D5.1
REQUIREMENT ANALYSIS, TECHNOLOGY REVIEW AND GAP ANALYSIS OF ENVIRONMENTAL RIs

Work Package	WP5
Lead partner	ICOS ERIC
Status	Draft
Deliverable type	Report
Dissemination level	Public
Due date	31-12-2019
Submission date	28-02-2020

Deliverable abstract

The overarching goal of ENVRI-FAIR is for all participating ENVRIs to improve their FAIRness and prepare the connection of their data repositories and services to the European Open Science Cloud (EOSC). With the development of FAIR implementations from the participating RIs and integrated services among the environmental subdomains, these data and services will be brought together at a higher level (for the entire cluster), providing more efficient services for researchers and policy makers.

This deliverable introduces the FAIR principles, describes the approach chosen for the FAIRness assessment, gives insights into the assessment results at the project/subdomain level (and for each RI in the protected project-internal Redmine environment) and discusses the requirements for achieving FAIRer data and services. It provides a summary of the identified gaps by the subdomains and gives an overview of the development plans. It further describes the current plan for the next steps in the project for this task.

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	Barbara Magagna	EAA	19-11-2019 16-12-2019 03-02-2020
Contributing Authors	Angeliki Adamaki Xiaofeng Liao Riccardo Rabissoni Zhiming Zhao	ULUND UvA INGV UvA	03-02-2020 10-10-2019 10-10-2019 16-12-2019
Reviewer(s)	Markus Stocker Peter Thijsse	TIB MARIS	06-02-2020 17-02-2020
Approver	Alex Vermeulen Andreas Petzold	ULUND FZJ	27-02-2020

DELIVERY LOG

Issue	Date	Comment	Author
V 0.1	19-11-2019	Approach and set-up of Deliverable 5.1	Barbara Magagna
V 1.0	16-12-2019	First Draft	Barbara Magagna
V 2.0	03-02-2020	Second Draft	Barbara Magagna, Angeliki Adamaki
V 2.0	06-02-2020	Comments from Reviewer 1	Markus Stocker
V 2.0	17-02-2020	Comments from Reviewer 2	Peter Thijsse
V 3.0	19-02-2020	Additions	Angeliki Adamaki
V 4.0	26.02.2020	Final version	Barbara Magagna Angeliki Adamaki
V 5.0	27-02-2020	Comments from Approver 1	Alex Vermeulen
V 5.1	28-02-2020	Editing, Template format	Angeliki Adamaki

OCUMENT 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 1. 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 that 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

REQUIREMENT ANALYSIS, TECHNOLOGY REVIEW AND GAP ANALYSIS OF ENVRIs

1 Introduction

This document summarizes the results and the progress made by the Environmental Research Infrastructure (ENVRI) cluster during the first year of the project. Starting in January 2019, Work Package 5 (WP5) of the ENVRI-FAIR project has been working closely with the other WPs to coordinate the required steps towards FAIRness among all involved RIs. The process started with an assessment of the FAIRness status in each RI, to identify their strengths as well as the gaps between their current status and what is expected based on the FAIR principles (see chapter 2 for an overview of the principles). The steps and concept of the applied methodology to perform the FAIRness assessment of the involved RIs is described in chapter 2, while the results of this process for the four subdomains of the ENVRI cluster (Atmosphere, Marine, Solid Earth, Ecosystem) are presented in chapter 3. Having this information in detail gives the opportunity to all participants to benefit from the existing technology and knowledge from other experts in the environmental cluster and plan their next steps in accordance with the subdomain (and as a result the cluster). A summary of the identified gaps within the framework of fulfilling the FAIR and EOSC requirements is given in chapter 4, together with a synopsis of the reported implementation plans at subdomain level and the current plan to introduce new thematic groups (joint task forces) which will be cross-cutting the four subdomains and will contribute to the harmonisation of the common solutions in the cluster. Finally, in chapter 5, the potential future of the FAIRness assessment of the ENVRIs and some improved techniques are discussed (as next steps for the FAIRness assessment).

The requirement collection and analysis approach

The present document (hereafter D5.1) provides an up-to-date and integrated analysis of the most common gaps that the ENVRIs have identified and need to bridge to meet the FAIR requirements during the development cycle (for data, metadata and services; see also chapter 4 where the requirements are further discussed). The starting point for the evaluation process was provided by maturity self-assessments prepared by the RI communities during the proposal preparation phase.

The basic structure of the requirement collection and analysis can be summarised in the following steps:

1. Guided self-assessment of the FAIRness level by means of a questionnaire
2. Harmonised analysis of the gaps identified in each RI
3. Harmonisation in a common plan for each subdomain, to derive the first set of requirements
4. Newly identified requirements tracked during the project and eventually included in the common development actions (e.g., via the joint task forces, use cases and other development activities).

To effectively coordinate the FAIRness assessment in all four subdomains and the overarching requirement analysis in WP5 and WP7, the Task T5.1 team has:

1. actively checked the latest progress from relevant initiatives, e.g., GO FAIR, RDA and EOSC,

2. defined a FAIR questionnaire together with the GO FAIR Convergence Matrix team[1] and provided customized templates for WP8-11 (i.e. the subdomain WPs) to perform their FAIRness assessment,
3. actively contributed to the workshops organized by the subdomain WPs to support the FAIRness assessment process,
4. reviewed the short-term development plans and prioritised actions proposed by each RI within their subdomain, harmonised the analysis within the subdomains and provided the cluster view on the FAIRness gaps and development plan. Figure 1 provides a graphic representation of the basic approach.

The focus of D5.1 lies on the FAIRness assessment and its results at the cluster, subdomain and RI level, providing also a summary of the identified gaps. The output of D5.1 also indicates important components for the Tasks T7.1 and T7.3, aiming to better plan the support that will be provided for the involved subdomains with e.g. common development activities.

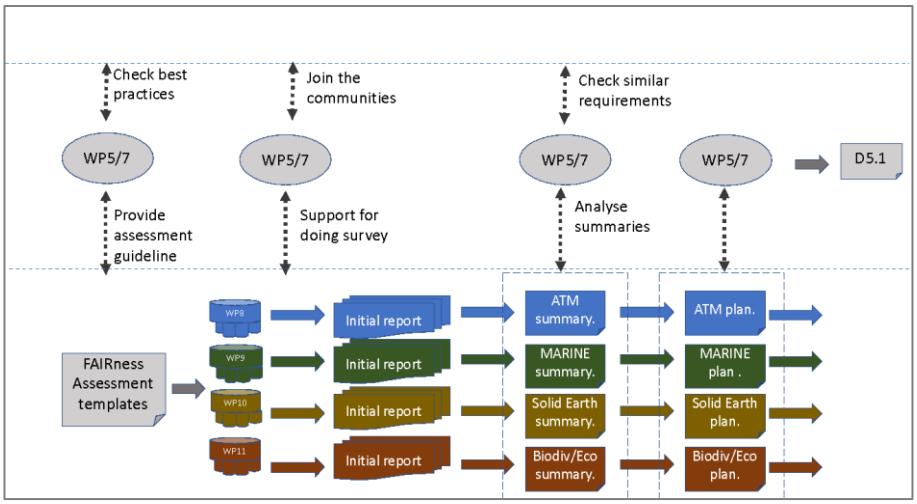

Figure 1: The first phase of the requirement analysis approach, focusing on the FAIRness assessment and resulting in D5.1.

[1] https://www.go-fair.org/today/FAIR-matrix

2 FAIR Assessment: An overview and the ENVRI approach

2.1 FAIR principles overview

The FAIR acronym and concept stands for: "Data and services that should be Findable, Accessible, Interoperable, and Re-usable, both for machines and for people". The FAIR principles have been published in 2016 [1] but the term FAIR was already conceived at the Lorentz conference in 2014 by the FORCE11 Group[2]. Yet much earlier, in 2007, some of these ideas were addressed in OECD's document 'Principles and Guidelines for Access to Research Data from Public Funding'[3] and later in 2013 in G8 Science Ministers' statement[4], saying that research data should be easily discoverable, accessible, intelligible, usable and if possible interoperable. These criteria were included (in the same year) in the data guidelines for the EU Horizon 2020 framework programme, and then picked up by the FORCE11 Group.

The Principles provide guidance on a general level expressing the kind of behaviour that researchers should expect from contemporary data resources. They describe aspirations for systems and services to support the creation of valuable research outputs and enable their reuse. Table 1 lists all 15 Principles. More details for each of the Principles follow in section 3, together with the FAIRness analysis for the ENVRIs.

The FAIR Guiding Principles article [1] had a remarkable resonance and stimulated broad adoption. On the other hand, because the paper did not specify how the FAIR principles should manifest in reality, there is space for diverging interpretations inducing partially incompatible implementations.

Some of the original authors of the FAIR principles intentionally clarified [2] ambiguities around the Principles to avoid further misinterpretations. The FAIR Principles should not be conceived as standards, which is per se restrictive, but only as guidelines with a permissive nature. Although the original paper underscores the machine-actionability of data and metadata, the Principles don't prescribe the use of RDF or linked data. While semantic technologies are currently a good solution to fulfil this requirement, other potentially more efficient approaches may appear in the future.

FAIR compliant data and services should be primarily machine actionable and on top of that also facilitate humans to find, assess and reuse data (and not vice versa). The time spent by researchers with 'data munging' (finding and reformatting data) should be reduced as much as possible by enabling computers to take over these tasks. FAIR should also not be considered as equal to open or free, because there might be good reasons (personal privacy, national security, etc.) to restrict access to data and services, even when generated with public funding. The 'A' in FAIR addresses only the need to describe clearly and transparently a process for accessing discovered data, which includes the presence of a machine-readable license.

There is also some uncertainty on how to assess the FAIRness level of digital objects. This has led to many different initiatives to design diverse evaluation tools to assess either qualitatively or quantitatively how far the principles are met. Some of the most representative methodologies are described in the following section.

---

[2] FORCE11 grow out of the FORC (Future of Research Communication) Workshop held in Dagstuhl, Germany in 2011

[3] https://doi.org/10.1787/9789264034020-en-fr

[4] https://www.gov.uk/government/news/g8-science-ministers-statement

Table 1: The FAIR guiding Principles [1]

To be Findable: F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource
To be Accessible: A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available
To be Interoperable: I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data
To be Reusable: R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards

2.2 FAIR assessment approaches

This section presents the different approaches (as tools or initiatives in the form of Working Groups) that were considered during this first phase of the ENVRI-FAIR project, with some comments on their strengths and weaknesses that explain the eventual choice made by the ENVRI subdomains. A good overview of different approaches is also given by FAIRsharing[5].

2.2.1 Quantitative approaches

FAIRmetrics.org is a group collaborating with a broad set of stakeholders to design a framework of FAIRness indicators for machines that can be objectively measured in a semi-automated process. Based on the assumption that it would be difficult for humans to perform the assessment objectively, the framework should allow to measure the degree to which a digital resource is findable, accessible, interoperable, and reusable without human intervention. The goal was to develop at least one metric for each of the FAIR sub principles

[5] https://fairassist.org/

that would be universally applicable to all digital resources in all scholarly domains [3]. In addition, FAIRness should be measurable, making the assessment in an objective, quantitative, machine-interpretable, scalable and reproducible manner. To ensure transparency, a template for creating metrics was developed. Each metric was represented by one or more questions, which in many cases requests from the respondent to provide, if available, a URL to a specific digital object, which would provide evidence for compliance to the metric in question. This approach is referred to as "generation 1" questionnaire-style Maturity indicator tests.

Further developments led to the design of a framework for the automated evaluation of metrics, the so-called "generation 2" automatable **FAIR Maturity Indicator (FMI)** tests. They are conceived as self-describing and programmatically executable web-interfaces using the smartAPI specification[6]. The execution of such a test returns a binary pass/fail result. Because the test tracks every action, the reasons for failure/success are transparently documented and thus helpful for improvement. Communities can decide which Maturity Indicators are relevant to them and create their own tests according to their specific requirements. The Indicator tests should not be interpreted as 'judgements' but rather as means to evaluate objectively if a resource successfully fulfills the FAIRness requirements which the community has established[7].

2.2.2 Qualitative approaches

The main aim of a qualitative approach is to increase awareness about the need for FAIRness.

The DANS FAIRdat assessment tool[8]

is an online prototype tool which guides the user through a set of questions to assess a specific dataset. Although this seems to be a properly documented and user friendly tool, the questionnaire is oversimplified. Some of the FAIRness requirements are not explicitly considered (e.g. Reusability). There are other issues that are not sufficiently addressed with this method (e.g. whether open data should score higher than closed data), some metrics are not clearly defined, while the final "FAIRness score" seems to be affected by subjectivity during the assessment of some of the FAIR principles (as reported by DANS[9]). To improve the self-assessment process, DANS proposed the "FAIR enough? checklist", described below.

The DANS FAIR enough? checklist[10]

is an assessment technique referring to the quality of a dataset and the trustworthiness of the repository (to which the dataset will be deposited). Thus, it covers 4 levels for each set of principles,

1. The data repository which is planned to be used
2. The metadata with which the dataset (to be deposited) is described
3. The dataset
4. The data files which consist the dataset

The checklist with 11 questions brings out two aspects of the FAIRness assessment, the FAIR data itself and the trustworthy repository. Taking the CoreTrustSeal (hereafter CTS) data

---

[6] https://smart-api.info/
[7] https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/
[8] Questionnaire: https://www.surveymonkey.com/r/fairdat, Specification Document: https://docs.google.com/document/d/1bRQDN_VFSPSMnsADLyzky-sbd6ZPArsHOcYhERdyrL8/edit?pli=1#
[9] More information: https://planeurope.files.wordpress.com/2018/11/08-peter-doorn-fair-enough-plan-e-workshop-ieee-doorn-amsterdam-october-2018.pdf
[10] The checklist can be accessed through this link: https://docs.google.com/forms/d/e/1FAIpQLSf7t1Z9IOBoj5GgWqik8KnhtH3B819Ch6lD5KuAz7yn0I0Opw/viewform

repository certification (see section 4.2.2) as an example, the repositories with such a certification are considered to be compliant to the FAIR principles to a big extent. Still, also this checklist is considered to be too generic and doesn't cover all FAIR sub principles.

As noted by DANS, the FAIR principles might make no claim to which level of granularity they pertain (repository, collection, data set, file, record, triple, etc). However, they often mention "(meta)data", which in the current analysis is interpreted as pertaining both to data and metadata. Specifically, for data in trustworthy (as the certified CTS; see 4.2.2) repositories, most of the FAIR principles are followed, for all data and metadata existing in the repository. It is also worth mentioning that the principles (see Table 1) F2, I2, I3, R1 (R1.2 and R1.3) can vary for the metadata in a CTS certified repository, while the I1, I2, I3, R1 (R1.3) can vary for data (sets and files) in a CTS certified repository.

The CSIRO 5-star data rating tool[11]

allows users to evaluate the FAIRness of their data based on the 4 sets of the FAIR principles, adding one more quality, namely whether the data is Trusted. For each of the 5 qualities, the corresponding questions allow the assessment tool users to rate (1-5 stars) their data according to its current state. The results include the final rating of the data (subject to assessment) for each of the 5 qualities.

Inspired by the CSIRO data rating tool and the FAIRdat tool, the Australian Research Data Commons (ARDC) developed a FAIR self-assessment tool, described next.

The ARDC FAIR self-assessment tool[12]

is an initiative of the Australian research community that aims to "build coherent national and collaborative research data commons". To contribute to the data management within their community, ARDC developed a self-assessment tool, initially designed for data librarians and IT staff to assess the 'FAIRness' of a dataset. The tool also gives tips to users (during the assessment process) which might help enhancing the FAIRness of the tested dataset.

The FAIR Data Maturity Model WG[13]

is a Research Data Alliance (RDA) initiative that develops as an RDA Recommendation a common set of core assessment criteria for FAIRness and a generic and expandable self-assessment model for measuring the maturity level of a dataset. Moreover, the WG will design a self-assessment toolset to improve the readiness and FAIR implementation level of datasets. The goal is to increase the coherence and interoperability of existing or emerging FAIR assessment frameworks.

FORCE11 FAIR Data Management Plans (DMP)

is an initiative of FORCE11, the international community/platform that hosted the open consultation for the definition of the 15 FAIR guiding principles in 2016 (see 2.1). FORCE11 has established the "FAIR DMPs" Working Group aiming to provide a simple set of principles, along with examples of domain-specific implementations and recommendations for best practices, that emphasize good data management, stewardship and machine-readability for making data FAIR.

RDA SHARC IG[14]

stands for the SHAring Rewards and Credit interest group of the RDA which focuses on the crediting and rewarding mechanisms in the sharing process of data and resources. Working with two assessment grids, scientists can first identify their data/services (mentioned as activities) compliance to the FAIR principles. Second, evaluators can assess in 2 levels (simplified/extensive) the sharing practice (which is being evaluated) over a period of time, using essential, recommended and/or desirable criteria and considering other factors, e.g. the

[11] https://research.csiro.au/oznome/tools/oznome-5-star-data/
[12] https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/
[13] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg
[14] https://www.rd-alliance.org/groups/sharing-rewards-and-credit-sharc-ig

available means and support. The goal of this method is to simplify the assessment grid focusing on essential criteria (for each set of the FAIR principles) which can be comprehended by the data providers. As noted by the SHARC IG, this grid cannot be used alone if the goal is a comprehensive assessment of the level of FAIRness regarding the sharing practices but can provide an initial evaluation.

RDA/FORCE11 FAIRsharing Working Group

is an initiative now called the 'FAIR sharing registry: connecting data policies, standards & databases WG'. It is a use cases-driven joint effort between RDA and FORCE11 to develop:

- a set of recommendations to guide users and producers of databases and content standards to select and describe them, or recommend them in data policies, and
- a curated registry[15], which enacts the recommendations and assists a variety of end users, providing well described, interlinked, and cross-searchable records on content standards, databases and data policies.

The GO FAIR Convergence Matrix

is an approach introduced by GO FAIR, the community-led and self-governed initiative that aims to coordinate the coherent development of the Internet of FAIR Data & Services. With the objective to accelerate broad community convergence on FAIR implementation options, the GO FAIR community launched in June 2019 the development of the FAIR Convergence Matrix[16]. Members of WP5 participate actively in this group. The matrix is conceived as a platform that compiles (for any community of practice) an inventory of their technology implementations needed to comply with the FAIR principles. This inventory is conducted by help of a **Research Database Management** (**RDM**) based questionnaire, developed by Peter Wittenburg and Kristina Hettne. The questionnaire comprises 53 questions around the used repositories, data, metadata and vocabularies, and closes with a self-evaluation on the FAIR compliance of the data. This approach is also discussed in the following section.

2.2.3 The ENVRI-FAIR assessment approach

A key task of the ENVRI-FAIR project is to assess and to monitor the FAIR maturity of participant RIs over the project lifetime. This assessment enables the identification of current gaps and thus informs the Information and Communications Technology (ICT) experts of the project to enable a customized consultation and support to RIs.

For the common FAIRness analysis methodology, it was decided at the ENVRI-FAIR Kick-Off meeting[17] in January 2019 that MARIS would seek cooperation with the GO FAIR initiative, as they develop analytical methods for assessing FAIRness of data and services that seemed best fit-for-purpose. As a follow-up it was agreed that the ENVRI-FAIR community would adopt the GO FAIR analysis tools and benefit from the experience already gained by GO FAIR with their "Implementation Networks", while GO FAIR would benefit and learn from the additional analysis activities that ENVRI-FAIR would (and will) undertake as an additional Implementation Network. The GO FAIR methodology includes completing survey questionnaires (see 2.2.3.1) that GO FAIR has developed for gathering information on the FAIRness level of an infrastructure.

2.2.3.1 The questionnaires

The first survey used the **RDM questionnaire** (see 2.2.2 for more on the GO FAIR initiative) of the FAIR Convergence Matrix with 53 questions.

This questionnaire had several purposes for the ENVRI-FAIR community, as to:

---

[15] https://fairsharing.org/
[16] https://www.go-fair.org/today/FAIR-matrix/
[17] D1.1 Organization of project Kickoff meeting, including a Steering Committee and a General Assembly meeting: https://envri.eu/wp-content/uploads/2019/10/ENVRI-FAIR_D_1-1_Organization-of-project-Kick-off-meeting.pdf

- Increase our understanding of the FAIR principles and their advantages for the RIs
- Assess the state of RI data and services in terms of FAIR requirements
- Compile a technology landscape of the RIs
- Detect information and implementation gaps
- Discover strengths
- Compare implementations by different RIs
- Evaluate possible technology take-ups for improvements
- Prioritize FAIR improvements
- Include chosen FAIR improvements in RI plans

The GO FAIR team provided a spreadsheet of these questions with explanations and example answers. In addition, references to the FAIR principles where appropriate were linked to the questions. In addition to this approach, it was decided to use the **FAIR Maturity Indicator** (hereafter **FMI**) 'generation 1' **questionnaire** (see also 2.2.1) with 25 questions. The purpose was to assess in a semi-automatic way the compliance with FAIR by the presence or absence of specific requested resources.

The surveys were distributed through the leads of the subdomains (WP8-11 leads) to representatives of the participating RIs. The survey was conducted in the period between March and May 2019 using Google Forms. All responses from the RIs were collected in a Google spreadsheet.

2.2.3.2 Information collection

The first analysis (by mid of April 2019) immediately revealed that the responses to the questionnaires by the RI representatives were not directly usable for downstream analysis without substantial post-processing and harmonization of the responses. The FMI questionnaire was not sufficiently understood, and many questions were not answered at all. Although a few RIs returned some URIs, most of them did not point to the requested resources. The FMI questionnaire was designed to test FAIRness indicators for machines, which relies on a high level of technical expertise from the respondent to provide the appropriate information. The FAIRmetrics.org group (see section 2.2.1) used the questionnaire "generation 1" as an input for the automated evaluation service which gets along without human intervention. The WP5/7 team decided to use this service in the future to get more accurate results.

The majority of questions from the RDM questionnaire were well understood and answered by the RI representatives. Nevertheless, some were wrongly interpreted and thus not answered sufficiently. In chapter 3 of this document, these ambiguous questions are highlighted and discussed. The questionnaire allows free text answers which cannot easily be used for comparison. The main problem of free text questionnaires is the one-to-many cardinality of certain questions when more than one answer is allowed. The follow up questions could relate to more than one resource described before, thus it becomes indistinct which resource is described further down.

At that stage of the project, the need to review the whole procedure and re-design the survey became evident. Ambiguous questions (e.g. where two questions were embedded in one) were re-structured into clear questions with only one specific meaning. Moreover, the two questionnaires (RDM and FMI) were merged into one, excluding the FMI questions which were not understood by the majority of the participants. This resulted in a questionnaire with a total of 78 questions, ordered in a new and more logical way. To resolve the multiple cardinality problem, it was requested that for each repository a new questionnaire would be filled. This led to a redistribution of the new questionnaire (hereafter **RDM+**; the questionnaire can be found in Appendix 2) to those RIs that had not yet provided their answers.

For the RIs that had already delivered the filled forms, it became evident that their answers had to be reviewed and post-processed to extract the key information needed, or to mark answers as insufficient (NULL) where applicable. This could only be done in close collaboration between the WP5 experts and the responsible RI representatives. The format chosen to enable such needed interaction was a 1-2 days face-to-face workshop, for each subdomain, followed by videoconferences when required (see Figure 2). The aim of these workshops was to achieve complete and quality-controlled answers, and a better understanding of the FAIR concepts.

In order to support the face-to-face interviews with the RI representatives to resolve the outstanding issues in the questionnaire responses (in post-processing), the answers (which were collected in spreadsheets-XLS as mentioned above) were converted, and the extracted key information was transformed into a structured form in YAML (Yet Another Markup Language) format, following a template also written in YAML. This format was chosen for its conciseness and readability as well as for the fact that it requires minimal extra information to encode answers. The sequence of the YAML attributes is aligned with the questions in RDM+. While making this conversion the answers were translated as much as possible from free text to reference lists (same label for same concept/responses), and, if that was not possible, to condensed answers. The original responses in XLS were kept (and used again in later stage for reviewing purposes), as they contain additional information about planned implementations and FAIRness gaps (note here that these are not used for the automated repeatable FAIRness check). This work supported the efficient editing of the inadequate answers during the face-to-face interviews.

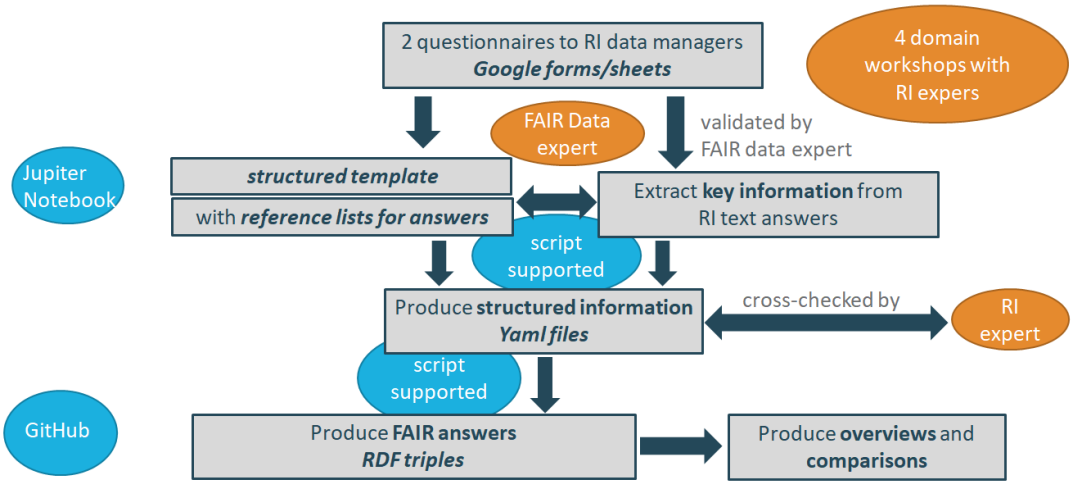

Figure 2. Approach from questionnaires to FAIRness overviews.

The structured YAML format of the information collected through questionnaire responses also served a second purpose. Compared to questionnaire answers in natural language text, the structured information is better suited as input to further processing of the information. Given the requirement to analyse the collected information efficiently, it was decided to build a database to include this information. Concretely, a knowledge base was implemented in the form of a triple store (see the below explanatory box) using RDF as the data model. Hence, as an additional step, the information in YAML was converted into RDF. The YAML documents were converted into an RDF document (data.trig file) using a fully automated script implemented in Python as a Jupyter notebook that can be executed on EGI Notebooks service.

A triple store or RDF store is a purpose-built database for the storage and retrieval of triples through SPARQL queries. A triple is a data entity composed of subject-predicate-object, e.g. "Bob is 35" or "Bob knows Fred". Much like a relational database, one stores information in a triple store and retrieves it via a query language. Unlike a relational database, a triple store is optimized for the storage and retrieval of triples. In addition to queries, triples can usually be imported/exported using Resource Description Framework (RDF) and other formats. RDF is a family of World Wide Web Consortium (W3C) specifications and is in use as a general method for conceptual description or modeling of information that is implemented in web resources. It is also used in knowledge management applications. SPARQL is a semantic query language for triple or RDF stores and facilitates retrieving and manipulating data stored in RDF. Information analysis can be conducted with Jupyter notebooks, which is an open-source web application that allows users to create and share documents that contain live code, equations, visualisations and narrative text.

2.2.3.3. Information Analysis

The resulting RDF documents have been loaded into a triple store. This enabled us to formulate declarative queries in SPARQL that implement user requests for information, not only about individual RIs but also across RIs. SPARQL queries have also been formulated in Jupyter. The results can be stored as Excel sheets and downloaded for further processing.

As explained above, the RI responses in XLS were converted into the structured YAML format. The basis for this process builds a template (see Appendix 3) and a related list of references that are allowed to be used in certain elements.

The YAML template (later referred to as YaT) clearly shows the structure of the information: a generic part describing the RI details, followed by the attributes for each repository. This means that for one RI there can be information about several repositories in a single YAML file. The repository attributes are grouped in different sections, which are themselves organized as nested attributes:

- Identifier
- Access mechanisms
- Data
- Metadata
- Vocabularies
- Data management plans
- Data processing
- Data fairness self-evaluation

Table 2 gives an overview of the delivered questionnaires for the repositories per RI and also shows the subdomain they belong.

The questionnaires were returned by all 14 ENVRIs. Note here that 3 ENVRIs are present in more than one subdomain:

- ICOS: marine, ecosystem and atmosphere
- LifeWatch ERIC: marine and ecosystem
- SIOS: ecosystem and atmosphere.

As the overview shows (Table 2), at the time of the survey there were 34 repositories, 30 of which are characterised as "Existing" while 4 were at a planning stage (EMSO, Danubius, DiSSCo and AnaEE). Because this analysis can only consider existing repositories, the planned repositories are not included in chapter 5 (where the FAIRness analysis results are presented). Regarding the respective RIs, the EMSO, Danubius and DisSSCo are not considered in the present analysis, whereas AnaEE is included with one existing repository. Thus, the present study refers to **11 ENVRIs,** with a total of **30 repositories** representing the **4 subdomains**. Table 3 lists **all the included repositories: 12 for the Atmosphere, 6 for the Marine, 8 for the Solid Earth and 8 for the Ecosystem Subdomain**.

The results of the FAIRness assessment are summarised in chapter 3. Note here that if no answers were given, the FAIRness level could not be examined. This means that only the repositories for which qualified answers were provided were assessed. In the main text of the present document, the results are presented at cluster and subdomain level. For detailed information on each RI, we refer the reader to the supplementary material which includes all answers given (including 'planned' or 'partially') and are available in the protected project-internal Redmine environment.

Table 2: Overview of the described repositories per RI (as they have been reported at the time of the FAIRness assessment analysed in the present document). The repositories are grouped into the 4 subdomains of the ENVRI cluster.

	Subdomains				Repositories	
RI	Atmosphere	Marine	Solid Earth	Ecosystem	Existing	Planned
EPOS			x		8	
EMSO		x				1
SDN		x			2	
Euro-Argo		x			1	
LifeWatch ERIC	x	x			3	
ICOS	x	x		x	1	
ACTRIS	x				6	
EISCAT-3D	x				2	
IAGOS	x				1	
eLTER				x	3	
AnaEE				x	1	1
DANUBIUS				x		1
DiSSCo				x		1
SIOS	x			x	2	
Total					30	4

Table 3: List of the (30) ENVRI repositories examined during the analysis presented in the present document. The first column refers to the environmental subdomain, the second column names the participating RIs and the third column lists the respective (and assessed for their FAIRness) repositories.

Sub-domain	*RI*	*repository name*
Atmosphere	ACTRIS	ACTRIS - In-Situ unit
		ACTRIS-ACCESS
		ASC
		CLOUDNET
		EARLINET Database
		GRES
	SIOS	Norwegian Meteorological Institute
		Norwegian Polar Data Centre
	IAGOS	IAGOS repository
	EISCAT	EISCAT Schedule
		Madrigal
	ICOS	Carbon Portal
Marine	SDN	SeaDataNet Central Data Products
		SeaDataNet Common DAta Index (CDI)
	Euro-Argo	Euro-Argo Data
	LifeWatch (marine)	EUROBIS
		Marine Data Archive
	ICOS	Carbon Portal
Solid Earth	EPOS	EPOS CSW
		EPOS INGV
		European Federated Data Archive
		local EU-EIDA
		MySQL
		RESIF (France)
		Terradue
		VERCE Seismic Forward Modeling Experimental Data
Ecosystem	AnaEE	ANAEE-France Metadata Catalog
	eLTER	DEIMS-SDR
		eLTER CDN
		EUDAT/FZJ B2SHARE
	LifeWatch	LifeWatch Italy Portal
	SIOS	Norwegian Meteorological Institute
		Norwegian Polar Data Centre
	ICOS	Carbon Portal

3 FAIR Assessment Analysis

This chapter presents the results of the FAIRness assessment performed in the first year of the ENVRI-FAIR project. As explained in the previous chapter, the results here are analysed at cluster and subdomain level, without specific information on the participating RIs (this information is available in the form of supplementary material in the protected project-internal Redmine environment). The chapter is organised in sections (3.1 to 3.4) which include the analyses performed for each group of the FAIR principles (as listed in Table 1). Each section consists of subsections where the reader can find some explanatory information per principle ("Descriptions"), which is compiled using the GO FAIR descriptions and relevant literature. The questions asked to the RI representatives are also displayed in relation to each principle, and the results (i.e. the summarised information gathered from the questionnaires) are presented at cluster and/or subdomain level.

3.1 Findable

F1. (Meta)data are assigned a globally unique and persistent identifier

a) Description

The **uniqueness** of an identifier is a necessary condition to enable unambiguous reference to one specific resource. This is essential for identifying, retrieving, linking and citing datasets. A World Wide Web address (URL – Uniform Resource Locator) is used to specify the online location of a resource but over time URLs can change, which leads to broken links to the data. Therefore, identifiers must also be **persistent**. Persistent identifiers (PIDs) provide **a permanent citable reference** to the dataset **no matter where** it is located online (UUID are not PIDs!). Persistence means also that the identifier continues to identify the same resource, **even if that resource no longer exists**. Another requirement for a PID is (Web) **resolvability**, a mapping between the PID name onto a PID landing page (URL). This is often realized using an **independent third-party** to generate an identifier that has **guaranteed longevity**. It should be noted that GO FAIR refers to the identifier complying to F1 as Global Unique Persistent and Referable Identifier, GUPRI. For simplicity and consistency with the previous ENVRI project(s) we continue to use the acronym PID, but actually refer to GUPRIs.

b) Questions to the ENVRIs

In RDM+ there were 2 questions which are relevant to that principle:

Which identifier do you use?
Which PID registration provider do you use?

c) Results

Below, the given (from the participating RIs) answers are summarised in tables and grouped at two levels, i) for the cluster (all ENVRIs) and ii) per subdomain (i.e. Atmosphere, Marine, Solid Earth and Ecosystem).

i) Cluster level

As mentioned earlier in this section, 11 RIs with 30 repositories have been analysed. The results show that **23 identifier systems are in place** (which are listed in Table 4), **while 4 RIs use 2 different persistent identifier systems each**. The information on the identifiers used in the ENVRI cluster is gathered in Table 5, while the answers on the PID providers the RIs have reported are summarised in Table 6. Note here that Tables 5 and 6 also include the "NULL" answers.

Regarding the implemented identifier systems, 48% of the repositories are using unique, persistent identifiers. The rest is using local systems, UUIDs or there is no information provided to assess this properly.

Table 4: The status of the ENVRI cluster regarding the F1 Principle is summarised here. The first row groups the answers in terms of the identifiers used by the ENVRIs, and the second row defines the PID provider. *Note: the green colored answers are the ones which are FAIR compliant.*

ID kind	DOI	Handle PID		UUID	URI	local ID
PID provider	DataCite	ePIC service	B2Handle	Python lib.	NULL	local service
	12	2	3	4	1	1

Table 5: The reported identifiers used by the ENVRIs. *Note: the green colored answers are the ones which are FAIR compliant. The NULL answers are also shown.*

DOI	Handle PID	UUID	URI	local ID	NULL
12	5	4	1	1	12

Table 6: As in Table 5, for the PID registration provider.

DataCite	B2Handle	ePIC service	Python lib.	local service	NULL
12	3	2	4	1	13

ii) Subdomain level

Here, the answers to the questions related to F1 (see earlier paragraph b) are grouped per subdomain, and the results are summarised in Tables 7 and 8. Note again that a) the green colored columns correspond to the identifiers which are FAIR compliant, b) the total number of the given answers in Tables 7 and 8 is not equal to the number of repositories neither the number reported earlier for the cluster, as there are repositories which are represented in more than one subdomain, as well as repositories which use more than one identifier system.

Table 7: Similar to Table 5, here per subdomain.

	Identifier kind					
Subdomain	**DOI**	**Handle PID**	**UUID**	**URI**	**Local ID**	**Null**
Atmosphere	5	1	2	-	-	5
Marine	6	1	-	-	1	1
Solid Earth	3	1	-	1	-	4
Ecosystem	2	2	2	-	1	3

Table 8: Similar to Table 6, here per subdomain.

	PID provider					
Subdomain	**DataCite**	**B2Handle**	**ePIC service**	**Python lib.**	**Local service**	**Null**

Atmosphere	5	-	1	2	-	5
Marine	5	1	1	-	2	-
Solid Earth	3	-	1	-	-	5
Ecosystem	1	2	1	2	1	3

F2. Data are described with rich metadata (defined by R1 below)

a) Description

F2 refers to the **ability to find a resource** for instance through search or filtering **with rich metadata**. The more detailed the information about a digital resource is, the more accurately findable it becomes. Generic as well as domain-specific metadata descriptors are required to enable both global and local search engines to locate a digital resource.

The metadata of a resource should be sufficiently rich that **a machine or a human user**, upon discovery, can **make an informed choice about whether or not it is appropriate to use** that data object in the context of their analysis.

The minimal 'richness' of the metadata will depend on the requirements of domain-specific community users in their discovery of the resource. It is considered a challenge for each community to define their own metadata schema and to create machine-actionable templates that facilitate capturing consistently uniform and harmonized metadata about similar data resources among all community stakeholders [4].

b) Questions to the ENVRIs

In the RDM+ questionnaire there is unfortunately no question that addresses the F2 principle directly, neither about the richness nor about the findability of the metadata specifically. One question from FMI relates to the availability of machine-readable metadata that describes a digital resource:

Please provide the IRI to a document that contains machine-readable metadata for the digital resource

c) Results

i) Cluster level

For **10 out of 30 repositories within the ENVRI cluster**, a working IRI to a machine-readable metadata document has been provided. This corresponds to **33.3% of the cluster**.

ii) Subdomain level

The answers for each of the 4 environmental subdomains are summarised in Table 9, both as absolute numbers and in the form of percentage. Note that one of the subdomains provided machine-readable metadata documents at 100% of the existing repositories at the time of the survey.

Table 9: Number of repositories providing machine-readable metadata for the available datasets. The first column names the 4 environmental subdomains, column 2 gives the absolute numbers and column 3 the corresponding percentage of the subdomain repositories.

Subdomain	**No of repositories with machine readable metadata**	**Percentage of repositories** (with machine-readable metadata)

	(out of No of repositories)	
Atmosphere	3 out of 12	25,0%
Marine	6 out of 6	100,0%
Solid Earth	1 out of 8	12,5%
Ecosystem	3 out of 8	37,5%

F3. Metadata clearly and explicitly include the identifier of the data they describe

a) Description

The **association between a metadata file and the dataset should be made explicit** by mentioning a dataset's globally unique and persistent identifier in the metadata.

An example of a technology that provides this link is FAIR Data Point, which is based on DCAT that provides identifiers for potentially multiple layers of metadata and a single searchable path through these layers down to the data object itself.

Another example would be the FAIR Digital Object technology[18].

b) Questions to the ENVRIs

Essentially what would be needed to ask is **which are the predicates** that associate the PID of the data with the PID of the metadata. We don't have any question related to the technology applied, but we ask for the presence of the identifier in the metadata:

Are PIDs included in the metadata description?

c) Results

Based on the answers received to the above question, **15 out of 30 repositories among the ENVRIs** include PIDs in their metadata description, which corresponds to **50% of the cluster**. The respective numbers per subdomain are listed in Table 10.

Table 10: Number of repositories with PIDs included in the metadata description. The first column names the 4 environmental subdomains, column 2 gives the absolute numbers and column 3 the corresponding percentage of the subdomain repositories.

Subdomain	**No of repositories with PIDs** (out of No of repositories)	**Percentage of repositories** (with PIDs)
Atmosphere	6 out of 12	50,0%
Marine	6 out of 6	100,0%
Solid Earth	3 out of 8	37,5%
Ecosystem	4 out of 8	50,0%

[18] See the FAIR Digital Framework https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects

F4. (Meta)data are registered or indexed in a searchable resource

a) Description

Identifiers and rich metadata descriptions alone will not ensure 'findability' on the internet. If the availability of a digital resource such as a dataset, service or repository is not known, then nobody (and no machine) can discover it. There are many ways in which digital resources can be made discoverable, including indexing.

b) Questions to the ENVRIs

Do you provide search on data?
Are your repositories registered in a registry? If so, which registry?

c) Results

The answers to the above questions are presented below, starting with the summarised results for the whole cluster.

i) Cluster level

19 repositories provide search on data. This corresponds to **63.3% of the ENVRI cluster**. More specifically for the subdomains, see Table 12.

Regarding the registries, the responses of the ENVRI representatives are shown in Table 11, where all given answers are listed and accompanied by the frequency of each answer. The registries which are FAIR compliant are also shown with a green color. Note here that the sum of registries might be bigger than the number of repositories in the respective subdomain (e.g. for Atmosphere and Marine), as **it is possible to have more than one registry reported per repository**.

Table 11: Summary of the repository registries for all ENVRIs. Note: a) the answer "none" means that the repositories are not registered in any registry, b) the green colored answers are the FAIR compliant ones.

re3data	GEOSS	European Geoscience Registry	IODE Ocean Data Portal	WIS	FAIRsharing	DataCite	local registry	none	planned	NULL
8	3	1	1	1	1	1	2	3	2	10

ii) Subdomain level

The availability of search data per subdomain is shown in Table 12.

The following Tables 13-16 show the registries per subdomain.

Table 12: Number of repositories which provide search on data. The first column names the 4 environmental subdomains, column 2 gives the absolute numbers and column 3 the corresponding percentage of the subdomain repositories.

Subdomain	**No of rep. with search on data** (out of No of repositories)	**Percentage of repositories** (with search on data)
Atmosphere	9 out of 12	75,0%
Marine	3 out of 6	50,0%
Solid Earth	5 out of 8	62,5%
Ecosystem	4 out of 8	50,0%

Table 13: As in Table 11, here for the Atmosphere subdomain repositories.

re3data	GEOSS	WIS	DataCite	none	NULL
6	1	1	1	5	3

Table 14: As in Table 11, here for the Marine subdomain repositories.

FAIRsharing.	GEOSS	IODE Ocean Data Portal	re3data	none
1	1	1	1	3

Table 15: As in Table 11, here for the Solid Earth subdomain repositories.

European Geoscience Registry	local registry	none
1	2	5

Table 16: As in Table 11, here for the Ecosystem subdomain repositories.

re3data	planned	NULL
5	2	1

3.2 Accessible

A1. (Meta)data are retrievable by their identifier using a standardised protocol

A1.1 The protocol is open, free, and universally implementable

a) Description

To maximise data reuse, the protocol should be free (no-cost) and open (-sourced) and thus globally implementable to facilitate data retrieval. Anyone with a computer and an internet connection can access at least the metadata.

Examples:

- HTTP, FTP, SMTP, …
- Telephone (arguably not universally-implementable, but close enough)
- A counter-example would be Skype, which is not universally-implementable because it is proprietary
- Microsoft Exchange Server protocol is also proprietary

b) Questions to the ENVRIs

A1.1 What is the major access technology supported?

c) Results

i) Cluster level

The answers provided by the ENVRIs were not always usable. As a result there are "NULL" answers which had to be interpreted but in 7 cases it was not possible to get more insights.

Most of the RIs (22 relevant answers) seem to have HTTP access protocols, and 1 of them answered with FTP.

ii) Subdomain level

The provided information concerning the 4 subdomains and their access technologies are given in Table 17. Note that in the case of the Solid Earth subdomain, half of the answers could not be interpreted ("NULL"). In general, for repositories which have open metadata, it is reasonable to assume that they are compliant with the principle A1.1.

Table 17: Summary of the major access technologies for all subdomains. Note: the green colored answers are FAIR compliant.

	Access Technologies		
Subdomain	**HTTP**	**FTP**	**NULL**
Atmosphere	10	1	1
Marine	5	-	1
Solid Earth	4	-	4
Ecosystem	7	-	1

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

a) Description

This is a key, but often misunderstood, element of FAIR. **The 'A' in FAIR does not necessarily mean 'open' or 'free'.** Rather, it implies that one should provide the exact conditions under which the data are accessible. Hence, even heavily protected and private data can be FAIR.

Ideally, accessibility is specified in such a way that a machine can automatically understand the requirements, and then either automatically execute the requirements or alert the user to the requirements. It often makes sense to request users to create a user account for a repository. This allows to authenticate the owner (or contributor) of each dataset, and to potentially set user-specific rights.

b) Questions to the ENVRIs

The RDM+ questionnaire included some questions related to A1.2, although the answers don't give insights about the compliance:

Do you make statements about access policies in your metadata?

How is authentication and authorization done?

c) Results

i) Cluster level

Regarding the statements on the access policies in the metadata, **46,7% of the ENVRI cluster** (14 out of 30 repositories) seem to be compliant. More specifically, the numbers for the repositories per subdomain which gave positive answers to that question are shown in Table 18.

ii) Subdomain level

The information on the authentication and authorization is gathered per subdomain and presented in the following Tables 19-22.

Table 18: Statements of access policies existing in the metadata, as reported per subdomain. The table shows the number of the corresponding repositories and the respective percentage per subdomain.

Subdomain	**No of rep. with statements on access policy in metadata** (out of No of repositories)	**Percentage of repositories** (with statement on access policy in metadata)
Atmosphere	5 out of 12	47,7%
Marine	4 out of 6	33,3%
Solid Earth	3 out of 8	37,5%
Ecosystem	6 out of 8	75,0%

Table 19: The responses of the Atmosphere RIs regarding authentication and authorisation. The columns colored with green indicate that they are FAIR compliant.

OAuth	Google	none-open	local method
3	1	4	4

Table 20: Same as in Table 19, here for the Marine subdomain.

OAuth	SSQL Service	none-open
3	2	1

Table 21: Same as in Table 19, here for the Solid Earth subdomain.

Certification Method	none-open	NULL
2	2	4

Table 22: Same as in Table 19, here for the Ecosystem subdomain.

OAuth	Liferay	none-open	local method	NULL
1	1	2	1	3

A2. Metadata are accessible, even when the data are no longer available

a) Description

Datasets tend to degrade or disappear over time because there is a cost to maintaining an online presence for data resources. When this happens, links become invalid and users waste time finding data that might no longer be available.

Storing the metadata generally is much easier and cheaper. Hence, principle A2 states that metadata should persist even when the data are no longer sustained. Another requirement is that metadata must be openly accessible without any barriers (authentication).

b) Questions to the ENVRIs

Please provide the URL to a metadata longevity plan
Are metadata openly available?

c) Results

None of the RIs has provided a documented longevity plan for the metadata. 6 RIs provided a link to an online resource, but no longevity information was found on any of these. **This is a clear gap over all ENVRIs.**

Regarding the second question, all but one RI from the atmosphere subdomain provide openly available metadata.

3.3 Interoperable

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

a) Description

Humans should be able to exchange and interpret each other's data. But this also applies to computers, meaning data that should be readable for machines without the need for specialised or ad hoc algorithms, translators, or mappings.

This principle is about data types, data formats, metadata formats, metadata exchange formats which allow for machine interoperation. In most cases data won't use any machine-interpretable language yet, but if the semantics of the datasets is described at the metadata level, machines can still figure out how to use them.

b) Questions to the ENVRIs

According to the above description, the focus lies here at the metadata formats:

Which are the export formats supported?

c) Results

i) Cluster level

With the exception of one RI, all RIs use at least one machine-readable metadata exchange format.

ii) Subdomain level

The answers of the RIs are grouped per subdomain and listed in the Tables 23-26 of the following paragraph (results at the RI level can be found in the supplementary material which is available in the protected project-internal Redmine environment).

Table 23: All supported formats (and the frequency of the respective given answer) as reported by the Atmosphere subdomain repositories. The numbers in the second row indicate the frequency of each answer (i.e. the number of repositories that report the listed answer). The green colored answers are the ones which are actually FAIR compliant.

XML	JSON	JSON-LD	turtle	RDF	NetCDF	HTML	CSV	ASCII	plain text
8	6	1	1	1	2	1	1	1	3

Table 24: Same as in Table 23, here for the Marine subdomain.

XML	JSON	RSS	NetCDF	turtle	HTML	plain text	CSV
5	3	2	1	1	5	1	1

Table 25: Same as in Table 23, here for the Solid Earth subdomain.

XML	JSON	JSON-LD	RDF	turtle	SEED	Atom	CSV	PDF	NULL
7	1	1	3	1	3	1	1	1	1

Table 26: Same as in Table 23, here for the Ecosystem subdomain.

XML	RDF	JSON	JSON-LD	turtle	HTML	plain text	PDF	NULL
6	2	2	1	1	2	1	1	2

I2. (Meta)data use vocabularies that follow FAIR principles

a) Description

It is critical to have commonly used controlled vocabularies, ontologies, and thesauri that are FAIR at least at the subdomain level. The controlled vocabulary used to describe datasets needs to be documented and resolvable using globally unique and persistent identifiers.

b) Questions to the ENVRIs

What is the name of the metadata schema? Indicate the vocabulary name

c) Results

The given answers to the above mentioned questions are quite diverse, thus the results are only presented here at subdomain level, where all answers are listed in the Tables 27-30. Note again here that the "NULL" answers are different than the "none", as "NULL" means that either the given answer (to the questionnaire) could not be interpreted or that there was no answer given, while "none" corresponds to an actual answer.

Table 27: The names of the metadata schemas reported for the Atmosphere subdomain are listed here. The green colored answers correspond to the ones that are FAIR compliant.

ISO 19115/19139	Nasa Ames ASCII	GeoDCAT	GCMD-DIF	local schema	none
7	1	1	1	1	2

Table 28: Similar to Table 27, here for the Marine subdomain.

ISO 19115/19139	EML 2.0	SDN community profile	GeoDCAT	Argo user namual	NetCDF CF
5	2	2	1	1	1

Table 29: Similar to Table 27, here for the Solid Earth subdomain.

FDSN StationXML	DCAT	ISO 19115/19139	SEED	S-PROV
7	2	1	1	1

Table 30: Similar to Table 27, here for the Ecosystem subdomain.

ISO 19115/19139	INSPIRE EF	EML 2.0	SensorML	GeoDCAT	NULL
7	2	2	2	1	2

The next group of Tables 31-34 list the vocabulary names, as those were reported per subdomain.

Table 31: The information on the vocabulary names for the Atmosphere subdomain is reported here. The green colored answers correspond to the ones that are FAIR compliant. *Note: Cedar is a Metadata schema editor and thus not a vocabulary.*

CF standard names	PROV-O	Darwin Core	ICOS	Cedar	none	NULL
9	1	1	1	1	1	1

Table 32: Similar to Table 31, here for the Marine subdomain.

NERC vocabulary service	WoRMS	Marine Regions	Marine species	EDMO	SeaDataNet CDI	PROV-O	Darwin Core	ICOS
4	2	2	2	2	2	1	1	1

Table 33: Similar to Table 31, here for the Solid Earth subdomain.

S-PROV	SEED	none	NULL
1	1	1	5

Table 34: Similar to Table 31, here for the Ecosystem subdomain.

LW vocab	EnvThes	CF standard names	Darwin Core	PROV-O	ICOS	ANAEETHES
6	3	2	1	1	1	1

I3 (Meta)data include qualified references to other (meta)data

a) Description

A qualified reference is a cross-reference that explains its intent. The goal therefore is to create as many meaningful links as possible between (meta)data resources to enrich the contextual knowledge about the data, balanced against the time/energy involved in making a good **data model**.

To be more concrete, one should specify if one dataset builds on another data set, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset. In particular, the scientific links between the datasets need to be described. Furthermore, all datasets need to be properly cited (i.e., including their globally unique and persistent identifiers).

b) Questions to the ENVRIs

Are all categories used in the schemas defined in open registries?

c) Results

i) Cluster level

The resulting answers to the above mentioned question for the I3 principle indicate that 40% of the cluster repositories (12 out of 30) report that all categories used in their schemas are defined in open registries. More specifically, per subdomain the results are summarised below.

ii) Subdomain level

The information on the categories defined in open registries are presented in Table 35, per subdomain. Here it becomes evident that some answers must be reviewed. Although 6 Solid Earth RIs did not indicate any vocabularies, 5 answered to use categories from vocabularies, which does not seem to be coherent. Also, in the ecosystem subdomain there is some incoherence between the two questions, as they all seem to have domain-specific vocabularies, but don't seem to use them.

Table 35: Information on the categories used in the schemas defined in open repositories, as reported per subdomain. The table shows the number of the corresponding repositories that gave positive answers, and the respective percentage per subdomain.

Subdomain	No of repositories with categories in open registries (out of No of repositories)	Percentage of repositories (with categories in open registries)
Atmosphere	1 out of 12	8,3%
Marine	5 out of 6	83,3%
Solid Earth	5 out of 8	62,5%
Ecosystem	3 out of 8	37,5%

3.4 Reusable

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

R1.1. (Meta)data are released with a clear and accessible data usage license

a) Description

R1.1 is about *legal* interoperability. What usage rights do you attach to your data? This should be described clearly. Ambiguity could severely limit the reuse of data by organisations that struggle to comply with licensing restrictions.

The conditions under which the data can be used should be clear to machines and humans.

a) Questions to the ENVRIs

Which specific licenses do you use for your data?

b) Results

i) Cluster level

4 different answers were found among the RI responses regarding the licenses, while 14 ENVRIs have not provided any information about their licenses (here listed as "NULL", Table 36).

Table 36: All reported licenses, used by the ENVRIs, are listed here. The green color indicates the ones which are FAIR compliant.

CC BY	CC BY NC	CC0	local license	NULL
8	2	2	5	14

ii) Subdomain level

Table 37 lists all the reported licenses per subdomain. The FAIR compliant ones are highlighted.

Table 37: As in Table 36, here per subdomain.

Subdomain	CC BY	CC BY NC	CC0	local license	NULL
Atmosphere	3	-	-	3	6
Marine	1	-	2	2	1
Solid Earth	3	1	-	-	5
Ecosystem	4	1	-	-	3

R1.2. (Meta)data are associated with detailed provenance

a) Description

For others to reuse the data, they should know where the data came from (i.e., clear story of origin/history, see R1), who to cite and/or how the data provider wishes to be acknowledged. Including a description of the workflow that led to the data is required. Ideally, this workflow is described in a machine-readable format.

b) Questions to the ENVRIs

Do you provide machine readable provenance information about your data (based on PROV-O or similar)?

c) Results

The machine readable provenance information about the data provided by the subdomain repositories is summarised in Table 38.

Table 38: Machine readable provenance information about the data provided by the RIs, summarised per subdomain. The answer with the green color is the only one which is FAIR compliant.

Subdomain	PROV-O	Simplified PROV-O	text only	NULL
Atmosphere	-	2	2	8
Marine	-	1	4	1
Solid Earth	1	-	-	7
Ecosystem	-	1	4	3

R1.3. (Meta)data meet domain-relevant community standards

Description

It is easier to reuse datasets if they are similar: same type of data, data organised in a standardised way, well-established and sustainable file formats, documentation (metadata) following a common template and using common vocabulary. If community standards or best practices for data archiving and sharing exist, they should be followed. FAIR data should at least meet minimal information standards agreed within the community.

This principle seeks for convergence and addresses all community specific standards applied, and is thus a summary of I2, I3 and R1.2.

4 Requirements

The goal of the FAIRness assessment of the ENVRIs and the interpretation of the results is to identify gaps between the current FAIR level of the ENVRIs and what is expected in order to achieve a FAIR environmental cluster, to efficiently plan and implement the required actions that will bring the ENVRIs closer to becoming FAIR. The relevant requirements are discussed in this chapter, where also a summarised overview of the identified gaps is presented together with an outline of the implementation plans which have been reported by the environmental subdomains.

4.1 FAIR requirements

4.1.1 Gap analysis

The gap analysis is derived from the FAIR assessment activity. Each of the subdomains used the analysis provided by WP5 to discuss gaps and strengths identified by each RI, to consolidate common implementation strategies and to compile their reports[19]. This deliverable aims at synthesizing the results from these reports to get an overview of the common requirements. These will build the basis for the WP7 facilitation of the implementation activities planned to be undertaken by the RIs in the upcoming years to achieve FAIRer data and services.

The gap analysis in this deliverable is thus taken from those reports and supplemented by observations from the assessment done in chapter 3 (text in italics in Tables 39-42).

Findability

As shown in Table 39, F1 is still a challenge for many RIs. PIDs are not yet fully implemented. Regarding F2 and metadata findability, it seems that the harmonization of the ENVRIs metadata for findability at a subdomain level is a common gap. Machine readable metadata and a single metadata catalogue for the datasets of each RI would also increase their findability. F4 requires that metadata are registered or indexed in a searchable resource and this seems to be a common need to address at subdomain or even at cluster level.

Accessibility

Generally A1 is the sub-principle which is mainly solved for all ENVRIs. But there is still room for optimization. Many RIs still need to mention access policies in their metadata. All RIs need to provide machine readable metadata longevity information (Principle A2). Table 40 gives an overview.

Interoperability

Table 41 reveals the need for convergence in the use of vocabularies at the subdomain level. According to the FAIRness analysis presented earlier in chapter 3, EPOS (in Solid Earth) needs to increase the use of vocabularies in general, but this was not addressed in the WP10 deliverable as a gap.

Reusability

None but one repository (reported by EPOS in Solid Earth) has implemented machine readable provenance. As shown in Table 42, most RIs still need to provide machine readable usage license and it would be useful to agree on relevant common strategies also at cluster level.

[19] D8.1, D9.1, D10.1, D11.1; List of published ENVRI-FAIR Deliverables: https://envri.eu/deliverables/

Table 39: Identified gaps by the RIs (after their first FAIRness assessment), regarding the Principles for Findability. The text in italics indicates gaps which were identified during the assessment described in chapter 3 of the present deliverable.

RI	F1	F2	F2	F2	F3	F4	F4
ACTRIS	DOIs	harmonize discovery metadata	single metadata catalogue	*machine readable metadata*			metadata indexed
EISCAT	PIDs		metadata registry	*machine readable metadata*	*PIDs included in metadata*	search on data	metadata indexed
IAGOS	PIDs			machine readable metadata	*PIDs included in metadata*	semantic search on data	metadata indexed
SIOS	PIDs/DOI					semantic search on data	metadata indexed
ICOS		discovery of metadata for data resources				develop search capability with elaborated filters	
EURO-ARGO			API cloud service for data discovery			local search engine	*register of repository in registry*
SDN						search on data via ERDDAP	register of repository in registry
LifeWatch (marine)	PIDs	harmonize discovery metadata					(MDA) register of repository in registry
EPOS	*PIDs*		single metadata catalogue	*machine readable metadata*	*PIDS included in metadata*		*register of repository in registry*
AnaEE		discovery of metadata for data resources		*machine readable metadata*		search on data	
eLTER	full provision of PIDs			*machine readable metadata*	*PIDS included in metadata*		
LifeWatch	DOIs			*machine readable metadata*			*register of repository in registry*

Table 40: As in Table 39, here regarding the Principles for Accessibility.

RI	A1.1	A1.2	A2
ACTRIS	standardized solution for access protocol	*access policies mentioned in metadata*	*metadata longevity plan*
EISCAT		authentication and authorisation	*metadata longevity plan*
IAGOS	RESTful services for data and metadata access	ORCID	metadata longevity plan
SIOS	OPeNDAP	ORCID	*metadata longevity plan*
ICOS		mention explicitely access policy statements	*metadata longevity plan*
EURO-ARGO			*metadata longevity plan*
SDN	(CDI) API for improved machine access	(Central Data Products) missing access to dataplots via WPS	metadata longevity plan
LifeWatch (marine)		*access policies mentioned in metadata*	metadata longevity plan
EPOS		*access policies mentioned in metadata*	*metadata longevity plan*
AnaEE	*standardized solution for access protocol*	*access policies mentioned in metadata*	*metadata longevity plan*
eLTER		*access policies mentioned in metadata*	*metadata longevity plan*
LifeWatch			metadata longevity plan

Table 41: As in Table 39, here regarding the Principles for Interoperability.

RI	I1	I1	I1	I2	I2	I2	I3
ACTRIS				use of standardized metadata			categories in metadata marked up with vocabularies
EISCAT	plain text			use of standardized metadata			*categories in metadata marked up with vocabularies*
IAGOS					wider use of vocabularies	register schemas in common	categories in metadata marked up with vocabularies
SIOS	OAI-PMH compliant interface		OGC CSW, OpenSearch	harmonise metadata standards	mapping of vocabulary		*categories in metadata marked up with vocabularies*
ICOS	NetCDF				develop ontology	register schemas in common	categories in metadata marked up with vocabularies
Euro-Argo		SPARQL endpoint	OGC WMS SOS		harmonise vocabulary	register schemas in common	
SDN	RDF, JSON	(CDI) SPARQL endpoint					categories in metadata marked up with vocabularies
LifeWatch (marine)	OAI-PMH compliant interface		OGC WMS SOS			register schemas in common	categories in metadata marked up with vocabularies
EPOS					*wider use of vocabularies*		
AnaEE			semantic pipeline				
eLTER							semantic integration

Table 42: As in Table 39, here regarding the Principles for Reusability.

RI	R1.1	R1.2	R1.2	R1.3
ACTRIS	usage license	machine interpretable provenance info	include metadata in provenance	
EISCAT	*usage license*	machine interpretable provenance info	include metadata in provenance	
IAGOS	usage license	*machine interpretable provenance info*	*include metadata in provenance*	compliance validation service
SIOS	usage license	machine interpretable provenance info	include metadata in provenance	improve dataset validation service
ICOS		machine interpretable provenance info	include metadata in provenance	compliance validation service
EURO-ARGO	usage license	machine interpretable provenance info	*include metadata in provenance*	
SDN	*usage license*	machine interpretable provenance info	include metadata in provenance	
LifeWatch (marine)		machine interpretable provenance info	include metadata in provenance	compliance validation service
EPOS				
AnaEE	*usage license*	*machine interpretable provenance info*	*include metadata in provenance*	
eLTER		machine interpretable provenance info	provenance chaining from different sources	
LifeWatch	usage license	machine interpretable provenance info for dataset	*include metadata in provenance*	

4.1.2 Implementation Plans

Following the gap analysis by each RI, the subdomains together with the RIs plan the actions each RI needs to take so that they meet the FAIR requirements. The challenge then for WP5 and WP7 is to support the common development targets and prepare the output ENVRI-FAIR catalogue of services, which will have to be properly defined. The cluster, with help from WP5 and more specifically by WP5 Task Forces, should design and provide the guidelines for the validation of these services, and work together with EOSC to formulate a strategic roadmap for future development. During the project, issues regarding the ENVRI-Hub and how this will be built (e.g. as a federated virtual hub) are discussed among all subdomains, aiming to define the necessary common solutions.

Based on the FAIRness analysis, the participating RIs have been found in a wide range of readiness. There are differences, but common characteristics as well. For example, it is common that more attention is required on the machine-to-machine (M2M) interfaces. Regarding the PIDs, their use can still be improved. It seems that many RIs plan the publication of their metadata through DataCite. There are also some common standards used. Subsetting seems to be a common solution, aiming to help the users working on the datasets of their interest, without transferring big amounts of data which are not required for their analysis. A set of APIs is required. AAI is also work in progress. In principle, there are technical solutions which are already available (and collected in the Knowledge Base), and the experts assigned to WP7 can help the RIs to implement them.

The cluster can already work on some common developments, with suggestions from WP5, but also using the experience with solutions followed by the different RIs. The ENVRI cluster can become a virtual organization, where e.g. the users but also all people involved in the RIs will need only one login, meaning that they will keep one identity which will give them access to all ENVRIs. This type of virtual organization will also have great advantage for M2M in the ENVRI domain. The experience from AARC and EPOS can contribute along these lines.

Regarding PIDs, those form a core component of FAIR repositories and there are several examples to follow (as in RDA, FREYA[20], FAIRsFAIR[21]). Because of the questioned sustainability of such services, one of the goals of the ENVRI community is to find common solutions. It is also important to work on the provenance, going towards DOIP systems where PIDs are used for direct access to the data, metadata etc. Another challenge is the issue of data storage, access and use. It is important that the data can be accessed across the RIs, with M2M to access data directly in the cloud. The EOSC could offer the required resources for storage. VREs are also of great importance. ICOS has now experience with Jupyter, which seems to be a very good solution to develop use cases.

The different ENVRI-FAIR WPs need to work on their tasks which require collaboration between the teams. The 4 subdomains (WP8-11) have already reported their initial implementation plans[22]. With help from the training team of WP6, there are now opportunities for training of personnel, which can be particularly useful for starting communities. There are also additional plans, with several stakeholders defining their components as e.g. the users of the ENVRI data and services. Taking all the above into account, ENVRI-FAIR has now formed the cross-cutting Task Forces, i.e. thematic teams which involve representatives from all ENVRIs and subdomains, to coordinate the work on specific topics which are considered important components of the required common solutions at the cluster level. The list below names the first Task Forces and their themes, as decided during the Workshop of WP5 with the subdomains (2019-10-30, Lund):

1. ENVRI Catalogue of services (FAIR)
2. ENVRI (VO) AAI implementation (A)
3. PIDs, identification, types and registries (FAIR)
4. Triple stores and data storage certification (FAR)
5. Licenses, citation and usage tracking (of data and VRE) (IR)
6. User oriented cross-domain demonstration cases in e.g. Jupyter (IR)

The Task Forces have been active since January 2020.

4.2 Other requirements

4.2.1 EOSC requirements

In November 2018, under the Austrian Presidency, the European Commission launched the European Open Science Cloud at the University of Vienna. The EOSC is not a dedicated infrastructure or a software package, it is a process of making research data and services in Europe accessible to all researchers under the same conditions for use. The initiative aims to give a strong push in Europe towards a culture of open research data that are findable, accessible, interoperable and reusable (FAIR), thereby allowing all European researchers to engage in data-driven science.

European researchers are faced today with data fragmentation and unequal access to quality information sets. This situation applies across scientific domains, countries and governance models with varying access policies. There is limited cross-disciplinary access to datasets, services (and data) are mostly non-interoperable while data is often closed. Based on the cost benefit analysis by the European commission, there is a huge cost to the European economy by not having FAIR data [5]. To promote interdisciplinary research across Europe, a Federating Core[23] is planned by the European Open Science Cloud (EOSC) to be widely used.

---

[20] https://www.project-freya.eu/en
[21] https://www.fairsfair.eu/
[22] List of published ENVRI-FAIR Deliverables: https://envri.eu/deliverables/
[23] Solutions for a Sustainable EOSC, A strawman report from the Sustainability Working Group, available here: https://www.eoscsecretariat.eu/sites/default/files/swg_-_solutions_for_a_sustainable_eosc_0.pdf

This will establish a seamless environment providing universal access to data, supported by data infrastructures like those in ENVRI-FAIR. The EOSC Service Portfolio will provide additional added-value services (common and thematic) to exploit the Federating Core, which will be discoverable through the EOSC Portal. Through EOSC, researchers will be able to find, access, share and (re)use data and services, or promote and support their research in the framework of open science. As service providers, they will have the chance (and also the responsibility) to use the Federating Core to publish their services by adding them to the EOSC Portal Catalogue and Marketplace[24], as well as identify the end-user needs.

The Federating Core and the EOSC Service Portfolio will be built on prerequisite Rules of Participation (defined by EOSC), technical and policy requirements to define the EOSC conformance preconditions for providers. During the first steps of EOSC, a preliminary set of rules of participation for service providers and users was suggested (EOSCpilot activities). Since then the rules have been adjusted, aligned with the Implementation Roadmap for the European Open Science Cloud (EU commission, 2018)[25]. Considering that all EOSC services will be registered in an EOSC compliant or compatible service catalogue visible to the global EOSC gateway, standardization, transparency and interoperability of the registered services is needed. To register their services, the providers will have to describe them according to the EOSC service guidelines. These instruct sufficient machine readable information (metadata) on availability, functionalities, operations, maturity, user support, interoperability, licenses for openness, GDPR[26] compliance regarding the privacy, terms of use, conceptual framework. Further detailed rules of participation for service providers are listed in Figure 3.

Rules of participation will be applied also for the use of the EOSC services. Within the Terms of Use, the EOSC users should be encouraged to share and deposit their data in community-agreed data repositories. The data provided on the other hand, such as data from the RI's should be FAIR and if possible also open. Users are also requested to acknowledge by means of citation the specific services accessed through EOSC.

To summarize, the Federating Core will deliver three capabilities:

1. Federating tier: this corresponds to a hub portfolio of services provided by multiple suppliers, for coordinated access and management of resources
2. Resource tier: shared resources to include e.g. data, applications, software, pipelines etc. (i.e. scientific outputs), storage and compute hosting platforms (to deposit, share and process the scientific outputs)
3. Regulatory tier: the Compliance Framework that defines the policies and processes for the demand and supply sides to engage with EOSC (e.g. the Rules of Participation, the Service Management System and related policies).

[24] https://marketplace.eosc-portal.eu/
[25] EOSCpilot: Rules of participation (2018). https://ec.europa.eu/info/sites/info/files/conferences/eosc_summit_2018/eosc_pilot_considerations_on_the_rules_of_participation.pdf, accessed 2019-12-07.
[26] The General Data Protection Regulation 2016/679 is a regulation in EU law on data protection and privacy for all individual citizens of the European Union and the European Economic Area, see https://eur-lex.europa.eu/eli/reg/2016/679/oj

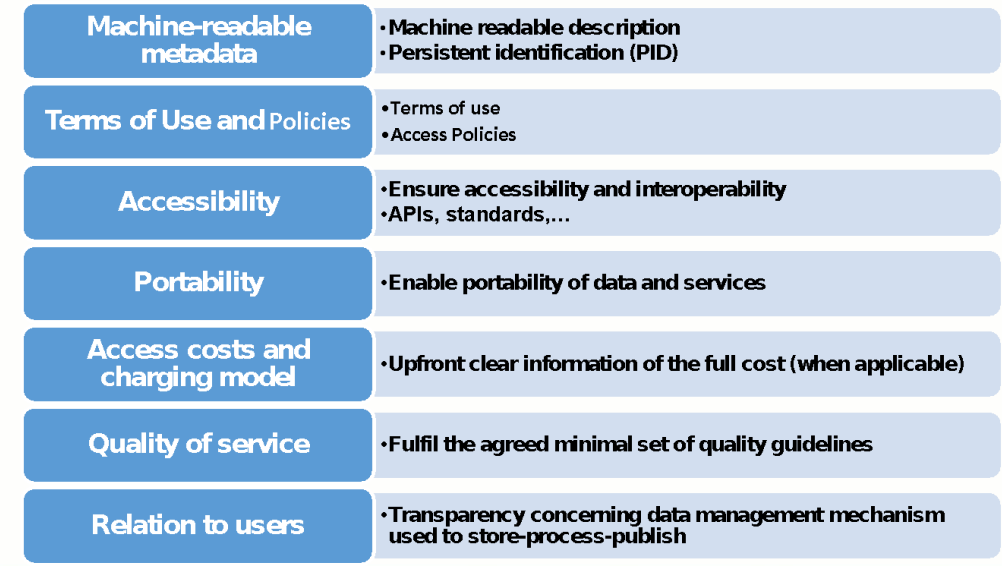

Figure 3: Detailed Rules of Participation for Service Providers.

There are technical components which will enable the federation, access and order/delivery of the services. Processes that include naming, locating, discovering and accessing data (and/or services) through EOSC will require the application of standard mechanisms, along with a common framework to manage a user's identity and access.

The EOSC pilot project performed an e-Infrastructure gap analysis and identified the main difficulties in overcoming these gaps. Those include:

- Diversity and incompatibility of the AAIs (Authentication and Authorisation Infrastructure)
- Missing network services
- Diversity of services and providers
- Diversity of access policy
- Low awareness of the e-infrastructures and services
- Lack of expertise, training, easy tools, human networks

Regarding interoperability, the EOSC pilot project has published some recommendations:

- The EOSC should propose specific and simple guidelines for the data and the technical solutions, but also information across subdomains on key operational metadata which are required for services
- All contributors should provide structured metadata
- Within EOSC, building on existing standards and formats is encouraged, using common practices across scientific domains
- An interconnected ecosystem of metadata (to facilitate data discovery) supported by EOSC
- Implementation of a monitoring service to validate standards and recommendations proposed by the EOSC
- Figure 4 visualizes the recommended bridges that the infrastructures will have to build, in order to overcome the identified gaps in interoperability.

Figure 4 visualizes the recommended bridges that the infrastructures will have to build, in order to overcome the identified gaps in interoperability.

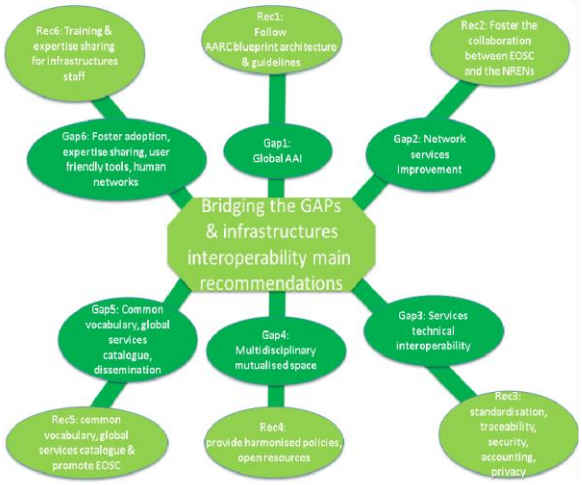

Figure 4: Main recommendations for infrastructure interoperability.

By promoting the sharing of FAIR data and services, EOSC can also be seen as the European contribution to the Internet of FAIR data and services (see GO FAIR[27]). It builds on minimal standards, lightweight international guidance with a large degree of freedom regarding practical implementation. To that extent, the ENVRI community is now working on bringing all the involved RIs to an appropriate level of FAIRness, to be prepared for participating in EOSC. Table 43 lists some of the points the ENVRI community has addressed, in relation to EOSC, from the community perspective.

Table 43: ENVRI-FAIR requirements[28]

What ENVRI-FAIR needs from the EOSC (user perspective)
Generic data and metadata services such as for AAI, PID, and provenance, for tailoring to specific Research Infrastructure needs and adoption by individual research infrastructures
Generic workflow management tools and services, for tailoring to specific Research Infrastructure needs and adoption by individual RIs
Access to shared resources such as repositories, HPC and data management tools
Common APIs to support remote data discovery, access, and sharing
Provision of notebook based environments which allow to access and integrate data services for the community
What ENVRI-FAIR can offer to EOSC (provider perspective)
Collective domain-specific knowledge and competencies that underlie all the data and other services provided by the European ENVRIs
FAIR-based tools and resources for easy and seamless access to environmental data and services provided by the European ENVRIs
ENVRI-hub – a virtual, federated machine-to-machine interface to access environmental data and services provided by the contributing ENVRIs

[27] https://www.go-fair.org/go-fair-initiative/, accessed 2019-12-07.
[28] ENVRI-FAIR EOSC Position Paper https://zenodo.org/record/3666806#.XlpWinsxnD4

4.2.2 CoreTrustSeal requirements

Although not in the main focus of this deliverable we consider it necessary to briefly include here considerations about standardization efforts for trustworthy data repositories, which are critical for the preservation of research data. These considerations should be taken into account for the next FAIR assessment runs.

A cooperation between the World Data System of International Science Council (WDS) and the Data Seal of Approval (DSA) under the umbrella of the Research Data Alliance lead recently (November 2019) to the development of a common CoreTrustSeal certification. This procedure is based on the DSA-WDS Core Trustworthy Data Repository Requirements catalogue and replaces existing DSA and WDS certifications. It will substantially support long-term access to reusable data, an objective shared also by the FAIR principles. There are quite some overlaps and complementarities between the goals of the CoreTrustSeal certification and the FAIR criteria, which are carefully examined in [5].

According to [6] the FAIR principles do not explicitly address the long-term preservation of data needed to ensure that this access endures. "Data should be stored in a trusted and sustainable digital repository to provide reassurances about the standard of stewardship and the commitment to preserve" (p. 22). If a data repository fulfills the CoreTrustSeal requirements, also the data it hosts, will in most cases meet many of the FAIR principles. The CoreTrustSeal certification requirements could therefore be used as a basis to assess the FAIR compliance of datasets at least for those principles that relate to attributes of the repositories holding the data. In addition, they address other very important aspects not covered by the FAIR principles such as maintaining the understandability and reusability of datasets over time. But at the time of the project's start these requirements were not yet developed, so this option can only be taken into account in future activities of WP5 tasks.

Although openness, the availability and the reusability of data is becoming more and more recognized as essential there are technical limitations to data sharing including systems not operating correctly, datasets not being complete or not containing what they claim and access not being guaranteed, but most importantly there are social limitations such as trust:

- data funders want reassurances that their investment in the production of research data is not wasted
- data providers want to be sure their data are safe and remain accessible with all associated meaning to be usable over time
- data users expect that data have been preserved properly and are of high quality[29]

The CoreTrustSeal certification process responds to these trust needs with a list of requirements[30], which comes along with a guidance text to assist applicants in providing sufficient evidence about their repositories. The self-assessment should provide indications about the compliance level for each of the requirements (0-4), and in case of not applicability the reason must be documented. This assessment should be repeated every three years.

Here we provide just a plain list of the requirements, further details can be found in the certification document.

R0. Please provide context of your repository (type, brief description, level of curation performed, outsource partners, other relevant information

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

---

[29] Rorie Edmunds: CoreTrustSeal Certification Cohort meeting 2018-10-30, https://www.coretrustseal.org/why-certification/requirements/, accessed 2019-07-12.
[30] CoreTrustSeal: https://www.coretrustseal.org/wp-content/uploads/2017/01/Core_Trustworthy_Data_Repositories_Requirements_01_00.pdf, accessed 2019-12-07.

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse, or external, including scientific guidance, if relevant).

R7. The repository guarantees the integrity and authenticity of the data.

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

R9. The repository applies documented processes and procedures in managing archival storage of the data.

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

R12. Archiving takes place according to defined workflows from ingest to dissemination.

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

5 Next steps for the FAIRness assessment

The assessment must be as efficient as possible, given the number of involved RIs, the fact that most RIs are distributed, and thus require different assessments for their centres, and that such assessment should be performed repeatedly during the course of the project in order to monitor the development. Not only should it be efficient in collecting the required information, but also in analysing the collected information. Making the assessment efficient has been and continues to be a challenge led by WP5.

5.1 Current developments

There is close contact with GO FAIR to share experiences with the questionnaire created to collect information about FAIRness. ENVRI-FAIR members are now part of the FAIR Convergence Matrix Working Group[31] to develop the next generation FAIR questionnaire and to bring in the knowledge of translating questionnaire answers into RDF and creating the SPARQL queries. The FAIR Convergence Matrix is a online platform that uses a core ontology[32] to compile for any Community of Practice ("columns" in the Matrix), an inventory of their FAIR implementation Choices and Challenges selected from a list of existing or proposed digital Resources ("rows" in the Matrix) for each of the FAIR principles. FAIR implementations are highly dynamic and undergo continuous development requiring regular updates to both the questionnaire and the responses given by the communities.

The Data Stewardship Wizard (DSW) [7] is used as the Convergence Matrix environment providing the possibility to capture the questions and the answers using semantically-enabled drop-down menus and auto-complete functions. The answer values are taken from FAIRsharing[33], which provides globally unique and persistent identifiers and metadata descriptions of FAIR–related standards, repositories and data policies. The questionnaire structure is captured as a machine-readable Knowledge Model in JSON format, easily editable and trackable as a FAIR resource. The answers are stored in a document database and are subsequently transformed into JSON-LD for an interoperable RDF representation [8]. It is planned to link the DSW Knowledge Model with the ENVRI Knowledge Base, to accommodate all implementation choices of the ENVRI communities as instances in a larger conceptual context.

Each column of the Matrix comprises a profile (FIP) characterizing how each community has chosen to implement FAIR and as such it is a unique signature representing each community. FIPs can be used as a powerful accelerator of convergence on FAIR standards and technologies [9].

FIPs can be represented as collections of Convergence Matrix nanopublications, forming a semantically enabled knowlet of all the choices and challenges declared by a community. FIPs can be themselves FAIR Digital Objects (FDOs)[34] having GUPRIs, type specifications and other FAIR metadata.

5.2 Future ENVRI-FAIR assessments

During the lifetime of the ENVRI-FAIR project the FAIRness level of the ENVRIs will be measured again, presumably in the middle period and before the end of the project. For this purpose the WP5 team is working closely with the GO FAIR Convergence Matrix team to

[31] https://www.go-fair.org/today/FAIR-matrix/
[32] Convergence Matrix Ontology: https://github.com/go-fair-ins/GO-FAIR-Ontology/blob/master/Diagrams/Matrix.pdf; GO FAIR Core Ontology https://github.com/go-fair-ins/GO-FAIR-Ontology/blob/master/Diagrams/Core.pdf
[33] https://fairsharing.org/
[34] See the FAIR Digital Framework https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects

prepare the new DSW (see preceding paragraph) questionnaire. In the meantime the responses from the first assessment will be converted into the new data model for comparison. We will also collaborate closely with WP6 to suggest training activities on specific aspects of the FAIR principles, to improve the RIs' representative knowledge on the subject. Moreover, we intend to actively contribute with suggestions and comments to the deliverable D4.1[35] of FAIRsFAIR, namely the "Draft Recommendations on Requirements for Fair Datasets in Certified Repositories" before the end of the review phase (July 2020).

We will also consider using the FMI Evaluator, when it is improved and ready to be tested by the communities. This will lead to FAIR assessed profiles (FIPs), which will allow us to track and visualize the improvement of the ENVRIs FAIRness during the project.

References

[1] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific data, vol. 3, 2016.

[2] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, and M. D. Wilkinson, "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud," Information Services & Use, vol. 37, no. 1, pp. 49–56, Jan. 2017.

[3] Wilkinson, Mark & Sansone, Susanna-Assunta & Schultes, Erik & Doorn, Peter & Bonino da Silva Santos, Luiz Olavo & Dumontier, Michel. (2018). A design framework and exemplar metrics for FAIRness. Scientific Data. 5. 180118. 10.1038/sdata.2018.118.

[4] A. Jacobsen et al., FAIR Principles: Interpretations and Implementation Considerations. In: Special Issue on Emerging FAIR Practices, MIT Press, 2019.

[5] P. O. of the E. Union, "Cost-benefit analysis for FAIR research data : cost of not having FAIR research data.," 16-Jan-2019. [Online]. Available: https://op.europa.eu:443/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1. [Accessed: 07-Dec-2019].

[6] M. Mokrane and J. Recker, "Enabling Findable, Accessible, Interoperable, and Reusable (FAIR) Data." 16th International Conference on Digital Preservation iPRES 2019, Amsterdam, The Netherlands.

[7] Dutch Techcentre for Life Sciences &Czech Technical University of Prague (2016). Data Stewardship Wizard (DSW). Available at https://ds-wizard.org/.

[8] H. Pergl Sustkova et al. , FAIR Convergence Matrix: Optimizing the Reuse of Existing FAIR-Related Resources. In: Special Issue on Emerging FAIR Practices, MIT Press, 2019.

[9] E. Schultes: Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence, (Pre-print v2.0) December 11, 2019, available at: https://mfr.osf.io/render?url=https://osf.io/c4fth/?direct%26mode=render%26action=download%26mode=render

[35] https://doi.org/10.5281/zenodo.3678715

Appendix 1: Glossary

AAI	Authentication and Authorisation Infrastructure
AARC	Authentication and Authorisation for Research Collaborations
API	Application Programming Interface
ARDC	Australian Research Data Commons
B2HANDLE	EUDAT minting, storing, managing and accessing persistent identifiers
CDI	Common Data Index (metadata format and data access system by SeaDataNet)
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CSW	Catalogue Service for the Web
CTS	Core Trust Seal
DANS	Data Archiving and Networked Services
DCAT	Data Catalogue Vocabulary
DMP	1) Data Management Plan 2) Data Management Platform (WP9)
DOI	Digital Object Identifier
DSA	Data Seal of Approval
DSW	Data Stewardship Wizard
EGI	European Grid Infrastructure
EMSO	European Multidisciplinary Seafloor and water column Observatory
ENVRI	Environment research infrastructures (in ESFRI level or upcoming) as a community
ENVRIplus	An environmental RI cluster H2020 project
EOSC	European Open Science Cloud
ERIC	European Research Infrastructure Consortium (legal entity type)
ESFRI	European Strategy Forum on Research Infrastructures
FAIR	Findable Accessible Interoperable Reusable
FIP	FAIR Implementation Profile
FMI	FAIR Maturity Indicator
FORC	Future of Research Communication
FORCE11	The Future of Research Communication and e-Scholarship (gre out from FORC Workshop in 2011)
FTP	File Transfer Protocol
GDPR	General Data Protection Regulation
GO FAIR	An international programme on FAIR implementation
GUI	Graphical User Interface
GUPRI	Global Unique Persistent and Referable Identifier
HTTP	HyperText Transfer Protocol
ICOS	Integrated Carbon Observation System

ICT	Information and Communications Technology
IRI	Internationalised Resource Identifier
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation for Linked Data
LW	LifeWatch
M2M	Machine-to-Machine
NetCDF	Network Common Data Format
OAUTH	Open Authorization (standard)
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OBIS	Ocean Biogeographic Information System
OECD	Organisation for Economic Cooperation and Development
ORCID	Open Researcher and Contributor ID
PID	Persistent Identifiers
PROV-O	Web Ontology Language encoding of the PROV Data Model
RDA	Research Data Alliance
RDF	Resource Description Framework
rdflib	RDF Python library
RDM	Research Database Management
RI	Research Infrastructure
SEADATANET	SeaDataNet pan-European infrastructure for marine data management
SHARC IG	SHAring Rewards and Credit Interest Group
SMTP	Simple Mail Transfer Protocol
SPARQL	SPARQL Protocol and RDF Query Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
UUID	Universally Unique IDentifier
VRE	Virtual Research Environment
WDS	World Data System of International Science Council
W3C	World Wide Web Consortium
YAML	Yet Another Markup Language
YaT	YAML Template

Appendix 2: Questionnaire

Questions in green were used in the analysis chapter, those in light green are related questions, all other information is considered to be integrated in the ENVRI Knowledge Base.

Table 1: RDM+ Questionnaire

	RDM+ questionnaire							
RDM + Nr	**RDM Nr**	**Question**	**Question explanation**	**YAML field**	**yaml line**	**list**	**type**	**FAIR principle**
1	0	Date of response	N/a	date	2		yyy-mm-dd	
2		Version:		version	3		number	
	General							
3	1	Contact name *	N/a	contact name	5		string	
4	2	Email*	N/a	email	6		IRI	
5		Research Infrastructure acronym		acronym	8		ref	
6	3	Research Infrastructure Name *	N/a	infrastructure name	9		string	
7		Research Infrastructure Website		website	10		URL	
8	3.1	Please indicate in which domain your RI is mainly working	marine, atmosphere, etc	domain	11	x	ref	
9	4	Please provide the URL of one of the datasets in scope for your answers	N/a	URL/IRI of dataset	12		URL	
10	5	Please provide the URL to the discovery portal in which the dataset can be downloaded	N/a	URL of discovery portal	13		URL	
	Repositories		Please specify answers for each single respository			x		
11	6	Please provide the URL of the repository you use		URL	15		URL	F4
12		Please provide the name of the repository		name	16		string	
13		Which kind of repository is this?	metadata, data, instruments, vocabularies, sites	repository kind	17	x	ref	
14		How is the	local, domain,	repository	18		ref	

		repository within your Research Infrastructure organised?	central, distributed	allocation				
15	7	Which repository software is being used?	In many cases specific repository software such as D-SPace, Fedora, etc are being used, these can also be home-made	software	19		ref	
				identifier		x		
16	9	Do you use persistent identifiers or local IDs?	PID or local ID	identifier kind	21		ref	F1
17		If you use PID's, which PID system do you use?	DOI, PURLs, Handle, EPIC	system	22		ref	F1
18		Do your identifiers resolve to a landing page?	boolean	landing page	23		bool	F1
19	10	Do you assign identifier manually or automatically?	Often DOIs are assigned manually, but more often PIDs are assigned automatically by scripts.	assigned	24		ref	F1
20	11	Which identifier registration provider do you use?	Popular provider is DataCite for DOIs, and for general Handles local or ePIC services are being used.	provider	25		ref	F1
21	12	Is the identifier described with metadata? According to which schema?	Often repositories use the PID record to store properties about the data and refer to other information such as metadata. Indicate if and how you are using it	includes metadata schema	26	x	ref	F2
22	13	Is the repository certified? If so,	Some centres are applying	certification methods	27	x	ref	

		which methods are used?	Data Seal of Approval certification for example, but many repositories still do not apply certification					
23	14	Are repository policies mentioned at the website? If so, indicate the major ones.	Repositories maintain different kinds of policies such as about persistency, number of automatic copies, openness levels, metadata creation, PIO creation, etc.	policies	28	x	ref	R1.1
24	15	Are your repositories registered in a registry? If so which registry?	An example is the re3data.org registry service.	registries	29	x	ref	F4
25	16	Which persistency guaranties are typically given?	This is much very varying, often no explicit guarantees are given.	persistency-guaranty	30		ref	R1.1
	Access mechanisms							
26	33	How is authentication done?	What are the methods of authentication which are supported by your RI? Examples are Shibboleth paired with eduGain, OAuth, but many other methods are in use.	authenication method	32		ref	A1.2
27	FMI13	Please provide a URL to the description of the Acess Protocol		access protocol URL	33		URL	A1.1
28	FMI14	Does the protocol allow open access?	yes/no	access without costs	34		bool	A1.1
29	34	Do you maintain an own user	In many cases local user	own user database maintained	35		bool	A1.2

		database?	databases are used by repositories to store identities and to pair them with authorization information.					
30	35	Do you use a person identification system in your AAI? Which one?	ORCIDs are widely used for publication purposes, do you make use of ORCID for AAI purposes?	person identification system	36		ref	A1.3 change to R1.2
31	36	What is the major access technology supported?	N/a	major access technology supported	37		ref	A1.1
32	37	How is authorization done?	N/a	authorization technique	38		ref	A1.2
33	FMI16	Authorization is required to access the content of my RESOURCE ID	yes/no	authorization for accessing content needed	39		bool	A1.2
34	38	Which specific licenses do you use for your data?	Do you use for example Creative common licenses or similar?	data licenses in use	40	x	ref	R1.1
35	FMI22	Please provide the IRI for your usage license regarding the content returned from RESOURCE ID (be that data, or metadata):		data license IRI	41		IRI	R1.1
36	39	Are metadata openly available?	Are your metadata openly accessible via some access mechanism?	metadata openly available	42		bool	A2
	Data		Focus here purely on the DATA, next section will be on METADATA.					
37	17	Which are the most popular data types used?	This can be a rich set of types ranging from text, to media recordings, to specific scientific/domain formats, only	data type	44	x	ref	I1, R1.3

			the most relevant should be indicated.					
				preferred formats:		x	ref	
38	18	Which are the preferred data formats?	Probably a large number of formats are being used, simply indicate some major ones	format name	46		ref	I1, R1.3
39	19	Do those formats include metadata headers? if so, which?	Formats such as dicom, jpg, NetCDF, etc. store some metadata in headers which can be extracted.	metadata types in data headers	47	x	ref	I1, R1.3
40	21	Did you register your schemas in a common registry?	Some schemas/formats such as MPEG media files are standardised and well described and are typed by the MIME type registry, other schemas are well described and point to open web-pages maintained by large organisations.	registered data schema	48		ref	I1
41	20	Do you provide search on data?	Some RIs store structured data or texts on which search is being supported.	search on data	49		bool	F4
	Metadata		Now the focus is on the METADATA.					
				schema:		x		
42	22	Please provide the URL of the metadata schema used	Could be databases for metadata, but often XML schemas are defined within the community. Please provide the URLs for	metadata schema URL	52		URL	I2, R1.3

			this.					
43		What is the name of the metadata schema?	And the name of the schema used.	metadata schema name	53		ref	I2, R1.3
44	24	How is provenance included?	Most metadata schemas use some categories or text describing provenance without using standards, but others may use separate W3C PROV categories in separate provenance descriptions. Please indicate what you use.	provenance fields included	54	x	ref	F2
45		Do you provide machine readable provenance information about your data (based on PROV-O or similar)?	yes/no	machine readable provenance	55		bool	R1.2
46	23	Are all categories used in the schemas defined in open registries?	Are the semantic concepts used in the metadata schema well-defined and openly registered so that others can point to them and/or reuse them. Think "Vocabularies" for this question.	categories defined in registries	56		bool	I2
47	25	Are PIDs included in the metadata description?	In general if PIDs are assigned they should be findable in the metadata as a separate field. If possible, provide an example.	PIDs included	57		bool	F3
48	26	What is the primary storage format for	Sometimes spreadsheets and relational	primary storage format	58		ref	I1

		metadata?	databases are in use to store metadata, but others use XML files or are using RDF stores.					
49	27	Which are the export formats supported?	Examples are HTML, XML, JSON, RDF etc.	export formats supported	59	x	ref	I1
50	FMI11	In which searchable resources is your metadata indexed?	yes/no	search engine indexing	60		bool	F4
51	28	Which metadata exchange/harvesting methods are supported?	Common protocol is OAI-PMH, others may already make use of resourceSync or other methods.	exchange/harvesting methods	61	x	ref	I1
52	29	Do you have a local search engine?	Many repositories build their own metadata search engine. Please provide the URL (same as on page 1).	local search engine URL	62		URL	F4
53	30	Do you support external search engines?	Do you publish your metadata to community or higher level search engines?	external search engine types supported	63	x	ref	F4
54	31	Do you make statements about access policies in your metadata?	Do you provide access and license information in your metadata, or is this information available elsewhere?	access policy statements included	64		bool	A1.2
55	FMI18	Please provide the URL to a metadata longevity plan		metadata longevity plan URL	65		URL	A2
56	32	Is your metadata machine actionable?	Do you believe that all your metadata can be processed by machines? For example is "license	machine actionable	66		bool	F2

			information" encoded in a formal language?					
57	FMI7	Please provide the IRI to a document that contains machine-readable metadata for the digital resource	n/a	IRI of machine readable metadata of dataset	67		IRI	F2
	Semantics							
A2	47	Please provide the URL of the semantic vocabulary in use		vocabulary IRI	69		IRI	I2
59		Indicate the vocabulary name		vocabulary name	70		ref	I2
60		What type of vocabulary is it (taxonomy, thesaurus, ontology)?		vocabulary type	71		ref	I1
61		Indicate the vocabulary topic (generic, domain-specific, project-specific)		vocabulary topic	72		ref	I1
62	FMI19	Which vocabulary language is used?		specification language	73		ref	I1
	Data Management Plans							
63	40	Do you use or provide specific DMP tools? If so, which DMP tool are you using or advocating in your community?		specific DMP tools used	75		ref	
64	41	Do you apply special data publishing steps?	Often specific data curation steps are taken before publishing data. Provide specific metadata as required for example by DataCite and create DOis.	data publishing steps applied	76	list	ref	
65	FMI25	Do you use a community compliance validation service for data?	yes/no	compliance validation service	77		bool	R1.3
	Data processing							
66	42	Do you apply	Duplicate	special data	79	x	ref	

		special data [processing] steps?	question?	processing steps applied				
67	43	Do you apply workflow frameworks for processing your data?		workflow frameworks applied	80	x	ref	
68	44	Do you use distributed workflow tools? if so, which?		distributed workflows tools used	81	x	ref	
69	45	Do you offer other type of support or analytics services?		other analysis services offered	82	x	ref	A1
70	46	Do you offer data products in your RI?		data products offered	83	x	ref	
	FAIRness							
				data findability				
71	50	Do you believe that your data is Findable (F)?	See the FAIR specifications. https://www.go fair.org/fair-principies/	data findable	86		bool	
72		Indicate where you see major gaps.		data findability gaps	87	x	ref	
				data accessibility				
73	51	Do you believe that your data is Accessible (A)?	See the FAIR specifications. https://www.go fair.org/fair-principies/	data accessible	89		bool	
74		Indicate where you see major gaps.		data accessibiltiy gaps	90	x	ref	
				data interoperability				
75	52	Do you believe that your data is interoperable (I)?	See the FAIR specifications. https://www.go fair.org/fair-principies/	data interoperable	92		bool	
76		Indicate where you see major gaps.		data interoperability gaps	93	x	ref	
				data reusabilty				
77	53	Do you believe that your data is re-usable (R)?	See the FAIR specifications. https://www.go fair.org/fair-principies/	data reusable	95		bool	
78		Indicate where you see major gaps.		data re-usability gaps	96	x	ref	

Appendix 3: The YAML template

The template provides guidance on:

- The field type:
    - free: free text (literal)
    - URL/IRI: website address (requires to start with 'http://')
    - bool: Boolean (yes/no)
    - date: yyyy-mm-dd

  The cardinality: 'lists' indicates that more than one answer is allowed. The syntax of lists is different from non-list answers (where only one answer is allowed). Lists need indentation and a hyphen at the beginning of the added values for each line e.g. also in case if only one answer is provided.

```
export formats supported:
- NetCDF
- Nasa Ames
```

- The linked FAIR principle (F1, R1.1 etc) to the different questions/attributes (note that not all questions are FAIR related, but sometimes just giving context information). This is used later for analysis purposes in SPARQL queries enabling the grouping of answers.
- ref: request to use a reference list value from the reference list
- The compilation of reference lists was done by respondents of the questionnaire according to their needs in a google document which was quality controlled by EAA and included after some consolidation in the github repository. This on the fly compilation represented a major challenge as it was not always used as requested and revealed to be not very user-friendly. In fact, it should have been provided a priori together with the survey.

Allowed answers:

- If no answer is given: NULL
- If no answer can be given: VOID
- If the answer is 'planned to provide a solution': planned (this is to give the possibility to reflect the status of ongoing developments in RI)
- If a reference list value should be used, but the answer is negative: none
- If the attribute field type is defined 'bool', it is also possible to use 'partially'. This option has been introduced, because often the situation is more complex, than just a clear yes no answer due to the often very distributed structure of RI.

To be able to refer to specific questions in chapter 5, the question number is directly attached to the questionnaire acronym (e.g. RDM4). The reference of the question from the merged questionnaire (RDM+) comes always along with the correspondent YAML attribute (marked with the line number in the template), for instance RDM+17/YaT22.

```
survey:
  date: yyyy-mm-dd
  version: number
  creator:
    name: free
    email:
infrastructure:
  acronym: ref
  name: free
  website:
  domain: list ref
  URL/IRI of dataset:
  URL of discovery portal:
  repositories: list
    URL: F4
    name: free
    kind: list ref
    allocation: ref F4
    software: list ref F4
    identifier: list
      kind: ref F1
      system: ref F1
      landing page: bool F1
      assigned: ref F1
      provider: ref F1
      includes metadata schema: list ref F2
    certification methods: list ref
    policies: list ref R1.1
    registries: list ref F4
    persistency-guaranty: ref A1
    access mechanisms:
      authentication method: ref A1.2
      access protocol URL: A1.1
      access without costs: bool A1.1
      own user database maintained: bool A1.2
      person identification system: ref A1.2
      major access technology supported: ref A1.1
      authorisation technique: ref A1.2
      authorization for accessing content needed: bool A1.2
      data licenses in use: list ref R1.1
      data license IRI: R1.1
      metadata openly available: bool R1.1
    data:
      type name: list ref I1
      preferred formats: list
        format name: ref I1
        metadata types in data headers: list ref I1
      registered data schema: ref I1
      search on data: bool F4
```

```
    metadata:
      schema: list
        URL:
        name: ref I1
        provenance fields included: list ref F2
      machine readable provenance: bool R1.2
      categories defined in registries: bool I2
      PIDs included: bool F3
      primary storage format: ref I1
      export formats supported: list ref I1
      search engine indexing: bool F4
      exchange/harvesting methods: list ref I1
      local search engine URL: F4
      external search engine types supported: list ref F4
      access policy statements included: bool R1.1
      metadata longevity plan URL:
      machine actionable: bool I1
      IRI of machine readable metadata of dataset: F2A
    vocabularies: list
      IRI: I2
      name: ref
      type: ref I1
      topic: ref I1
      specification language: ref I1
    data management plans:
      specific DMP tools used: ref
      data publishing steps applied: list ref
      compliance validation service: bool R1.3
    data processing:
      special data processing steps applied: list ref
      workflow frameworks applied: list ref
      distributed workflows tools used: list ref
      other analysis services offered: list ref A1
      data products offered: list ref
    fairness:
      data findability:
        data findable: bool
        gaps: list ref
      data accessibility:
        data accessible: bool
        gaps: list ref
      data interoperability:
        data interoperable: bool
        gaps: list ref
      data re-usability:
        data reusable: bool
        gaps: list ref
```

Figure 1: The YAML template.