D7.2
Design documents describing FAIR data services provided by selected RIs V2

Work Package	WP7
Lead partner	TIB
Status	Final
Deliverable type	Report
Dissemination level	Public
Due date	30-06-2022
Submission date	10-08-2022

Deliverable abstract

We summarize the current state of the ENVRI-FAIR approach to ICT consultancy for research infrastructure managers and developers, aimed at technology convergence across the community of environmental research infrastructures (ENVRIs) in order to ensure the production of FAIR (meta)data across the community. Of particular focus are the recent developments in FAIR assessment and showcasing how the advanced approach leads to improved insights into FAIR principle implementation in ENVRIs. Furthermore, we review the progress along the range of instruments for consultancy at our disposal and how we have been using them to support ENVRIs in implementing the FAIR principles.

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	Markus Stocker	TIB	30-05-2022
Contributing Authors	Barbara Magagna Keith Jeffery Markus Fiebig Peter Thijsse Daniele Bailo Marc Portier Siamak Farshidi Zhiming Zhao	EAA UKRI/BGS NILU MARIS INGV VLIZ UvA UvA	30-06-2022
Reviewer(s)	Margareta Hellström Christian Pichot	ULUND INRAE	05-07-2022
Approver	Andreas Petzold	FZJ	10-08-2022

DELIVERY LOG

Issue	Date	Comment	Author
V 1.0	23.05.2022	First Draft	Markus Stocker
V 2.0	31.05.2022	Second Draft	Markus Stocker Barbara Magagna Keith Jeffery
	05.07.2022	Comments from Reviewer 1	Margareta Hellström
	05.07.2022	Comments from Reviewer 2	Christian Pichot
V 3.0	30.06.2022	Final version	Markus Stocker

DOCUMENT AMENDMENT PROCEDURE

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

GLOSSARY

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

PROJECT SUMMARY

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

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

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

TABLE OF CONTENTS

D7.2 - Design documents describing FAIR data services provided by selected RIs V2

1 Introduction

With the aim to increase the FAIRness of their (meta)data, the ENVRI-FAIR subdomains - Atmosphere, Marine, Biodiversity/Ecosystem, Solid Earth - and their RIs have each developed implementation plans (Fiebig et al., 2020; Thijsse et al., 2019; Papale, 2020; Rabissoni et al., 2019) that present a roadmap towards FAIR assets and services following the principles outlined in Wilkinson et al. (2016).

ENVRI-FAIR WP7 has been supporting such implementations by closely interacting with RIs to provide targeted support for designing, recommending and implementing solutions to overcome gaps identified by WP5, and for testing and validating the development results beyond RIs at subdomain, cluster and EOSC levels. The focus of T7.1 has been on the interaction with RIs and targeted consultation for selecting technologies, integrating solutions for missing functionality and establishing realistic development plans for FAIR (meta)data services.

As a key aim, WP7 has been working on technology convergence to ensure FAIR (meta)data at subdomain and ENVRI cluster levels. This deliverable describes how ENVRI-FAIR has worked toward harmonising FAIR implementations, specifically the instruments used to support this goal. Based on 3-year FAIR assessment data, we evaluate if technology convergence across the ENVRIs has been attained.

In our earlier deliverable (Stocker et al., 2020) we suggested that to achieve FAIR (meta)data among the ENVRIs, in three key areas for effective sharing and reuse of digital assets, RIs should work on common solutions, i.e. (individual) solutions should converge. The three areas are introduced as follows.

Table 1: The three critical areas and recommendations for common solutions.

Nr	Area	Short Name	Recommendation
1	Catalogue	Cat	ENVRIs must have the ability to present their metadata to a common catalogue through which (meta)data published by each individual ENVRI can be found.
2	Syntax	Syn	(Meta)data published by ENVRIs must have a machine readable format (syntax) compliant with domain-relevant community standards.
3	Semantics	Sem	(Meta)data published by ENVRIs must have unambiguous interpretation (semantic consistency).

In order to find (meta)data published by an RI, the (meta)data must be catalogued, most obviously in a **Catalogue** operated by the RI itself. In order to find (meta)data published by any ENVRI, a common (meta) catalogue is needed. Such a common catalogue has to physically or virtually (through distributed querying) harvest individual RI catalogues. The ENVRI Hub Catalogue addresses this goal by leveraging DCAT.

The second area, **Syntax**, entails that it must be possible for machines to process (meta)data published by RIs using approaches commonly used in the domain. For instance, if in a particular community it is common to publish observational data in NetCDF format then it is effective for RIs to publish such data in this format. Publishing data in another open format would be less effective and it would not be effective to use a proprietary format. This goal is addressed by technology harmonisation, overall in ENVRI and in particular for the same data types (across ENVRI subdomains).

The third area, **Semantics**, implies that (meta)data that ought to be interpreted in a given way is indeed understood in its original meaning by everyone, either humans or machines. For instance, if two RIs each publish an observation about water temperature then the values should be unambiguously interpreted as

measurement of water temperature, by both humans and machines. This implies that the RIs must have common vocabularies or at least use concepts that are mapped to each other through vocabulary alignments. Similar to Syntax, this goal is addressed by technology harmonisation in ENVRI.

As we suggested in our earlier deliverable, in these three areas - (meta)data Catalogue, Syntax and Semantics - RIs should work on common solutions, i.e. aim at technology convergence. The deliverable also discussed the FAIR sub principles in light of these three areas, and proposed recommendations for technology convergence. As an example, for R1.2 - (Meta)data are associated with detailed provenance - we suggested that "ENVRI-FAIR should ensure convergence on how provenance information is represented across ENVRIs and related to (meta)data catalogues." In Section 2, we present the instruments used for consultancy aimed at supporting technology convergence in ENVRI-FAIR. In Section 3, we evaluate whether such convergence can be seen in FAIR assessment evaluation data.

2 Instruments

Six instruments of different nature and positioned on complementary perspectives have been developed and implemented: FAIR Assessments, FAIR Dashboard, Technology Demonstrators, ENVRI-FAIR Task Forces, Subdomain Working Groups, and Training Events.

2.1 FAIR Assessments

ENVRI-FAIR had to provide FAIR assessments of the data and services within the scope of T5.1. 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 requirements collection and analysis can be summarized 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 TFs, use cases and other development activities).

To effectively coordinate the FAIR assessment (Step 1) in all four subdomains and the overarching requirement analysis in WP5 and WP7, the 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 and provided customized templates for WP8-11 (i.e. the subdomain WPs) to perform their FAIR assessment
3. Actively contributed to the workshops organized by the subdomain WPs to support the FAIR 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 FAIR gaps and development plan.

2.1.1 Year 2019

The FAIR questionnaire evolved over time. The *first FAIR assessment* survey (March to May 2019) involved two questionnaires using Google Forms – the RDM questionnaire provided by GO FAIR with 53 questions and the FMI questionnaire with 25 questions based on FAIR Maturity Indicators to assess the compliance with the FAIR principles.

The first analysis immediately revealed that the responses to the questionnaires by the RI representatives were not directly usable for downstream analysis without substantial post-processing and harmonisation of the responses. Thus, we decided to employ a YAML-based approach with more structured data, which is described in D5.1.

Meanwhile, a closer collaboration with GO FAIR led to the establishment of the FAIR Convergence Matrix Working Group[1]. The goal was to create a tool that makes it easier for communities to reuse existing solutions when implementing FAIR or to spot gaps and organize new development projects where necessary (see Sustkova et al., 2020). The Matrix should be a declaration / registration service to make it easy for communities to publicly announce their commitments to the reuse of existing, or to develop novel, FAIR resources (Jacubson et al., 2020; Sustkova et al., 2020).

2.1.2 Years 2020-2021

This collaboration led to the creation of two ontologies: One defined the FAIR vocabulary[2] and the other evolved the matrix approach into the conceptualization of FAIR Implementation Profiles[3].

A **FAIR Implementation Profile** (FIP) (Schultes et al., 2020) is a list of declared technology choices intended to implement each of the FAIR Guiding Principles, made as a collective decision by the members of a particular community of practice. FIPs are captured by means of a questionnaire that supports collecting answers that are themselves machine-readable and FAIR. The tool used for this approach is a customized Data Stewardship Wizard (DSW) instance, named the FIP Wizard[4]. The tool development was first financed by CODATA (in 2020) and later by the ENVRI-FAIR project (in 2021). The FIP Wizard prompts a representative of the community (the Community Data Steward) to provide answers that explicitly profile the FAIR implementation approach of that community. FIPs are published by the FIP Wizard as FAIR (machine-readable) and Open Data, which can then serve as a reference for practical FAIR data stewardship activities conducted by members of that community. FIP publication also encourages FIP reuse and repurposing by other communities, which saves time by avoiding "reinventing the wheel" and simultaneously drives convergence on FAIR implementation choices. Over time, FIPs need to be revised to reflect the evolving needs of the community and the ongoing development of FAIR technologies. The FIP Wizard supports versioning for systematic revisioning, which in turn can provide insight into FAIR-related technology trends.

Implementing the FAIR Principles is based on numerous choices concerning the use of FAIR Enabling Resources, in particular commitments to either domain-relevant standards or infrastructure technologies. These collective decisions compose the FIP, and are made on behalf of that community of practice.

A **FAIR Implementation Community** (FIC) is defined as a voluntary association of people and organizations that agree to adhere to the same FIP. The FIC can be large or small, formal or informal, they can be defined around a research infrastructure or a repository. The FIC is fundamentally important to FAIR and defining the FIC is the beginning of any FAIRification effort.

The FIP Wizard is essentially a questionnaire with 21 questions that ask respondents to identify one or more FAIR Enabling Resources for each of the FAIR Principles, mostly for metadata and data separately. Only F2 and I2 address only metadata and F3 requests a resource that links metadata to data. The questionnaire does not include R1.3 since we interpret the FIP as community specific metadata. A **FAIR Enabling Resource** (FER) is an artifact or service that can contribute to the implementation of the FAIR principles. The FERs can be available or in development. We identified 12 different FER types each serving a Sub-Principle (with the exception of R1.3 which is interpreted as the FIP as a whole, see Figure 1).

[1] https://www.go-fair.org/today/fair-matrix/
[2] https://peta-pico.github.io/FAIR-nanopubs/principles/index-en.html
[3] https://peta-pico.github.io/FAIR-nanopubs/fip/index-en.html
[4] https://fip-wizard.ds-wizard.org/

FAIR Sub-principle	Type of FAIR Enabling Resource	Definition
F1	identifier service	A **service** that provides for any digital object (1) algorithms guaranteeing global uniqueness, (2) policy document that guarantees persistent and (3) resolution of the identifier to machine-actionable metadata describing the object and its location.
F2	metadata schema	A **specification** that defines metadata fields describing attributes of data or other digital objects.
F3	metadata-Data linking schema	A **specification** that provides a unique, persistent, (ideally) bi-directional, machine-actionable link between metadata and the data they describe.
F4	registry	A **service** that indexes metadata and data and provides search over that index.
A1.1	communication protocol	A **specification** that defines how messages are structured and exchanged.
A1.2	authorization and authentication service	A **service** that mediates access to digital objects according to specifed conditions.
A2	metadata preservation policy	A **document** that describes the conditions under which metadata are to be provisioned in the future (maybe part of a data management plan).
I1	knowledge representation language	A language **specification** whereby knowledge can be made processible by machines.
I2	structured vocabulary	A **specification** of uniquely identified and unambigous concepts with their definitions represented preferably using web standards.
I3	semantic model	A **specification** that defines qualified relations between entities describing data or other digital objects using structured vocabularies.
R1.1	usage license	A **document** that describes the conditions under which a digital object can be legally used.
R1.2	provenance model	A **specification** that defines metadata fields describing the origin and lineage of data or other digital objects.
R1.3	FIP as a whole	

Fig. 1: FAIR enabling resource types associated with each FAIR sub principle.[5]

[5] https://osf.io/dsvnk

		2019	2020	2021
AIR	ACTRIS_DVAS	☑	☑	☑
	ACTRIS-Gres FIP	☑	☑	☑
	ACRIS-inSitu	☑	☑	☑
	ACTRIS_ARES	☐	☑	☑
	ACTRIS_CLU_FIP	☑	☑	☑
	ACTRIS-ASC	☑	☑	☑
	IAGOS	☑	☑	☑
	EISCAT_FIP	☑	☑	☑
WATER	ARGO	☑	☑	☑
	EMSO ERIC FIP	☐	☐	☑
	LW marine	☑	☑	☑
	SeaDataNet-CDI	☑	☑	☑
	SeaDataNet-Sextant	☑	☑	☑
LAND	EPOS	☐	☐	☑
LIFE	AnaEE	☑	☑	☑
	AnaEE-Crea	☑	☑	☑
	Danubius	☐	☐	☑
	DiSSCo_FIP	☐	☑	☑
	eLTER-RI	☑	☑	☑
	LWERIC Ecosystem	☑	☑	☑
multi-domain	ICOS FIP	☑	☑	☑
	SIOS FIP	☑	☑	☑
	Total count: 57	17	19	22

Fig. 2: FAIR assessments conducted per year and community.

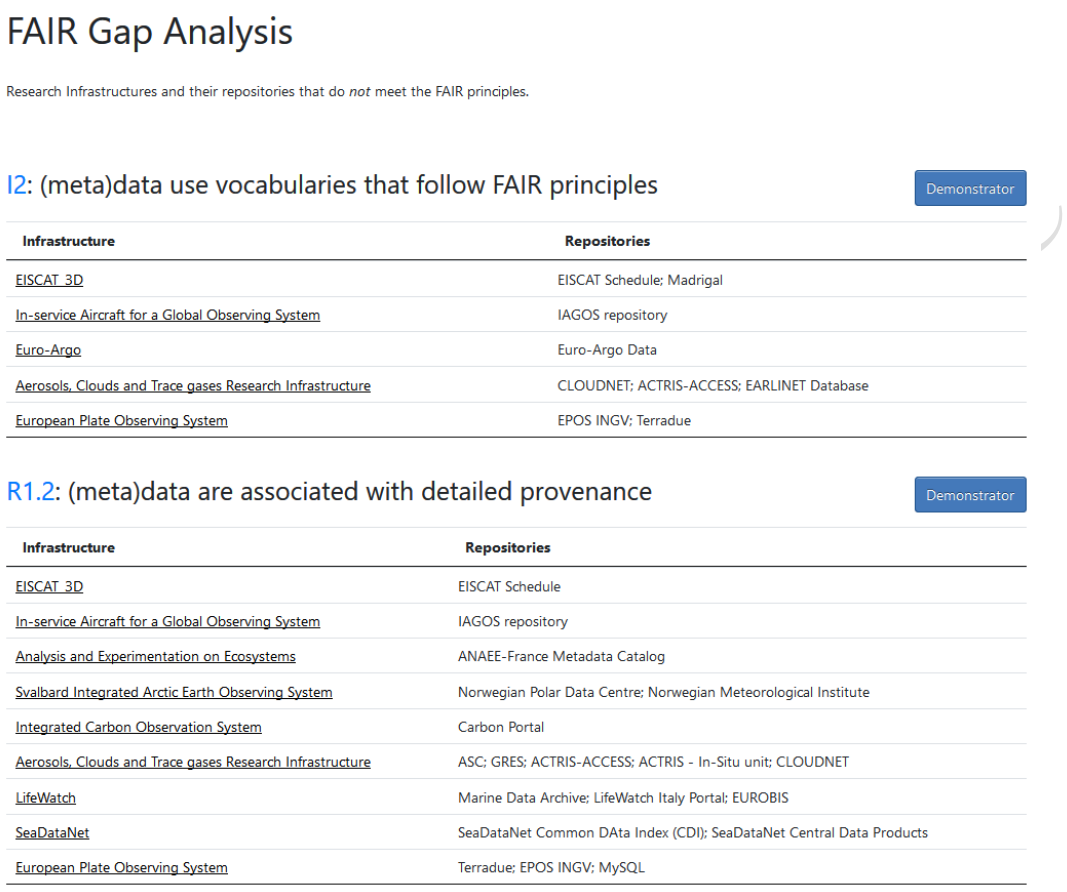
The *second FAIR assessment* of the ENVRI-FAIR project was conducted during the Pre-Symposium Convergence Workshops[6] (October/November 2020) with the involvement of 25 communities, with most of the ENVRIs included. Subsequently GO FAIR started a process to qualify the FERs according to expert judgements and the GO FAIR Foundation (GFF) **qualification** criteria document[7] which includes its own set of explicit interpretations and implementation considerations for each of the FAIR principles.[8] All the FERs are represented and identified as RDF **nanopublications**[9] which are fully expressed in a formal and machine-interpretable manner. The DSW-nanopub-API[10] provides access to all GFF qualified things. The FIP Wizard allows selecting a pre-curated list of FER nanopublications per question, which show a GFF badge in case they were qualified by GO FAIR. A data steward can mint new FERs as nanopublications where needed within the FIP Wizard via specified templates, which are small questionnaires designed to gather the essential metadata about a FER. In addition to FERs, also FICs and FIPs are represented as nanopublications. FIPs describe which resources are used by communities. It is also possible to express the future use of a resource (available/or in development) and the replacement of a resource with a new resource in future.

The *third* round of *FAIR assessment* was conducted early 2022 in the 'FIP for purpose through FAIR Convergence'. The results from the first FAIR assessment were replicated using the FIP questionnaire so that after the third FAIR assessment in January/February 2022 the total number of available FIPs is 57 (Figure 2). A forth FAIR assessment is planned at the end of 2022/beginning of 2023.

[6] https://conference.codata.org/FAIRconvergence2020/sessions/258/
[7] https://www.gofair.foundation/criteria
[8] https://envri.eu/event/fip-workshops-making-the-envris-fip-for-purpose-through-fair-convergence/
[9] https://nanopub.org/wordpress/?page_id=65
[10] https://peta-pico.github.io/tapas/tapas.html?api=peta-pico/dsw-nanopub-api&op=/find_gofair_qualified_things

2.2 FAIR Dashboard

Building on the 2019 FAIR assessment and the resulting data, WP7 had developed a first prototype for a specialized Web-based user interface that supports the discovery of gaps in FAIR principle implementation at the granularity of RI repositories and the discovery of possible technology solutions (Technology Demonstrators) to address such gaps. The user interface is available at the following address https://envri-fair.github.io/knowledge-base-ui/ and Figure 3 is a screenshot of the application.



FAIR Gap Analysis

Research Infrastructures and their repositories that do *not* meet the FAIR principles.

I2: (meta)data use vocabularies that follow FAIR principles — Demonstrator

Infrastructure	Repositories
EISCAT_3D	EISCAT Schedule; Madrigal
In-service Aircraft for a Global Observing System	IAGOS repository
Euro-Argo	Euro-Argo Data
Aerosols, Clouds and Trace gases Research Infrastructure	CLOUDNET; ACTRIS-ACCESS; EARLINET Database
European Plate Observing System	EPOS INGV; Terradue

R1.2: (meta)data are associated with detailed provenance — Demonstrator

Infrastructure	Repositories
EISCAT_3D	EISCAT Schedule
In-service Aircraft for a Global Observing System	IAGOS repository
Analysis and Experimentation on Ecosystems	ANAEE-France Metadata Catalog
Svalbard Integrated Arctic Earth Observing System	Norwegian Polar Data Centre; Norwegian Meteorological Institute
Integrated Carbon Observation System	Carbon Portal
Aerosols, Clouds and Trace gases Research Infrastructure	ASC; GRES; ACTRIS-ACCESS; ACTRIS - In-Situ unit; CLOUDNET
LifeWatch	Marine Data Archive; LifeWatch Italy Portal; EUROBIS
SeaDataNet	SeaDataNet Common DAta Index (CDI); SeaDataNet Central Data Products
European Plate Observing System	Terradue; EPOS INGV; MySQL

Fig. 3: Screenshot of the earlier Web-based user interface for discovery of FAIRness gaps at the granularity of RI repositories and corresponding Technology Demonstrators. Listed are repositories that do not meet the corresponding FAIR principle.

With the additional 2020-2021 FAIR assessments conducted more recently as well as the consolidated resulting data, WP7 developed a second version of the specialized Web-based user interface, now called FAIR Dashboard and available at https://envri-fair.github.io/fair-dashboard/.

While the overall aim has not changed, the new version allows for a temporal analysis of FAIR implementation of RI repositories (now FIP Communities). Figure 4 is a screenshot for principle F1 and the assessed infrastructures. FAIR implementation maturity is color-coded red-to-green, whereby solid green indicates highest maturity. The color is a function of the fraction of declared (FAIR Enabling) Resources that are currently in use (as opposed to their use being merely planned). The higher the fraction the more green and, thus, mature is the implementation of the given principle in the respective infrastructure. The 3-year trend should tend to be green, but as an RI may declare the planned use of additional resources to meet the requirements of a given principle it is entirely possible for a trend to worsen before it improves again (see, for instance, IAGOS in Figure 4). In contrast to the previous version, the newer FAIR Dashboard now includes an overview for all FAIR sub principles, independently of whether or not any Technology Demonstrators are available for the principle.

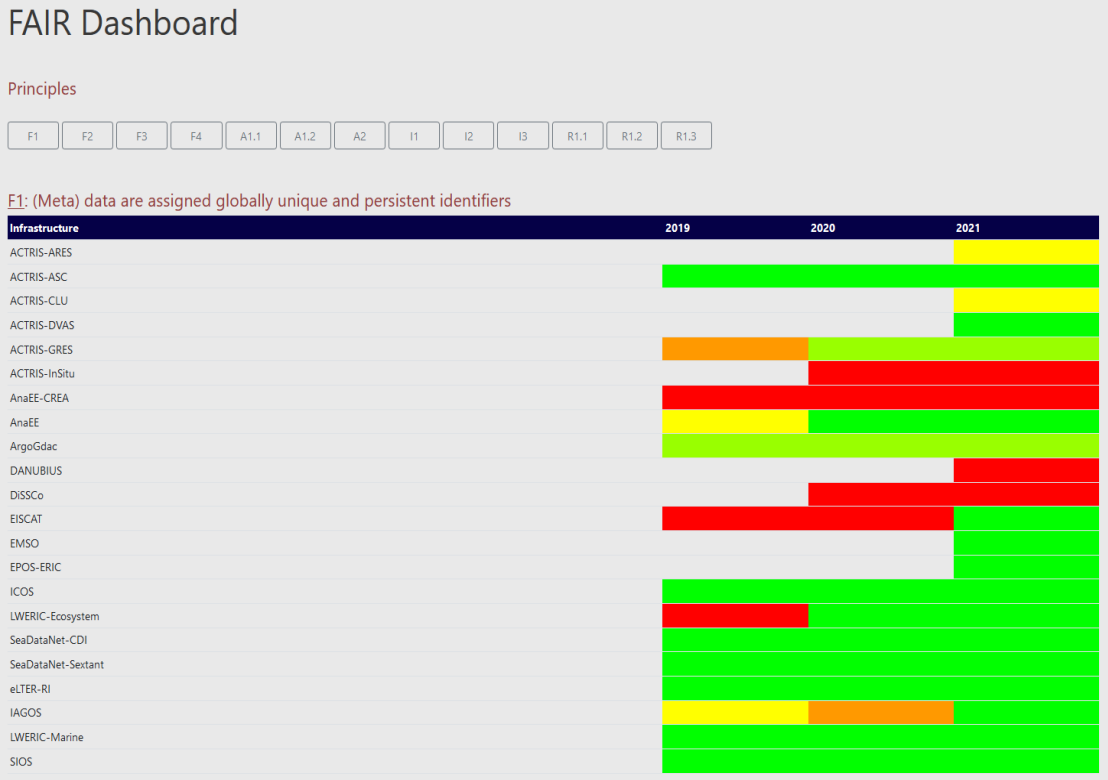


Fig. 4: Overview of the more recent Web-based user interface for FAIR implementation assessment at the granularity of RI repositories for principle F1.

Preliminary feedback from the ENVRI community suggested possible improvements to the FAIR Dashboard, specifically: Including a color legend, using ENVRI colors instead of red-to-green, showing the FIP profile e.g., on hover over the infrastructure and year, as well as testing more compact visualizations (to avoid having to scroll). We will consider these suggestions in future development. Moreover, we plan to evaluate the usefulness of the FAIR Dashboard for RIs by means of a questionnaire.

2.3 Technology Demonstrators

Technology Demonstrators are designed to be low-barrier demonstrations for how selected technologies can support the implementation of FAIR sub principles. As online digital resources, they provide a possible technology solution to address a FAIR principle or possibly an aspect of a principle. They are perhaps the most important, tangible and scalable instrument for consultancy. While Technology Demonstrators can take on various forms, including textual documents, they should allow for execution and hands-on experimentation with the technology. Demonstrators should be intuitive, self-explaining, and easy to follow for anyone interested in the ENVRI community. As such, Technology Demonstrators are primarily implemented as Jupyter Notebooks and should be executable on mybinder.org. They are discoverable via the FAIR Dashboard (https://envri-fair.github.io/fair-dashboard/).

In the context of our earlier deliverable, we had developed two Technology Demonstrators as a proof-of-concept for how this instrument can be implemented and used by the community. The Provenance Demonstrator shows how provenance information can be represented using PROV-O (Lebo et al., 2013) on three example datasets by NILU. The Vocabularies Demonstrator was contributed by BODC and shows how to use SPARQL to query vocabulary published by the NERC Vocabulary Server.

More recently, these two Technology Demonstrators were complemented by demonstrators by IAGOS and SIOS. IAGOS provided an additional and alternative demonstration of provenance implementation using PROV-O and SIOS demonstrates how NetCDF-LD can be used as formal, accessible, shared, and broadly applicable language for knowledge representation for NetCDF formatted data.

In total, we currently have gathered from the community Technology Demonstrators for the following FAIR sub principles: I1, I2, and R1.2. While some principles are easier to demonstrate with simple code snippets than others (say, F1, which relies on access to a PID infrastructure), there is still ample room for improvement and we will continue encouraging the community to deliver demonstrators for other sub principles.

In the context of the ENVRI Hub, WP7 has delivered a demonstration of a CKAN-implementation for the ENVRI Hub Catalogue at the ENVRI week in February 2021. The CKAN prototype was also presented and discussed at the ENVRI-FAIR meeting on March 22, 2021. During this early phase for the ENVRI Hub development, the demonstrator played an important role in shaping future development directions. The prototype catalyzed the discussion and was contrasted with the proposal by Daniele Bailo et al. (EPOS and TF1).

2.4 Task Forces

In 2019, ENVRI-FAIR initiated Task Forces (TFs) to advance shared understanding and development around six themes. From the WP7 point of view, the TFs are understood as instruments for consultancy, platforms for knowledge transfer and for identifying viable technology solutions, and thus support convergence. Of particular interest here is the work and development within TF1 - ENVRI Catalogue of Services; TF2 - ENVRI (VO) AAI implementation; and TF4 - Triple stores and data storage certification. In the following, we summarize the activities in TFs since Q3 2020.

TF1 has continued with its plan as described in the white paper[11]. Essentially, the metadata ingestion pipeline from any native metadata format converted to EPOS-DCAT-AP and thence to CERIF for the actual ENVRI Hub catalog has been utilized. Currently the ingestion pipeline has been tested with 10 RIs providing 28 example digital assets. The rich metadata format ensures FAIRness. A milestone report (M20) provides a template for testing of which the third criterion covers FAIRness.

On the basis of the objectives and key actions described in its white paper[12], **TF2** has continued its work of landscaping and design. TF2 is not aiming at building a centralized ENVRI-FAIR identity provider, which has implications in terms of sustainability and risk of using an external supplier balanced by saving of resources and wider compatibility. Rather, the goals is to negotiate among RIs a protocol that allows RIs to federate access to shared resources. A convergence in this sense has already been found, as Oath2 and OpenID Connect were selected as key standards. TF2 has decided to leverage on previous work done at the European level, in particular by considering the AARC2 Blueprint architecture[13] . TF2 is now tackling the final challenge, namely to set up a proof of concept of AAI integration amongst the most advanced RIs, namely ICOS, EMSO and EPOS.

TF3 is concerned with persistent identification of research objects, including not only data and metadata, but also digital representations of physical entities, such as instruments and samples. With the aim of providing the ENVRI-FAIR members with good examples, the task force has compiled a list of existing "best practices" -- as defined by RIs and projects working in the Earth and environmental sciences -- for assigning persistent identifiers (PIDs) to data. Another ongoing activity is mapping out, by means of recurrent surveys, the trends in usage and application of PIDs in the ENVRI community. Thirdly, TF3 acts as a platform where project member RIs can bring up more general PID-related topics for discussion.

For both triple stores and data storage certification schemes, most importantly Core Trust Seal (CTS), the **TF4** continued to collect experience reports about these technologies from RI representatives who have already used them in production. On triple stores, BODC, ICOS and ANAEE have demonstrated experiences and delivered short experience reports. On certification, IFREMER, BODC and IAGOS have compiled experience reports. These experience reports are useful documentation for those RI representatives that are only planning to employ the corresponding technologies and can leverage the experiences of others in their decision making, possibly learning and ultimately deciding to adopt similar approaches and thus converge on common technology solutions. At the time of writing this deliverable, the following RIs have deployed triple store based services:

[11] https://docs.google.com/document/d/1RiCTzqqmOGvotJHzKhfEqyFAxjSHuv3110Ow9BDov-Q/edit
[12] https://drive.google.com/file/d/1HMb6IhwwsuFG9UFUXo4hPnOpLsN8ww9U/view?usp=sharing
[13] https://aarc-project.eu/architecture/

- IFREMER released the SPARQL endpoint for Cruise Summary Reports[14]
- SIOS released their VocPrez-based Vocabulary Service[15], technologically closely aligned with the NERC Vocabulary Service by BODC
- MARIS released a DCAT-AP SPARQL endpoint for SeaDataNet CDIs[16] (Common Data Index), and IFREMER a Sextant SPARQL endpoint[17] for SeaDataNet products. An SDN product consists of many CDIs and the two sparql endpoints greatly demonstrate how Linked Data works through federated queries can be used for proving the provenance of the products.

Furthermore, the following RIs plan to employ triple store technologies:

- IFREMER a VocPrez-based Vocabulary Service, technologically closely aligned with those by BODC and SIOS
- IAGOS and ACTRIS for provenance metadata.

On certification, IFREMER/Sismer obtained CTS-certification while ICOS and IAGOS have submitted their application and LifeWatch, ANAEE and EISCAT_3D are evaluating whether to submit their applications.

During this phase of the ENVRI-FAIR project, TF4 has also extended its scope. First, it has been actively supporting the development of certain aspects of the ENVRI Hub, in particular with respect to SPARQL Endpoints and DCAT.

TF5 is entitled Licences, Citation, and Use Tracking of Data. It covers several aspects of FAIR enabling technologies in order to achieve the ultimate goal of machine-actionable citation of data, more precisely these aspects:

- Improved metadata schemas for licenses for both data and metadata that indicate entities to receive attribution (license connected) or citation (not license connected) explicitly in a machine-actionable way.
- A concept that enables both competing objectives of data PID identification: 1) accounting of data use, down to the granularity of the individual principal investigator; 2) easy citation of large assembled data collections.
- A concept allowing research infrastructures to do accounting of data use using the research graphs of indexing providers.

Thus, TF5 works on implementing the FAIRness objectives of Re-usability (machine actionable license metadata) and Interoperability (data identification). TF5 has provided a written report, and is presently looking for external partners to implement its recommendations.

In **TF6**, we are designing and developing a novel knowledge management system called ENVRI-KBS to meet the ENVRI research community requirements and make the research assets FAIR for the community. The ENVRI-KBS is a Knowledge-as-a-Service (KaaS) for ENVRI-FAIR research communities to document the development and operation processes of RIs and support them with their engineering and design decisions. In general, the ENVRI-KBS should (1) ingest technical results from ENVRIplus, FAIR assessment, the key sub-domains, and other tasks using a formal language for knowledge representation and proven semantic technologies; (2) provide services and tools to enable RI developers and data managers to browse, search, retrieve and compare RI technical statuses and technical solutions to development problems via available content; (3) provide content management tools for specialists in the ENVRI community to ingest new knowledge and control the quality of content; (4) also provide interfaces to other existing semantic resources, e.g., the service catalog of a future ENVRI Hub, to enhance knowledge discovery and cross-RI search, between knowledge services and the online presence of ENVRI resources. ENVRI-KBS provides an interface for VRE users to search assets (e.g., data sets, Web API, or notebooks) from the community and select relevant ones in a basket. The search engine provides an API for Jupyter users to retrieve research assets from the basket in the cells of a notebook.

[14] https://csr.seadatanet.org/sparql/
[15] https://vocab.met.no
[16] https://cdi.seadatanet.org/sparql
[17] https://sextant.seadatanet.org/sparql/

2.5 Subdomain Working Groups

ENVRI-FAIR subdomains operate their own WGs to advance shared understanding and development around themes of particular relevance to the subdomain. Such WGs are an additional instrument for consultancy to catalyze convergence on technology solutions, especially where convergence matters most.

A good example can be found in WP9 in which the SeaDataNet RI has set up a WG for the FAIRness developments for the CDI Data Access System as well as the Sextant Data Product Catalogue. In both cases the developments concern similar actions, in particular the development of the RDF data model, SPARQL endpoint, and metadata updates (e.g. vocabulary improvement). This work was supported by expertise coming from TFs brought in via the BODC experts and enabled tuning SPARQL endpoint developments. This approach leads to harmonised solutions, in line with the development in the other ENVRIs.

Another such example is the vocabulary WG of the atmospheric subdomain WP8. The WP8 vocabulary WG has not only resulted in a subdomain specific recommendation for common reference vocabularies for observed variables, observation platforms, and quality control metadata, but has also made significant progress towards implementing these recommendations by completing the reference vocabularies. Several WP8 RIs have used the capacity building by WP7 for setting up their own FAIR compliant vocabulary servers:

- ACTRIS: https://vocabulary.actris.nilu.no/, also implementing the InteroperAble Descriptions of Observable Property Terminology (I-ADOPT) RDA recommendation
- IAGOS: https://skosmos.aeris-data.fr/en/
- SIOS: https://vocab.met.no/nb/

2.6 Training Events

Webinars, face-to-face meetings, e.g. at ENVRI weeks, are additional important instruments for consultancy used also to catalyze convergence on technology solutions. COVID-19 has had a severe impact on physical meetings, which have not yet rebounded in ENVRI-FAIR. WP7 contributed to the following with T7.1 consultancy:

- November/December 2020 training on provenance
- ENVRI week February 2021 training on persistent identification of instruments, based on the outcomes (Stocker et al., 2020; Krahl et al., 2021) of the corresponding RDA WG
- ENVRI-FAIR workshop 'Bring your own Provenance Issues' on February 18, 2021
- Technical lectures on services, cloud and DevOps in the ENVRI summer/winter school 2021

3 Analysis

In this section, we briefly analyze the FAIR assessment data for two questions: (1) What level of FAIR implementation maturity has the ENVRI community achieved; and (2) To what extent has technology convergence occurred over the past years.

The first question can be addressed by means of the FAIR Dashboard. Given the 3-year data and the updated functionality of the FAIR Dashboard, now covering all FAIR sub principles, it is now possible to more comprehensively analyze FAIR implementation maturity in ENVRIs. Given that it summarizes the FIP, principle R1.3 is particularly useful for an overall analysis (Figure 5). With only a few exceptions, the data suggest an overall advanced FAIR implementation maturity across ENVRIs and their infrastructures. Moreover, we can observe an increasing maturity trend over the years 2019-2021. These are encouraging results, which suggest that the ENVRI community has made considerable progress over the past years and is well on its way to meeting the important objective of FAIR principles implementation in ENVRIs.

Looking at the FAIR sub principle more granularly, we note that while for many sub principles the implementations are very good, a few principles are lagging behind, most importantly I2 - (Meta)data use vocabularies that follow the FAIR principles and R1.2 - (Meta)data are associated with detailed provenance. This is an expected result, since these sub principles are notoriously difficult to implement.

More surprising is the overall excellent result for sub principle I1 - (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation, which is also non-trivial to implement. Arguably, this principle heavily relies on what exactly is understood as a language for knowledge representation. In a stringent sense, only few technologies fall into this category, most importantly RDF Schema and the Web Ontology Language (OWL). As FAIR enabling resources, RDF Schema is in current use by various infrastructures while OWL is in current use only by one infrastructure. A less stringent interpretation, which is indeed required to apply in practice the sub principle I1 for Earth and Environmental Sciences *data*, allows for other technologies to be considered FAIR enabling resources in regard to principle I1. Indeed, XML Schema is particularly popular for metadata while NetCDF is very popular for data, and both are implemented in a mature manner across ENVRIs. Hence, while non-trivial to implement, ENVRIs show a high level over implementation maturity for principle I1.

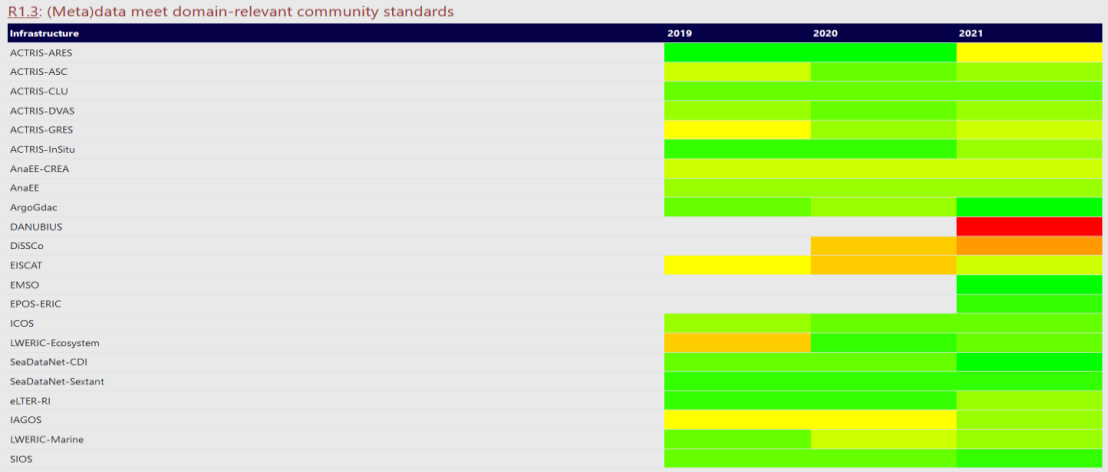


Fig. 5: Overview of FAIR implementation for principle R1.3 which provides an aggregate representation of all other FAIR sub principles across ENVRIs and, thus, an overview of the overall state of FAIR implementation in the ENVRI community.

The second question, namely to what extent has technology convergence occurred over the past years, is also one for which we aim to obtain an answer. The question is important because high technology convergence (harmonisation) is arguably beneficial toward the goal of ensuring interoperability across ENVRIs and, thus, a high degree of interoperability in the ENVRI community. Naturally, interoperability can be achieved also in a heterogeneous ecosystem, but it heavily relies on conversions and alignments. In practical terms, it is easier to retrieve provenance information from multiple ENVRIs if they all expose their provenance information in PROV-O by means of a SPARQL endpoint compared to a situation where the RIs publish provenance information in various formats and data exchange protocols.

To address this question, we used, once again, the FAIR assessment data. We define convergence as the fraction of distinct FAIR enabling resources in current use in ENVRIs for a given year and principle to the total number of FAIR enabling resources in current use for the same year and principle. This is computed for each FAIR sub principle and year. We then compute the difference between the years 2021 and 2019. A negative resulting number indicates a convergence over time. The more negative the stronger the convergence. We can therefore rank the sub-principles according to the strength of technology convergence, with the following result: F3, A2, F4, R1.2, R1.1, R1.3, A1.1, F1, I2, I1, I3, A1.2, F2. Hence, the strongest convergence can be observed for sub principle F3 while for F2 we observe the least convergence.

For principle F4 - (Meta)data are registered or indexed in a searchable resource, with relatively strong convergence, Figure 6 contrasts the number of infrastructures reporting current use of FAIR enabling resources in years 2019 and 2021. The figure suggests that there has been substantial more use of the 7 most used resources in 2019. Particularly notable are Google Dataset Search and DataCite.

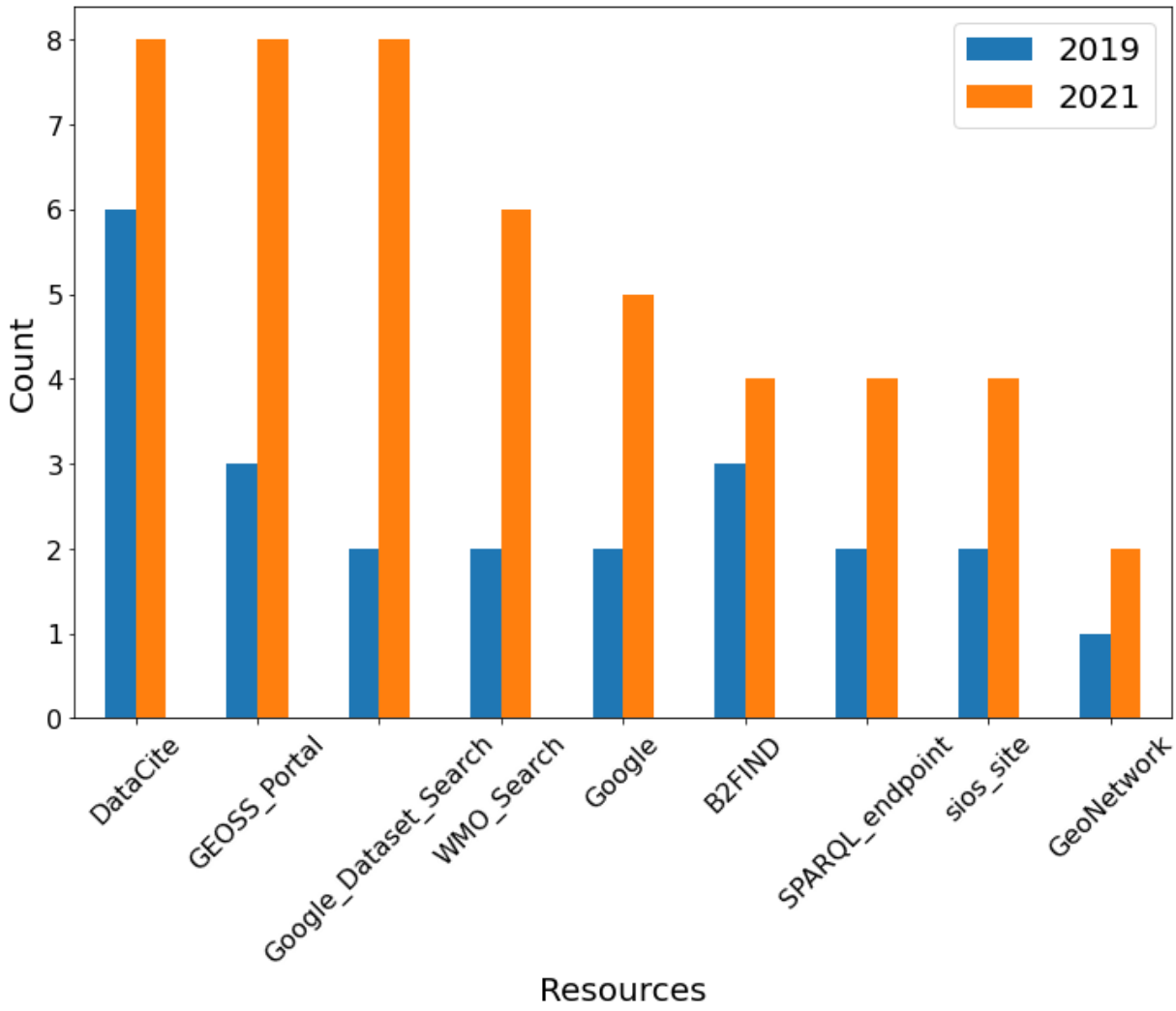


Fig. 6: Contrasting years 2019 and 2021 for the number of infrastructures reporting current use of resources enabling FAIR sub principle F4.

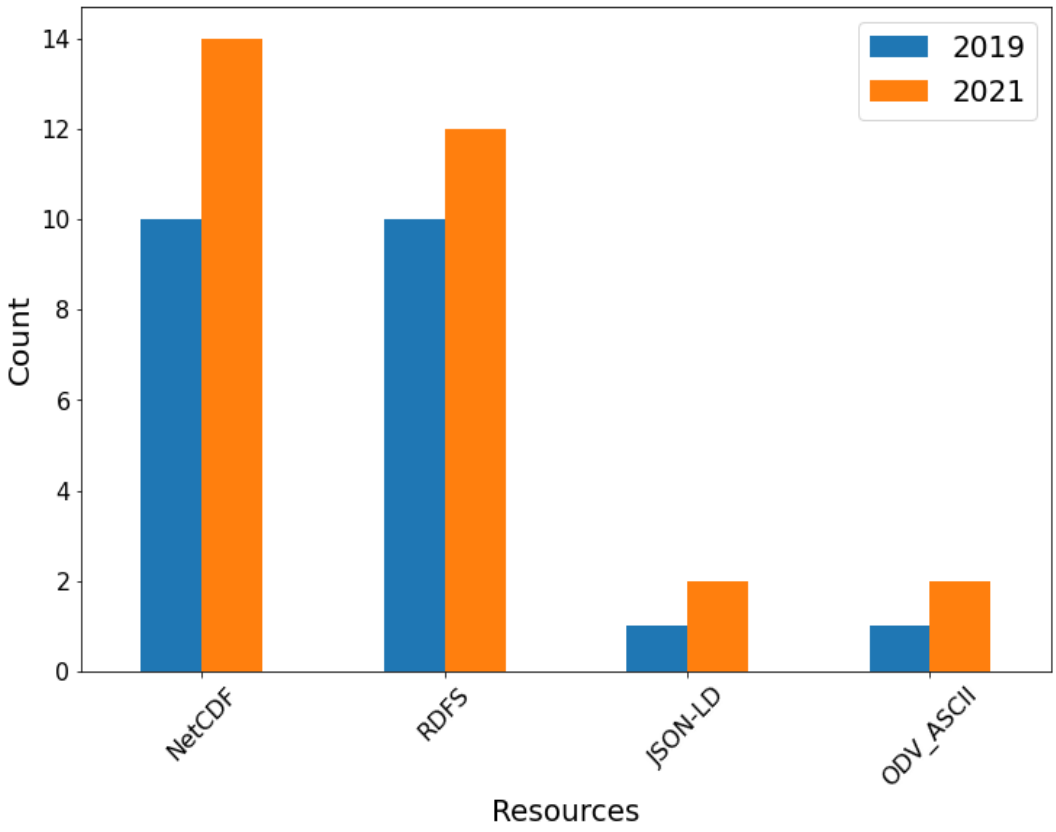


Fig. 7: Contrasting years 2019 and 2021 for the number of infrastructures reporting current use of resources enabling FAIR sub principle I1.

In contrast, for principle I1 - (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation, with weaker convergence, Figure 7 suggests that the use of the resources in use during 2019 has increased less.

Finally, it is interesting to note that while the diversity of FAIR enabling resources in use has increased for all principles except for F3 and I1, only 5 resources that were in use in 2019 are no longer in use in 2021 (EML, INSPIRE, GeoDCAT-AP being the most well known).

The preliminary analysis of convergence reported here was implemented using Jupyter and the code is available at Github[18]. The analysis is executable in the cloud using mybinder.org[19]. Depending on interest from the ENVRI community, we will consider extending the analysis of FAIR assessment data to provide quantitative answers to other questions.

4 Conclusions

We provided an update on the ENVRI-FAIR approach, activities and results of consultancy by ICT experts with research infrastructure managers and developers, aimed at technology convergence across the cluster of environmental research infrastructures (ENVRI) in order to ensure FAIR (meta)data across the cluster.

Building on the earlier work reported in D7.1, we have now been able to assess the FAIR implementation maturity as well as the technology convergence that has occurred in ENVRI-FAIR during the years 2019-2021. The results suggest that ENVRIs have achieved good FAIR implementation maturity over these years which was accompanied by a substantial technology convergence across almost all sub principles.

[18] https://github.com/envri-fair/fair-analysis
[19] https://mybinder.org/v2/gh/envri-fair/fair-analysis/HEAD?labpath=analysis.ipynb

Based on these results, we conclude that the ENVRI community has made considerable advances over the past three years in FAIR implementation as well as in ensuring a high degree of interoperability across ENVRIs.

5 References

Fiebig, M., Lund Myhre, C., Boulanger, D., Rivier, L., Vermeulen, A., Häggström, I., … Tukiainen, S. (2020). ENVRI-FAIR D8.3 Atmospheric subdomain implementation plan (Version 1). Zenodo. https://doi.org/10.5281/zenodo.3885239

Jacobsen A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., Kaliyaperumal, R., Kersloot, M. G., Kirkpatrick, C. R., Kuhn, T., Labastida, I., Magagna, B., McQuilton, P., Meyers, N., Montesanti, A., van Reisen, M., Rocca-Serra, P., Pergl, R., Sansone, S.-A., Bonino da Silva Santos, L.-O., Schneider, J., Strawn, G., Thompson, M., Waagmeester, A., Weigel, T., Wilkinson, M. D., Willighagen, E.-L., Wittenburg, P., Roos, M., Mons, B. & E. Schultes; FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2020; 2 (1-2): 10–29. https://doi.org/10.1162/dint_r_00024

Krahl, R., Darroch, L., Huber, R., Devaraju, A., Klump, J., Habermann, T., Stocker, M., & The Research Data Alliance Persistent Identification of Instruments Working Group members. (2021). Metadata Schema for the Persistent Identification of Instruments. Research Data Alliance. https://doi.org/10.15497/RDA00070

Papale, D. (2020). ENVRI-FAIR D11.1 Biodiversity and Ecosystem subdomain implementation short term plan (Version 1). Zenodo. https://doi.org/10.5281/zenodo.3885360

Rabissoni, R., Jeffery, K., Roquencourt, J.-B., Bailo, D., Grellet, S., Feliachi, A. (2019). ENVRI-FAIR D10.1 Technical analysis and definition of implementation components for FAIR implementation of RIs in the Solid Earth subdomain (Version 1). Zenodo. https://doi.org/10.5281/zenodo.3885335

Schultes, E., Magagna, B., Hettne, K. M., Pergl, R., Suchánek, M., & Kuhn, T. (2020). Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. In Lecture Notes in Computer Science (pp. 138–147). Springer International Publishing. https://doi.org/10.1007/978-3-030-65847-2_13

Stocker, M., Magagna, B., Jeffery, K., Fiebig, M., Thijsse, P., Bailo, D., Liao, X. (2020). ENVRI-FAIR D7.1: Design documents describing FAIR data services provided by selected RIs (V1) (Version 1). Zenodo. https://doi.org/10.5281/zenodo.4061712

Stocker, M., Darroch, L., Krahl, R., Habermann, T., Devaraju, A., Schwardmann, U., D'Onofrio, C., & Häggström, I. (2020). Persistent Identification of Instruments. In Data Science Journal (Vol. 19). Ubiquity Press, Ltd. https://doi.org/10.5334/dsj-2020-018

Sustkova, H. P., Hettne, K. M., Wittenburg, P., Jacobsen, A., Kuhn, T., Pergl, R., Slifka, J., McQuilton, P., Magagna, B., Sansone, S.-A., Stocker, M., Imming, M., Lannom, L., Musen, M., & Schultes, E. (2020). FAIR Convergence Matrix: Optimizing the Reuse of Existing FAIR-Related Resources. In Data Intelligence (Vol. 2, Issues 1–2, pp. 158–170). MIT Press - Journals. https://doi.org/10.1162/dint_a_00038

Thijsse, P., Schaap, D., Exter, K., Vermeulen, A., Pfeil, B., Carval, T., … Rodero, I. (2019). ENVRI-FAIR D9.2 Marine subdomain implementation plan (Version 1). Zenodo. https://doi.org/10.5281/zenodo.3885326

Wilkinson, M., Dumontier, M., Aalbersberg, I. ... (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

6 Appendix 1: Glossary

AAI	Authentication and Authorisation Infrastructure
AARC	Authentication and Authorisation For Research and Collaboration
ACTRIS	Aerosols, Clouds, and Trace gases Research InfraStructure network
ANAEE	Analysis and Experimentation on Ecosystems
API	Application Programming Interface
BODC	British Oceanographic Data Centre
CDI	Common Data Index (SeaDataNet)
CERIF	Common European Research Information Format
CKAN	Comprehensive Knowledge Archive Network
CTS	Core Trust Seal
DCAT	Data Catalog Vocabulary
DSW	Data Stewardship Wizard
EGI	European Grid Infrastructure
EISCAT	EISCAT Scientific Association
EML	Ecological Metadata Language
EMSO	European Multidisciplinary Seafloor and Water Column Observatory
ENVRI	Environmental Research Infrastructures
EOSC	European Open Science Cloud
EPOS	European Plate Observing System
ESFRI	European Strategy Forum on Research Infrastructures
FAIR	Findable, Accessible, Interoperable, Reusable
FMI	Finnish Meteorological Institute
FER	FAIR Enabling Resource
FIC	FAIR Implementation Community
FIP	FAIR Implementation Profile
GFF	GO FAIR Foundation
IAGOS	In-service Aircraft for a Global Observing System
ICOS	Integrated Carbon Observation System
ICT	Information and Communications Technologies
IFREMER	Institut Français de Recherche pour l'Exploitation de la Mer

INSPIRE	Infrastructure for Spatial Information in Europe
KBS	Knowledge-Based System
MARIS	Marine Information Service
NERC	Natural Environment Research Council
NetCDF	Network Common Data Form
NILU	Norwegian Institute for Air Research
OWL	Web Ontology Language
PID	Persistent Identifier
PROV-O	Provenance Ontology
RDA	Research Data Alliance
RDF	Resource Description Framework
RDM	Research Data Management
RI	Research Infrastructure
SDN	SeaDataNet
SIOS	Svalbard Integrated Arctic Earth Observing System
SPARQL	SPARQL Protocol And RDF Query Language
TX.Y	Task X.Y
VRE	Virtual Research Environment
WPX	Work Package X
XML	Extensible Markup Language
YAML	YAML Ain't Markup Language