D8.7
The FAIRness of ICOS

Work Package	WP8
Lead partner	ICOS ERIC
Status	Final
Deliverable type	Report
Dissemination level	Public
Due date	28 February 2023
Submission date	17 May 2023

Deliverable abstract

This report presents the extended implementation of the FAIR principles by ICOS Atmosphere during the ENVRI-FAIR project.

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	Alex Vermeulen	ICOS ERIC – Carbon portal	11 April 2023
Contributing Authors	Leo Rivier	CEA – ICOS ATC	
	Oleg Mirzov	Lund University- Carbon Portal	
	Maggie Hellström	Lund University- Carbon Portal	
	Angeliki Adamaki	Lund University- Carbon Portal	
	Ute Karstens	Lund University- Carbon Portal	
	Andre Bjärby	Lund University- Carbon Portal	
	Jonathan Thiry	Lund University- Carbon Portal	
	Klara Broman	Lund University- Carbon Portal	
	Claudio D'Onofrio	Lund University- Carbon Portal	
	Lynn Hazan	CEA – ICOS ATC	
	Jerôme Tarnowiecz	CEA – ICOS ATC	
Reviewer(s)	Cathrine Lund Myhre	ACTRIS/NILU	8 May 2023
Approver	Andreas Petzold	FZJ	17 May 2023

DELIVERY LOG

Issue	Date	Comment	Author
V 0.1	11 April 2023		Alex Vermeulen
V 0.2	8 May 2023		Alex Vermeulen

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

D8.7 - The FAIRness of ICOS

1 Introduction

The ENVRI-FAIR project's objective is to implement "FAIRness" for data produced in the European Research Infrastructures (RIs) organized in the Environmental Research Infrastructures (ENVRI) community, in order to make them ready for connecting to the European Open Science Cloud (EOSC). In this context, "FAIR" is an acronym comprising the aspects of "Findable", "Accessible", "Interoperable", and "Reusable" as specified by the FORCE11 community.

ENVRI-FAIR WP8 organises and conducts this implementation work for the community of ENVRI RIs in the atmospheric subdomain, comprised of the RIs ACTRIS, EISCAT, IAGOS, ICOS (Atmosphere), and SIOS (atmosphere). Of these Research Infrastructures, ICOS is involved with three sub-domains, i.e. Atmosphere, Ocean and Ecosystems. This document focuses on updates performed in the context of the subdomain Atmosphere, though many of the changes implemented in this subdomain Work Package also are relevant the other subdomains and vice versa.

D8.3 Atmospheric subdomain implementation plan describes the implementation plan of FAIRness for the atmospheric sub-domain and was ready March 2020. Later a revised version of this plane was produced, taking the implementation of the ENVRI-hub into account. This 2nd version was finalized September 2021, and is the most recent implementation plan for WP8.

Towards the end of the project, a set of deliverables are due in Month 50 describing the implementation of FAIRness, for each RI.

- D8.4: The FAIRness of ACTRIS;
- D8.5: The FAIRness of EISCAT_3D;
- D8.6: The FAIRness of IAGOS;
- D8.7: The FAIRness of ICOS-atm;
- D8.8: The FAIRness of SIOS;

This deliverable provides the "The FAIRness of ICOS Atmosphere" (D8.7) and refers to the status of the implementation in relation to the updated, 2nd version of the implementation plan from September 2021 until the end of the project (January 2023).

Furthermore, a FAIRness assessment will be performed in the first part of 2023 to monitor the complete progress over the project period using FAIR Implementation Profiles (FIPs). This is ready in April 2023 and will be reported in the Deliverable "*D8.13 Atmospheric subdomain FAIRness assessment report*" due May 2023.

2 Background and the starting point for ICOS Atmosphere

2.1 Gap analysis

The RIs of the ENVRI atmospheric subdomain conducted a comprehensive analysis of the FAIRness of their data centre, data curation and management late 2018.
ICOS is the Integrated Carbon Observation System that produces Greenhouse gas and related observations relevant for the European carbon cycle from more than 140 measurement stations from 16 member countries, divided over three domains: Atmosphere, Ecosystem and Ocean. ICOS RI became an ERIC in 2015.
The ICOS Atmosphere domain has about 40 stations whose data is gathered and centrally processed by the Atmospheric Thematic Centre.

Compared to the other atmospheric RIs, ICOS as ESFRI landmark research infrastructure already had a relatively advanced data portal that had been designed as a F.A.I.R. and open repository from the ground up.

At the start of the ENVRI-FAIR project the ICOS data portal (Carbon Portal, CP) was already fully operational, with emphasis on machine-to-machine communication and automated metadata and data ingestion and access, topped with a user-friendly interfaces for data discovery, preview and access and higher level data services.

All ICOS data, including collections of data, is identified using Persistent Identifiers (PIDs) using the Handle system. PIDs contain part of the SHA256 checksum of the dataset or the list of PIDs in the data collection, to ensure data integrity. The ICOS data portal is based on semantic web technology. Metadata on the data is stored in a versioned triple store using a custom ontology, defined in owl. All metadata can be queried through an open Sparql endpoint. All Level 2 data (final quality controlled data) and Level 3 elaborated data sets or collections of these data are minted Datacite DOIs, next to the Handle PIDs.

Based on the initial atmosphere sub-domain and other project gap analyses and the 1st FIP, a detailed implementation plan was developed for ICOS Atmosphere, further details can be found in the next paragraph.

2.2 The FAIR implementation plan for ICOS Atmosphere

The gaps identified for improvement of the fairness of the ICOS data repository concern mainly relatively small details and refinements that nevertheless are important for implementing future machine to machine interoperability. Until now the main emphasis in the development had been on pure functionality through a robust, flexible, and modern core (back-end), and achieving a high scalability and performance of the system. Next steps are on core functionalities for the users, for access, preview and attribution of the data providers.
Planned improvements

1. Metadata provision:
    - Metadata tags on landing pages following schema.org to enable for Google data search harvesting
    - Implement OAI-PMH export of metadata to feed higher level data collection metadata into GEOSS
    - Publish the ICOS ontology as controlled vocabulary
    - Implement a ISO19115 landing page (implementing WIS at the same time as WIS is a weak subset of ISO19115)
    - Extend the metadata with related documents, datasets, images, software and other resources
    - Add license information as resolvable URL on the landing pages
    - Extend the provenance metadata to the full ICOS processing chain
2. Data access
    - Implement a ENVRI standard RESTful API for data subsetting
    - Implement the export of ICOS timeseries as NetCDF-CF files

- Updating the current THREDDS (possibly ERDDAP for compatibility with the marine subdomain) service to the latest version and connecting this to the ICOS metadata and data objects

3. User interface
   - Increased search capabilities for spatial extent, conditions and free text keywords

Most of these improvements are further detailed in the following sections of 2.2.

2.2.1 Links to external repositories

The ICOS repository is fully searchable through an open SPARQL endpoint that also allows to access the complete ontology in OWL. Furthermore, there are several graphical web interfaces for human search and access to the data.

Currently there is no connection to WIS for greenhouse gas observations. WMO collects greenhouse gas observations through the World Data Centre for Greenhous Gases (WDCGG) and maintains station metadata trough the OSCAR database. The WDCGG does not have a machine-to-machine interface for metadata and data exchange. It requires now that individual PIs enter manually the metadata and data into their portal and then the data is processed in several manual steps for publication. The WDCGG is still far from being a FAIR repository and therefore publication on WDCGG has little added value.

A request to allow ICOS in the role of recognized WMO GAW contributing network to submit station metadata to WMO OSCAR on behalf of the station PIs is pending.
Another big potential would be to annotate all metadata in schema.org and have it indexed by Google Data Search. Another useful mapping would be to map metadata to ISO 19115 and DCAT V2.0, the latter for inclusion in a future ENVRI Hub data catalogue and GEOSS.

2.2.2 Authorisation

ICOS already supports SAML, Edugain and OAuth2 for login (e.g. OrCID), next to a local login on the basis of the user email address. All of these can be mixed and matched. When the choice for a common AAI scheme is made, it will be integrated in the current system.

2.2.3 Provenance

ICOS provides data provenance metadata based on a simplified Prov-O ontology. The ambition is to include full provenance information in the metadata from observed property to sensor to instrument to site to dataset. ICOS provenance information links each data set to station and operators to enable attribution of the data. The ambition is to include here contributor information including identification (ORCID), institute (ROR) and funding information (project ID).

2.2.4 Licenses

ICOS decided in 2013 to use as its data licence Creative Commons Attribution 4.0 International. Ambition is to also support other licenses, like CC0, moratoria, and support this through machine readable metadata, and choose a data licence for the metadata (CC0).

2.2.5 Semantic search

The Carbon Portal sparql endpoint allows to use semantic search to identify datasets by querying for example by contributor, station, variable name, description or keywords. The faceted search main data portal at https://data.icos-cp.eu/portal/ is a perfect example, where users can select search terms that are listed through sparql queries, to filter down to a selection of datasets from the more than one million datasets in the ICOS portal that fulfil the search criteria. Ambition here is to extend this to allow for free text search.

2.2.6 Graphical User Interface Development

Producing a user friendly and efficient data discovery service that can be used by both new and experienced users is a big challenge. The ambition is to improve the usability of the main data portal GUI Based on user feedback, and add functionalities like visualisation of more data types, two-way interaction between selection on the map and in the faceted search, and integration of collections and DOI information.

Another major upgrade would be to introduce a GUI for the creation of collections, upload of data and manual maintenance of dataset metadata. This will also have to allow for metadata updates and upload of new versions of data and deprecation in a user-friendly way. At the same time this should allow easy transfer of ICOS metadata into the Datacite DOI records associated with the data, also using a graphical user interface for completed and correction of ICOS minted DOI records.

2.2.7 Common APIs

As interface to the ICOS data users can use the linked open data approach and sparql queries, and standard http protocol for metadata and data exchange from machine to machine. Access to metadata and data through http for up- and download is described in the ICOS CP Github repository (https://github.com/ICOS-Carbon-Portal/meta#icos-carbon-portal-metadata-service and https://github.com/ICOS-Carbon-Portal/data#icos-carbon-portals-data-service).

Building on these mechanisms ICOS also provide a python library that provides higher level access to data and metadata. This library is available through pip (https://pypi.org/project/icoscp/).
Through the ICOS Jupyter services (https://jupyter.icos-cp.eu and https://exploredata.icos-cp.eu) this library gives easy and efficient access to ICOS data, but of course data access will work from any machine connected to the internet using the icoscp Python library.

Together with the other atmospheric RIs we decided on developing further the ICOS python library and build on top of this as compatibility layer an interface unit to establish a common PythonAPI endpoint service that when combined with similar interface units from the other RIs provides access to all relevant metadata on stations, variables and data for the subdomain. This endpoint service has been used as the main building block of the Atmosphere Demonstrator.

3 Progress on FAIRness for ICOS Atmosphere

This section includes the progress reports of implementation of FAIRness within ICOS Atmosphere in relation to the implementation plan and identification of potential open issues and need for changes in the implementation plans for each RI.
The next table represents the status of the main tasks identified in the D8.3 Atmospheric subdomain implementation plan. The status is indicated for three consecutive years (end of 2020, end of 2021 and end of 2022).

In the first year of the project ICOS focused on the first priorities for the improvements, which were extension of the metadata profiles to support schema.org and ISO19115 and improved provenance information such as documents and instrument metadata. A major improvement established was an updated metadata exchange mechanism between ICOS Thematic Centres and the Carbon Portal, where in the case of Atmosphere metadata on stations, instrument deployment and person roles is pulled regularly (every few hours) from the ATC server to CP and then synchronised with the metadata in the triple store, keeping track of all changes together with the time of each change.

In the second year we focused on further improvements of the provenance, to include detailed information on instruments linked to the variables and stations. Also a lot of attention has been paid to updates of the graphical user interfaces, like the interactive web in the data discovery portal, data previews and support for netcdf spatial data.

The ICOS General Assembly was asked to decide to define all ICOS metadata as licensed under the CC0 public domain licence, which happened in autumn 2021. We implemented support in the ICOS ontology for other data licenses and now also support CC0, data licence information is now explicitly modelled in the ontology and provided in the data landing pages. We also implemented an optional moratorium for data access limitation.

We changed from Thredds to ERDDAP because of the requirements for the Ocean subdomain demonstrator. All Ocean Level 2 and NRT L1 dataset metadata and data is also available through the erddap service. If users need would arise, also atmosphere data could be published through this Erddap server, but the disadvantages of Erddap next to low performance are that there is no way that it supports the data licence check, data usage counting, citation provision, dynamic metadata support, provenance metadata etc. We added support to add GCMD keywords to dataset metadata.

In the third year we focused on the remaining open issues and further improvements of the graphical user interfaces. We further improved the integration of our Handle PID system with the DOI minting of collections and Level 3 data products in the data upload application, now extended with an improved GUI for data upload and automatic minting of equivalent Datacite metadata of the upload, including geolocation. We also upgraded the metadata model to the latest update of the Datacite metadata model and API and extended support for geo-location and sponsor and project information, using ORCID and ROR persistent identifiers.

We implemented a mapping of our metadata to DCAT V2 in the framework of a demonstrator in Task Force 4. Mapping of metadata to ISO 19115 XML metadata was added to all landing pages, available by content negotiation and a link in the HTML version of the data object landing pages.
The ICOS ontology was published at LOV: https://lov.linkeddata.es/dataset/lov/vocabs/cpmeta

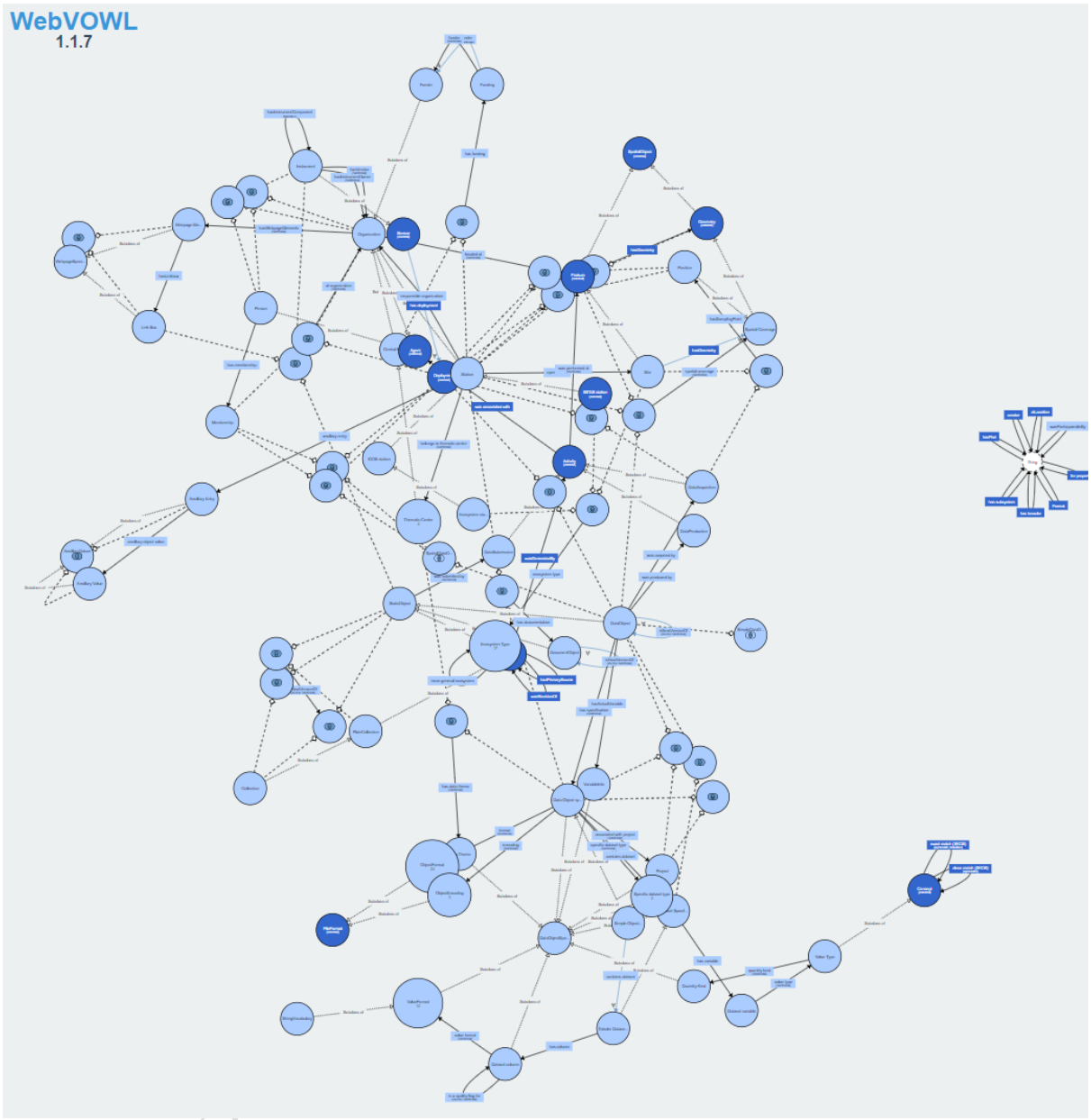

Figure 1 Visualisation of the ICOS cpmeta ontology (viewed through https://service.tib.eu/webvowl/#iri=http://meta.icos-cp.eu/ontologies/cpmeta/)

Table 1: ICOS Atmosphere summary of implementation of FAIRness

ICOS				
Progress on implementaion of FAIRness	**April 2023**			
Completed				
In progress				
Not started, according to plan				
Delay				
Task and Milestones under implementation	**Due month Implementation**	**2020**	**2021**	**2022**
1.1 Use of PIDs throughout workflow				
ORCID & ROR, integration DOI and Handle PID				
1.2 Standard interfaces for (meta)data access				
ICOS: metadata for Google search published	3rd QTR 20			
ICOS: OAI-PMH metadata server implemented	4th QTR 20			
ICOS: publish ICOS controlled vocabulary	4th QTR 20			
ICOS: implement WIS metadata provision	4th QTR 20			
ICOS: implement link to external documents	2nd QTR 21			
ICOS: implement license as resolvable URL	4th QTR 21			
ICOS: full provenance metadata implemented	1st QTR 22			
ICOS: REST API for data subsetting implemented	4th QTR 21			
ICOS: NetCDF-CF data export implemented	4th QTR 20			
ICOS: THREDDS server updated	1st QTR 22			
ICOS: increased search capabilities in interface	3rd QTR 20			
1.3 Data indexing in WIS and GEOSS				
ICOS: link to GEOSS established	n.a.			
1.4 Common authentication schemes				
ICOS: common authentication implemented	4th QTR 21			
Task where implementation plan is needed				
2.1 Domain vocabulary / ontology	3rd QTR 22			
2.2 Documentation of provenance	1st QTR 21			
2.3 Recommendations for licenses	1st QTR 21			
2.4 Semantic search	1st QTR 21			
2.5 Graphical user interfaces	3rd QTR 22			

4 Remaining tasks

In progress:
Adoption of a (sub)domain vocabulary/ontology proved difficult as current available vocabularies are not suitable to describe the ICOS variable metadata rich enough, as they lack the necessary detail, are too specific or are just plainly wrong. We considered several options like the NVS from NERC and the GCMD vocabulary. We applied for a change in cfconventions.org to support the most important ICOS variable through standard name mole_fraction_of_carbon_dioxide_in_dry_air (instead of mole_fraction_of_carbon_dioxide_in_air), equivalent to the total column variable dry_atmosphere_mole_fraction_of_carbon_dioxide, but this change is pending now for two years, despite support by the committee. This might seem like nit-picking, but when we set up a network that is supposed to obtain a network compatibility of 0.1 ppm for CO_2, one cannot live with a variable definition that allows for deviations of several percent. Also many of the other ICOS measurements in all three subdomains do not have a correct corresponding exact match in the cf standard names.

Need revision of plans:
As we see no actual use of WIS and GEOSS in the Atmosphere domain this has received lower priority. The new metadata model from WMO is WMDR2, based on ISO 19156 and extended with i-Adopt, both are draft specifications and under quite heavy development. Mapping from ISO19156/WMDR2 will not be difficult when compared with ISO19115 but should wait until the standard is final. WIS will be replaced by WIS 2.0 in 2025, but ICOS will perform a pilot implementation in 2023-2024.
Next stage is adding a OAI-PMH service that lists our Datacite DOI and ISO19115 metadata records for syncing our data repository to GEOSS, B2FIND and OpenAire (https://explore.openaire.eu/search/dataprovider?pid=10.25504%2FFAIRsharing.p9SgT0).

5 Main achievements and improvement of FAIRness for ICOS Atmosphere

5.1 What can we do now which we could not before advancing of the FAIRness?

The implementation and increase of the FAIRness for ICOS can be demonstrated by many results:

- The interoperability between ICOS data and other Atmosphere RIs data has been demonstrated with the FAIR demonstrator developed by IAGOS. It allows to cross search, visualize, compare and download data from all the RIs from one place.
- Provenance metadata information has been improved considerable, linking dynamically observed properties to instruments and institutes, persons and roles to stations, including also information on supportive documentation, sponsors and funding information.
- Data upload/metadata update now has a user friendly GUI that also provides pre-minted DOI information in Datacite metadata schema for use in the DOI minting GUI app.
- ICOS metadata are now licensed under a clear CC0 licence.

5.2 What would we like to do / achieve within the next 5 years?

Some tasks still need to be completed by ICOS within the next years:

- Implement Keycloak and connect the ICOS Single Sign On system to ENVRI and EOSC AAI
- Implement iAdopt ontology to better describe the observed properties in community vocabularies and publish metadata through mapping to WMDR2 and data through WIS2
- Implement sustainable cross RI services