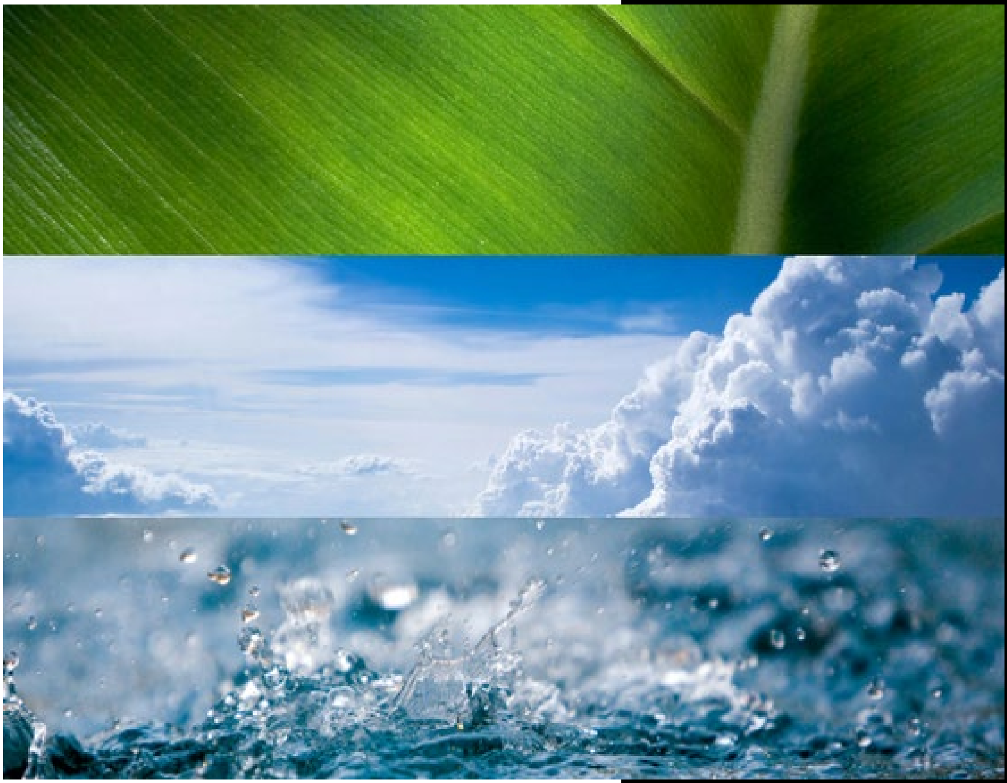
RINGO | Readiness of ICOS

Readiness of ICOS for Necessities of integrated Global Obse tions

D5.5

ICOS improved data lifecycle



RINGO (GA no 730944)

Public Document

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 730944

Deliverable: D5.5 ICOS improved data lifecycle

Author(s): Alex Vermeulen, Oleg Mirzov, Harry Lankreijer, Maggie Hellström, Claudio D'Onofrio, Eija Juurola, Dario Papale, Leo Rivier, Lynn Hazan, Jerôme Tarniewicz, Benjamin Pfeil, Steve Jones
Date: 30 November 2020
Activity: WP5 Task 1
Lead Partner: ICOS ERIC
Document Issue:
Dissemination Level: Public
Contact: alex.vermeulen@icos-ri.eu

	Name	Partner	Date
From	Alex Vermeulen	ICOS ERIC	30/11/2020
Reviewed by	Ville Kasurinen	ICOS ERIC	16/12/2020
Approved by	Elena Saltikoff	ICOS ERIC	22/12/2020

Version	Date	Comments/Changes	Author/Partner

Deliverable Review Checklist

A list of checkpoints has been created to be ticked off by the Task Leader before finalizing the deliverable. These checkpoints are incorporated into the deliverable template where the Task Leader must tick off the list.

Appearance is generally appealing and according to the RINGO template. Cover page has been updated according to the Deliverable details.	x
The executive summary is provided giving a short and to the point description of the deliverable.	x
All abbreviations are explained in a separate list.	x
All references are listed in a concise list.	x
The deliverable clearly identifies all contributions from partners and justifies the resources used.	x
A full spell check has been executed and is completed.	x

DISCLAIMER

This document has been produced in the context of the *project* Readiness of ICOS for Necessities of integrated Global Observations (RINGO) . The Research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730944. All Information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors view.

Amendments, comments and suggestions should be sent to the authors.

ABSTRACT

The Integrated Carbon Observation System (ICOS) provides long term, high quality observations that follow (and cooperatively set) the global standards for the best possible quality data on the atmospheric composition for greenhouse gases (GHG), greenhouse gas exchange fluxes measured by eddy covariance and CO_2 partial pressure at water surfaces. The ICOS observational data feeds into a wide area of science that covers for example plant physiology, agriculture, biology, ecology, energy & fuels, forestry, hydrology, (micro)meteorology, environmental, oceanography, geochemistry, physical geography, remote sensing, earth-, climate-, soil- science and combinations of these in multi-disciplinary projects.

As ICOS is committed to provide all data and methods in an open and transparent way as free data, a dedicated system is needed to secure the long term archiving and availability of the data together with the descriptive metadata that belongs to the data and is needed to find, identify, understand and properly use the data, also in the far future, following the FAIR data principles. An added requirement is that the full data lifecycle should be completely reproducible to enable full trust in the observations and the derived data products.

As part of the RINGO project we defined and started to implement a comprehensive unified metadata flow from Thematic Centres to the Carbon Portal. The design criteria of this system were to integrate as much as possible the operational (legacy) database systems at the TCs with the data portal, thereby preserving the investments in the robust and proven QA/QC and database systems at the TCs and combining these with the benefits of a linked open data system with connected data licence check, usage tracking and dynamic machine operable data and metadata based on a versioned RDF triple store.

Also in the framework of this project we developed a connected DOI minting system, implemented the generation of data collections and a linked system for versioning of the data, all connected to the ontology driven single point of ingestion, optimised for machine to machine communication. This has been used incrementally in full operational mode over the last years and is now in place and used by all ICOS domains for all data streams, from raw data through near-real-time to final quality controlled data, and by the external users that provide elaborated products.

The licence check and data usage tracking has been implemented in a completely unobtrusive way and is flexible enough to be started to interoperate with major data portals like those of FLUXNET, NEON, SOCAT and WMO WDCGG. The use of DOIs increases the exposure of the ICOS data to global and European data portals like the future EOSC portal and current OpenAIRE portal and Google Dataset Search. As will be shown in chapter 4 the ICOS data is already finding it way to many users and through the developments ignited with RINGO, the growing length of the ICOS timeseries in all domains and the interoperation with the global portals this data use of ICOS data can now grow further optimally.

Contents

1 Background of the ICOS data lifecycle

1.1 About ICOS – Knowledge through measurements

The Integrated Carbon Observing System (ICOS) is a Research Infrastructure that has been developed by a community of European scientists to study the carbon cycle in the framework of Climate Change including the influence of human activity and the changing climate on the balance of greenhouse gases of Europe and its surroundings. From a series of partly overlapping national and European projects this has developed into a full blown so called Landmark European Research Infrastructure Consortium (ERIC, established in 2015). ICOS ERIC is now the legal representation for the whole of ICOS Research Infrastructure. ICOS ERIC now has 14 member countries and encompasses more than 150 observation stations distributed over (mostly) Europe and organised in three domains: Atmosphere, Ecosystem and Ocean.

ICOS provides long term, high quality observations that follow (and cooperatively set) the global standards for the best possible quality data on the atmospheric composition for greenhouse gases (GHG), greenhouse gas exchange fluxes measured by eddy covariance and CO_2 partial pressure at water surfaces. Next to these main data products ICOS also produces observations of many ancillary variables, using the same highest quality standards. All measurement methods follow published common specifications and protocols. The data is quality controlled and processed at dedicated central Thematic Centres, one for each domain, using open and published processing chains. The Central Calibration labs provide the stations with working standards and analyse flask samples for greenhouse gas concentrations and radiocarbon ($^{14}CO_2$). All data from raw data up to the final quality controlled (averaged) data is openly accessible from the ICOS Carbon Portal with minimal delays.

Overall, the direct ICOS community involves close to 100 institutions and close to 1000 scientists and technicians. ICOS ERIC is a relatively small entity that governs the complete infrastructure from the ICOS Head Office and takes care of the resulting data in the ICOS Carbon Portal, each involving around 15 persons.

The ICOS observational data feeds into a wide area of science that covers for example plant physiology, agriculture, biology, ecology, energy & fuels, forestry, hydrology, (micro)meteorology, environmental, oceanography, geochemistry, physical geography, remote sensing, earth-, climate-, soil- science and combinations of these in multi-disciplinary projects. Scientists from the ICOS community actively contribute themselves to these scientific works that use ICOS data together with other data to improve our understanding of the carbon cycle. Also the institutes and universities that participate in ICOS RI are involved in these science fields and direct colleagues of the people involved in ICOS are developing models and scientific data products that inform other scientists and policy makers and also use the knowledge acquired to pass on to scholars. This wider community of ICOS data users is also addressed by ICOS (co-)organising and participating to workshops and conferences and coordination of and participation in (inter)national projects. ICOS also provides and gathers feedback from the value chain by taking part in global networks, like FLUXNET, WMO GAW and by representation in initiatives like GEO and UNFCCC.

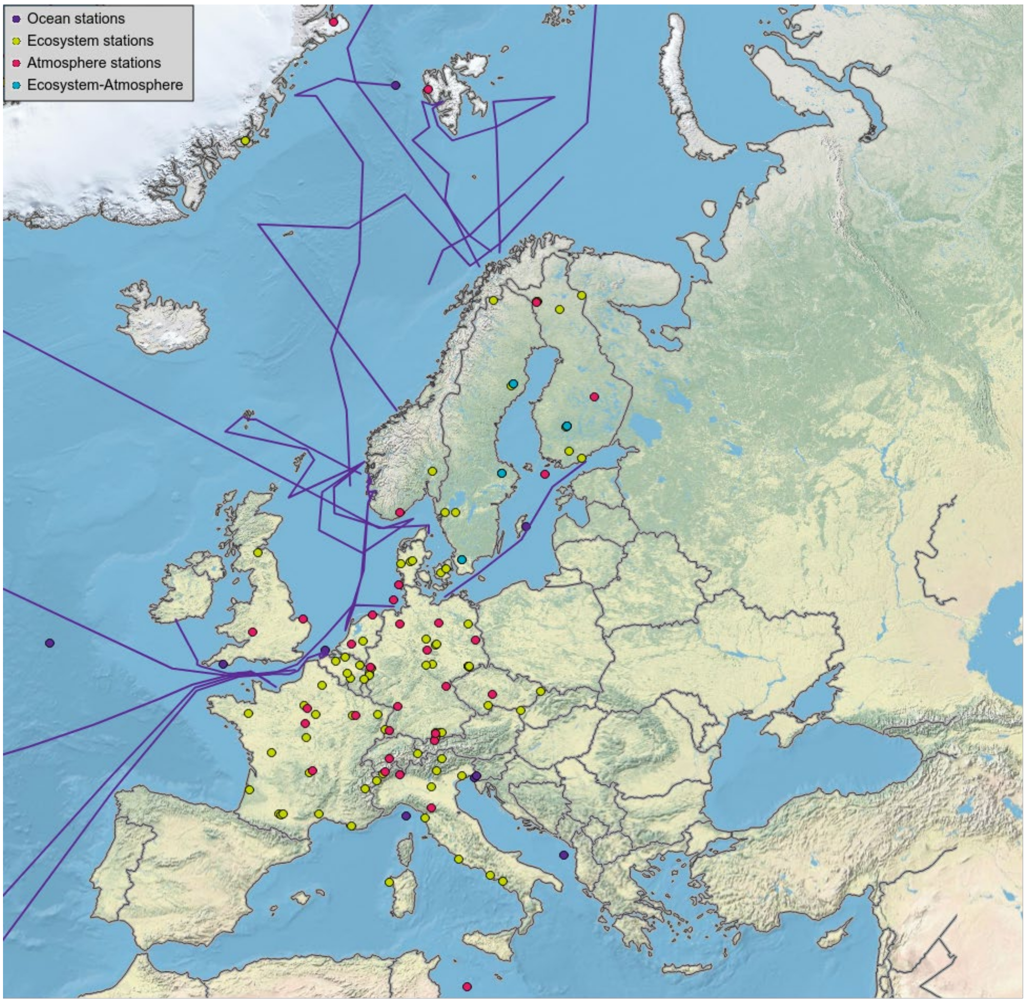


Figure 1 The ICOS observational network (mid 2020) with more than 140 stations, not all are visible in this map. From 2021 onwards the new member countries Spain and Poland plan to add another 15 stations.

1.2 Overview of the three ICOS-RI networks

1.2.1 Atmospheric Observational Network

Each ICOS Atmospheric Station (ICOS AS) is an observatory established to measure continuously the concentration variability of greenhouse gases (CO_2, CH_4) and other trace gases (e.g. CO) due to regional and global fluxes. A site chosen for installing an atmospheric station is typically representative of a footprint area of more than 100 km^2. Additional stations, with a more local footprint, for instance located in areas of high local emissions, are also part of the network. These stations meet the same precision requirements as the main ICOS stations. The ICOS AS is equipped with standardized and approved instruments associated into an "integrated" measurement system, computer controlled with custom-made software. ICOS AS's modular character will allow for different configurations. Two classes of stations exist: Class-1 ICOS AS are equipped with the complete equipment for measuring the full set of ICOS AS core parameters and Class-2 ICOS AS are equipped for measuring a pre-defined subset of the ICOS AS core parameters.

The addition of novel instruments into the existing structure for measuring additional gas species (e.g. N_2O, SF_6, etc.) or replacements of the existing instruments with more advanced ones at a later date may occur.

1.2.2 Ecosystem Observational Network

The ICOS Ecosystem Stations (ICOS ES) are based on instrumentation, partly commercial, embedded in an integrated system for ecosystem monitoring. As the ecosystem monitoring involves human intervention on field activities (such as plant and soil sampling), the ICOS ES follows a set of rigorously standardized protocols developed for the field ecosystem measurements.

The ecosystem network adheres to the monitoring principles of the Global Climate and Terrestrial Observing Systems (GCOS, http://gcos.wmo.int and GTOS, http://www.fao.org/gtos/). These consist of an established set of principles (GCOS Climate Monitoring Principles) that need special attention in their practical execution. They concern detailed measurement protocols, quality control and data management plans for secure long-term operation. The instrument setup and measurement protocols of the ICOS ES follow these guidelines ensuring that the instrumentation yields the observations of comparable accuracy and that the changes in setups are documented and traceable.

The ecosystem network includes two classes of the ICOS ES, referred to as Class-1 (complete) and Class-2 (basic) stations, which differ in the costs of construction, operation and maintenance due to a reduced number of variables measured at the Class-2 stations. This strategy will enhance flexibility and ensures a high level of participation. Nonetheless, as a major characteristic of ICOS RI is standardization and data quality, all ICOS ES, either Class-1 or Class-2, are characterized by a strict standardization of instrumentation and procedures, and consequently the same level of data quality.

There is a possibility to establish ICOS ES Associated sites. The data from these sites are processed within the ETC database. The requirement is to submit at least one full year of data that must include a set of key variables with full description and meta-information, with the acceptance of the ICOS data policy. The associated sites receive ICOS-Associated status.

1.2.3 Ocean Observational Network

The marine network is based on instrumented "Ships of Opportunity (SOOP)", fixed stations like buoys and repeat sections. The VOS are usually commercial ships operating regularly repeated routes, e.g. ferry routes on European shelf and marginal seas, and cargo vessels on open marine routes. For the fixed time series, observations are recorded by means of moorings. These platforms need visits of well-equipped research vessels (like those in EUROFLEET), preferentially 4 to 12 times a year.

The ships and fixed stations are equipped with a suite of automated instrumentation to measure atmospheric and surface ocean pCO_2, sea surface temperature, salinity and related variables. On VOS-lines, measurements are repeated along the same transects at intervals of days to months; they only cover the marine surface.

Repeat-section measurement campaigns are performed on research vessels based on the cruise plan of the respective ICOS participating country, depending on the dynamics of the area. All sections are measured to full depth to follow the changes also in the marine mixed layer, the general mixing as well as the ecosystems and marine current CO_2 transport.

The ocean data lifecycle scheme is optimized for automated sampling from VOS and fixed time series stations. During repeat-section campaigns data is manually sampled and data is being produced, quality controlled in a delayed mode therefore data transmission will be treated slightly different.

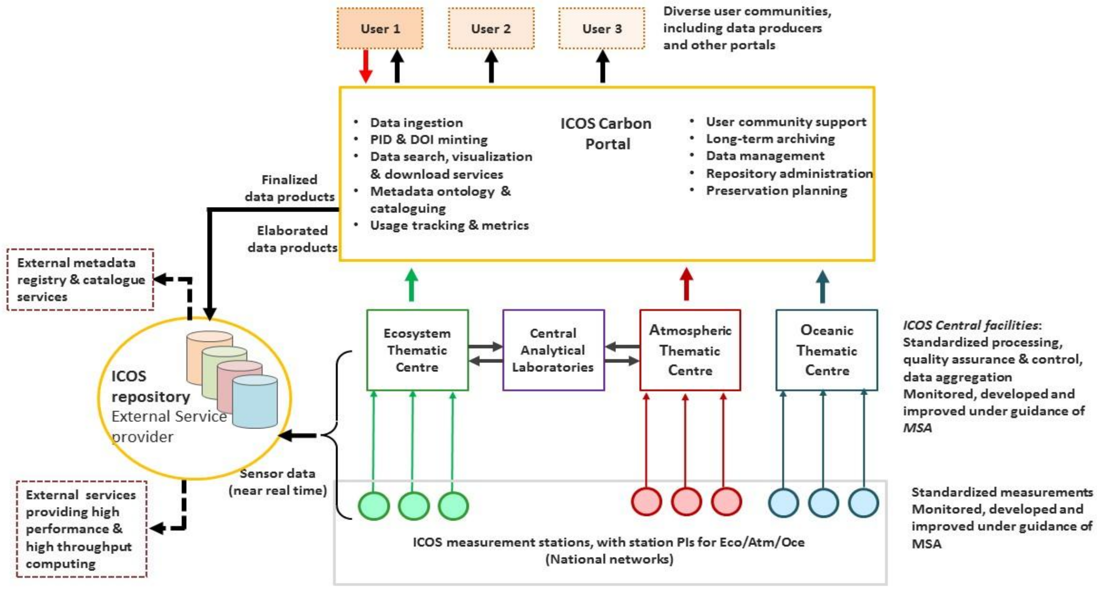

User 1
User 2
User 3
Diverse user communities, including data producers and other portals
Data ingestion
PID & DOI minting
Data search, visualization & download services
Metadata ontology & cataloguing
Usage tracking & metrics
ICOS Carbon Portal
User community support
Long-term archiving
Data management
Repository administration
Preservation planning
Finalized data products
Elaborated data products
External metadata registry & catalogue services
ICOS repository External Service provider
Ecosystem Thematic Centre
Central Analytical Laboratories
Atmospheric Thematic Centre
Oceanic Thematic Centre
ICOS Central facilities: Standardized processing, quality assurance & control, data aggregation Monitored, developed and improved under guidance of MSA
Sensor data (near real time)
External services providing high performance & high throughput computing
ICOS measurement stations, with station PIs for Eco/Atm/Oce (National networks)
Standardized measurements Monitored, developed and improved under guidance of MSA

Figure 2 Overview of the ICOS RI data flow. Arrows indicate the transfer of data and metadata objects between or within sub-communities of the ICOS -RI. Black arrows indicate both data and describing metadata, red indicates only data and broken line arrows only metadata objects that include a pointer to the corresponding dataset

1.3 ICOS data policy and (meta)data flow

The general ICOS data policy principles are described in the **Data Policy Document** (DPD, https://doi.org/10.18160/JDEE-PZNY). The aim of this data policy is to guarantee a smooth data flow from the ICOS RI National Networks, Central Facilities and Carbon Portal to the users, in order to provide open, transparent and easy access to ICOS data, and to develop user-driven services for user communities (e.g. to scientists, national and international agencies, local authorities, general public). To achieve the ICOS RI objectives, the General Assembly of ICOS ERIC has adopted this data policy.

The Data Policy Document defines what *"ICOS Data"* is (see below) and clarifies that *the ICOS ERIC has the right and responsibility to manage the use of ICOS Data that are sub-licensed to the Data Users via Carbon Portal. The ICOS Data management and archiving will be ensured by the Carbon Portal. ICOS ERIC shall oversee that ICOS Data is available with minimal delay, preferably in near real-time, to maximize the value for the Data Users."*

1.3.1 Definition of "ICOS data" and different data levels

"ICOS Data" are quantitative or qualitative attributes of variables or sets of variables that have been gathered by using ICOS RI recommended sensors at validated ICOS stations in an ICOS ERIC member or observer country. ICOS stations are operated at the national level and together they form an ICOS National Network in each ICOS member country. (DPD)

The measurements are standardized due to protocols mutually agreed on by TC and MSA. The responsibility to make the station run according the protocols lies with the station PI: *The principal investigators (PIs) of the ICOS stations are responsible for quality assurance (QA) at the station and the first order quality control of the data. QA protocols developed by the ATC, ETC and OTC in cooperation with the associated MSAs must be used.* (TSD)

1.3.1.1 Raw data

Raw data are information or objects directly obtained from human measurements or automated sensors and having received no transformation since. They can provide a quantitative or qualitative information about physical variables of the environment, and may be of different forms like images, text files, human activities recordings, or physical samples.

1.3.1.2 Level 0 data

Level 0 data are data in physical units either directly provided by the instruments or converted from engineer units (e.g. mV, mA, Ω) to physical units at the TC. They may have been filtered by a quality check (e.g. thresholds).

1.3.1.3 Level 1 data

Level 1 Near Real Time Data (L1_NRT)

NRT data are generally developed for fast distribution using automated quality control within 24 hours after the measurement. NRT data is defined as a high-quality data set that will be distributed in the default way. These datasets have their own provenance metadata that describe the raw data used, versions of the software and scripts, settings, and the results of the automatic quality control.

Level 1 Internal Working data (L1_IW)

Internal Working (IW) data is data that is generated as intermediate steps in the data processing for NRT or Level 2 data preparation and for this reason is not handled as persistent data and not shared outside ICOS RI. The level 1 data is used for internal quality checks like in communication between CF and the PIs. During the production of IW data and following quality checks important provenance information is generated that needs to become part of the provenance metadata of Level 2 data.

1.3.1.4 Level 2 data

Level 2 data is the final quality checked ICOS RI data set, published by the CFs, to be distributed through the Carbon Portal.

1.3.1.5 Level 3 data

All kinds of elaborated products by scientific communities that rely partly or completely on ICOS data products are called Level 3 data. The CP will provide resources to integrate and disseminate L3 products that will be provided by the research community on a voluntary basis and/or, if agreed upon, from collaborative projects.

1.3.2 Rules and responsibilities for data dissemination

1.3.2.1 General rules

As a consequence of the strict requirements on data quality (including the distribution of the best and most recent data version available, and a clear indication and tracking of the versions produced, in order to avoid dissemination of different data versions), data citation and usage tracking (including voluntary user authentication), L1_NRT, L2 and L3 data provision will be coordinated by the CP in collaboration with TCs and MSA where applicable. In principle only L1_NRT, L2 and L3 data will be made available for external users, but at the moment the agreement between all domains is that Level 0 data can only be provided to users on request. L1_IW data are not foreseen for publication.

The basic principle for processing ICOS data is that all data are identified and preserved, together with the information (metadata) that describes the data and the processing steps (curation) applied to the data, so that the whole chain of provenance is traceable, transparent and reproducible. This starts as soon as the raw data is acquired up until the delivery of the final quality assured data to the users. In ICOS the central facilities take care of the essential processing steps between acquisition until the final data, thereby taking care of the central requirements regarding data quality. The Carbon Portal takes care of the commonly shared tasks of the data provenance with regards to data identification, preservation, and external access to the data and metadata, including tracking the data use and acceptance of the license.

Changes in the processing steps will result in different versions of a dataset. Reasons for updating a dataset into a different version will be stored in the metadata for transparency. How versioning of datasets is applied is a subject of international research for standards in with the use of persistent identifier (PID). More details on this are given below on the use of PIDs.

Data access will require the acceptance of a disclaimer and data usage license as well as a voluntary registration/authentication in case a user is interested in information about new versions of the data or the availability of NRT data for a certain period in an equivalent L2 version of higher quality. The disclaimer will explain the differences between NRT and L2 data since they have different quality, to ascertain that they are meeting the needs of the different usage types. It will furthermore explain the requirement for the user to accept that he will use the citation requested when publishing results that depend on the data and to acknowledge that it is suggested to not redistribute (NRT) ICOS data, as they are incremented with time and will be superseded later by L2 data.

NRT data is archived and is minted PIDs (a collection of these is minted a DOI) through the Carbon Portal. Also, the tracking of the NRT data access through downloads and visualization is a task of CP, together with the checking of the ICOS Data License for downloads (through the automated registration, disclaimer, and licensing (RDL) system). The services from CP that enable this allow also for transparently serving of NRT data for download from the TC portals.

After final processing and approval of the data by the TC (and MSA where applicable), the L2 data will be transferred from the TC to the CP. The provision frequency of L2 data is set to at least once per year with the ambition of more frequent updates. The provision of data will differ per thematic network but can also differ within networks as delivery is depending on the measurement platform. The delay for L2 data should be minimized. CP is responsible for providing a digital object identifier (PID), visualisation and providing download facilities including user identification, disclaimer, and licensing system. L2 data will be stored in a trusted external repository.

1.3.2.2 Data licensing and intellectual property rights (IPR)

ICOS data is licensed under Creative Commons Attribution 4.0 International license (CC BY 4.0). In short this means that the user is free to share, copy, adapt, transform and redistribute the material with the conditional term that the user gives appropriate credit, provide a link to the license, and indicate if changes were made. The user 'may not apply legal terms or technological measures that legally restrict others from doing anything the license permits'.

Before ICOS data can be downloaded the user has to accept the data license. The user can choose to register him/herself but can also download the data anonymously. Registered users only have to accept the data license once and can receive information regarding updates on the data.

1.3.3 Persistent identifiers in ICOS - what, why and how?

Identification of data (and associated metadata) throughout all stages of processing is central in any RI. This can be achieved by allocating unique and persistent digital identifiers (PIDs) to data objects throughout the data processing life cycle. The PIDs allow unambiguous references to be made to data during curation, cataloguing and support provenance tracking. They are also a necessary requirement for correct citation (and hence attribution) of the data by end users, as this is only possible when persistent identifiers exist and are applied in the attribution. In short, in today's expanding "open data world", PIDs are an essential tool for establishing clear links between all entities involved in or connected with any given research project

There are a range of different types of persistent identifiers available. ICOS has chosen to primarily work with those built on the Handle system[1], and DOIs (Digital Object Identifiers) from DataCite[2.] For people in ICOS, the Open Researcher and Contributor ID system (ORCID[3]) is strongly recommended. For organisations IDs can be assigned through the ROR system[4] , that is still under development and for example also used by DataCite. Internally CP uses URIs to identify the organisations, which will be mapped to ROR IDs in the DataCite metadata.

1.3.4 Identifying data objects

In principle, one can say that if one can think of at least one situation in which someone will have to make a reference to a given digital object, then it needs to be registered and assigned its own unambiguous and unique identifier. The associated information in the registry needs to contain at least information about where the data object is located (can be a direct pointer to the storage location, or to a so-called landing page), who created it and when it was created. Other useful metadata includes size, checksum and (mime) type of the object.

[1] http://www.handle.net/

[2] https://www.datacite.org/

[3] http://orcid.org

[4] https://ror.org/

DataCite DOIs are identifiers that have been used for scientific articles and reports for over a decade and are therefore well known in the researcher communities. In ICOS, DOIs will be assigned to all collections of published data objects and in some cases to single data objects (NRT, Level 2 and Level 3 - "citable data" in the Figure below), since these are the ones most likely to be referred to, or cited, in scientific contexts. All other data objects that are stored in the ICOS repository, including sensor data, can be considered as "raw data" or "referable data" in the Figure. These will instead be assigned Handle PIDs. These PIDs are just as unique and persistent as the DOIs and could very well be used for citing data in articles and reports - but they are primarily used for referencing data objects in workflows and provenance records.

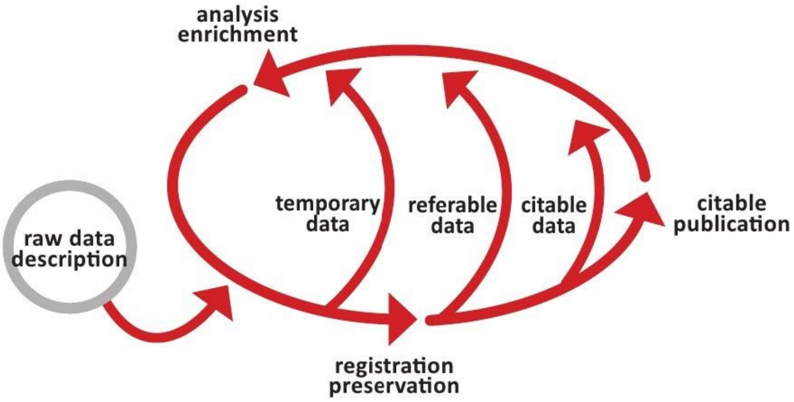


Figure 3 Figure from Ulrich Schwardmann, see http://doi.org/10.5281/zenodo.31785

1.3.5 Versioning of datasets

How to use PID's in relation to different versions of a dataset, or dynamic datasets where new data is continuously added, is described in deliverable 6.1 of the ENVRIplus project, available at the ENVRI website[5].

1.3.6 Identifying people: ORCID

ICOS is built on a large number of distributed observational or experimental sites, run by hundreds of scientists and technicians, financially supported and administrated by a large number of institutions. Since the data from ICOS is shared under an open access policy it becomes therefore very important to acknowledge the data sources and their providers. This process is greatly simplified when also people (PIs & site personnel, and thematic centre staff) and their organizations can be unambiguously identified.

1.3.7 Trusted Repositories

Storage of data and being searchable through metadata is developed in cooperation with so called Trusted Repositories (TR). The development was concentrated in the EUDAT 2020 project, where systems and services have been developed that can be used by a variety of research infrastructures. For the moment ICOS makes use of the services from the EUDAT CDI (now part of the EOSC offerings

[5] http://www.envriplus.eu/wp-content/uploads/2015/08/D6.1-A-system-design-for-data-identifier-and-citation-services-for-environmental-RIs.pdf

from the e-infrastructures) known as B2SAFE and B2FIND. B2SAFE is the system for storage of data and B2FIND for searchable metadata pointing to the datasets, see chapter 4.

Observations are performed in the measurement stations, following the published protocols and descriptions defined by the community and the Thematic Centres. All the methods are based on and partly set the global standard for these long-term high-quality observations. The raw data (Level 0) is transferred to the Thematic Centres that quality control the data and process it to produce Near Real Time (with maximum 24 h delay) Level 1 data and after inspection by the data providers process the data to final quality controlled (Level 3) data. The raw data is transferred either directly to Carbon Portal or through the Thematic Centre. At Carbon Portal all data files from all sources are ingested at real time and minted persistent identifiers, while at the same time streaming the data to the trusted repository. Working copies of the data are kept at stations, Thematic Centres, and Carbon Portal. The Central Analytical laboratory (meta)data always flows through the Thematic Centres and is combine with the data flow from the TCs.

1.4 Data portal design criteria

As ICOS is committed to provide all data and methods in an open and transparent way as free data, a dedicated system is needed to secure the long term archiving and availability of the data together with the descriptive metadata that belongs to the data and is needed to find, identify, understand and properly use the data, also in the far future. An added requirement is that the full data lifecycle should be completely reproducible to enable full trust in the observations and the derived data products. Another important requirement is to make sure that the ICOS data is licenced to the users and that this licence is passed to all users in the value chain. Part of the licence is that the ICOS data should be properly attributed to the data providers.

The requirements for the data portal were gathered by and from the ICOS community and laid down in the Carbon Portal Concept paper[6] that was approved by the ICOS Interim Infrastructure Committee in 2014. This was even before and partly in parallel with the formulation of the FAIR data principles by FORCE11 (see section 2.1), but in fact many of the concepts and implemented details of the Carbon Portal are well in line with the FAIR principles.

Design requirements:

- Support a completely transparent data life cycle
- Open access
- User friendly user interfaces
- Rich metadata model that support the community standard(s)
- Metadata included attribution information with regards to the data providers
- Licence condition check at every download
- Data usage is tracked
- Use trusted repositories for long-term storage

Implementation features:

- Persistent identification of all data objects
- Maximum granularity of the data
- Support for collections

[6] Carbon Portal concept paper

- Full versioning of data
- Linked open data system
- Ontology based metadata store in RDF
- Services are designed as container modules, ready for cloud deployment, upscaling and (fail-) safe deployment with clear separation of functions and dependencies

Further considerations

- Open source software only, with minimal dependencies on proprietary code and a minimal dependence on proprietary protocols
- Using only mainstream and modern application frameworks and libraries with good (community) support, well maintained code and broad use
- Object oriented programming with strict typing
- Strict separation of back-end and front-end
- Regular perform refactoring of code
- Regular update of code libraries to keep them up to date and safe
- All code from ICOS Carbon Portal is published and versioned in an open code repository (GitHub) with GPL v3 licence
- The ICOS ontology of the metadata will describe the ICOS data and elaborated data products with minimum dependencies on external vocabularies to avoid reasoning errors and logical conflicts.
- The ICOS ontology will be mapped to the community and worldwide standards by mapping the equivalences. This more flexible and targeted approach will avoid the 'pushing the elephant through a straw' dilemma and will be more efficient and pragmatic than implementing (different) standards from the beginning.

Development of the code is following in general the agile methodology, with step by step improvements and implementation of functionalities based on user feedback and product owner specification. As the developer team is small (3 FTE) and the ambitions are high, a pragmatic development path has to be chosen, with clear priorities for an operational back-end metadata store and data ingestion and open data access backbone, followed by a user-friendly front-end user interfaces on top.

2 FAIRness and trust

2.1 The FAIR principles and ICOS

The FAIR acronym and concept stand 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[7] but the term FAIR was already conceived at the Lorentz conference in 2014 by the FORCE11[8] Group. Already 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'[9] and later in 2013 in G8 Science Ministers' statement[10], 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.

The FAIR Guiding Principles article (7) had a remarkable resonance and stimulated broad adoption. On the other hand, because the paper did not specify how the FAIR principles should manifest, there is space for diverging interpretations inducing partially incompatible implementations.

Some of the original authors of the FAIR principles intentionally clarified 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 do not 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

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

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

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

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

quantitatively how far the principles are met. Some of the most representative methodologies are described in the following section.

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

Table 1 The FAIR guiding principles as published by FORCE11

2.2 CoreTrustSeal

National and international funders are increasingly likely to mandate open data and data management policies that call for the long-term storage and accessibility of data.

If we want to be able to share data, we need to store them in a trustworthy data repository. Data created and used by scientists should be managed, curated, and archived in such a way to preserve the initial investment in collecting them. Researchers must be certain that data held in archives remain useful and meaningful into the future. Funding authorities increasingly require continued access to data produced by the projects they fund and have made this an important element in Data Management Plans. Indeed, some funders now stipulate that the data they fund must be deposited in a trustworthy repository.

Sustainability of repositories raises several challenging issues in different areas: organizational, technical, financial, legal, etc. Certification can be an important contribution to ensuring the reliability and durability of data repositories and hence the potential for sharing data over a long period of time. By becoming certified, repositories can demonstrate to both their users and their funders that an independent authority has evaluated them and endorsed their trustworthiness.

Nowadays certification standards are available at different levels, from a core level to extended and formal levels. Even at the core level, certification offers many benefits to a repository and its stakeholders.

Core certification involves a minimally intensive process whereby data repositories supply evidence that they are sustainable and trustworthy. A repository first conducts an internal self-assessment, which is then reviewed by community peers. Such assessments help data communities—producers, repositories, and consumers—to improve the quality and transparency of their processes, and to increase awareness of and compliance with established standards. This community approach guarantees an inclusive atmosphere in which the candidate repository and the reviewers closely interact.

In addition to external benefits, such as building stakeholder confidence, enhancing the reputation of the repository, and demonstrating that the repository is following good practices, core certification provides several internal benefits to a repository. Specifically, core certification offers a benchmark for comparison and helps to determine the strengths and weaknesses of a repository.

CoreTrustSeal offers to any interested data repository a core level certification based on the DSA–WDS Core Trustworthy Data Repositories Requirements catalogue and procedures. This universal catalogue of requirements reflects the core characteristics of trustworthy data repositories and is the culmination of a cooperative effort between DSA and WDS under the umbrella of the Research Data Alliance to merge their data repositories certifications.

In 2019 ICOS Carbon Portal applied and was selected by the FAIRsFAIR project to receive support to apply for the CoreTrustSeal certificate. It is expected that the application will be submitted at the end of 2020.

In the CoreTrustSeal application the repository is supposed to provide information and documentation with regards to 16 questions in 4 categories:

2.2.1 Organizational Infrastructure

Mission/Scope

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

Licenses

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

Continuity of access

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

Confidentiality/Ethics

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

Organizational infrastructure

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.

Expert guidance

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

2.2.2 Digital Object Management

Data integrity and authenticity

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

Appraisal

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

Documented storage procedures

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

Preservation plan

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

Data quality

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.

Workflows

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

Data discovery and identification

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

Data reuse

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.

2.2.3 Technology

Technical infrastructure

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.

Security

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

After submission of ICOS ERIC application to CoreTrustSeal the answers formulated on the questions above will be placed in an appendix to this document.

3 Unified ICOS metadata and data ontology at ICOS CP

3.1 A brief introduction to ontologies, RDF and OWL

In the last decades, the use of ontologies in information systems has become more and more popular in various fields, such as web technologies, database integration, multi-agent systems, Natural Language Processing, etc.

There are several types of ontologies. The word "ontology" can designate different computer science objects depending on the context. For example, an ontology can be:

- a thesaurus in the field of information retrieval or
- a model represented in OWL in the field of linked-data or
- an XML schema in the context of databases
- etc.

It is important to distinguish these different forms of ontologies to clarify their content, their use and their goal. It is also needed to define precisely the vocabulary derived from the word ontology. In eScience often the terms "controlled vocabulary" and "community standards" are used. Ontologies are a perfect way to formally describe these and make them available in a machine interpretable and interoperable way.

The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web. RDF is a recommendation from the W3C for creating meta-data structures that define data on the Web. RDF is used to improve searching and navigation for Semantic Web search engine (Web 3.0 applications).

RDF is composed of Triples: (1) the subject (the web page), (2) a property or predicate (an attribute name) and (3) an object (the actual value of the attribute for the web page).

1. The subject is a resource. Resource is anything that can have a Unique Resource Identifier (URI); this includes all the world's web pages, as well as individual elements of an XML document.
2. The property is a resource that has a name. For example the Dublin Core Metadata Initiative propose to use the name "dc:creator" to represent the author property. Property can be associated to a property type defined in an RDF Schema (RDFS). RDFS defined a RDF vocabulary composed of property type and resource type.
3. The object can be a URI, a literal (a string of character representing a number, a date, a noun etc...) or a blank node.

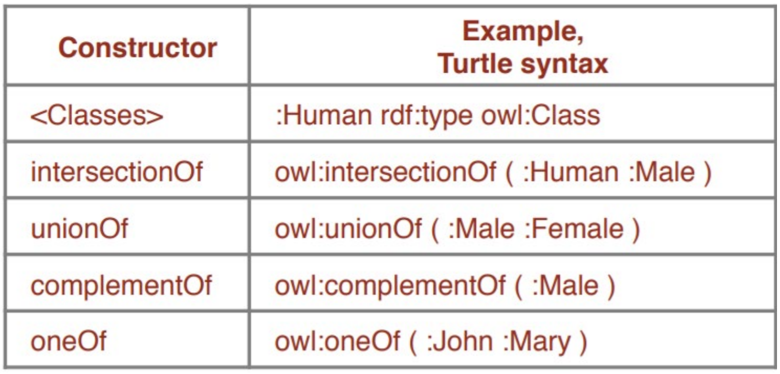
[see http://www.w3.org/RDF/ for more details].

The OWL Web Ontology Language is a standard recommended by the W3C. It is designed for use by applications that need to process the content of information instead of just presenting information to humans. OWL facilitates greater machine interpretability of Web content than that supported by XML, RDF, and RDF Schema (RDF-S) by providing additional vocabulary along with a formal semantics. The OWL is intended to provide a language that can be used to describe concepts and relations between them that are inherent in Web documents and applications. OWL language is used for:

1. formalize a domain by defining concepts called classes and properties of those classes,
2. define instances called individuals and assert properties about them,

3. reason about these classes and individuals to the degree permitted by the formal semantics of the OWL language.

One of the most powerful features of OWL is that it can be represented in RDF and thus can be stored and exposed in the same way as the ontology that it describes. This way a complete RDF database and its description as a formal ontology can be made machine readable and interpretable.



Constructor	Example, Turtle syntax
<Classes>	:Human rdf:type owl:Class
intersectionOf	owl:intersectionOf (:Human :Male)
unionOf	owl:unionOf (:Male :Female)
complementOf	owl:complementOf (:Male)
oneOf	owl:oneOf (:John :Mary)

Figure 3.1 Example of some OWL constructors

3.2 The ICOS CP data and metadata handling

3.2.1 Design and philosophy

The philosophy of CP is to treat all data objects equal and preserve the complete integrity of all data objects, so the actual data is never touched or changed up to the bit level. This goes for all data levels, i.e. from raw data, NRT data, final data quality-controlled data up to elaborated data products. CP strives for the maximum granularity of Data Objects.

The metadata that accompanies the data objects is maintained in a versioned so-called RDF triple store, following the Web 3.0, linked open data approach.

3.2.2 Data object handling

Before ingestion CP requires the uploader to calculate the SHA256 checksum of the data object. All ingestion data transport uses standard http(s) put and get methods and can be invoked by for example using the curl program. In the first stage of ingestion the uploader informs through a small metadata packet in JSON format of the object specification and the checksum of the data object together with some minimal provenance metadata that informs on the uploader, the spatial and/or temporal coverage that the data relates to for as far as applicable and depending of the object specification also on other important information like station, measurement level and instrument ID.

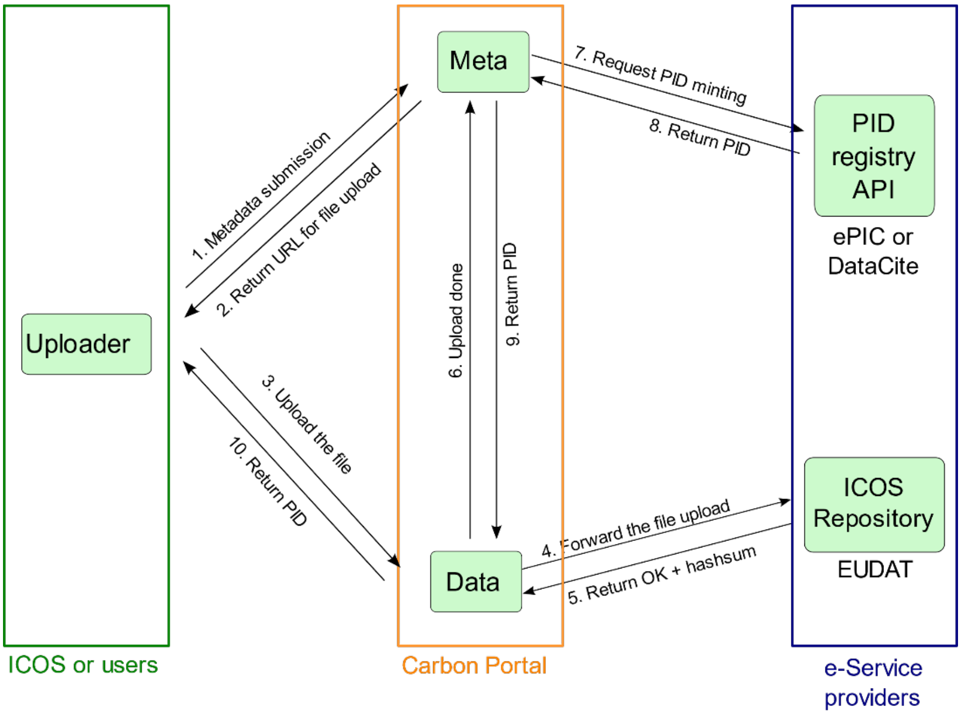


Figure 3.1 Schematic diagram of the data ingestion process at Carbon Portal. In 10 steps the new data is registered, ingested, minted a PID and stored together with the relevant metadata in the ICOS repository and the trusted repository.

Only objects with a known and registered Object Specification type are accepted. After successfully registering in this first step the user can start uploading the data object. While the uploader streams the data to CP, the data is forked and streamed at the same time to the B2SAFE trusted repository.

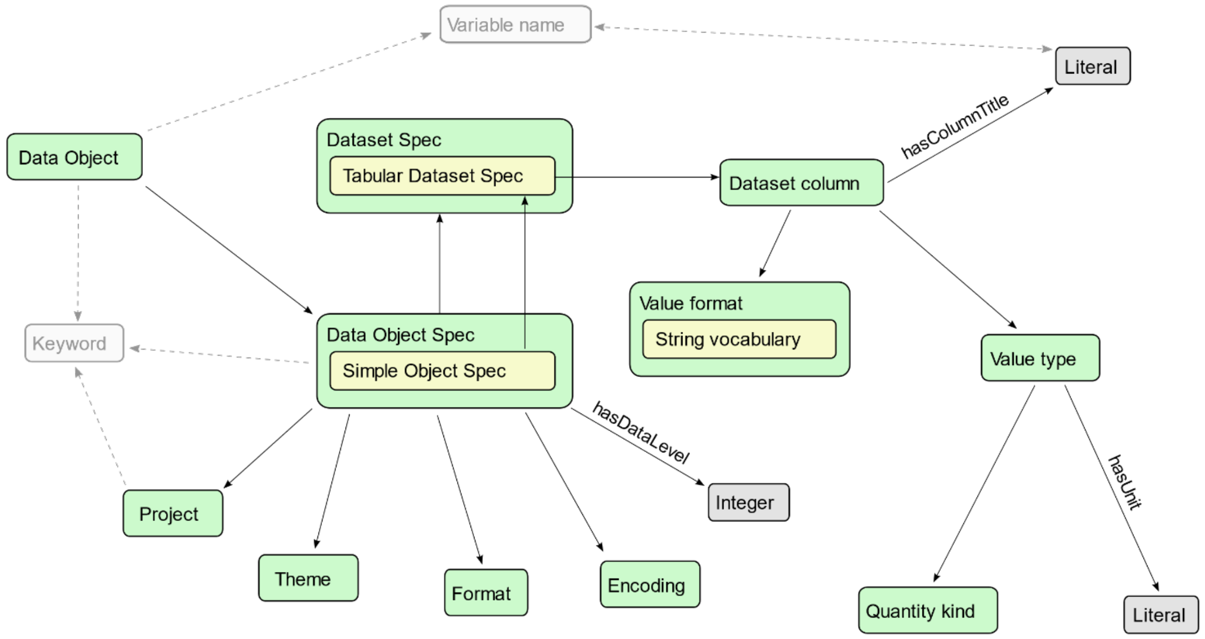


Figure 3.2 Relationship diagram of the ICOS data object specification and the metadata elements that describe the format, value type and quantity kinds of the different variable in an (ICOS) data file and relation to project and theme to which the data belongs.

When the object specification defines the data format of the file, a check is performed after the complete upload, to check the compliance to the data format and even possibly the validity of the data columns and spatial and temporal coverage as contained in the data file. Any deviation from the definition or prescribed metadata results in refusal of the file and abortion of the ingestion. The successful parsing of the data for text files also results also in the generation of binary CP-internal representations of the data that are used for the visualisation of time series in the data preview.

After upload completion, the checksum of the upload is compared with the registered checksum and when ok, a handle PID is minted for the data object and returned to the user. The metadata from the metadata packet is then added to the metadata repository and enriched with information on the PID, the checksum and other Object Specification dependent metadata. The suffix of the data object PID consists of the first 18 characters of the checksum of the data object and is thus unique for the data object. Later the PID suffix can at any time be compared with the SHA256 checksum of the data object to ensure that the data is up to the bit and exact copy of the original data object.

3.2.3 The metadata system

The metadata database can be queried using an open SparQL endpoint at https://meta.icos-cp.eu/sparql/?query=. The metadata store fully supports data versioning and data collections. It is machine actionable through standard http(s) protocol. The metadata store is fully described by the underlying ontology, that again itself is defined in RDF through the OWL language.

The design of the metadata system is fully configurable to act with a single or multiple portal frontend(s) using a single or multiple metadata stores. This means that for example multiple infrastructures can have their own differently styled data portal and use one single metadata store, or that one infrastructure has one portal that uses several external metadata stores, or that several infrastructures use one common portal that relies on a set of federated metadata stores, one per infrastructure. All completely transparent to the outside user.

The ICOS CP metadata store is for example shared with the Swedish SITES national infrastructure that has its own dedicated and styled portal.

3.2.4 Data discovery

The main entry point for data discovery for humans is https://data.icos-cp.eu. Here a set of filters can be easily set to filter to the data sets that the user is looking for. The list of data objects that fulfils the set of filters is display dynamically. Changing the filters also dynamically updates the remaining options for the other filters that comply with the other filter settings. Filters can at will be added, removed and applied incrementally. From the results page the user can view the most relevant information on the data object and/or drill down to the data object landing page for all relevant metadata. Most data objects can be previewed, see data visualisation. Most data objects can also be added to the user's data cart for easy download, see data access.

An example of the powerful possibilities of open access to metadata trough SparQL is the overview of time coverage from L0 data for all stations from the heatmap tool[11]. This is a small python program that collects through one single SparQL call the metadata of all available raw data for a domain and then plots per station the availability of data in percentage per week. These heatmaps give a fast overview of the performance in sending the required raw data from the stations through TCs to Carbon Portal. Gaps identify periods with problems in either the measurements or the data transfer.

[11] https://github.com/ICOS-Carbon-Portal/python-tools/tree/master/heatmap

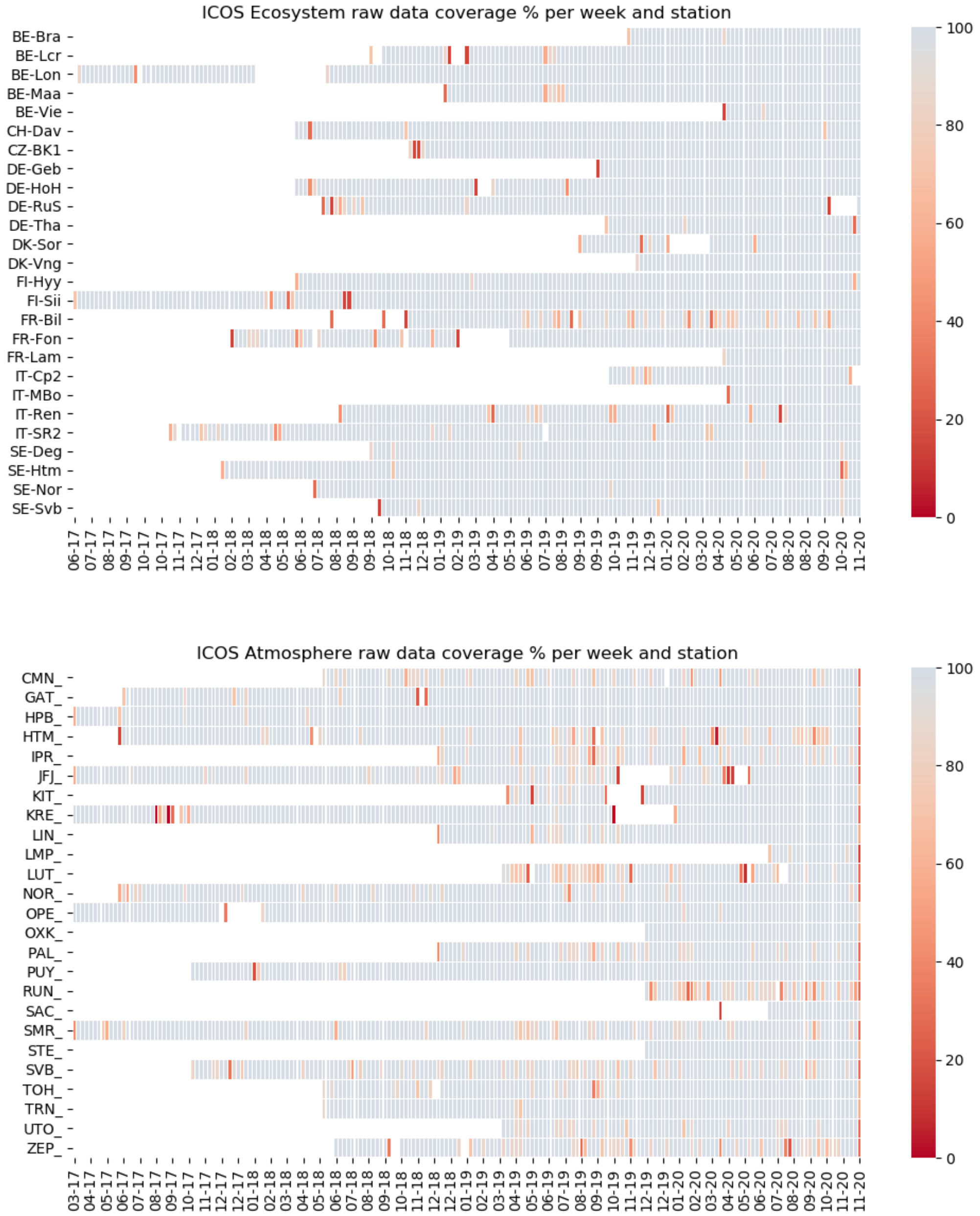


Figure 3.3 Heat maps of data coverage of ICOS raw data for the domains Atmosphere and Ecosystem from March 2017 up to now. Time coverage over each week is indicated from blue (100%) to red (0%).

3.2.5 Data access

Data access is provided through the PID (or DOI) of the data objects. Resolving this PID through the Handle or DataCite DOI system leads normally to a landing page that contains a link to the data object(s). In case of non-ICOS data objects this link can point to another data portal due to data license restrictions. Raw data objects are currently also not directly downloadable but require contact with the relevant thematic centre.

The data discovery tool allows to add selected data objects to the user's data cart from where the collected objects can be downloaded in one batch into a single zip archive.

3.2.6 Trusted data repository usage (B2SAFE)

While the data is ingested as depicted in Figure 6.1, simultaneously a copy is streamed to the EUDAT B2SAFE server at CSC in Finland. At the end of the transfer also at their server the checksum of the received file is calculated and compared with the SHA256 checksum at CP. Only when all checksums are ok, the transfer is considered successful and the provenance metadata is finalized in the CP RDF store. In all other cases the uploading client is returned an error and transfer should be retried at a later stage. The EUDAT B2SAFE system will transfer a second copy to the EUDAT B2SAFE server in Jülich later. Both the CSC and the Jülich B2SAFE system are built using redundant data services that also have independent backup systems that allow to restore the data in case of a data storage failure. This adds to the data security already in place through similar systems at ICOS CP, the Thematic Centres and/or data providers and extends beyond the lifetime of these facilities. The ICOS data can be easily identified using the ICOS PIDs and the independently minted Handle PIDs that EUDAT minted for the data files using the B2Handle system. Also, if required, the ICOS data can be exposed through the EUDAT B2SHARE service. At any time ICOS CP can access and retrieve the ICOS data objects stored at B2SAFE through the ICOS PIDs.

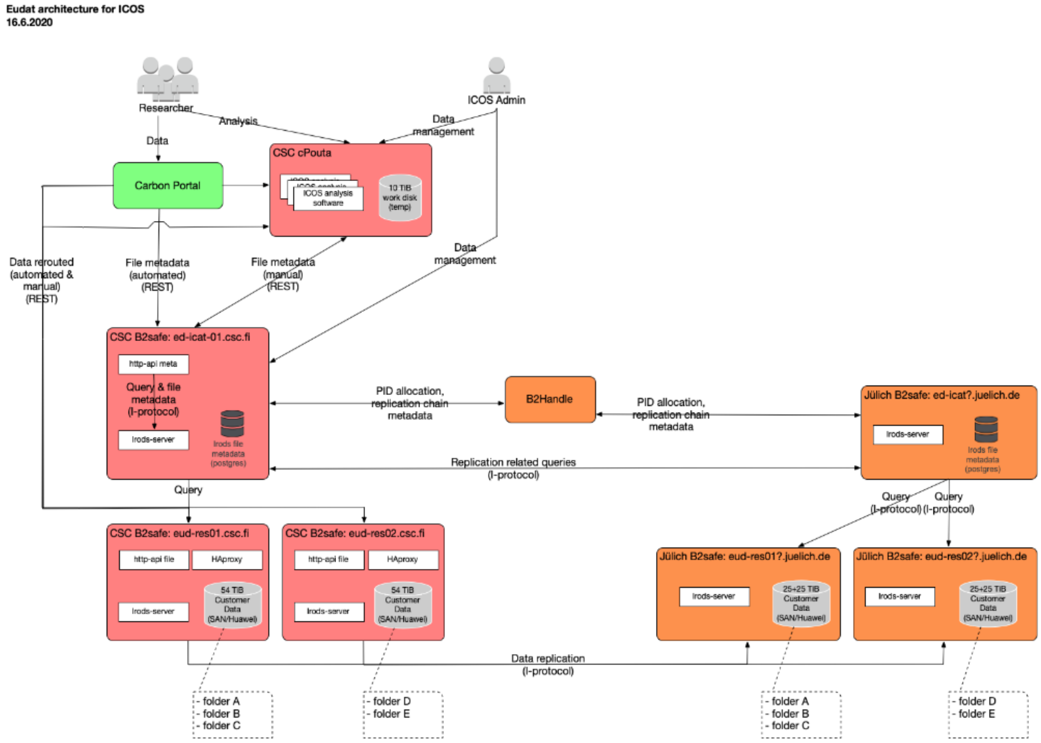

Eudat architecture for ICOS
16.6.2020
Researcher
ICOS Admin
Analysis
Data
Data management
Carbon Portal
CSC cPouta
ICOS analysis software
10 TIB work disk (temp)
Data rerouted (automated & manual) (REST)
File metadata (automated) (REST)
File metadata (manual) (REST)
Data management
CSC B2safe: ed-icat-01.csc.fi
http-api meta
Query & file metadata (I-protocol)
Irods-server
Irods file metadata (postgres)
PID allocation, replication chain metadata
B2Handle
PID allocation, replication chain metadata
Jülich B2safe: ed-icat?.juelich.de
Irods-server
Irods file metadata (postgres)
Replication related queries (I-protocol)
Query
Query (I-protocol)
Query (I-protocol)
CSC B2safe: eud-res01.csc.fi
http-api file
HAproxy
Irods-server
54 TiB Customer Data (SAN/Huawei)
CSC B2safe: eud-res02.csc.fi
http-api file
HAproxy
Irods-server
54 TiB Customer Data (SAN/Huawei)
Jülich B2safe: eud-res01?.juelich.de
Irods-server
25+25 TiB Customer Data (SAN/Huawei)
Jülich B2safe: eud-res02?.juelich.de
Irods-server
25+25 TiB Customer Data (SAN/Huawei)
Data replication (I-protocol)
- folder A
- folder B
- folder C
- folder D
- folder E
- folder A
- folder B
- folder C
- folder D
- folder E

Figure 3.4 Architecture of the data transfer from ICOS to the B2SAFE service and replication at the two N2SAFE instances at CSC (Finland) and FZ Jülich (Germany).

ICOS keeps its metadata following an ontology-based RDF store. The ontology relations are modelled in OWL. The whole metadata store is available through the SparQL endpoint, including the OWL definitions, so that the complete metadata and its relations can be read for machine to machine communication. The ontology models the relations between the person, stations, instruments, measurements, and data processing actions as shown in Figure 3.4. At ingestion the digital objects are based on their data object specification and provenance information in the objects specification JSON package linked to the ontology and stored in the repository and streamed to the trusted repository, while performing the checksum comparison between source and target to ensure the data integrity. The RDF store is versioned, this is to say that every entry in the RDF store is serialized and kept together with the timestamp of the change. This means that the state of the ontology can be restored to its previous state of any point in time.

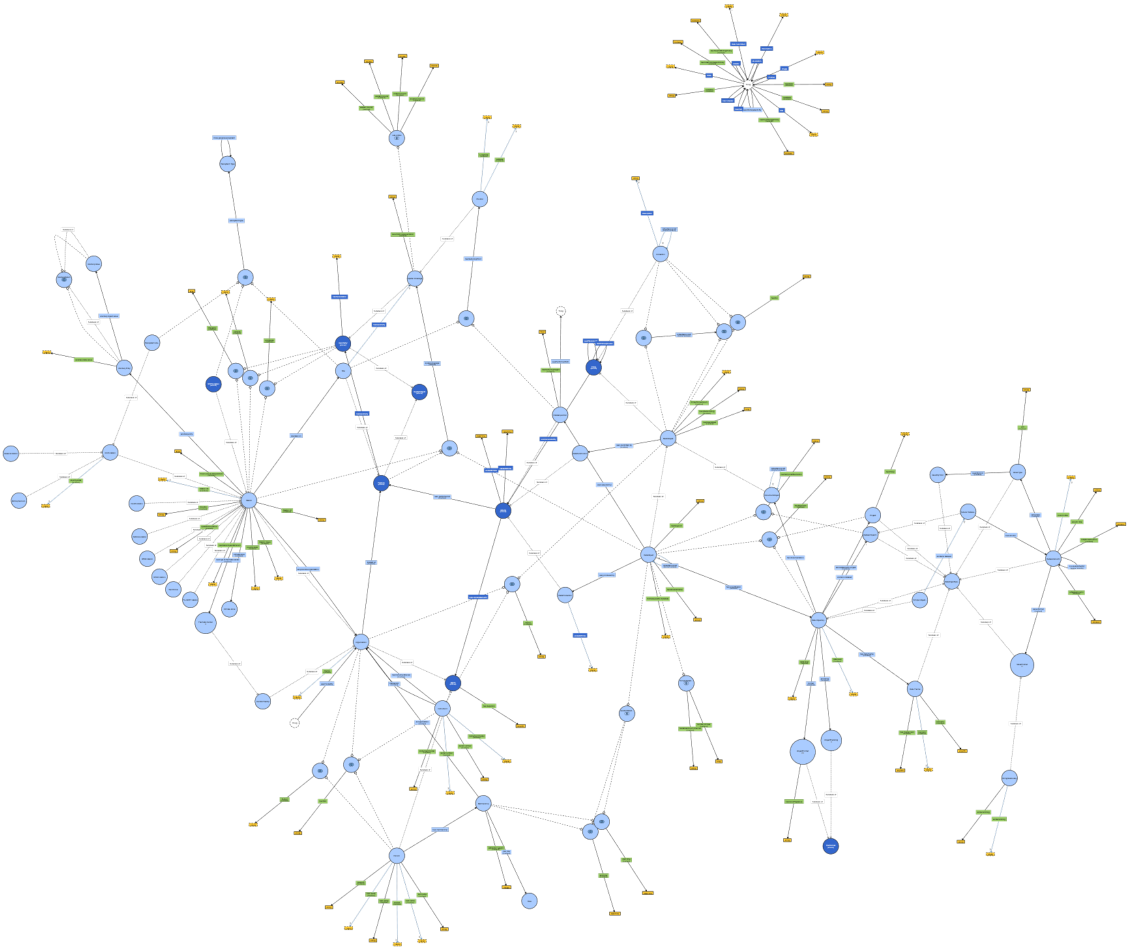


Figure 3.5 Overview of the ICOS CP ontology, modelled in OWL (http://www.visualdataweb.de/webvowl/#iri=http://meta.icos-cp.eu/ontologies/cpmeta/) as of June 2020.

When data is discovered through the portal app or a SparQL query the PID of the data object will resolve through the Handle system into a landing page, which contains either human and/or machine readable metadata that is gathered on the fly by following the relations defined by the ontology. That way always the most accurate and up to date version of the metadata available can be shown that corresponds with the start and end dates of acquisition, submission or processing contained in the metadata belonging to the object.

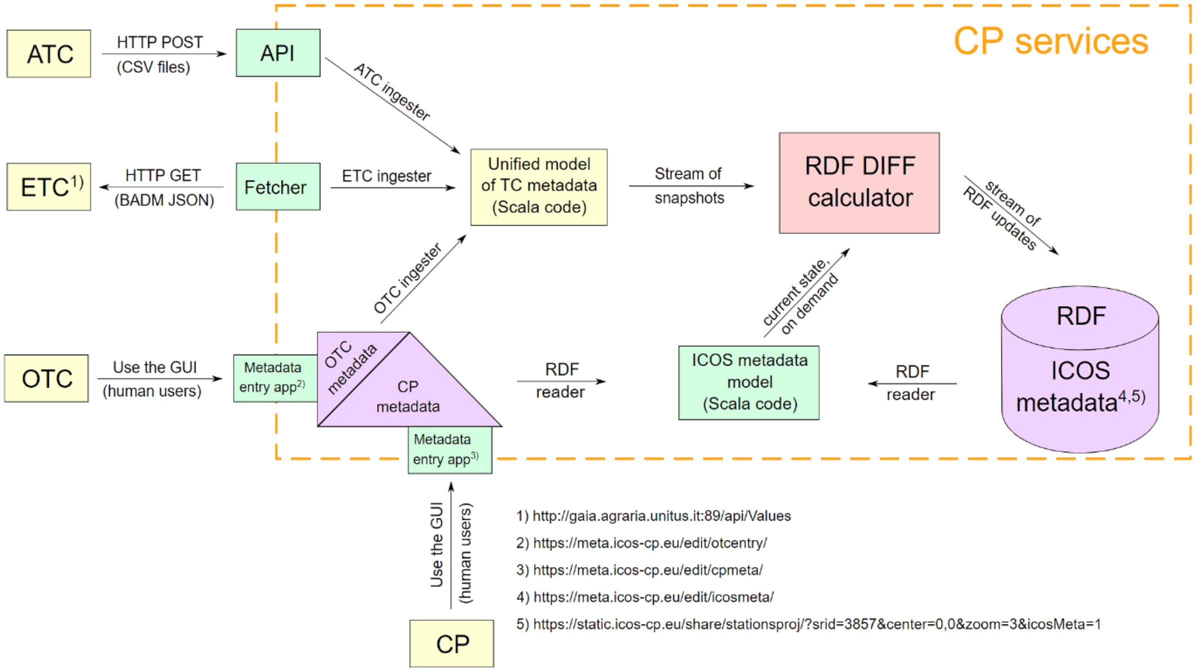
This is practically the most efficient, flexible and consistent way to associate time dependent metadata with the data objects while avoiding duplication of data and metadata and reducing the risk of metadata getting out-of-sync or not properly adjusted after corrections.

It also allows to model complex relations between metadata elements that are hard to efficiently model in traditional relational database management systems. A disadvantage of the flexibility and complex relations is that querying the ontology for even quite simple questions like: 'give me the list of all L2 data objects' can become a time-consuming operation due to the required traversal of a complex tree of a large amount of RDF triples. At Carbon Portal we invested quite some effort in designing a caching and query optimisation system using 'magic' indexing that makes sure that the queries that are used in our apps and data portal all execute within say 50 milliseconds.

3.3 Unified metadata gathering

As discussed in the chapter 2 the ICOS Thematic Centres and Calibration Labs perform independent data processing and offer services to the national networks. However, all data, from raw (L0) to Near Real Time (L1) and final quality controlled and calibrated data (L2) is ingested at Carbon Portal at the time of the generation of the data. Also, the networks maintain the metadata describing the contributors, the measurement systems and observations through the IT systems present at the Thematic Centres. The TC's also offer software clients that allow the measurement performing persons to enter provenance data and flag the measurement data as part of the Quality Control, to mark periods with problems in instrumentation for example. Also, the TC's add quality information to the data by for example performing analyses of instrument precision, conditions like low turbulence that impact the data quality, calibration uncertainties, drift and spike detection. In general, this QC information is added to the time series data files as separate columns with estimates of the different kinds of uncertainty per parameter or as columns with information in the form of data quality flags.

As part of the RINGO project we defined a comprehensive unified metadata flow from Thematic Centers to the Carbon Portal. The design criteria of this system were to integrate as much as possible the operational (legacy) database systems at the TCs with the data portal, thereby preserving the investments in the robust and proven QA/QC and database systems at the TCs and combining these with the benefits of a linked open data system with connected data licence check, usage tracking and dynamic machine operable data and metadata based on a versioned RDF triple store.



1) http://gaia.agraria.unitus.it:89/api/Values
2) https://meta.icos-cp.eu/edit/otcentry/
3) https://meta.icos-cp.eu/edit/cpmeta/
4) https://meta.icos-cp.eu/edit/icosmeta/
5) https://static.icos-cp.eu/share/stationsproj/?srid=3857&center=0,0&zoom=3&icosMeta=1

Figure 3.6 Diagram of the implemented metadata transfer scheme between Thematic Centers and Carbon Portal. TCs either post or make available the agreed tables with information wrt persons, roles, stations and instrumentation and the information is

routinely synced with the metadata ontology at CP by comparing current and available information and storing the differences. OTC adds the metadata directly into the CP RDF database.

All relevant metadata is synced with the CP repository using the methods shown in Figure 3.5. The Atmospheric TC provides the information as CSV tables that are transferred at regular intervals (daily) to a receiving service at the CP end. The Ecosystem TC provides a service that is polled by the CP at regular intervals (hourly). The Ocean TC uses the CP RDF store and metadata entry GUI to maintain the metadata for the Ocean domain. All three metadata exchange mechanisms rely on standard internet protocols and have been implemented, TCs are free to switch between these free choices to further harmonise the operations across the domains and reduce the complexity of the ICOS metadata handling.

The data from those three sources is then converted to a unified model of the TC metadata and compared with the current metadata in the CP final metadata store. Any difference is then translated in an RDF update statement that is logged in the CP metadata repository, just like any other metadata update.

In the first stage TCs provide the relevant station and person (plus role) metadata to make the attribution and dynamic citation generation work. This part is now working for Ocean and Ecosystem. As the corresponding Atmosphere metadata synchronisation only started in June 2020 the integration of this part was completed by the end of Summer 2020. In the next stage instrument information will be integrated and merged with the ICOS ontology as an important pre-condition to be able to complete the provenance information of the observational data sets.

The roles acknowledged in the thematic centres and stations, mapped to the roles that are differentiated in Carbon Portal and DataCite and the weights that determine the order in the citation string at CP and DataCite are detailed in Table 3.1 and Table 3.2 below. Not all roles result in a mention in the author list in the citation string. But all roles will result in a entry in the contributor list for the relevant data objects. Ocean and Ecosystem personnel are acknowledged in the citation string as ICOS RI. In Atmosphere one can specify roles for the Thematic Centres that will become part of the citation string. Also other Thematic Centres could adopt rolesfor their contributors.

Table 3.1 Station roles mapped to Carbon Portal and Datacite. Weights determine the order in the citation, higher means higher weight, no value means the role is listed in the contributor info, but not added as author to the citation string.

Station roles					Order		
Ecosystem	**Atmosphere**	**Ocean**	**Carbon Portal**	**DataCite**	**ETC**	**ATC**	**OTC**
Affiliated	Other		Other	RelatedPerson			
Scientist		Researcher	Researcher	Researcher	3		2
Scientist Flux			Researcher	Researcher	3		
Scientist Ancillary			Researcher	Researcher	3		
Principal Investigator	Principal Investigator	Principal Investigator	Principal Investigator	Contact Person	4	1	3
Co-PI	Deputy PI		Principal Investigator	Contact Person	4	1	
Manager	Station Supervising PI		Administrator	Supervisor	5	2	
	Species PI		Principal Investigator	Contact Person		1	
Technician	Instrument responsible	Engineer	Engineer	DataCollector	1		1
Technician Flux			Engineer	DataCollector	1		
Data manager	Data controller		Data manager	DataManager	2		
	Tank configurator		Engineer	ProjectMember			
Thematic Center Roles							
as ICOS ETC	Data manager	as ICOS OTC	Data manager	DataManager		0	
	ATC staff member		Engineer	Other		0	
	Checker		Data manager	DataCurator		0	
	Data Analist		Data manager	DataManager		0	
	Calib. centre responsible		Researcher	Researcher		0	
	Supervisor		Administrator	Supervisor		0	

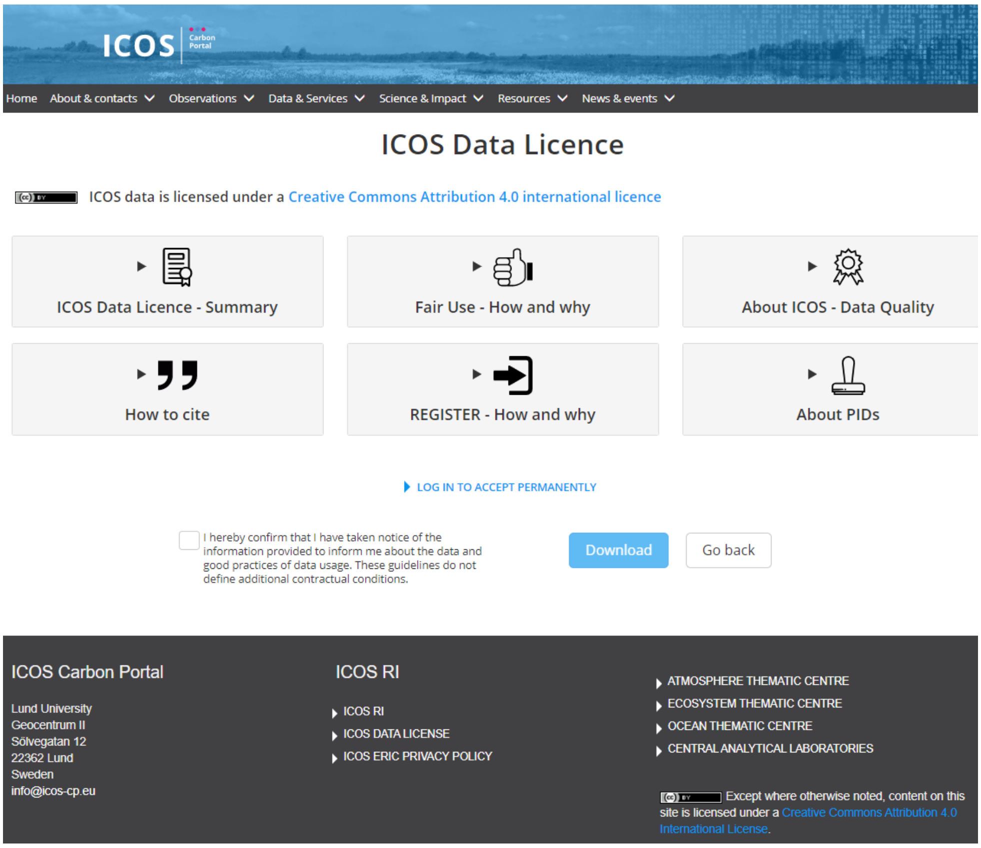
Table 3.2 Mapping of the DataCite to the Carbon Portal contributor roles. Only roles indicated with a start are added to the author list and included in the citation string.

Contributor roles Datacite	Author	Contributor roles Carbon Portal
ContactPerson	*	Administrator
DataCollector	*	Data manager
DataCurator		Engineer
DataManager	*	Principal Investigator
Distributor		Researcher
Editor		Other
HostingInstitution		
Producer		
ProjectLeader		
ProjectManager		
ProjectMember		
RegistrationAgency		
RegistrationAuthority		
RelatedPerson		
Researcher	*	Researcher
ResearchGroup		
RightsHolder		
Sponsor		
Supervisor	*	Administrator
WorkPackageLeader		
Other		

4 Increases in data flow and improved data usage tracking in global data streams

4.1 Download counts and usage tracking in the ICOS data portal

When users access an ICOS data object this will always go through using the persistent identifier of the data object. In most cases the user does not have to and will not be aware of this, because this process is hidden through a user interface. But when the data object has been found and the PID of the data object is resolved this will point through the metadata to the access URL of the data object, which again contains (part of) the checksum. Through the access URL the download of the data object can start, but the central proxy first checks the acceptance of the ICOS data license. For normal users this means that the download first is redirected to the ICOS data licence acceptance page where the user is fully informed of the CC4BY licence and its conditions and has to check the box for acceptance of the licence before the download will start.



ICOS Data Licence

ICOS data is licensed under a Creative Commons Attribution 4.0 international licence

□ I hereby confirm that I have taken notice of the information provided to inform me about the data and good practices of data usage. These guidelines do not define additional contractual conditions.

ICOS Carbon Portal

Lund University
Geocentrum II
Sölvegatan 12
22362 Lund
Sweden
info@icos-cp.eu

ICOS RI

- ICOS RI
- ICOS DATA LICENSE
- ICOS ERIC PRIVACY POLICY

- ATMOSPHERE THEMATIC CENTRE
- ECOSYSTEM THEMATIC CENTRE
- OCEAN THEMATIC CENTRE
- CENTRAL ANALYTICAL LABORATORIES

Except where otherwise noted, content on this site is licensed under a Creative Commons Attribution 4.0 International License.

Figure 4.1 ICOS CC4BY data licence acceptance page, shown and to be accepted before a download can take place

When the user has created a user account at CP, has logged in to that account and checked the relevant option, the user has signalled his acceptance and can bypass this licence acceptance at every next download. For machine to machine data transfer several mechanisms exist to flag the acceptance of the data licence to make the data transfer go ahead without interruptions.

The portal proxy will register a complete download in the download registration database together with the IP of the user, time and an extract of the data object metadata. Through the data download statistics app[12] at the portal one can generate a report of the number of downloads with time and per country, while selecting the data for which one needs to see the count by a faceted search like in the main portal app. Also previews of data are counted per dataset. The app allows to view the download count per domain, station, data provider and member state. By querying the database itself, the portal admin can extract information on for example what data types and from which domain data are accessed per country of the downloaders.

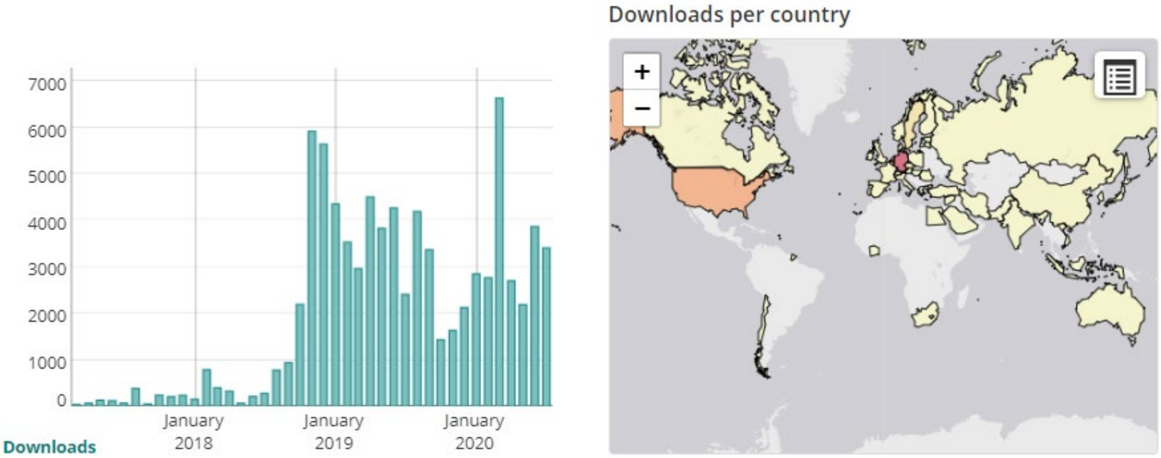


Figure 4.2 Example output of the download statistics application, showing the number of downloads of all ICOS Level 2 products per month and per country. Status of 12 July 2020; at that moment 2372 Level 2 ICOS products were offered.

4.2 Data flow increases in the framework of RINGO

The challenge of reaching the maximum amount of users for ICOS data can be established by making the data discoverable and accessible with minimum friction through multiple access points, both from the ICOS data portal as other global portals. It is desirable to avoid replication of data and to make sure that the usage is tracked by ICOS and that the downloaded data keeps on being traceable to ICOS. As mentioned before the PID minting based on the SHA256 checksum of the data is crucial for this.

Discoverability is ensured by the easy access to the metadata through linked open data and the open SPARQL endpoint and the easy to use data portal app with its faceted search. By also sharing the metadata on (global) discovery portals like GEOSS, WMO WDCGG and the coming ENVRI Hub and EOSC portal the use base can be enhanced. In the framework of ENVRIFAIR it is planned to extend the data landing pages with (json-ld) schema.org tags, next to the already available JSON, HTML, XML and Turtle content negotiable machine readable metadata formats. This will make data discoverable in Google dataset search.

4.2.1 Global Obspack integration

In the framework of RINGO we worked on the inclusion of the ICOS Atmosphere data into the Globalview Obspack product, originally developed by NOAA. Obspack is in essence a generic repackaging of the Greenhouse Gas atmospheric composition data into a self-descriptive netCDF format, by gathering the data from a multitude of stations into one comprehensive data file. The Obspack products have become very popular for use in education and inverse modelling studies and

[12] https://data.icos-cp.eu/stats/

the Obspack format has become a de-facto standard in the Greenhouse Gas atmosphere community. Together with NOAA we worked on an automated data flow where ICOS L2 releases and NRT data can flow automatically into the Obspack new releases. NOAA relies for this on the DOI and collection info for the data release as generated by ICOS and published at the CP. Also we implemented an API[13] that is called for each download of an Obspack product containing ICOS data from the NOAA server which registers the download inclusive the download time and IP from the downloader and this is registered as download of the respective ICOS dataset. Obspack products that contain ICOS data are adding an additional 400 downloads per year in the period starting 2018 to now.

4.2.2 DataCite DOI minting and production of collections

DOIs are used extensively since the 1990's to describe the attribution and provenance data of scientific literature and its unique and persistent identifiers allow to easily identify scientific publications. Since 2008 DataCite, a non-profit organisation, has taken the task of adapting the DOI framework to keep track of scientific data. DataCite is now the leading global provider of DOIs (Date Object Identifiers) for research data.

DOIs are persistent identifiers that are based on the same Handle PIDs that ICOS mints for the individual data objects. In ICOS the main use of DOIs is to mint additional citable DOIs for collections of data that are also called data products. Data products are packages of data objects that together as collection make up for example a data release for a domain or cover a certain period of ancillary data like gridded fossil fuel emissions.

The URL to which the DOI resolves is set when the DOI is minted but is mutable. DOIs differ from Handle in that there is a specific set of metadata attached to each DOI that must be provided during the minting process, but again this metadata can be changed afterwards. In this sense one could say that the ICOS PIDs are ICOS specific DOIs where the underlying metadata model is custom made by ICOS and modelled in the ontology. An important difference is that the normal DOI DataCite metadata is fixed and can only be updated as a whole, while the ICOS metadata is completely dynamic including the content of the landing pages. For example if the PI information for a station for a certain period is corrected this will automatically be reflected in the attribution information on the landing pages of data products from that station in that period and the citation strings for the relevant data sets. Same as when a new version of a data object arrives this will automatically refer the landing page of the old version to point the user to the updated version (and vice versa).

Currently the metadata schema that is attached to the DataCite DOI is Metadata Scheme v4.314 (16 Aug 2019). A bit in contrast to the FAIR principles the (DataCite) DOIs do not resolve to landing pages that reflect its own metadata model, but to the referral URL provided by the minting instance. This is heritage from the original DOI model from the International DOI Foundation (IDF), where publishers of scientific articles required that the DOIs landing pages would point directly at the publishers' websites or paywall and for example give direct access to the article or its abstract.

[13] https://github.com/ICOS-Carbon-Portal/data#reporting-data-object-downloads-for-partners-distributing-icos-data

[14] https://schema.datacite.org/meta/kernel-4.3/

As part of the work for RINGO, ICOS CP produced a user friendly graphical user interface[15] that not only allows the user to browse the complete DataCite metadata for DOIs minted by ICOS but also serves as a way to completely enter a completely new DOI and link this to ICOS data objects. DOIs were further integrated completely in the metadata workflow. The DOI associated with an ICOS data object or collection can be easily linked to its metadata and this way will be shown on the landing page including the data citation that can be generated using the DOI metadata.

ICOS mints DataCITE DOIs for important data products, that can be single data objects or a collection of data objects. The advantage of DataCite DOIs is the wide acceptance of these DOIs for use in citations and the connection to the globally dominant DOI system. By registration of DataCite DOIs the ICOS data and it metadata becomes findable and accessible through the global DOI system and can be found for example using the DataCite Search[16] engine that again is harvested by Google Dataset Search[17] and other search engines like OpenAIRE Explore[18]. This all widens the findability of ICOS data and increases it use, while making sure that still all access takes place through the ICOS portal, including usage tracking.

As part of the Drought-2018 initiative the ICOS facilities assisted the community in the effort and produced new releases of historic datasets for both atmosphere and ecosystem fluxes that complement the now growing Level 2 and NRT datasets from ICOS to really homogeneous long-term and high quality datasets that are needed for model studies. These datasets were published as collections and minted DOIs and are now exposed as Main Data Products at the ICOS website. This data is fully citable and is cited in the 16 papers of the special issue of Philosophical Transactions of the Royal Society B that will appear summer 2020. Figure 6 shows the number of downloads per month and the locations of the downloaders since the publications of the data sets in July 2019.

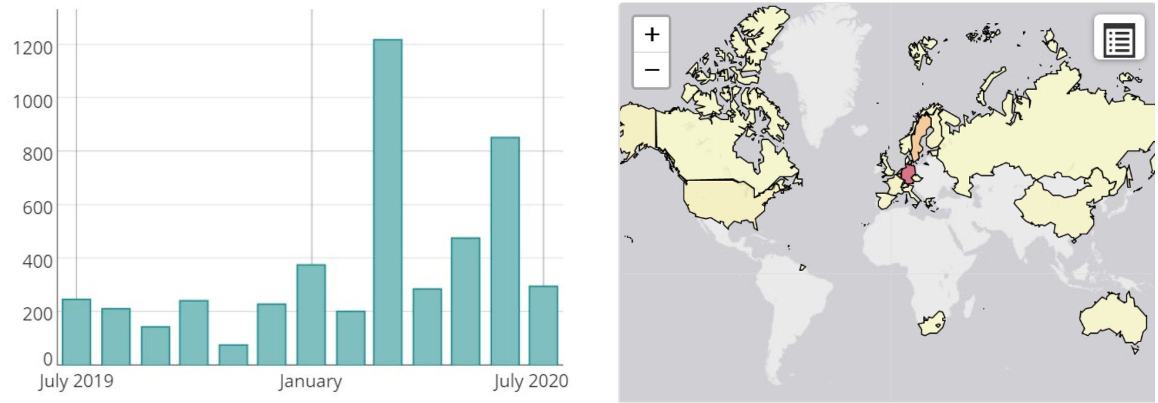


Figure 4.3 Downloads of Drought-2018 observational and elaborated products per month and by location of the downloader

4.2.3 Curation of elaborated data products

Almost any publisher of scientific articles now demands that associated data to an article is deposited in an open data repository, and this can be either a general data repository like Pangaea or Zenodo or a community repository like ICOS Carbon Portal. Also funding schemes like Horizon 2020 and other

[15] https://doi.icos-cp.eu/

[16] https://search.datacite.org/

[17] https://datasetsearch.research.google.com/

[18] https://explore.openaire.eu/search/find

public funding agencies now demand that data generated through public funds is published openly through a, preferably FAIR, data repository that takes care of the long term and trusted storage beyond the project lifetime. In RINGO we further improved our interfaces to support the community with this in serving of ICOS Carbon Portal as a community data repository.

The wider ICOS community that uses ICOS data to produce its science also generates so called elaborated data products that need to be published. The ICOS Carbon Portal also allows the community to cooperate through the cloud storage and cooperation tool Nextcloud[19] (>300 users in July 2020). From there or from other data locations the final products can be uploaded through the Upload GUI20 developed within RINGO and/or through expert assistance from the CP personnel. The Upload GUI also allows to mint collections and attach DOIs to the data objects or collections. The data experts at ICOS assist the users in the whole process and evaluate and curate all data and metadata before the data is published. The Upload GUI and the underlying backend is the same as is used for the ICOS automated data streams which is built for reliability and performs automated data consistency, integrity, and sanity checks as part of the ingestion process.

Very successful and important projects like the Global Carbon Project now use the ICOS data portal to publish their data sets, like the GCP Global Methane Budget21, GCP Global Carbon Project22 and the recent GCP COVID-19 shutdown study for temporary reduction of global daily CO_2 emissions due to the COVID-19 forced confinement23. But for example also the EUROCOM project and the Drought-2018 study publish data products with ancillary data sets and inversion model results[24] for biogenic fluxes through ICOS Carbon Portal.

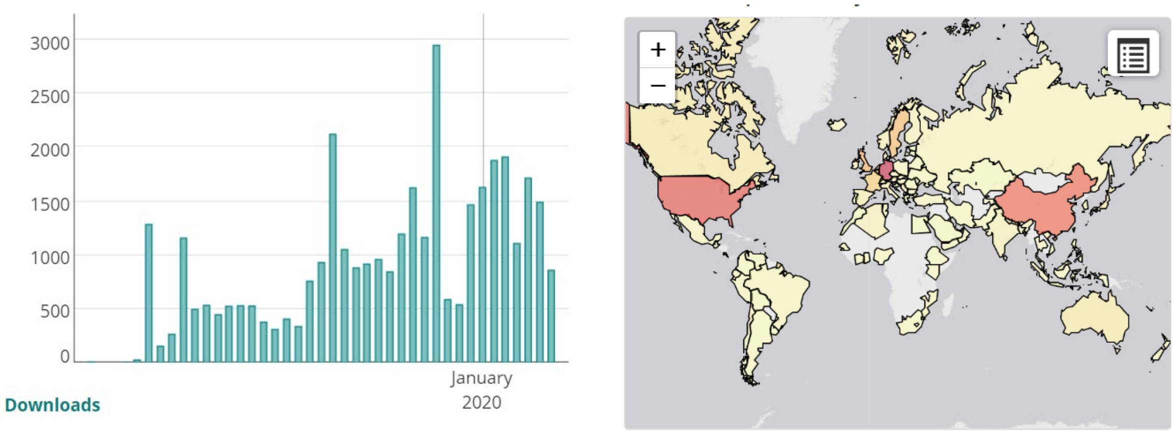


Figure 4.4 Downloads of elaborated products per month and by location of the downloader

[19] https://fileshare.icos-cp.eu

20 https://meta.icos-cp.eu/uploadgui/

21 https://doi.org/10.18160/GCP-CH4-2019

22 https://doi.org/10.18160/GCP-2019

23 https://doi.org/10.18160/RQDW-BTJU

[24] e.g. https://doi.org/10.18160/E72F-D093

The elaborated data sets attract a large number of users to the ICOS website and on average result in more than 1000 downloads per month from all over the world, as can be seen in Figure 4.4.

4.2.4 Collaboration with NEON

ICOS' focus is set to provide high quality greenhouse gas measurements for Europe. The National Ecological Observatory Network (NEON[25]) is an equivalent Research Infrastructure for the USA, operated by Battelle Inc. and designed to collect long-term open access ecological data to better understand how the U.S. ecosystems are changing. NEON operates 81 field sites, quantifying ecological processes over time and across the continent. The collected data and observations overlap between ICOS and NEON for atmospheric greenhouse gas concentrations, meteorological observations, and Eddy Covariance fluxes (vertical turbulent fluxes within atmospheric boundary layers). Both networks are working with the FLUXNET community to standardise and unify these EC data products.

Since both Networks are open linked data repositories, a collaboration has started to connect both repositories with a machine actionable system. Several remote sessions were held between NEON and CP and Claudio D'Onofrio spent a working visit to NEON HQ in Boulder (CO) from 13-27 November to exchange ideas and draft a first plan. The plan outlined between NEON and ICOS is to represent specified data products from the NEON portfolio within the ICOS Carbon Portal. The representation is achieved by mapping the NEON meta data to the ICOS meta data standard and an automatic ingestion into the triple store. As a result of this, all meta data for the ingested products are automatically available at the ICOS SparQL endpoint.

The ICOS Carbon Portal has developed a unique way of data ingestion. On one side the original datafile is stored as provided by the creator. This step will be omitted for NEON data products. Additionally, a low-level binary file representation is created for the dataset to provide a performant and easy way to access data for visualisations. This step does not provide means of downloading the data but to have an instant glimpse into the dataset. Download of data products is provided at the NEON data portal and hence any statistics of download or license information remains in the authority of the providing Research Infrastructure.

This approach and the collaboration between the RI's will showcase an exemplary case of FAIR data. The findability of data is increased, and accessibility unified. The automated exchange of metadata is achieved through the interoperability of both data portals and should ultimately lead to a reuse of data. Overall a lot of steps will be 'behind the scene', but ease the way for the user to find data have a quick look what is inside and find a persistent and citable link to the data.

4.2.5 Integration with SOCAT and GLODAP

For the Ocean domain the OTC made available the historical Level 2 datasets from ICOS stations that have been taken up in SOCAT and published them as ICOS OTC on behalf of SOCAT, consisting of 934 data objects[26.] Starting 2020 the ICOS Level 2 datasets released will be minted ICOS DOIs and these will be used for attribution in the SOCAT database, together with the implementation of the download information forwarding API developed for Obspack. This to make sure that each download

[25] https://www.neonscience.org/

[26] https://data.icos-cp.eu/portal/#%7B%22filterCategories%22%3A%7B%22project%22%3A%5B%22socat%22%5D%7D%7D

of SOCAT data where ICOS is included increases the data usage count for these data objects at the ICOS portal. More details on the integrations of SOCAT and GLODAP can be found in RINGO Deliverable D5.2.

4.2.6 Integration with FLUXNET

FLUXNET released in July 2020 an update of the FLUXNET2015[27] dataset, the main DOI for the dataset is now the citation of the Nature Scientific Data reference paper by Pastorello et al[28.] Most (206 out of 212) individual sites of this global dataset now also have a CC4.0BY data licence like ICOS and each datafile has a dedicated DOI minted for citation. Machine readable metadata describing the collection of files and their DOI is available as JSON file at figshare[29.] Just like in ICOS, each DOI in the collection redirects to a (in the FLUXNET case not machine-readable) landing page that contains a generic button for download that in this case redirects to the FLUXNET login page that requires a login and manual textual input. After logging in the whole selection process has to be started all over again, after which a download directory is prepared with the desired files. For each download the user has to provide the reason and planned use of the data and each PI of downloaded datasets will be informed of this by email.

In the new FLUXNET setup in the near future it is proposed that the ICOS Ecosystem Level 2 data files produced by the ETC will be stored and transmitted using their ICOS minted DOIs and downloads will be either re-directed straight to the Carbon Portal without going through the FLUXNET login and/or will use the download information forwarding API developed for Obspack. This to prevent wherever possible the replication of data and make sure that each download of FLUXNET data where ICOS is concerned increases the data usage count for these data objects at the ICOS portal. Just like with the NEON collaboration ICOS can offer full discoverability of the FLUXNET dataset and offer quick previews to the data and redirect actual downloads of non-ICOS FLUXNET data to the FLUXNET data portal. Implementation of these features is still in the planning phase.

4.2.7 Integration with WMO GAW, WIGOS and WDCGG

ICOS Atmosphere provides data files that follow the WMO GAW (World Meteorological Organisation, Global Atmosphere Watch) specification and can be readily submitted to the WMO GAW World Data Centre for Greenhouse Gases[30] (WDCGG). However, the metadata system of WDCGG is not machine accessible and the system relies on manual upload and metadata entry. Only PIs are allowed to enter data and metadata for their own station using their personal account. For each data submission a complete manual metadata submission is necessary.

There is no persistent identification of data in the WDCGG system, though currently the minting of DOIs for individual data files and collections per station are planned for 2021, but the granularity is still under discussion. From the plans it is already clear that they will accept also the DOIs minted by the data providers. Versioning of the data is foreseen, and the DOI minted by WDCGG will point to the latest version of a data set. Other planned improvements are provision of the data in netCDF

[27] https://fluxnet.org/data/fluxnet2015-dataset/

[28] https://doi.org/10.1038/s41597-020-0534-3

[29] https://doi.org/10.6084/m9.figshare.12295910

[30] https://gaw.kishou.go.jp/

format and most importantly to follow the WMO WIGOS[31] (WMO Integrated Global Observation System) metadata standard WMDS[32.] Discussions have been started with the operators of WDCGG to start cooperation using the same setup for metadata and data exchange as used for NEON, SOCAT, FLUXNET and NOAA Obspack but progress has been very slow thus far.

Integration with WMO WIGOS through OSCAR surface/GAWSIS[33], the station metadata database of WMO GAW has progressed as precondition for access to provision of ICOS station metadata and data. ICOS is now registered as contributing network of WMO GAW for greenhouse gases and been given the authorisation to modify the relevant metadata for stations belonging to the ICOS network. As soon as the automated synchronisation of metadata between ATC and CP is functioning the dynamic exchange of mutations in stations, instrumentation and operators will be arranged through exchange of WIGOS metadata in the WMDS XML standard. Next steps will be the transfer of NRT data into the WIGOS system and update of Level 2 data in the WDCGG database.

The previous, more open, WDCGG data portal was used in 2015 to demonstrate the interoperability of the Carbon Portal concept. The complete content of the WDCGG hourly time series for all greenhouse gases was at that time ingested by Carbon Portal and could be searched and previewed using the prototype of the current data portal app. For download of WDCGG data the app would link the user to the relevant data set at the WDCGG portal. This prototype is with limited functionality is still available at CP and the 2015 WDCGG data can still be previewed using the current portal app. This allows for example to link the historic WDCGG data to the STILT Footprint tool.

4.3 Python library for access to ICOS data and metadata

The icoscp python open source library[34] provides an easy access to data and associated metadata hosted at the ICOS Carbon Portal. By using this library one can load data files directly into memory. The approach of this library is to free the user from downloading and maintaining a local copy of data files and if one uses the ICOS Jupyter Hub services (described in chapter 5), one does not even need computational power.

The library is developed with Python 3.7.x and can be installed using the standard installation program pip. The library contains a set of modules that allows to discover data and get access to the most relevant metadata for ICOS data timeseries. All data and metadata is then provided as python native objects and dataframes. The station module provides a search facility to explore ICOS stations and find associated data objects and data products. The collection module supports to load a collection of digital objects. The SparQL module lies at the basis of the previous modules and gives access to the ICOS metadata triple store. Finally the Dobj module gives access to the timeseries data by indexing the data with the persistent identified found through the other python modules, and allows the user to discover the column names and other info on the data like value type and unit.

[31] https://www.wmo.int/pages/prog/www/wigos/index_en.html

[32] https://www.wmo.int/pages/prog/sat/meetings/documents/IPET-SUP-1_INF_05-02_WIGOS-Metadata-Standard-V0.2.pdf

[33] https://gawsis.meteoswiss.ch/GAWSIS/

[34] https://github.com/ICOS-Carbon-Portal/pylib

Part of the metadata discovered for each data object or collection is the citation string that should be used according to the ICOS data licence to properly attribute the data providers.

In three lines of code the user can now open an ICOS data object and plot the time series:

```
dobj = Dobj('https://meta.icos-cp.eu/objects/lNJPHqvsMuTAh-3DOvJejgYc')
dobj.data.plot(x='TIMESTAMP', y='ch4', grid=True)

plt.show()
```

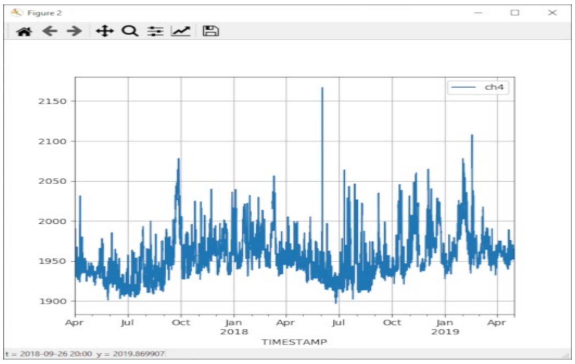


Figure 4.6 With just three lines of code the user can plot the ICOS L2 methane data from station Norunda and find the citation of this dataset: Lehner, I., Mölder, M., ICOS RI, 2019. ICOS ATC CH4 Release, Norunda (59.0 m), 2017-04-01–2019-04-30, https://hdl.handle.net/11676/lNJPHqvsMuTAh-3DOvJejgYc

The access to the data through the python library is counted and logged separately by the data usage database.

The icoscp python library is extensively used in the Jupyter hub services and examples at the ICOS website and to a large extent simplifies and enables the fair use of ICOS data in reproducible scientific workflows.

4.4 Gathering of data citation statistics using PIDs

The assignment of Handle PIDs and DOIs allows the users to easily attribute the data providers according to the ICOS data licence CC4BY. Whenever they use ICOS data in their publications or when they base new data on ICOS data they should include references to the respective PID(s) or DOI(s) of the data used. The high granularity of the ICOS PID information and the provision of DOIs for collections enables the desired detail when a small amount of data objects is used or aggregation when a large number of data is used.

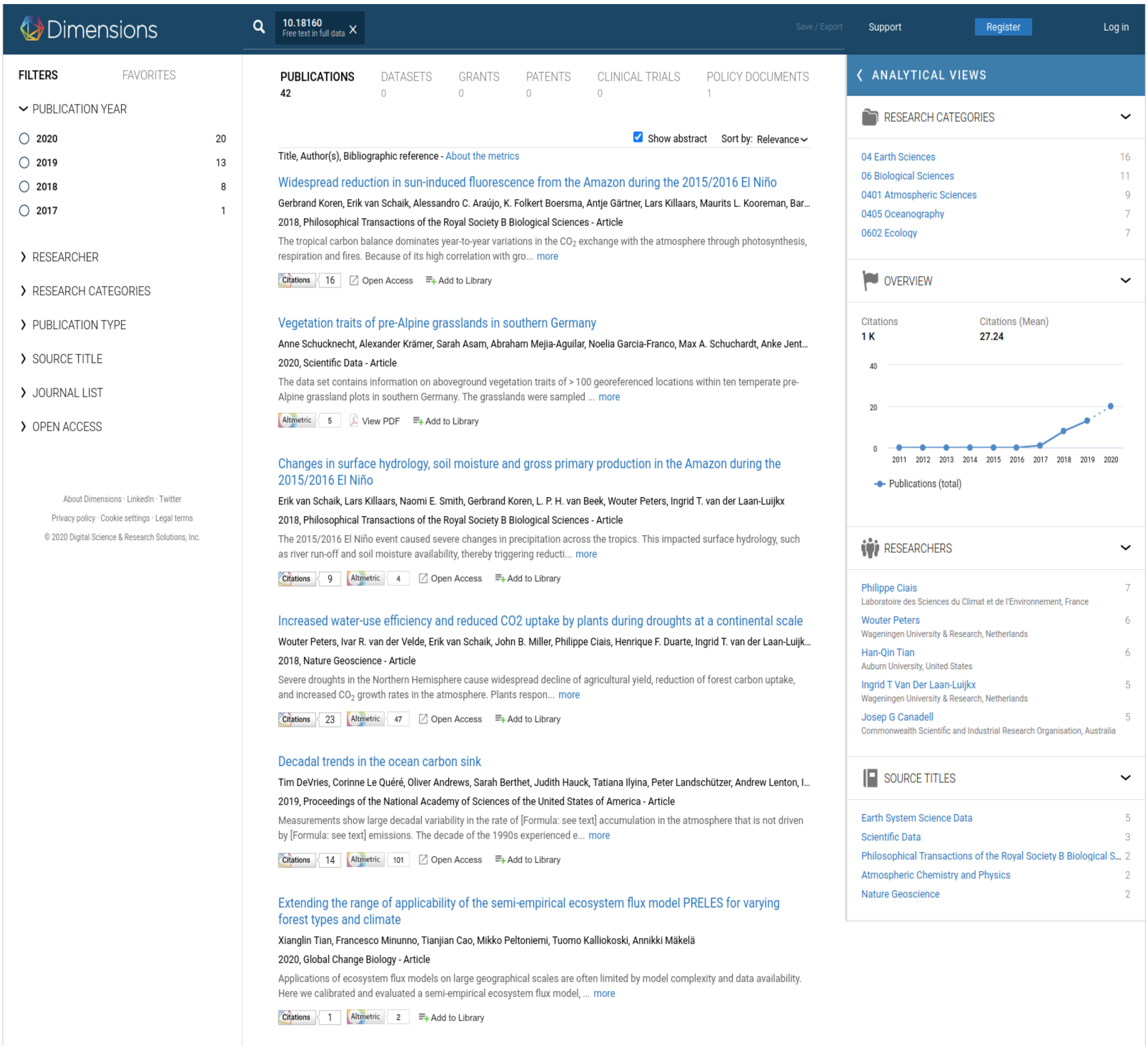
In the future we anticipate that automated workflows will further simplify this task by minting a DOI for a collection that entails all the data (objects and/or collections) that have been used in the workflow and that can be used to cite that elaborated data product. When this DOI then is created using an ICOS prefix this data usage is easy to track and further follow in its use and citation. However at the moment there is no (global) system that would ping the ICOS repository when a data set is minted a DOI that refers to a collection that also includes ICOS data objects or collections and it does not look like such a system will come soon.

Another possibility would be that the citation statistics providers would walk the tree inside each collection and that way count to some first or second degree the citation of data through another data collection. This would require a solid standard on how to enclose a collection in a DOI and also the willingness of the citation statistics provider to perform this work, where both conditions do not seem to exist at this moment and are not likely to arise any time soon.

However, even the existence of first order citation of data through PIDs and DOIs is already a large improvement over the current practice of either not citing data at all or only through non-standard acknowledgements in journal articles. This allows to already track now through Google Scholar and citation trackers like Dimensions.ai[35] the usage of ICOS DOIs, PIDs through citations in scientific literature, media and mentions in social media. And example of the citations of ICOS DOI prefixes is show in figure 4.7. As ICOS only mints DOIs since mid-2018, the fact that the community still has to get used to the practice of data citation through DOIs and as the lead time of scientific publications is a year or more, one cannot expect a large number of data citations yet, but the results of the search demonstrates the principle and we can only expect that we are in the very young part of an exponential rise.

The only alternatives for the moment to get citation statistics now is the painstakingly bibliometric search where each reference found has to be checked through reading of abstract, acknowledgements of the whole article whether ICOS data was actually used, or the fast but too unrestrictive search where all keywords that reference to networks of which ICOS takes part and of which the result is shown in Figure 4.8. The current approach of gathering the ICOS related references through the community is for the moment a good approach that balances effort and accuracy, the citations per year resulting from that are shown in figure 4.9.

[35] https://app.dimensions.ai/discover/publication



10.18160
Free text in full data ×

FILTERS FAVORITES

⌄ PUBLICATION YEAR

○ 2020	20
○ 2019	13
○ 2018	8
○ 2017	1

❯ RESEARCHER

❯ RESEARCH CATEGORIES

❯ PUBLICATION TYPE

❯ SOURCE TITLE

❯ JOURNAL LIST

❯ OPEN ACCESS

About Dimensions · LinkedIn · Twitter
Privacy policy · Cookie settings · Legal terms
© 2020 Digital Science & Research Solutions, Inc.

PUBLICATIONS	DATASETS	GRANTS	PATENTS	CLINICAL TRIALS	POLICY DOCUMENTS
42	0	0	0	0	1

☑ Show abstract Sort by: Relevance ⌄

Title, Author(s), Bibliographic reference - About the metrics

Widespread reduction in sun-induced fluorescence from the Amazon during the 2015/2016 El Niño

Gerbrand Koren, Erik van Schaik, Alessandro C. Araújo, K. Folkert Boersma, Antje Gärtner, Lars Killaars, Maurits L. Kooreman, Bar...

2018, Philosophical Transactions of the Royal Society B Biological Sciences - Article

The tropical carbon balance dominates year-to-year variations in the CO_2 exchange with the atmosphere through photosynthesis, respiration and fires. Because of its high correlation with gro... more

Citations 16 Open Access Add to Library

Vegetation traits of pre-Alpine grasslands in southern Germany

Anne Schucknecht, Alexander Krämer, Sarah Asam, Abraham Mejia-Aguilar, Noelia Garcia-Franco, Max A. Schuchardt, Anke Jent...

2020, Scientific Data - Article

The data set contains information on aboveground vegetation traits of > 100 georeferenced locations within ten temperate pre-Alpine grassland plots in southern Germany. The grasslands were sampled ... more

Altmetric 5 View PDF Add to Library

Changes in surface hydrology, soil moisture and gross primary production in the Amazon during the 2015/2016 El Niño

Erik van Schaik, Lars Killaars, Naomi E. Smith, Gerbrand Koren, L. P. H. van Beek, Wouter Peters, Ingrid T. van der Laan-Luijkx

2018, Philosophical Transactions of the Royal Society B Biological Sciences - Article

The 2015/2016 El Niño event caused severe changes in precipitation across the tropics. This impacted surface hydrology, such as river run-off and soil moisture availability, thereby triggering reducti... more

Citations 9 Altmetric 4 Open Access Add to Library

Increased water-use efficiency and reduced CO2 uptake by plants during droughts at a continental scale

Wouter Peters, Ivar R. van der Velde, Erik van Schaik, John B. Miller, Philippe Ciais, Henrique F. Duarte, Ingrid T. van der Laan-Luijk...

2018, Nature Geoscience - Article

Severe droughts in the Northern Hemisphere cause widespread decline of agricultural yield, reduction of forest carbon uptake, and increased CO_2 growth rates in the atmosphere. Plants respon... more

Citations 23 Altmetric 47 Open Access Add to Library

Decadal trends in the ocean carbon sink

Tim DeVries, Corinne Le Quéré, Oliver Andrews, Sarah Berthet, Judith Hauck, Tatiana Ilyina, Peter Landschützer, Andrew Lenton, I...

2019, Proceedings of the National Academy of Sciences of the United States of America - Article

Measurements show large decadal variability in the rate of [Formula: see text] accumulation in the atmosphere that is not driven by [Formula: see text] emissions. The decade of the 1990s experienced e... more

Citations 14 Altmetric 101 Open Access Add to Library

Extending the range of applicability of the semi-empirical ecosystem flux model PRELES for varying forest types and climate

Xianglin Tian, Francesco Minunno, Tianjian Cao, Mikko Peltoniemi, Tuomo Kalliokoski, Annikki Mäkelä

2020, Global Change Biology - Article

Applications of ecosystem flux models on large geographical scales are often limited by model complexity and data availability. Here we calibrated and evaluated a semi-empirical ecosystem flux model, ... more

Citations 1 Altmetric 2 Add to Library

❮ ANALYTICAL VIEWS

RESEARCH CATEGORIES ⌄

04 Earth Sciences	16
06 Biological Sciences	11
0401 Atmospheric Sciences	9
0405 Oceanography	7
0602 Ecology	7

OVERVIEW ⌄

Citations	Citations (Mean)
1 K	27.24

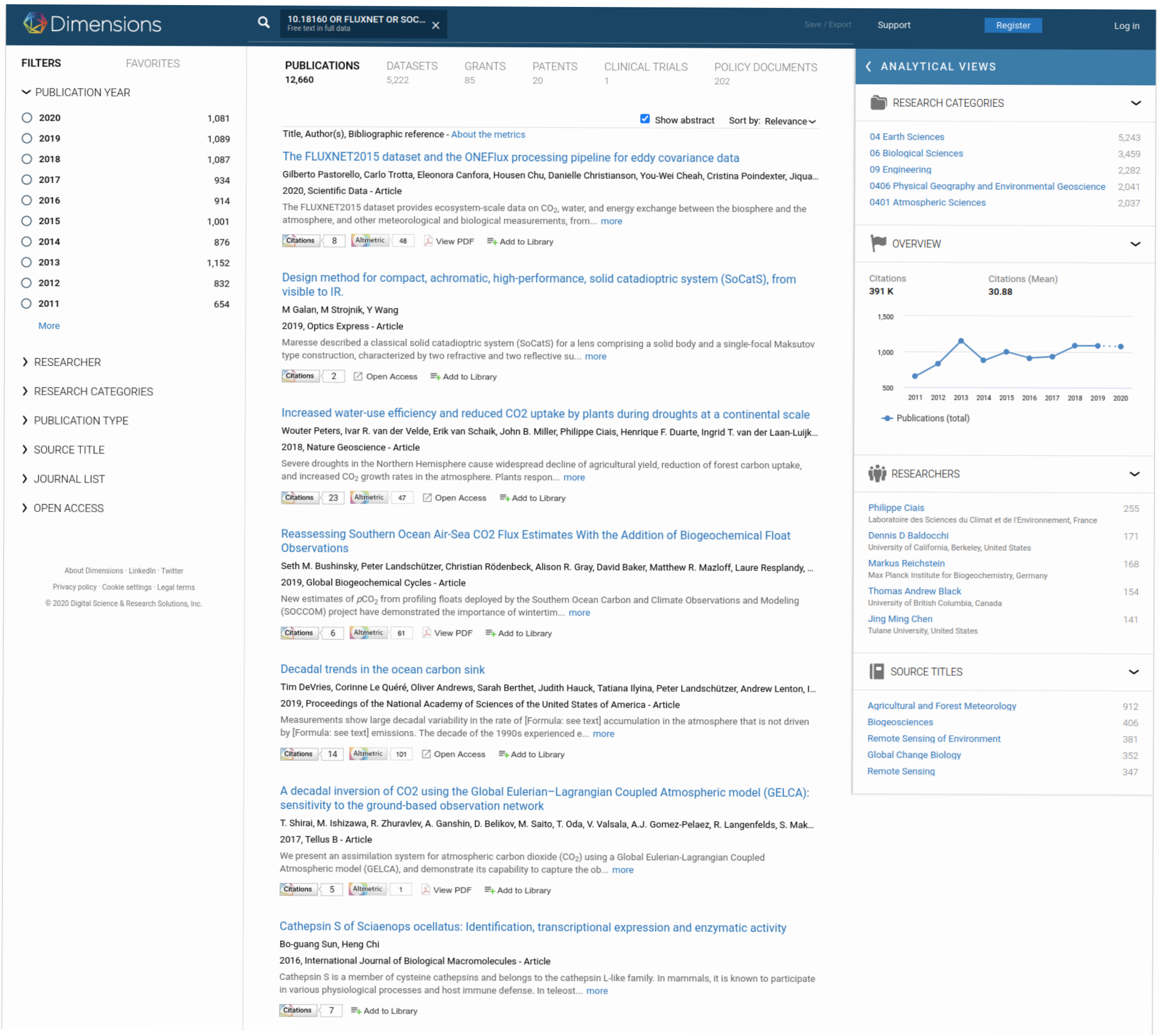
RESEARCHERS ⌄

Philippe Ciais Laboratoire des Sciences du Climat et de l'Environnement, France	7
Wouter Peters Wageningen University & Research, Netherlands	6
Han-Qin Tian Auburn University, United States	6
Ingrid T Van Der Laan-Luijkx Wageningen University & Research, Netherlands	5
Josep G Canadell Commonwealth Scientific and Industrial Research Organisation, Australia	5

SOURCE TITLES ⌄

Earth System Science Data	5
Scientific Data	3
Philosophical Transactions of the Royal Society B Biological S...	2
Atmospheric Chemistry and Physics	2
Nature Geoscience	2

Figure 4.7 Result of a query to track citations of data with ICOS DOI prefixes at Dimensions.ai. Numbers are relatively low for the reasons indicated in the main text.



FILTERS FAVORITES

PUBLICATION YEAR

○ 2020	1,081
○ 2019	1,089
○ 2018	1,087
○ 2017	934
○ 2016	914
○ 2015	1,001
○ 2014	876
○ 2013	1,152
○ 2012	832
○ 2011	654

More

RESEARCHER

RESEARCH CATEGORIES

PUBLICATION TYPE

SOURCE TITLE

JOURNAL LIST

OPEN ACCESS

About Dimensions · LinkedIn · Twitter
Privacy policy · Cookie settings · Legal terms
© 2020 Digital Science & Research Solutions, Inc.

PUBLICATIONS	DATASETS	GRANTS	PATENTS	CLINICAL TRIALS	POLICY DOCUMENTS
12,660	5,222	85	20	1	202

☑ Show abstract Sort by: Relevance

Title, Author(s), Bibliographic reference - About the metrics

The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data
Gilberto Pastorello, Carlo Trotta, Eleonora Canfora, Housen Chu, Danielle Christianson, You-Wei Cheah, Cristina Poindexter, Jiqua...
2020, Scientific Data - Article
The FLUXNET2015 dataset provides ecosystem-scale data on CO_2, water, and energy exchange between the biosphere and the atmosphere, and other meteorological and biological measurements, from... more
Citations 8 Altmetric 48 View PDF Add to Library

Design method for compact, achromatic, high-performance, solid catadioptric system (SoCatS), from visible to IR.
M Galan, M Strojnik, Y Wang
2019, Optics Express - Article
Maresse described a classical solid catadioptric system (SoCatS) for a lens comprising a solid body and a single-focal Maksutov type construction, characterized by two refractive and two reflective su... more
Citations 2 Open Access Add to Library

Increased water-use efficiency and reduced CO2 uptake by plants during droughts at a continental scale
Wouter Peters, Ivar R. van der Velde, Erik van Schaik, John B. Miller, Philippe Ciais, Henrique F. Duarte, Ingrid T. van der Laan-Luijk...
2018, Nature Geoscience - Article
Severe droughts in the Northern Hemisphere cause widespread decline of agricultural yield, reduction of forest carbon uptake, and increased CO_2 growth rates in the atmosphere. Plants respon... more
Citations 23 Altmetric 47 Open Access Add to Library

Reassessing Southern Ocean Air-Sea CO2 Flux Estimates With the Addition of Biogeochemical Float Observations
Seth M. Bushinsky, Peter Landschützer, Christian Rödenbeck, Alison R. Gray, David Baker, Matthew R. Mazloff, Laure Resplandy, ...
2019, Global Biogeochemical Cycles - Article
New estimates of pCO_2 from profiling floats deployed by the Southern Ocean Carbon and Climate Observations and Modeling (SOCCOM) project have demonstrated the importance of wintertim... more
Citations 6 Altmetric 61 View PDF Add to Library

Decadal trends in the ocean carbon sink
Tim DeVries, Corinne Le Quéré, Oliver Andrews, Sarah Berthet, Judith Hauck, Tatiana Ilyina, Peter Landschützer, Andrew Lenton, I...
2019, Proceedings of the National Academy of Sciences of the United States of America - Article
Measurements show large decadal variability in the rate of [Formula: see text] accumulation in the atmosphere that is not driven by [Formula: see text] emissions. The decade of the 1990s experienced e... more
Citations 14 Altmetric 101 Open Access Add to Library

A decadal inversion of CO2 using the Global Eulerian–Lagrangian Coupled Atmospheric model (GELCA): sensitivity to the ground-based observation network
T. Shirai, M. Ishizawa, R. Zhuravlev, A. Ganshin, D. Belikov, M. Saito, T. Oda, V. Valsala, A.J. Gomez-Pelaez, R. Langenfelds, S. Mak...
2017, Tellus B - Article
We present an assimilation system for atmospheric carbon dioxide (CO_2) using a Global Eulerian-Lagrangian Coupled Atmospheric model (GELCA), and demonstrate its capability to capture the ob... more
Citations 5 Altmetric 1 View PDF Add to Library

Cathepsin S of Sciaenops ocellatus: Identification, transcriptional expression and enzymatic activity
Bo-guang Sun, Heng Chi
2016, International Journal of Biological Macromolecules - Article
Cathepsin S is a member of cysteine cathepsins and belongs to the cathepsin L-like family. In mammals, it is known to participate in various physiological processes and host immune defense. In teleost... more
Citations 7 Add to Library

ANALYTICAL VIEWS

RESEARCH CATEGORIES

04 Earth Sciences	5,243
06 Biological Sciences	3,459
09 Engineering	2,282
0406 Physical Geography and Environmental Geoscience	2,041
0401 Atmospheric Sciences	2,037

OVERVIEW

Citations
391 K

Citations (Mean)
30.88

RESEARCHERS

Philippe Ciais Laboratoire des Sciences du Climat et de l'Environnement, France	255
Dennis D Baldocchi University of California, Berkeley, United States	171
Markus Reichstein Max Planck Institute for Biogeochemistry, Germany	168
Thomas Andrew Black University of British Columbia, Canada	154
Jing Ming Chen Tulane University, United States	141

SOURCE TITLES

Agricultural and Forest Meteorology	912
Biogeosciences	406
Remote Sensing of Environment	381
Global Change Biology	352
Remote Sensing	347

Figure 4.8 Result of a query to track citations using keywords of ICOS related networks at Dimensions.ai. Numbers are very high as for the reasons indicated in the main text.

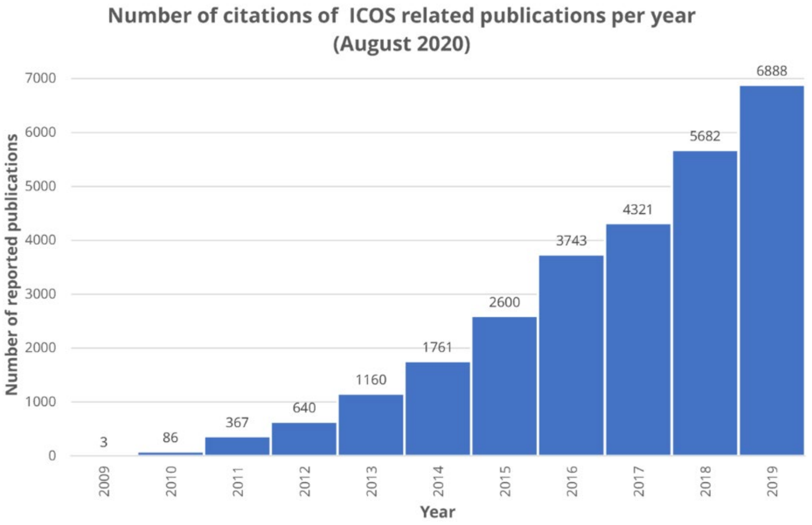


Figure 4.9 Bibliometric result of the number of citations of ICOS related publication gathered by the ICOS community.

ABBREVIATIONS and ACRONYMS

AS	Atmosphere Station
ATC	Atmosphere Thematic Centre
B2FIND	EUDAT CDI service to search for data object metadata
B2SAFE	EUDAT CDI service to store data
CC4BY	Creative Commons - Attribution 4.0 International data license
CDI	Common Data services Infrastructure
CO_2	Carbon Dioxide
CP	Carbon Portal
CTS	CoreTrustSeal
DOI	Digital Object Identifier system
DSA-WDS	Data Seal of Approval-World Data System
ENVRI	ENVironmental Research Infrastructure
ERIC	European Research Infrastructure Consortium
ES	Ecosystem Station
ETC	Ecosystem thematic Centre
FAIR	Findable, Accessible, Interoperable, Reusable
GCOS	Global Climate Observing System
GTOS	and Terrestrial Observing Syste
GEO	Group on Earth Observations
GHG	GreenHouse Gas
GCP	Global Carbon Project
ICOS	Integrated Carbon Observing System
IW	Internal Working (data, Level 1)
JSON	JavaScript Object Notation – a lightweight (self-descriptive) data-interchange format
MSA	Monitoring Station Assembly
NEON	National Ecological Observatory Network (USA)
netCDF	network Common Data Form
NOAA	National Oceanic and Atmospheric Administration (USA federal agency)
NRT	Near Real Time
ObsPack	Observation Package
OTC	Ocean Thematic Centre
OWL	Web Ontology Language
RDF	Resource Description Framework
RDL	Registration, Disclaimer and Licensing
RI	Research Infrastructure
SHA	Secure Hash Algorithms
SHA-256	Hash algorithm using a fixes 256 bits hash size
SOOP	Ships of Opportunity
TC	Thematic Centre
TR	Trusted Respository
PID	Persistent Identifier
SOCAT	Surface Ocean CO_2 Atlas
STILT	Stochastic Time-Inverted Lagrangian Transport (atmospheric transport model)
SparQL	recursive acronym for SPARQL Protocol and RDF Query Language
UNFCCC	United Nations Framework for the Convention on Climate Change

URI	Universal Resource Identifier
VOS	Voluntary Observing Ship
WDCGG	World Data Centre for Greenhouse Gases
WIGOS	WMO Integrated Global Observation System
XML	eXtended Markup Language