. This means also that all of the RI policies should be clearly visible in their webpages and include the necessary metadata.
- **Feasible,** meaning that the policies developed are realistic to implement in the RI, considering the resources (people, funds, time) available, institutional and stakeholder acceptance, and long term policies. Evaluation of the feasibility needs clear consideration of the short- and long-term resource commitment at the very least.
- **Monitorable,** listing clearly defined ways to monitor the policy implementation and adherence in the RI operations. There may be specific monitorable activities, or selected well-defined key performance indicators (KPIs). These should be also reportable to outside of the organization.

Note the similarity to the FAIR principles, and in the general sense one should require that all ENVRI Policies are also FAI(R).

2.2.3 Interoperability goals

As the PWG (and ENVRI-FAIR in general) is aiming towards interoperable ENVRI services, there is a need to define what interoperability means in this context. There are many types and levels of service interoperability, which can be defined from very general to extremely detailed technical levels. In the loosest sense, interoperability does already exist in practically all available research services as, with enough (manual) work, reading of manuals and communication with authors, one can (in theory) join almost any service with another. However, this is not usually considered to be interoperability in the context of the PWG.

In the strictest sense, interoperability would mean that every service provided by the RIs (data, visualizations, metadata, analysis tools) would be completely interchangeable, using the same formats, metadata, dictionaries, terminologies, interfaces, access methods and terms, licences etc. This would mean that all of the services could be "dropped-in" to compatible virtual research environments (VREs) or similar platforms. This level of interoperability can currently only be achieved in a single RI product catalogue or, in some rare cases, over a smaller subfield of the ENVRI services, e.g. within a single sub-domain. Due to practical considerations, the full interoperability described above is unlikely to be feasible for the ENVRI community. For PWG is aiming towards determining the realistic level of interoperability achievable and adjust the target level of the PWG work accordingly.

The ENVRI-FAIR mission can be considered - from the perspective of user communities - to be creating an environmental research catalogue of services, which are usable with each other with *minimal additional effort* from the user communities.

This means that we do not directly aim for full interoperability of all services, with a full canonical integration. Instead we aim towards ***well documented service (and policy) interfaces, with (minimal set of) common elements needed for service integration***. From the policy point of view, it suggests that **each individual policy does not need to be identical** – they just have to be **identical from the service interaction interface** perspective, i.e. where they are critical to the interaction between individual services.

2.3 From requirements to the interview questions

Analysis of the FAIR principles and the associated descriptions were used to derive an initial list of policy requirements for FAIR service provision. This list was then analysed to produce a reduced list of questions, which is based on the identified potential policies, and the information which can be realistically collected.

Due to impacts of the COVID-19 crisis and the need to consolidate this work with that of ENVRI-FAIR WP5 and specifically deliverable D5.1, the interviews have been severely delayed. Furthermore, it has been recognised that to maximise the potential benefit of this landscape analysis, rather than capturing a single 'snap shot' of the current status of policy adoption and implementation within the participating RIs, it should be a living document that is regularly updated throughout the ENVRI-FAIR project. This will then help to understand how the adoption of these policies with the ENVRI community has evolved during the lifetime of the project. This version of the landscape analysis (and deliverable 4.2) presents the current situation as of 1 July

2020, and includes most of the participating RIs. Only ACTRIS, EMSO, and EPOS are not included in this iteration of the analysis.

The next section documents the interview questions, together with the responses and a preliminary analysis for each question.

3 Results

3.1 Licence policies

3.1.1 Licence policy. Do you have chosen a licence for your data? If so, which? Are multiple licences applied?

3.1.1.1 Background and purpose

Licences are a crucial part of the data provision, as they describe to the user communities what they are able to do with the (meta)data provided. Licences are important from the ENVRI-FAIR perspective because this information is needed to create an interoperability layer for combining the services. The requirement is driven mostly from the definition of “Accessible” in FAIR, but is crucial for “Interoperable” (especially for machine-readable form). A notable additional point is that all of the interviews specifically concentrated on the “core data” (if defined) for the RI, and not so much on additional data stored in the RI systems, which can be either legacy datasets, campaign-type data repositories, or other ancillary data.

3.1.1.2 Preliminary analysis

Overall, open licences seem to be the norm in all those RIs interviewed. However, convergence on a single licence type is still challenging. Almost all RIs mention the Creative Commons (CC) licences, particularly the Attribution (BY) version. In the discussions, the existence of national and local licences (sometimes for tighter use constraints) was also mentioned, particularly for those RIs dependent on the data collection from their national institutions. In some cases use of such licences were required by national laws or regulations, making harmonisation even more challenging. Heterogeneity of licences seem to be higher for RIs with less centralised data collection procedures, and it seems that at least a few of those interviewed are working on an analysis of legal interoperability of the partner RIs. Not all of the RIs have actual policies (i.e. documented practices) regarding the licences as yet, and thus have a degree of flexibility on the choice of licence for the future. In some cases, the licence is completely defined by the contributing RIs, but (at least for AnaEE) double licencing for CC is being considered.

Based on this analysis, there is the potential for ENVRI-FAIR to recommend a common policy which requires the RI to work for a limited set of different licences, although the prevalence of individual national licences can be challenging to map (particularly in context of their legal interoperability). The potential for double licencing should be investigated for the RIs not able to consolidate to one licence. Convergence on a single licence type is unlikely, but a recommendation for CC or CC-BY licences could be achievable.

3.1.2 Is the licence machine readable?

3.1.2.1 Background and purpose

Machine readability is a crucial part of the interoperability and findability criteria for data. Determination of the applicable licence for each dataset is therefore a key criterion for identifying those datasets which are interoperable and compatible. The technical solution for implementing machine readable licenses vary, but the main aspect in this question (as with others) is to map the potential for such policies. Inclusion of already existing systems is important if we consider machine readable licenses to become a more common policy.

3.1.2.2 Preliminary analysis

Some RIs report that their licence information is available in the metadata. However, this was rarely mentioned as a "policy", but it is a clearly a practice that is being widely implemented. This suggests that this practice is potentially a common policy, however any technical solution would need to be harmonized with other work packages, and especially WP5, which serves as the overseeing work package for the implementation on RI level.

3.1.3 Is the license also applicable for your metadata? If not, what license is used for your metadata?

3.1.3.1 Background and purpose

Metadata does not always have the same attention to licences and explicit policies as data does. If mentioned at all, it is often considered to be an integral part of the data, and therefore has the same licence. However, in practice, the metadata licence can differ from that of the data where their origins are not the same. For example, within distributed infrastructures the majority of the data comes from distributed nodes, but the metadata is at least partly completed by the RI itself. Furthermore, because the metadata and data can be used separately an explicit policy (and communication) for the metadata licence can be useful.

3.1.3.2 Preliminary analysis

The metadata does not yet appear to be explicitly licenced by the RIs in most cases, and where it is, the licence is often considered to be intrinsic to the data policy. Overall, in practice the metadata is considered to be open (even if not explicitly licenced) but some RIs are developing separate metadata licence policies.

3.1.4 Do you have a definition of what is a dataset in your RI?

This question refers to both scientific (i.e. variables involved) and practical (temporal, experiment, etc.) separation of each dataset (data granularity).

3.1.4.1 Background and purpose

The FAIR principles include several mentions of the concept of data and dataset. However, the definition of these terms is not always clear. This creates a problem because a clear definition of a (domain dependent) dataset is required in order to assign persistent identifiers PIDs, provide usage information and track use of individual datasets. This is key for the Findability element of the FAIR principles and a perquisite for many other aspects of interoperability.

3.1.4.2 Preliminary analysis

Definition of datasets, particularly for distributed and heterogeneous RIs are not commonly available. Some, especially operational or otherwise centralized RIs, do have a clear definition of a dataset with respect to their subdomain requirements. However, the nature of the RIs operations clearly does not support a single way to do dataset selection: The main concern in most cases is the applicability for the scientific use, and as such the “phase direction” of selecting the dataset boundaries vary widely. For some RIs, the selection criteria are geographic and/or time based, while for others it is a single experiment, or a specific species, etc. However, the common requirement is for a set of standard solutions, even if there several different definitions used in the same RI. For the purposes of interoperability, WP4 could potentially develop a framework for listing the different options for dataset definitions and create a mechanism for sharing this information with the user communities.

3.1.5 Do you define a data version? What are the policies/practices (if any) regarding versioning?

3.1.5.1 Background and purpose

Scientific outputs often require correction, re-calibration, and refinement as the methods for analysis improve. However, clear policies on what is defined as a “new” version vary, and methods and strategies to communicate changes in the data are not currently very common despite this being critical for reuse. Ideally, the RI data systems should manage the versioning of datasets, including directing users of earlier versions to the new ones (via the metadata or otherwise), and provide clear provenance for all changes to the (meta) data.

3.1.5.2 Preliminary analysis

The strategies for dealing with versioning of data differ between RIs, and are in some cases divergent (e.g. allocating a PID to the new or old version of the data? How are the versions differentiated?). Data levels do exists in several of the RIs, but have usually varying definitions and official status with respect to e.g., raw data, preliminary data and final data. The LifeWatch RI approach using blockchain-based versioning and provenance is a unique attempt to address this issue. Suggestions for a common policy will potentially include a recommendation to more fully map these versioning strategies, and define a standardised approach to describing them in the metadata available to users.

3.2 Persistent identifiers

3.2.1 Do your datasets have PID? Are there exceptions?

3.2.1.1 Background and purpose

Many of the FAIR principle definitions include the need for each dataset (see section 3.2) to have a Persistent (unique) Identifier. However, RI policies on this requirement are often non-existent or unclear. Having a clear PID policy is advantageous for RIs, especially if it is also clearly documented for users.

3.2.1.2 Preliminary analysis

Although most of the RIs involved in ENVRI-FAIR require use of some sort of PID for all datasets (often specified as Digital Object Identifiers - DOIs), this is not yet generally available policy, and some of the RIs specifically do not have them at the moment, but plan to implement them. Most RIs are considering use of PIDs (an exception is EISCAT_3D), but the degree of implementation varies between RIs. The responses from the participating RIs demonstrate the need for a clear policy on PIDs (both for each dataset, and when and how they are minted). This should form the basis of a recommendation to all RIs.

3.2.2 Which PID(s) do you use? Are they internal or external?

3.2.2.1 Background and purpose

Having a single policy defining the types of PIDs being used is important for users and the internal consistency of the RI. This will also avoid having multiple different ways to point to the same dataset. Internal PIDs have more flexibility, but external (especially widely used DOIs) have a much broader application area.

3.2.2.2 Preliminary analysis

The use of DOIs seems to be prevalent in most of the RIs, and there is a clear potential for harmonization. Furthermore, the number of other types of PID in use is still manageable, and they are employed due to very clear domain-specific requirements.

3.3 Metadata policies

3.3.1 Does your data have included metadata?

3.3.1.1 Background and purpose

Metadata is central to the FAIR principles, and to the European Open Science Cloud (EOSC). Without a clear policy to include metadata, and other associated policies, the FAIRness of the services is hard to evaluate, and interoperability is more difficult to achieve.

3.3.1.2 Preliminary analysis

All participating RIs do require metadata, at least for their core products.

3.3.2 Does your RI's metadata follow some standard(s)?

3.3.2.1 Background and purpose

Metadata standards are vital for the findability and reusability of RI data. However, many of the applicable standards are quite general, but there are more detailed (i.e. use and interoperability) related metadata elements that are highly domain specific. To achieve some of form of high-level interoperability between the RI data products requires collecting and (if possible) harmonizing or translating these standards.

3.3.2.2 Preliminary analysis

There a number of common and closely related standards used by the RIs for primary discovery metadata (ISO, INSPIRE, EML, Darwin core etc.), and their adoption and use varies widely. Many of the RIs have their own metadata model built on these standards, which are expanded according to subdomain-specific requirements.

3.3.3 Does your RI use controlled vocabularies for metadata?

3.3.3.1 Background and purpose

Connected to the previous question, a policy to use controlled vocabularies is a crucial part of a successful and interoperable metadata system. A related issue is the selection of these vocabularies, including how they are updated.

3.3.3.2 Preliminary analysis

Most of the RIs involved in ENVRI-FAIR are using relevant controlled vocabularies, with a reasonable amount of overlap, particularly in the atmospheric RIs.

3.3.4 Are these vocabularies controlled by your RI?

3.3.4.1 Background and purpose

If the vocabularies are controlled by the RI, they are generally more flexible and easier to update. However, such internal vocabularies also require maintenance and have long-term sustainability, and they can be harder to integrate into other systems. Common external vocabularies might be more challenging to implement, and are outside the governance of the RI.

3.4.5.2. Preliminary analysis

The majority of the RIs have some elements which are outside the control of the RI, but most keep some of the key vocabularies inside their own system. Policies on how to adjust common external vocabularies would be potentially useful in WP4.

3.3.5 Does your RI have a policy/practice for quality control of metadata?

3.3.5.1 Background and purpose

Metadata quality is crucial, and QC/QA processes are part of most analyses of the FAIR principles. RIs should be well prepared to produce quality controlled metadata (and data), and any associated policies are crucial.

3.3.5.2 Preliminary analysis

Not all of the RIs have policies or best practices related to the QC of metadata. However, some RIs do have varying levels of quality control, ranging from simple automated completeness checks to extensive peer review. Identifying a single process that is applicable across all RIs would not be feasible, but requiring some level of quality control would be a potentially useful common policy. Quality information should also be made available to the users.

3.3.6 Do you share your metadata?

3.3.6.1 Background and purpose

The FAIR principles, particularly the Findability criteria strongly suggest that metadata catalogues are openly available for external searches.

3.3.6.2 Preliminary analysis

Most of the catalogues are either already available or in the process of being enabled for general access, but there are some limitations and restrictions that need to be considered.

3.3.7 Do you have a policy on authorship of the data?

3.3.7.1 Background and purpose

Data authorship (usually described in the metadata), is a crucial element for following the data usage, and also offers an incentive for the national data producers to share their data. However, policies on who is actually identified in the metadata vary, and are sometimes inconsistent even within a single RI. Additionally, the metadata can also include references to the RI data curation work.

3.3.7.2 Preliminary analysis

There is significant variability in practice and policy between RIs with respect to data authorship and who is identified in the metadata. There is currently no single process with individual RIs having different levels of authorship for data.

3.4 Retention

3.4.1 Do you have a policy for retention of data and metadata?

3.4.1.1 Background and purpose

Data and metadata retention are key features of data repositories, and having related plans and policies are critical for provenance, sustainability, usefulness, and trustworthiness of data.

3.4.1.2 Preliminary analysis

Perhaps surprisingly, not all RIs have documented retention policies, and this is definitely an aspect which could be considered within ENVRI policy frameworks.

3.4.2 Do you have a described data deletion / reduction process?

3.4.2.1 Background and purpose

In some cases, there is a need for reducing and or deleting data volumes. The original basis of this question was more focused on managing resource limitations (i.e. too large space), but other reasons for data deletion could be also important for RI policies.

3.4.2.2 Preliminary analysis

Most RIs do not reduce their datasets, and have no such policies in place (or even plans for them). However, there could be exceptions for reasons associated with the size of the datasets (EISCAT), security and ethics (DISSCo), or legislation e.g. GDPR. These exceptions will potentially be explored more fully as this document evolves during the lifetime of the project.

3.4.3 Do you have a policy/plan for data/metadata availability in the long-term (e.g. closure of the RI)?

3.4.3.1 Background and purpose

Metadata availability in the event of a data system closing down due to financial or other reasons is a minimal requirement for FAIR data. Also having a long-term preservation policy for the data itself would be preferable, especially in the environmental sciences where most observations are not repeatable. A clearly stated policy on this issue would also be desirable for the user communities.

3.4.3.2 Preliminary analysis

Most RIs have considered long-term sustainability of data and metadata, but not all have a consistent policy or solution as yet. This issue was clearly recognised by the RIs during the interviews, and should be part of the future developments for those RIs that do not have policies and/or solutions already in place.

3.5 Data access

3.5.1 Do you have an access policy for the data?

3.5.1.1 Background and purpose

Access services is a core RI function. A clear, available, transparent, and documented access policy is essential for “Accessibility” in the context of FAIR data, as well as for any operational system contributing to the EOSC.

3.5.1.2 Preliminary analysis

The strategies differ between the RIs, but most have clear and written data access policies that are often set-out in the statutes, binding agreements, or work plans for the individual RIs. There could be clear benefits to further analysis of RI access policies, including categorisation of key aspects of the access mechanisms to develop a set of selected common policies.

3.5.2 What are the access formalities for data (or different kinds of data)?

3.5.2.1 Background and purpose

If the data is not available anonymously, the access protocols (even light ones) should follow a clear and transparent process, preferably documented in a policy.

3.5.2.2 Preliminary analysis

Those facilities which require some form of access system usually have defined processes in place at some level, even though the actual policy does not seem to be clearly defined in some cases, or is being developed. Clearer documentation of these processes would be advantageous for users and the purposes of the ENVRI-FAIR project.

3.5.3 Is there a formal process for accessing restricted data (if relevant)?

3.5.3.1 Background and purpose

In the case of restricted data, having a clear access policy with a transparent review process, appropriate ethical guidelines, and the potential for corrective actions is clearly good practice. This question was mostly aimed at collecting relevant information on potential solutions from existing RIs.

3.5.3.2 Preliminary analysis

Not all RIs have restricted data, and those that do refer such cases to the national nodes, without having a specific local RI policy.

3.5.4 Do you allow external authorization (i.e. registration via another trusted source)?

3.5.4.1 Background and purpose

The capability to use results across different RIs would benefit from only needing a limited set of authorisations. A single-sign-on is not always possible, but there are clear advantages to sharing the authorization or identification information between the ENVRI RIs.

3.5.4.2 Preliminary analysis

Most of the RIs either have such systems or are investigating their use. The only exception is EURO-ARGO which only provides anonymous access due to specific legislation that prevents any tracking of users. EDUGAIN is clearly most widely used authorisation system, and could potentially be a good technical solution to share between RIs.

3.5.5 Do you require users e.g. citation or reference to your data if used? If so, do you provide a clear definition of how this is done?

3.5.5.1 Background and purpose

Most of the ENVRI RIs have reporting requirements for usage of their services. Distributed RIs are also dependent on the provision of their national nodes, which have their own challenges for demonstrating their impact. Having a clear policy on how to attribute the RI and the data producers is beneficial for measuring impact of the RI. A more consistent approach to citation/referencing would make use of results across RIs easier.

3.5.5.2 Preliminary analysis

These policies vary significantly between RIs and range from co-authorship to simple recommendations. Not all of these recommendations are even easily (or automatically) visible to the users.

3.5.6 Is the access policy available in machine readable form?

3.5.6.1 Background and purpose

Unless the access policy is available in a machine readable form, the access procedures (even if automated), typically require some form of human interaction.

3.5.6.2 Preliminary analysis

A limited number of RIs have their access mechanisms available in the metadata. As the technological solutions are also not clear, suggesting such a policy can be challenging.

3.6 Ownership and rights

3.6.1 Do you have agreement with your data providers (if relevant) on the ownership and licencing of data?

3.6.1.1 Background and purpose

Sharing data and metadata via the RI platform requires the necessary permissions. In many cases this is via implicit agreements, which carries inherent risks such as questions on what can be done with the (meta)data in the future, or withdrawal of some of the data assets. Clear agreements with the data providers should be the norm, unless the data provision is completely covered by the internal ENVRI RI processes.

3.6.1.2 Preliminary analysis

Many of the ENVRI RIs already have this type of agreement, and others are either considering or building a set of licence agreements. How these are documented is still unclear and will need further investigation.

3.7 Service availability

3.7.1 Does your data service have a service level agreement (or other way to define uptime goals, etc.)?

3.7.1.1 Background and purpose

Service level is critical for several aspects of potential data use. The FAIR principles require that at least the metadata is available (see question 3.8.2), but more importantly from the EOSC point of view is that new services can be built on those that already exist. In this respect, having a clear (and documented) policy on expected service level is crucial for the downstream users.

3.7.1.2 Preliminary analysis

Almost none of the ENVRI RIs have this type of service agreement, or follow (and publish) such indicators on the services they provide. In most cases “best effort” is considered acceptable because many of the services are not considered to be operational (instead they are “science services”). However, monitoring and documenting historical levels of performance could be useful for statistical purposes.

3.7.2 Is there a policy for keeping data and metadata constantly available?

3.7.2.1 Background and purpose

This question is connected to an earlier section (3.8.1), but is more directly focused on the FAIR principles, where many of the interpretations specifically require that, as a minimum, the metadata is constantly available for users. As many services depend on federated metadata searches, such a requirement is understandable.

3.7.2.2 Preliminary analysis

Very few ENVRI RIs have a direct policy on maintaining constant access to metadata, but this is clearly either an intent or expectation for most of them. Only EISCAT_3D clearly stated that such constant metadata access is not in their operation plans at this point in time.

3.8 Policy availability

3.8.1 Generally, do you have the RI policies/practices published in a findable and accessible way? Do they have PIDs?

3.8.1.1 Background and purpose

A key aspect of policies is that they are available. However, not all of them can necessarily be included in the metadata for the services or data, but they should be FAIR. Having relevant policies even as a pointer to the policy document implemented makes human interface possible. Implementing policies as pointers to the policy documents makes human interface possible.

3.8.1.2 Preliminary analysis

Many of the ENVRI RIs either have a repository for policies, or are planning to create one The EURO-ARGO approach of publishing their policies in a best-practices journal is an interesting approach.

4 Conclusions and next steps

As the landscape assessment process is ongoing, and some of the RIs have not yet provided a response, the analysis is currently incomplete[7]. However, the RI responses received so far already show some common trends:

- There is clearly strong need for policy harmonization and development. Many of the aspects of RI operations (mostly driven directly from the FAIR principles) are not considered at the policy level, and even if there is a common practice, it is not transparently communicated to the users
- Many individual solutions are yet to be shared at the overall decision making level, especially between the ENVRI RI subdomains.
- There is clear understanding of the need for relatively formal policies, particularly in service access and definition, but the various forms are not compatible and need human interpretation

This work will continue to be developed as part of the ENVRI-FAIR project and in association with ongoing policy development work in the other EU-funded projects, e.g. FAIRsFAIR.

5 Appendix

5.1 Interview information package shared with each interview target (including questions)

ENVRI RI DATA POLICY INTERVIEW

Background and introduction

What is requested from you?

We wish to discuss with you (and any other RI experts you seem important) a short (max 1h) web interview on the current situation of your policies (see below) related to data issues. This is intended to be non-official discussion, and to probe the current situation regarding the different ways this kind of policies and best practices are currently handled in the ENVRI RIs.

We do not expect that all these policies do necessarily exist yet in all (or for some, any!) RI. Thus there is no reason not to answer if you cannot find such policy in your RI.

Process

We will agree on a Zoom (or other platform) meeting, and send you the questions (below) once more before the meeting. During the meeting the interviewer and your representatives will discuss each question, based on the different aspects and organisational situation, plans and developments regarding each policy. During the meeting the responses are on-line added to the common document. The document is sent to you for fact checking after the meeting. This document is (along any e.g. links

[7] By 2020-07-01 there is still missing answers from EMSO, ACTRIS and EPOS

or other shared documents from your end) used to prepare the landscape document for policies in ENVRI-FAIR.

Why is this necessary?

The FAIR principles, and by extension, EOSC, will require some level of interoperability of the services provided. Other parts of the ENVRI-FAIR will work on creating the needed technological developments, but the technology is not the only issue in interoperability. As, or even more important is that the human and organisational layer of the RI data services are compatible enough for ease of access and use. This requires some level of policy harmonisation, and policy creation, as well as suggesting model policies for RIs to use for different aspects of the service provision.

Before suggesting any kinds of changes in the actual RI policies and practices, we need some sort of understanding of the current policies, standards and internal procedures of the RIs. There is no point of suggesting a policy framework which is strongly against the established practices, at least without a strong reason for such change. Thus we need your help on finding out what is already decided, what is practical, and what would you think is realistic.

What we want this survey to inform us:
What is the current state of existing policies
Who can do such decisions in your RI? Which level of the organisation should agree on this (is this “European” or “Central facility” or “national” or “RPO” decision - is it made by executive decision, or by general assembly?)
What is the willingness and realistic expectations we could expect from the ENVRI RIs.

What do we mean by ”policy”?

Term “policy” is used to convey a set of rules, procedures and best practices used in the Research Infrastructure to describe the expected operations. The actual definition is a little more involved, but in this survey, we are trying to find out

a) what are actually currently available as internal rules and
b) what is considered to be possible to do within the few years.

For the purpose of this survey, we can consider the following spectrum of “policies”

Written official policies, such as data policy documents, agreed within the proper RI decision methods; No*te these are NOT always called “policies”!*
Unwritten practices, which are used as “best practices” in the organisation - used in practice, even not available as actual decisions or rules;
General “principles” which do not necessarily yet exist in actuality (e.g. decision on using PIDs for data, but not implemented yet)

If something is not yet available or decided yet, it is also useful knowledge, as are potential plans for policies.

Are there any additional aspects we are interested in?

Existence (or not) of a policy is interesting, but there are many aspects which (if the policy exists) could be interesting for this survey. These aspects are then considered for each question.

Task definition Are the persons (or organisational roles) related to this policy well characterised? This means in practice, there is actually person(s) in the RI who have a defined task to complete these actions.

Authorized Is this policy/practice officially accepted?

Machine actionable Is this policy/practice communicated via Machine-to-machine interface? (This is not really expected yet for almost any of the policies)

Monitoring Is the policy/practice followed in the RI effectively? I.e. do the KPIs or quality control mechanisms exist?

Universality Are there policies universally applied, or on a variable basis? As an example, “free and open access to data” might only be considered for some of the data sets. Or that e.g. data set definition varies within the RI, or that only some data is intended to have a PID assigned.

Language In which language are the policies available?

Availability Are the policies documented and available? For whom? Are they catalogued?

External dependencies Are the policies referring to outside (non-RI controlled) source? This means in practice things like vocabularies, standards, other community (living or changing) documents?
What next?

The data is used to draft the deliverable “policy landscape analysis” in WP4. The personal data involved in these questions is not stored, and as the policy landscape document is ONLY used to prepare realistic policy framework options. Your personal commentaries during the meeting will not be included unless you specifically give consent for them. All interview data outside of the landscape deliverable is stored securely for the duration of the project only accessible to the Policy Working Group active members, and will be destroyed at the end of the ENVRI-FAIR project. GDPR issues are considered appropriately.

Background documents

D5.1 for current technical situation - this is provided with the interview
FAIR principles and their interpretations at https://iagos-comm.iek.fz-juelich.de/dmsf/files/4111/view (for project partners)

Policy questions

Licence policy Do you have chosen a licence for your data? If so, which? Are multiple licences applied?

1.1. Is the licence machine readable?

1.2. Is the licence also applicable for your metadata? If not, what licence is used for your metadata?

Dataset definition Do you have a definition of what is a “dataset” in your RI? This refers to both scientific (i.e. variables involved) and practical (temporal, experiment, etc) separation of each dataset? (*data granularity*)

2.1. Do you define a data version? What are the policies/practices (if any) regarding versioning?

Persistent identifiers Do your datasets have PID? Are there exceptions?

3.1 Which PID(s) do you use? Are they internal or external? NOTE: This has been already asked in D5.1, but here the intent is to know if there is an actual policy regarding this choice.

Metadata Does your data have included metadata? (exceptions ?)

4.1 Does your RI metadata follow some standard(s)? NOTE: Asked also in D5.1

4.2. Does your RI use controlled vocabularies for metadata?

4.2.1. Are these vocabularies controlled by your RI?

4.3 Does your RI have a policy/practice for quality control of metadata?

4.4 Do you share your metadata (i.e. give access to it to outside searches, or share copies for some external service; Is the metadata openly accessible)?

4.5. Do you have a policy on authorship of the data (i.e. which roles are included as authors?)

Retention Do you have a policy for retention of data and metadata?
5.1. Do you have a described data deletion / reduction process?
5.2. Do you have a policy/plan for data/metadata availability in the long-run (e.g. closure of the RI)?

Data Access Do you have an access policy for the data?

6.1. What are the access formalities for data(or different kinds of data)?
6.2. Is there a formal process for accessing restricted data (if relevant)?
6.3. Do you allow external authorization (i.e. registration via another trusted source)?

6.4. Do you require users e.g. citation or reference to your data if used? If so, do you provide a clear definition of how this is done?
6.5. Is the access policy available in machine readable form?

Ownership and rights Do you have agreement with your data providers (if relevant) on the ownership and licencing of data?

Service availability Does your data service have a service level agreement (or other way to define uptime goals, etc)?

9.1. Is there a policy for keeping data and metadata constantly available?

Policy availability Generally, do you have the RI policies/practices published in a findable and accessible way? Do they have PIDs?

5.2 Interview results of contributing RIs

Table 1. Raw answers to the question *Licence policy. Do you have chosen a licence for your data? If so, which? Are multiple licences applied?* [*]

IAGOS AISBL	*CC4.0 (subtype unknown) is intended to be used universally, and consortium has agreed to use CC, but the details are still pending, (in English, available website) BUT,* ***currently a ad-hoc licence which requires co-authorship*** *(in the data policy)*
ICOS ERIC Carbon Portal	*Yes,* ***CC-BY 4.0. Official policy***, *universal* *Except for the raw data (no licence yet), property of the data providers*
EISCAT_3D	*"Rules of the road" are available for the data download.* ***Internal licence***. *Official, and practically agreed.*
EURO-ARGO ERIC	*Open and free of data policy.* *IOC resolution XX-6 Official policy, Agreed on the member state level. Data is free to use but acknowledgement is needed* *See* ***IOC resolution XX-6***: *http://argo.jcommops.org/IOC_Resolution_XX-6.html* ***CC-BY license***: *https://creativecommons.org/licenses/by/4.0/*
SIOS	***Complex***. *SIOS documentation* ***recommends CC4.0BY***, *SIOS depends on existing data which can have different licenses. recommendation of any dataset created from SIOS data to have CC4.0BY, but* ***no enforcement***.
eLTER	*No [generally agreed one], there are several. Open as possible, no concrete license has been agreed.* ***Draft licence was developed***. *Generally* ***releases have used CC0 & CC-BY***. *Central data products will have separate eLTER licence (intention CC0).* *This will be revised in the current PPP.*
LifeWatch ERIC	***Not at moment***. *But will be selected from the open source licences.* ***One of the CC licences*** *[most likely]. Many of the countries participating in LifeWatch have their own policies: Most are CC compliant (CC0 or CC-BY). Some cases include* ***embargo*** *(e.g. campaigns etc).*
DISSCo	*No* ***explicit choice yet***. ***Community norm on CC licences***. *Openness & interoperability. "As open as possible, as closed as legally necessary"* *Majority CC* (***often CC-BY***). *MoU required commitment on openness.* ***Software licences***: *Recommendation permissive licence* (***MIT, apache***, *etc),* ***not e.g. GPL*** *(although some components might have such licences).*
AnaEE	***Recommended licence CC-BY-SA***, *but can accept wide range of open data licence. Some platforms require e.g. Italian open data licence. AnaEE makes sure that all these national licences are allowed. Best practice: if not compatible licence with CC-BY, then use double-licences.* *Clearly stated in the DMP.*
DANUBIUS	*No licence yet, no decision yet. Discussions (CC mentioned).*

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 2. Raw answers to the question *Is the licence machine readable?* [*]

IAGOS AISBL	included in the metadata, not really Machine Readable
ICOS ERIC Carbon Portal	Not yet, working on it. Own ontology does not have it, DOI metadata has.
EISCAT_3D	No
EURO-ARGO ERIC	Yes (CC-BY mentioned in the metadata) Argo (2020). Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). SEANOE. https://doi.org/10.17882/42182 The licence is CC-BY https://creativecommons.org/licenses/by/4.0/
SIOS	SIOS harvests from existing data repositories. New information model can transfer licence data to users M2M1, but not implemented yet. If the data is not there from the existing data centre, this will of course not work.
eLTER	Only by reference to CC templates (not all would have it) Metadata includes data policy reference (but can very heterogeneous, M2M depends on metadata, e.g. data policy reference)
LifeWatch ERIC	Once selected, it will be machine readable, national level exists already some licences which are already machine readable.
DISSCo	DISCCo service catalogue has not been yet selected. Specimen records have machine readable licences associated with the data. This is the vast majority of DISSCo data assets. Some additional datasets are not yet on this form. Machine readability is the the goal, and the community is well behind this.
AnaEE	Yes, CC-BY-SA machine readable. For national licences, not all are machine readable, but then there is a link to the licence description.
DANUBIUS	N/A

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 3. Raw answers to the question *Is the license also applicable for your metadata? If not, what license is used for your metadata?* *

IAGOS AISBL	**No licence** for metadata
ICOS ERIC Carbon Portal	Most likely. Data in the files has the licence of the data. Includes the author information & affiliation **Other metadata is CC0** **Not an official policy** at the moment.
EISCAT_3D	**No official, but in practice open metadata**
EURO-ARGO ERIC	Each netcdf contains data & metadata - **metadata is open and free**
SIOS	Not specifically. **Metadata is thought as open, but licence not really considered yet**. WMO has discussions on metadata openness and licensing. For data which has metadata included (NETCDF, etc) the metadata has the same licence as the data.
eLTER	Open in practice **Data policy defines that metadata is freely and openly available**.
LifeWatch ERIC	Not yet (in the ERIC level). **National hub level metadata is free** (even as a policy in some countries, agreed when submitting data).
DISSCo	**Not likely to have different licence** of the metadata, as strongly connected to the data, they should have similar licence. Blurry boundary between the two.
AnaEE	Yes, **official DMP policy** mentions that the **metadata has the same licence as data**.
DANUBIUS	N/A, same direction of discussion as above [DANUBIUS answer to 1](CC).

* The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 4. Raw answers to the question *Dataset definition. Do you have a definition of what is a dataset in your RI? This refers to both scientific (i.e. variables involved) and practical (temporal, experiment, etc.) separation of each dataset? (data granularity)?* [*]

IAGOS AISBL	**NO written definition**. Data levels are defined. List of datasets exists (some are subsets of others). Practical: Variety internal practical standards for different purposes. Documented in the data portal. included in the metadata. Will be answered in the DMP (first version probably end of 2020)
ICOS ERIC Carbon Portal	Yes, Data objects and data collection, both have a PID. **Well defined in documentation and ontology.** "ICOS Data " well documented. But there are pre-ICOS data & project data. which does not follow the same standards.
EISCAT_3D	Yes, **definition exists: Experiment. Practical agreement**
EURO-ARGO ERIC	By platform, WMO number, trajectory, metadata. **One float time series is a dataset.** Euro-Argo is developing other PID solutions for higher resolutions (Cycle and Variable) to facilitate traceability (activity within ENVRI-FAIR Task Force 3)
SIOS	**No strict definition, but it is discussed**. Data policy has a definition of dataset (weak). Discussion on "user oriented" definition of dataset instead of "data producer oriented", Data set*: A SIOS data set is a discrete collection of scientific data, museum objects or samples which can be described by metadata, compliant to the SIOS data policy, and related to scientific efforts in and around Svalbard within the SIOS framework.* https://sios-svalbard.org/sites/sios-svalbard.org/files/common/SIOS_Data_Policy.pdf
eLTER	**Formal data set does not exist**. **Practical solutions are being developed** in the PPP. Likely particular set of measurements in temporal / spatial space.
LifeWatch ERIC	**Not formally decided, but depends on the nature of the data collected**. Multidimensional, space, variables, taxonomies, data type (real time, etc.). National hub level there are some definitions, and on treatments, etc. Partially solved in the national level, and will be used to prepare the LifeWatch level policy.
DISSCo	Yes, primary digital asset (**digital specimen**). Publication connected. Datasets will be varied, but the primary assets are the specimens. There will be **other products with varying dataset definitions in future**. Policy or policy type document. Exchange standard on this. Well defined.
AnaEE	**Dataset definition exists, in the DMP**: "*Datasets are the central objects of AnaEE Data Management Policy: they are self-contained sets of information that include data and all the metadata required to re-use that data. Examples of dataset are: a database of historical and georeferenced observations, an archive of field observation, an archive of laboratory analysis results, a set of field or laboratory images, a set of system logs. A project dataset is composed by the "Foreground Data" and the "Background Data" over the project period."* Self-contained set e.g. on one experiment one dataset. Full replicability on the data (incl. all background data included).
DANUBIUS	**Heterogeneous datasets, from many sources. observation data, monitoring, remote sensing etc. Each will require their own decision.** Data types (**being developed**): • By Collection method • By structure • By format

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 5. Raw answers to the question *Do you define a data version? What are the policies/practices (if any) regarding versioning?* [*]

IAGOS AISBL	**Versioning procedure is documented. The PID stays constant even if the version different**. Provenance system being developed, which will improve the versioning and documentation of it. This will also include more detailed PID system to track data versions.
ICOS ERIC Carbon Portal	**Levels exists**, (Raw, L1, L2). Full versioning for all data. Landing page has link to other versions. It is machine readable. Gap filling. **Reprocessing can lead to new versions, mentioned in the metadata**. Thematic centres responsible.
EISCAT_3D	**Two levels of data** (preliminary and final), both are published (**preliminary will be replaced by final**).
EURO-ARGO ERIC	**Monthly snapshots of the whole datacenter with each have their DOI**. Previous Monthly version can be returned with. History of the processing step available in the Netcdf Files in the History section
SIOS	Encourage data versioning, but no **policy or enforcement**. No single way of expressing this to users. Recommend to data centres a new identifier for new. versions, but not always followed.
eLTER	**Not centrally.** Individual national networks have their own practices. Versioning will be developed for the eLTER level products.
LifeWatch ERIC	As above. Some fields crucial for the LifeWatch. Identifying the best practices from the national level hubs. **LifeBlock blockchain will be used to help the traceability and versioning and origin of the data**. General assembly agreed the prototype, and encouraged to implement. (also collaborating with DISSCO and GBIF, probably others, e.g. ICOS). The technology chosen will guarantee implementation of a policy.
DISSCo	**Digital specimen design includes versioning policies** (including ways to follow earlier snapshots). **PID points to latest versions**. Some providers are versioning aggregated datasets. **Not a single unified policy**, but well advanced in many data providers. This will be unified in the DISSCo. Central requirement for us. Some do this via DOIs.
AnaEE	Yes, **two policies**: Datasets from **experiments** (spot datasets, foreground datasets) **should not require versions**, but there is a way to have dataset revision. Revision must be re-started for each new version (including the metadata). Exceptional case, but **there is process**. **For continuous observations: There is a continuous revision cycle**. **New PID for new version**. All versions stay in the system (Git). Metadata is also versioned.
DANUBIUS	Data levels are being discussed, **will be finalised the DMP**. Versioning will be also discussed in the DMP, no solutions implemented yet.

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 6. Raw answers to the question *Do your datasets have PID? Are there exceptions?* [*]

IAGOS AISBL	**All [datasets] have a PID**. Data center AERIS provides them. Implicit in the AERIS contract that they are responsible. Datacite DOIs. (agreement with Datacite) For citation, but not for management.
ICOS ERIC Carbon Portal	**Yes**, **official policy**. Carbon portal assigns. No exceptions
EISCAT_3D	**No. No PID** system is used, but there are internal identifiers.
EURO-ARGO ERIC	**Yes, At least at the level of a float [observation platform] with the WMO number that is unique** For the whole dataset the PiD is : http://doi.org/10.17882/42182
SIOS	Depends, every dataset should have a **Unique Identifiers, UUIDs (sometimes internal), some are using DOIs**. There are many identifiers as well. Increasing awareness for DOIs. SIOS is **recommending DOIs**, but the the documentation is being developed at the moment. Not likely to end up with a single PID system
eLTER	**Not a policy**. Preference to have a DOI (Eudat), unless there are local solutions. **Not a centralised system** (yet). Discussion on having data cite DOIs for at least eLTER level products. PIDs are being developed for the individual site definitions.
LifeWatch ERIC	**Will be implemented** by following GEDE-RDA biodiversity WG recommendations (together with DISSCO, etc.). **Not agreed policy on the PIDs** yet. Not a policy, but a practice and a principle
DISSCo	Well understood in the community **- actively discussed** for a long time. Intent in DISSCo is to unify the current practices. Specimen level PID vs. data set PIDs. Internal PIDs might not all be globally resolvable. Some secondary data/metadata might need more work on PIDs. DISSCo works in the global level, which requires discussion outside of Europe.
AnaEE	**DMP mentions each datasets must have a DOI**. AnaEE data centre will assign on in the dataset revision system.
DANUBIUS	**Data policy (officially approved) requires use of DOI for the datasets**. This reflects to future (DANUBIUS) data, not historical data. DMP will have more on the PIDs for traceability & security perspective. The organisation assigning the DOI/PIDs is not yet decided, but most likely will be the DANUBIUS data portal (but a challenge). Data centre should be in control of the data ingest and publication in the approved workflow principles.

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 7. Raw answers to the question *Which PID(s) do you use? Are they internal or external?* [*]

IAGOS AISBL	**Usually DOIs (Datacite**)- soon EPIC (for internal PIDs, still planned but very likely)
ICOS ERIC Carbon Portal	**External (Datacite), DOIs**. (mainly collections) PIDs (handle) for all (including raw) data Carbon portal concept paper defines.
EISCAT_3D	**internal**
EURO-ARGO ERIC	**DOI (datacite), WMO number (World Meteorological Organization) - both external** Decided at international level. Locally controlled at European level.
SIOS	See above [recommending DOIs]
eLTER	**Heterogeneous** (depends on centre). Typically **internal for the catalogue**. **External DOIs for the data**. Ambition to PIDs to semantics for linking parts of the RI together (people, models, etc.)
LifeWatch ERIC	Now used **PIDs which are compliant with communities** (different ones in marine (GBIF/OBIS) vs. molecular data). We are implementing test cases in invasive marine species. **Internal references to LifeBlock** to follow the data (see q.2.1). External PIDs you are capability for **Datacite DOIs**.
DISSCo	**DataCite** for datasets used in the community Specimen & collection level still requires harmonisation. FAIR digital object architecture & EOSC are the background policies.
AnaEE	**DOIs** (provider being negotiated) Internal pointers, URLs, which try to be as persistent as possible.
DANUBIUS	**DOI** (externally)

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 8. Raw answers to the question *Does your data have included metadata?* [*]

IAGOS AISBL	**Yes, policy status unknown but likely**. In practice, but will be included in the DMP for policy. Responsible: Data centre adds the metadata.
ICOS ERIC Carbon Portal	Yes. L0 metadata from PI, L1 metadata from Thematic centres, L2 from carbon portal. **Official policy**. Metadata profile for each data object type + ontology information. Data policy is most likely being updated (raw data) in some time relatively soon.
EISCAT_3D	**Metadata is included for all data.** Mainly automatic, some manual adjustments needed (e.g. authorship), done by the EISCAT data centre.
EURO-ARGO ERIC	**NetCDF has included metadata**. National data centres create the metadata, transferred on the International Argo Data Central portals (For security reasons there are 2 central portals one in USA one in France that synchronise their data holding every hour)
SIOS	Depends on the data. Guidelines are being developed. **Ambition is to harmonise the SIOS core data**, but the generally there are many data sources outside of it, with differing metadata. SIOS will only harvest the metadata provided by the sources, and transform it to SIOS metadata model (only for the discovery metadata).
eLTER	**Yes, there is a policy**. At least the central catalogue provided data will need to have metadata. Data is provided both centrally and on the local level. Landing pages of the DOIs are not controlled by the eLTER, and not part of the catalogue. B2SHARE includes LTER template, with reference to more descriptive metadata recording the eLTER catalogue.
LifeWatch ERIC	**Yes, official policy**. Task force on metadata. Annotated metadata. Corresponding catalogues. (1) Definition and implementation of good practices associated to metadata, controlled vocabularies, taxonomies and semantics in general terms.
DISSCo	**Yes, official policy** and key requirement for operations. Originates from the national/institutional level, required to follow standards (see below)
AnaEE	**Yes, part of DMP**. No exceptions.
DANUBIUS	**Yes this is required for new (DANUBIUS) data**

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 9. Raw answers to the question *Does your RI's metadata follow some standard?*[*]

IAGOS AISBL	**ISO standards** (19115), also **WMO standards** being implemented. no DMP at the moment (in preparation). Metadata provided will depend on the user group (e.g. WMO will get WMO compliant MD).
ICOS ERIC Carbon Portal	**W3C**, **own standard** mapped to required metadata standards.
EISCAT_3D	**Madrigal standard** (for the data which is published there).
EURO-ARGO ERIC	**Yes**, all developments are shared in the ENVRI-FAIR. The metadata are documented in "Argo User's manual" http://dx.doi.org/10.13155/29825 This document is published on Ocean Best Practices https://www.oceanbestpractices.org/
SIOS	**Recommendation** **CF convention**, ACDD - for NETCDF Darwin core archive for biodiversity data (lots of other data as well) **INSPIRE** directive. Data policy refers to European legislation. Same in the data management plan.
eLTER	Based on the ISO, internal standard is an evolution. Local metadata model, wide range of information. This is mapped to **ISO19115** (**INSPIRE**), **EML**. Metadata elements were combined from the both. Intention for harvesting the standards. There will a revision of the metadata model in the PPP, but will be still available in many flavours in the API. Community profiles will define the core set, and will be refined in the current PPP.
LifeWatch ERIC	**Yes. Best practices for metadata** (2) Collecting, Analysing and Opening-up existing standards, and then proposing a set of own standards which may be agreed with different ENVRI Communities-of-Practice (CoPs). **EML**, **Darwin core**, **INSPIRE**, ...
DISSCo	**Darwin Core** **ABCD** (access to biological collection data) **ABCDEFG** (extension for geology) Other geographical, etc. additional standards Well a developed practice in the field. Work to do: minimum information for collections & specimen standard, formalisation being done in DISSCo
AnaEE	**ISO metadata** (**INSPIRE**) - expected principle Platforms is given a broad range to fulfil the metadata. But in the end of revision cycle, the internal & external interoperability is assessed. Some criteria more harder than others.
DANUBIUS	There is a principle of **metadata model, but not yet implemented**. **INSPIRE**. Heterogeneous sources. ISO standards **19115**, **EML** standards. **OpenAire** standards. A final set of potential candidates have been prepared. Also discussed with representatives with other ENVRI RIs.

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 10. Raw answers to the question *Does your RI use controlled vocabularies for metadata?*[*]

IAGOS AISBL	**CF convention for metadata**, **GCMD** (global common... NASA). Practical AERIS decision. Can change in future.
ICOS ERIC Carbon Portal	Yes, own controlled vocabularies (ontology is one), **not follow ISO11759**(?). Content agreed internally with thematic centres. Netcdf files follow **CF conventions** (not INSPIRE actually)
EISCAT_3D	Yes, **Madrigal**
EURO-ARGO ERIC	Yes, **own vocabularies** (Argo-specific, but links with existing vocabularies, using existing if possible) They contain **subsets of CF standard names**. **SeaDataNet vocabularies** as well used.
SIOS	Discovery metadata: **OSGEO URL purpose** **CF standard** names **GBIF** keywords For use metadata, depends from the data source (usually NO).
eLTER	Yes. Recommendation to use **ENVTHES**, but no enforcement.
LifeWatch ERIC	**Yes**, semantics, **general terms**.
DISSCo	Yes. **Multiple are required in the standards definitions**. Describing the changing names of species.
AnaEE	**AnaEE thesaurus -** more detailed and tracks provenance and categorises the data Core presentation metadata - required minimum
DANUBIUS	Not selected yet, but **DMP will require some set of common vocabularies**.

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 11. Raw answers to the question *Are these vocabularies controlled by your RI?* [*]

IAGOS AISBL	CF **no**, GCMD no
ICOS ERIC Carbon Portal	**Own one is.**
EISCAT_3D	**EISCAT is a part of the joint venture which is responsible of the vocabulary**
EURO-ARGO ERIC	The **vocabularies are controlled by Euro-Argo RI and Argo data** management team. The CF and SeaDataNet subsets of vocabularies are controlled by these 2 authorities.
SIOS	**Principle use existing vocabularies, and recommends changes there.**
eLTER	**ENVTHES controls the terms used in the central catalogue**. Describes the parameters, but there are number of other concepts will need extensions. The ontology would then pick the relevant ones. http://vocabs.lter-europe.net/edg/tbl/EnvThes.editor#http%3A%2F%2Fvocabs.lter-europe.net%2FEnvThes%2F10000
LifeWatch ERIC	**Developed some, other ones are connected**. Collecting semantic resources from the from the **Ecoportal**.
DISSCo	Most are connected on the standards (also being developed by DISSCo members). Some might be needed for additional rich metadata inclusions. **Use of existing if possible.**
AnaEE	**Yes**
DANUBIUS	**not yet selected**. Metadata templates will be findable (inc. vocabularies).

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 12. Raw answers to the question 6. *3 Does your RI have a policy/practice for quality control of metadata?* [*]

IAGOS AISBL	**No for the metadata**
ICOS ERIC Carbon Portal	Levels progression & labelling has some level of QC for observation data. **Consistency is checked automatically in the ingestion**. Metadata concerning the stations regularly reviewed between HO and stations for the ICOS Handbook.
EISCAT_3D	**Not really**
EURO-ARGO ERIC	**Yes, the international data centre does metadata QC**. Official agreement
SIOS	**Discovery metadata: Not a policy, but in practice there is a quality check,** (strict), title abstract, temporal duration, station.. etc. **Use metadata: no check (currently)**. CF compliance (subset) required for some services.
eLTER	Main QC is from the campaigns, for integration in the DEIMS. There is a legacy of metadata not being QC. Catalogue still has such metadata. **DEIMS only checks completeness**. No content quality check. DEIMS is the historic catalogue (not only eLTER). **The subset of eLTER might need to be separately considered**. Policies would then be different for the eLTER only products than to the whole DEIMS. **Policy being considered**. Long term vision being considered for the eLTER level. Site descriptions is more a practice. There are update cycles, data centre will check this metadata QC. "As inclusive as possible", policies must reflect this.
LifeWatch ERIC	This metadata task force has a **periodical review control quarterly**. Lifewatch has **benchmarking process** for external sources and internal processes. Official practice (will be policy when agreed in the General Assembly) - one of the WGs is working on the metadata. based on the collected best practices.
DISSCo	Standards conformity will be most **probably done in the DISSCo level**. For **some partner organisations** (GBIF, etc) **have this already implemented**. Content quality are being considered in the community projects, which could be additional services provided by DISSCo. This is uncertain and will require discussion with the data providers. Also access to original (meta)data is going to be needed. These are community practices, but **not yet on the policy level**.
AnaEE	**Yes revision cycle is defined in the DMP**. (Peer review of data/metadata, supported by automated tools).
DANUBIUS	**Yes, there is an approved ingest process**. Metadata is required and complete. There will be a metadata **QC from the data provider**, verified in the ingest process (on/off), i.e. second level QM in the data centre to make sure they match the DANUBIUS requirements.

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 13. Raw answers to the question *Do you share your metadata?* *

IAGOS AISBL	Yes, openly available. No licence yet. Metadata can be harvested. **Catalogue openly available**. **Policy status not clear**, but could be handled in DMP
ICOS ERIC Carbon Portal	Handbook is shared for the stations. Stations also have external metadata shared via e.g. WMO GAW, Fluxnet, etc. This leads to harmonisation in field. **Observation data metadata is openly harvestable**. Linked open data, SPARQL interface.
EISCAT_3D	Have been part of projects, **not operational at the moment**. Could be in the future.
EURO-ARGO ERIC	**Openly available via central node**.
SIOS	**Not from the SIOS data centre**. In theory yes, but the performance now is not usable. But **soon available**. From the SIOS point of view, it is available as a practice (**not policy**). Not re-sharing things without appropriate identifiers and without the originator data centres' permission.
eLTER	Yes, **available and shared** (e.g. GEOSS). Decision made on this on LTER Europe level. https://deims.org/pycsw/csw (CSW GetCapabilities Link) https://deims.org/pycsw/csw.py?mode=oaipmh&verb=Identify (OAI-PMH OpenSearch site) INSPIRE requirements are not fulfilled via eLTER, but via their own catalogue. **API interface** (https://deims.org/api) https://deims.org/geoserver/ows?service=wms&version=1.3.0&request=GetCapabilities (WMS 1.3.0 GetCapabilities Link) https://deims.org/geoserver/ows?service=wfs&version=2.0.0&request=GetCapabilities (WFS 2.0.0 GetCapabilities Link) to access the information on the sites
LifeWatch ERIC	Yes, they **will be available** by following an open access policy by default, but implementing **proper embargo periods** (few cases for e.g. endangered species). Practice, not policy
DISSCo	**Part of the expectation of DISSCo, Machine & human accessible**. There can be some of the metadata is **not available from the institution level for security etc. reasons**. (example: rhino horns). Basic premise is that the partners share their collection information: some are not mentioned e.g. from legal etc. reasons. There is no community unified exception list, but could be a future DISSCo activity.
AnaEE	The **metadata is accessible**, trying to create as much as **possible machine readable**. Core metadata is well machine actionable. The metadata is open to external searches.
DANUBIUS	**Proposed in the DMP to be available to everyone (no decision yet).** Metadata catalogue in the data portal.

* The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 14. Raw answers to the question *Do you have a policy on authorship of the data?* *

IAGOS AISBL	**Yes, metadata has PIs included** (all RI connected personnel - especially 2 first are important). Mapping between and ISO and Datacite. AERIS is included as an author. DMP will fix this in a formal way.
ICOS ERIC Carbon Portal	**There is a policy on this**. **PI** (complicated) + **Sponsor** (organisation). No indication for RI curation work. Roles are defined, roles define he citation string.
EISCAT_3D	"**EISCAT" is the author**. "PI" is Ingemar [tech. director] (almost all data)
EURO-ARGO ERIC	Different levels, **Authors are included from PI level to data centre personnel.** Who-has-been-doing-what. Institution vocabulary. Mentioned in the user manual (official document). The Argo dataset DOI list its 250 contributors (preferably with their ORCID), see https://doi.org/10.17882/42182
SIOS	Comes from the data centres. **Depends on the contributing data centre**. Current data model has changed the way to theoretically connect persons to data sets. No contributions from the data centres yet though. Information model has room for data centre role (not much used).
eLTER	**No policy, but the standard would be DataCite data authorship template**. But not yet decided for the eLTER level data. For the cite provided data is defined by the data providers. Some recommendations exists from the earlier projects.
LifeWatch ERIC	Lifewatch is not a exclusively data providing RI, but a mostly modelling and an analysis and annotated RI. **New data is created also from the Lifewatch modelling, with author LifeWatch**. Authorship is mentioned in the followed metadata standards. **LifeBlock will be able to follow the original data and the authors in it (provenance information).** Practice, not a policy.
DISSCo	Authorship chain. political level : institution. **Data originator are available as well in the metadata**. Attribution information is acknowledged problem. **No real community solution yet**, extremely complex and heterogeneous situation. Work in progress. Promoting use of ORCIDs, especially for collectors. Provenance is a large part of the digital object model, which could be a major part of such process.
AnaEE	**Complicated**. Revision process requires owner & curators. These are coming from the research platforms.
DANUBIUS	Datasets must have identifiers, which need to **include information on the authors**. **No policy on the RI level** (but data provider will provide). Data should include information on the data originator. But there is no decision yet.

* The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 15. Raw answers to the question *Do you have a policy for retention of data and metadata?* [*]

IAGOS AISBL	**No**. Perhaps part of the DMP
ICOS ERIC Carbon Portal	**No**, **sustainability is a value**, projects are offered sustainable storage of the data. **Data policy requires** to **store the data for the period of ICOS**, permanent repository afterwards. B2SAFE for security reasons.
EISCAT_3D	**Raw data is destroyed** (but store as much as resources permit). Data products are **intended to be kept indefinitely. This official policy**.
EURO-ARGO ERIC	**Everything is kept as much as possible**, as much technical data is included as possible. European level all data could be re-processed. **Practice, not policy.** **ARGO data management level includes long term preservation in US NCEI. Long term archives**. Data distribution and preservation are separated. https://www.euro-argo.eu/Activities/Data-Management/Argo-Data-System
SIOS	Interoperability requires OAI-PMS. OAI-PMH has support for incremental harvest and identification of deleted datasets. For CSW we have to wipe the catalogue on a regular basis. Metadata daily harvested. **Data deletion is not supported by many data centres. Full cleanup periodically. Flagging the metadata links if dead.** Still learning from the harvesting process and **trying to set up a policy**. Most of the data are scientific,, not deleted, but could be superseded. Some model output etc is dead after while. **Practice, not a policy**.
eLTER	**No policy on eLTER level products**. Not a written policy, practice to **keep at least the metadata** (as the data storage is not always in the central repository).
LifeWatch ERIC	**Not a policy. But nothing is ever thrown away**. However, proper agreements with data provider institutions are being agreed in order to guarantee long-term data preservation (LTDP). In fact, one of our outstanding references is the e-IRG working document on LTDP. In the particular case where LW ERIC is the original data provider, proper retention policies will be applied to (meta-)data.
DISSCo	**No a policy, but community of practice: specimen record is retained for perpetuity.** Some secondary external digital data might be problematic (e.g. size, raw data etc), especially outside connections. DataCite also requires this.
AnaEE	**Yes, all is kept part of the DMP (stated)** Big capacity for the storage needs.
DANUBIUS	**Discussion on this, but foreseen to be a long term (25-30yr).** Should be mapped at least as long and all data/metadata should be kept. No decision made.

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 16. Raw answers to the question *Do you have a described data deletion / reduction process?* *

IAGOS AISBL	See above.
ICOS ERIC Carbon Portal	No, nothing is ever thrown away. Official policy.
EISCAT_3D	**Yes, for raw data.**
EURO-ARGO ERIC	No, the dataset is continuously expanded
SIOS	see above
eLTER	No policy on this yet
LifeWatch ERIC	Not in place.
DISSCo	**No, but it could happen and would be necessary (e.g. geo information for species collection). Has not yet been considered in the DISSCO context, but could be an important aspect.**
AnaEE	**No, except if the data provider (PI) requests data removal (= GDPR compliance)**
DANUBIUS	No policy on deleting data

* The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 17. Raw answers to the question *Do you have a policy/plan for data/metadata availability in the long-run (e.g. closure of the RI)?* *

IAGOS AISBL	**Coretrustseal certification** -> this will be fixed as a policy during 2021. AERIS will be responsible of long term storage & availability (contractual)
ICOS ERIC Carbon Portal	**Yes, there is official exit plan (part of the statutes)**
EISCAT_3D	**There is a policy - stakeholders should take over the data handling.**
EURO-ARGO ERIC	**Mentioned above on long term preservation**
SIOS	**Data management plan suggest that the observation data will be available at least 10 years from the SIOS access point** (in the case that SIOS is closed etc). Partner institutions have longer time sustainability and mandate.
eLTER	**Data is locally held / EUDAT etc. Metadata is unknown at the moment.** Partner organisations take care of the metadata at the moment.
LifeWatch ERIC	**Handled by the statutes**. All the assets (inc. data) go back to the countries responsible (for **holding in perpetuity**)
DISSCo	Providers provide data in the national level, but central level aggregation would be lost. Some of the technical choices are chosen to avoid lock in to specific implementations. **Upcoming development in the future.**
AnaEE	**External** party making **long term storage solution**. (CNRS CCIN2P3 infrastructure in Lyon, France)
DANUBIUS	**Not yet.**

* The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 18. Raw answers to the question *Do you have an access policy for the data?* *

IAGOS AISBL	**Written procedure** during the database registration.
ICOS ERIC Carbon Portal	Yes, data is certifiably FAIR. This is also the ICOS mission (**statutes**). **All ICOS data is available for free. Anonymously (registration optional).** **Raw data on request**. (under discussion, different approaches internal)
EISCAT_3D	**"Rules of the road" - including embargo times. Officially agreed.**
EURO-ARGO ERIC	**Everyone can access, commercial, anonymous**. **Official ARGO policy**. No right to do statistics of individual users.
SIOS	**Free and open, some services require registration (resource limitation)**
eLTER	**For the LTER data products needs to be defined** (in PPP). For the datasets provided by the national sites, there are heterogeneous processes. Some all free with citation, some require more. No registration for some datasets, some require more authorisation. For the openly shared data no registration, anonymous access.
LifeWatch ERIC	**AAI + accounting and accountability. Identifying each user for each data.** Follow up of the users and the data they access. A policy which requires registration and confirmation and access. Authorisation of each user. National level also limitations on the volume of the data.
DISSCo	**Everything is based on openness for digital content. No barrier exists**. There might be practical reasons (e.g. large data) requiring email address. **No actual gatekeeping.** At least in practice level now. Coordinated physical access is handled separately in the institutions, and being integrated. SYNTHESIS access policy development.
AnaEE	**Anyone can access. Registration is optional. APIs require registration (**resource limitation). (data under revision is confidential only to actors, curators, reviewers) Official in DMP.
DANUBIUS	**Authorisation is required** (EDUGAIN). User strategy document requires registration for monitoring and feedback reasons. There is an access policy draft.

* The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 19. Raw answers to the question *What are the access formalities for data (or different kinds of data)?* *

IAGOS AISBL	**a PI will evaluate the application for access**. The access is then provided by the database manager. The process is described in the website transparently (not machine readably). Agreed practice, but not actual policy of IAGOS (but could be in a GA notes etc.). Could be also answered in DMP
ICOS ERIC Carbon Portal	**End results are anonymously freely available**. (CC4.0 BY) Raw data licence is not clear yet, under discussion. **Raw data access is unknown** (contact PI currently). Metadata available of course.
EISCAT_3D	**Openly available, anonymous**, but should follow the "rules of the road"
EURO-ARGO ERIC	**No formalities** (just go on the website).
SIOS	**Data no formalities, for services: the registration is up to SIOS data centre (to determine real users). Currently no authorisation to data.**
eLTER	**Vary**
LifeWatch ERIC	**Apply for access** (depending the data type and volume) .Lifewatch decides access.
DISSCo	**None**
AnaEE	**Most data no formalities.** Only for reviewing, API development versions etc, one needs to apply.
DANUBIUS	**No review process is yet considered**

* The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 20. Raw answers to the question *Is there a formal process for accessing restricted data (if relevant)?* [*]

IAGOS AISBL	n/a
ICOS ERIC Carbon Portal	Yes (informal), raw data is asked to contact the PI. (ICOS does not have the licence for raw data)
EISCAT_3D	Raw is restricted. Policy is being considered.
EURO-ARGO ERIC	no restricted data, even non-standardised data must be shared. As part of the IOC resolution XX-6 countries can ask to stop data transmission of the Central portal when a float is collecting information in their EEZ. This has never happened in past 20 years .
SIOS	No restricted data (restricted data is not in the catalogue), information model could handle it (including restricted metadata), not implemented yet.
eLTER	Multiple ways.
LifeWatch ERIC	Not yet, a practice.
DISSCo	These should be handled in the national/institutional level, details to be developed.
AnaEE	No restricted data (outside revision cycle).
DANUBIUS	There will be some restricted data (user categories, resource limited data products). Embargoed data may be in the data portal (for a limited data). The process have not yet been created.

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 21. Raw answers to the question *Do you allow external authorization (i.e. registration via another trusted source)?* *

IAGOS AISBL	**Not right now, but soon will be able to use ORCID & Edugain access is allowed (technical level might need clarification)**
ICOS ERIC Carbon Portal	**ORCID, Edugain, (facebook)** **CarbonPortal, OBSpack via NOAA**
EISCAT_3D	For the low level data this is **being considered**.
EURO-ARGO ERIC	**No**
SIOS	Not needed as now, but **has been considered**. (ENVRI-FAIR solutions has not been concrete enough yet).
eLTER	No. But **probably considered in the future**
LifeWatch ERIC	Not yet, but **under development** (in ENVRI-FAIR etc)
DISSCo	Some services need authorisation for workflow needs, e.g. for requiring physical access, or digitalisation on-demand. **SYNTHESIS is building ORCID/EDUGAIN based system.**
AnaEE	**Federated access across the AnaEE** platforms, Plans on federation in ENVRI. Infrastructure already there.
DANUBIUS	**Edugain is considered to be used.**

* The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 22. Raw answers to the question *Do you require users e.g. citation or reference to your data if used? If so, do you provide a clear definition of how this is done?* [*]

IAGOS AISBL	Registration requires approval of data protocol, including the citation requirements and **co-authorship is required for major data use**. **Official policy**
ICOS ERIC Carbon Portal	**This is part of the data policy. Instructions are given directly when downloading. M2M also includes a step for confirming the acknowledgement requirements**. One can also set this on personal settings (not asking after that). DOI has URL to CC-BY
EISCAT_3D	**Acknowledgement is needed. Example is given.**
EURO-ARGO ERIC	**Required, but no way of imposing**. **Example citation provided in the metadata** (machine readable) and in the DOI.
SIOS	**Recommended, but not really a policy**. New data model should include necessary data This is **not visible in the portal**, but "from data management plan: Users of data supplied through SIOS shall acknowledge in any publication or any other derived work, the contribution made by those who have created and worked up the data. If the data licence does not specify how best to do this, data should be formally cited using the citation text provided on the dataset's landing page or in its metadata. Those who retrieve data through SIOS shall acknowledge SIOS as follows: Contains data retrieved through SIOS (year). " (probably changing very soon)
eLTER	**For some the products (which have DOIs), this is provided. Not true for a lot of other datasets**. DEIMS datasets have a recommended citation, for each dataset separately.
LifeWatch ERIC	**Depends on the licence which the data is available. CC-BY require a reference.** This is **practice** that follows from the national data providers policies. Datasets with DOI should be quoted with it. Not yet communicated to the users in a single way.
DISSCo	**Work in progress.**
AnaEE	**Licences CC-BY require this. Definition in the metadata.**
DANUBIUS	**Citation will be required, but the details have not yet been decided. Recommendations are being drafted.**

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 23. Raw answers to the question *Is the access policy available in machine readable form?* [*]

IAGOS AISBL	No
ICOS ERIC Carbon Portal	**CC4.0 BY mentioned in the metadata, included in the landing page.**
EISCAT_3D	(not machine readable)-
EURO-ARGO ERIC	yes, see above
SIOS	Not really.
eLTER	No
LifeWatch ERIC	Not yet. (sometimes licences are in the metadata). Most of the data provided with the CC licence.
DISSCo	It is possible but not yet implemented.
AnaEE	Metadata includes must include at least one citable item as reference.
DANUBIUS	N/A

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 24. Raw answers to the question *Do you have agreement with your data providers (if relevant) on the ownership and licencing of data?* *

IAGOS AISBL	Ownership and rights are defined in **the AISBL statutes**. The IAGOS has the licensing rights of this data from there on (with the approved scheme).
ICOS ERIC Carbon Portal	**Yes, official policy,** by contracts with national networks and thematic centres.
EISCAT_3D	**EISCAT owns the data.**
EURO-ARGO ERIC	Network requires the data becomes "ARGO data", ownership is given to ARGO. Implicit, and **based on IOC resolution**.
SIOS	**SIOS has no licence for the data**. All is owned by the data providers. In the DMP or data policy section on ownership.
eLTER	**No** **(when there is a legal entity this will be considered)**
LifeWatch ERIC	**No ownership is claimed for data**. **We are in the process of defining those licensing agreements**. And both aspects will be addressed following best practices applied by European RIs, compulsorily following ENVRI ones.
DISSCo	**DISSCo consortium will have agreements** between the nodes and the hub.
AnaEE	**Platform or PI owns the data**. Each platform should provide description of their licensing policy. Expected to be compatible, but if not, there is **ongoing discussions to harmonise** the policies.
DANUBIUS	**Discussion on this**. Ownership is not yet decided, but most likely preserved on the data providers.

* The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 25. Raw answers to the question *Does your data service have a service level agreement (or other way to define uptime goals, etc)?* [*]

IAGOS AISBL	**Right now no**, but probably certification process will clarify.
ICOS ERIC Carbon Portal	Probably related to FAIR certification. **The current situation does not have guaranteed uptimes,** etc, but in practice 99,5%
EISCAT_3D	**No**
EURO-ARGO ERIC	**two global data central synchronised data centres, at least one is guaranteed to work**. **Monitoring of the data centres is done, and statistics are followed annually**, 99.x% availability. Independent monitoring from IOC JCOMMOPS center (http://www.jcommops.org/board?t=argo) .
SIOS	**No formal agreement**, but in reality 98.5 availability for SIOS services
eLTER	**For the DEIMS best effort basis**. No defined agreements at the moment.
LifeWatch ERIC	**LW ERIC is defining those SLAs** with our providers to be able to define aggregated SLAs to our (LW ERIC) final clients . There will be, mainly based on computed aggregation of SLAs from our providers, as duly applied by any large-scale (distributed) RI.
DISSCo	**Not yet developed**, but planned.
AnaEE	**Based on large commercial cloud provider, with related SLAs**, uptimes, disaster management, security, etc.
DANUBIUS	**24/7 is intended**. So decision on the actual service KPIs etc. But there are internal KPIs for individual service use and users satisfaction.

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 26. Raw answers to the question *Is there a policy for keeping data and metadata constantly available?* [*]

IAGOS AISBL	**Should be, but no decision yet.**
ICOS ERIC Carbon Portal	Yes, carbon portal, raw data straight from the field to CP, and always available from the CP. Probably related to FAIR certification
EISCAT_3D	**No**
EURO-ARGO ERIC	**yes**
SIOS	**It is a practice, but not a policy**. A science system, but not operational system. But running on the same infra as in the operational system.
eLTER	**Not yet.**
LifeWatch ERIC	**Not a policy on this yet.**
DISSCo	**Yes, at least in practice**, being developed.
AnaEE	If a platform as a long term storage, they can use. **These can not be necessarily guaranteed by AnaEE**, but are guaranteed by them [member platforms]. Cloud data/metadata guaranteed by AnaEE.
DANUBIUS	**Is agreed on the overall level**.

[*] The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.

Table 27. Raw answers to the question *Generally, do you have the RI policies/practices published in a findable and accessible way? Do they have PIDs?* *

IAGOS AISBL	Yes, DOIs for products. **The documents in the website are not such that would have PIDs.** Data policy does not have a PID. Standard operating procedures are available. There is no general rule or policy to have the policies available in the website.
ICOS ERIC Carbon Portal	**In general this is the goal**, management plan is work in progress. Most important things available from the website (some have DOIs). **Own repository (openly available, findable).**
EISCAT_3D	**No**
EURO-ARGO ERIC	**Each significant document has a DOI, are published in Ocean Best Practices(https://www.oceanbestpractices.org/).**
SIOS	**Not yet**. Not SIOS DMP and Data policy are in the website. Interoperability guidelines in GitHub. https://sios-svalbard.org/Documents
eLTER	Deliverables are available in the website for the projects done. Governance data management deliverable are taken as a starting point for the PPP - this **could come as a policy** for the eLTER.
LifeWatch ERIC	**Publish every decision transparently in website**. There is the internal PID. This is a requirement in the LifeWatch ERIC statutes.
DISSCo	**Working on the DISSCo knowledge base**, will come alive in few months. Will have PID.
AnaEE	**Yes, but not yet approved**. SEISM application this far used internally - this is supposed to publish all these when AnaEE is approved.
DANUBIUS	**Should be available on the portal** (when it comes).

* The bolding and comments with square brackets are by the deliverable team to help with analysis below, and NOT part of the original interview document.