ICOS collated requirements

Introduction to ICOS

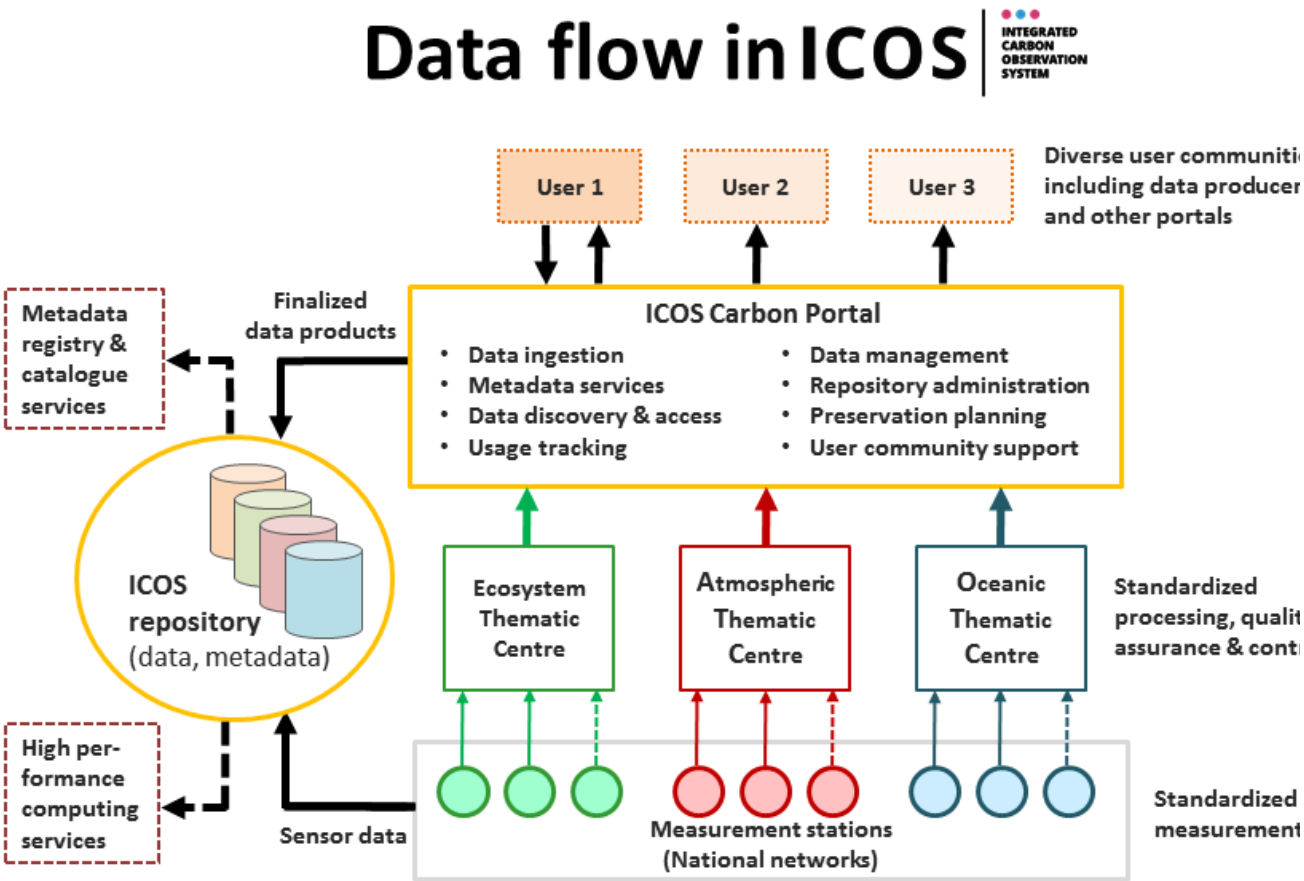
What is ICOS?

ICOS (Integrated Carbon Observation System) is a pan-European research infrastructure for observing and understanding the greenhouse gas (GHG) balance of Europe and its adjacent regions. The major task of ICOS is to collect and make available high-quality observational data from its state-of-the-art measurement stations operated with a long-term (20+ years) perspective. These data will contribute to research aiming to describe and understand the present state of the global carbon cycle, as well as help predict future behavior of GHG emissions. Importantly, all data produced and distributed by ICOS will be openly available to everyone wishing to use them, under a license similar to Creative Commons 4 Attribution-ShareAlike.

At its inception as an ERIC in November 2015, ICOS RI has 9 member countries: Belgium, Finland, France, Germany, Italy, the Netherlands, Norway, Sweden and Switzerland. The ICOS organization is quite distributed, with the Head Office located in Finland, the Carbon Portal data center in Sweden, and a number of central facilities (thematic centers and central analytical laboratories) hosted by Belgium, France, Germany, Italy and Norway.

Data flow in ICOS

The figure below schematically illustrates the data flow in ICOS.



The measurement station networks of ICOS span three themes - atmosphere, ecosystem and ocean - together, these provide information on greenhouse gas concentrations and exchange, meteorological and other environmental variables. Measurements are carried out on ecosystem sites, in tall atmospheric towers and on oceanic platforms and vessels. The stations are operated by the ICOS member countries. The collected data are then processed at common thematic centers, one for each main branch (atmospheric, ecosystem, ocean). The thematic centers each operate local computing centers, where data are processed. In addition, the centers offer both ICOS observation station personnel and end users of ICOS data products expert advice and support in technical matters.

Quality-assured and -controlled data products from the thematic centers are transferred to the Carbon Portal, which stores the data & associated metadata in the ICOS community data repository. The Carbon Portal acts as a “one-stop shop” for ICOS data products, featuring advanced search, visualization and downloading services.

The portal is also responsible for the central ICOS functions of curation, cataloguing, assigning identifiers & facilitating data citation, data usage tracking and long-term archiving, as well as for providing user community support. Finally, it will also manage “elaborated” data products from external users.

High diversity of data producers, products and users

A general characteristic of ICOS is diversity: diversity among data producers, data products and data users. In the following, we briefly describe these three aspects, and the challenges they bring to ICOS.

Data producers: In ICOS, data are produced at several levels: “Raw” observation data are collected at ICOS measurement stations, which are operated on a national level by research institutes or similar organizations. Next, the ICOS thematic centers take over to process and refine the raw sensor data in a standardized manner. Finally, expert users (mainly external to ICOS) make use of ICOS observations to produce various kinds of “elaborated products”.

Although many aspects of data management can and will be harmonized throughout the RI, there exists a broad range of tools and practices that are in parallel use, especially when comparing the details of the different thematic centers’ work flow. The challenge is to bring together all outcomes under a common data curation scheme.

Data products: These consist mainly of three kinds: 1) raw sensor data collected at the measurement stations associated with ICOS RI; 2) aggregated and quality-controlled observational data that are produced by ICOS expert centers based on the sensor data; and 3) so-called “elaborated” data produced by researchers external to ICOS, but based (in part) on ICOS observational data. The latter are typically results from calculations modeling global or regional greenhouse gas budgets.

The data products differ not only in content but also encompass a range of different sizes, release frequencies, file formats etc. This introduces a need for unique persistent identifiers at all levels of the data life cycle, and a common “data object metadata database” which can act both as a catalog and as a knowledge repository with respect to e.g. data types.

Data users: ICOS expects to serve users from a broad spectrum of communities and categories, including “experts” (with background in atmospheric, ecosystem, climate and environmental sciences), “other scientists” (with background in other fields, like medicine, geosciences, geography etc.), “educational” (teachers wanting to use data in courses, students needing data for reports & theses), “policymakers” (including governmental agencies), “companies” (wishing to use data for services, or interested in developing new measurement techniques), and “general public”. Each of these groups has quite different needs and interests.

Requirements gathered

General requirements

General requirements for ICOS

ICOS data management centers around four aspects: trust & transparency, versioning support, discoverability & ease of use, and long-term availability.

Firstly, (observational) data related to the environment, the climate system and greenhouse gases are of great global importance, both scientifically and “politically”. As such, they are subject to intense scrutiny from many interested parties. It is therefore essential that trust, transparency and verifiability are maintained throughout the entire data lifecycle. Methods for unambiguous identification of the data objects and related metadata must be combined with tools to check authenticity and fixity. At the same time, a consistent application of PIDs also offers solid support for proper data citation, which is a prerequisite for ensuring reproducibility (both of (RI-internal) work flows and in the scientific research process). In addition, citability facilitates the tracing of data usage, (evaluation of bibliometric statistics), and ensures consistent assignment of credit to data producers down to individuals (observation station personnel, thematic center experts, data curators, etc.)

Secondly, much of ICOS data consist of time series of e.g., atmospheric, ecosystem-related and meteorological variables, some of which are evaluated from measurements using complex algorithms. In some senses, the time series are open-ended - new data are continuously added as time progresses, which adds a dynamic aspect to the data. In addition as the scientific understanding of exchange processes between the Earth’s surface and the atmosphere deepens, new analysis methods become available, necessitating re-evaluations of existing sensor data. Together, these circumstances make a strong case for storing ICOS data in *database structures* that contain both the latest and previous sets of values for each parameter - and therefore may be considered as *fully versionable*.

Thirdly, an efficient cataloguing service, allowing searches both for datasets and their contents, is a pre-requisite for the functionalities of the ICOS data center. Users must be able to locate and pin-point the data of interest to them, obtain and view all relevant metadata, visualize the data values and of course download it. Access to complete and relevant metadata, including provenance tracking, will be central to most, thus requiring comprehensive curation.

Fourth, to ensure long-term sustainable access to ICOS data, the RI intends to set up and operate its own community data repository. The design of this ICOS Repository will be based on the Open Archival Information System (OAIS) reference model. In OAIS terms, the repository functionality will include most of the main functions of a data archive: Ingestion, Management and Access, as well as relevant parts of the Administration, Preservation Planning and Management layers. The only function that is envisaged to be (partly) outsourced is the long-term archival storage, which is foreseen to take place at an external trusted data center operating the EUDAT B2SAFE service. The intention is to apply for Data Seal of Approval (DSA) status for the ICOS repository.

Central to all these is the ability of the RI to operate a comprehensive and continuously updated *metadata database that describes all ICOS data objects* - including sensor data, aggregated data products, observation station information and measurement protocols. This database will be the backbone of the ICOS cataloguing service, serving the data discovery functionalities of the Carbon Portal, and supporting the long-term repository archiving. The data object metadata database (DOMDB) design must be flexible in order to both handle (merge) the various ICOS-internal metadata schemas, as well as allowing efficient interfacing with other data portals and cataloguing services.

Read more at the "General requirements for ICOS" page!

Identification and citation

Identification and citation in ICOS
ICOS will implement the allocation of unique persistent identifiers to all data objects that should be referable, both in the ICOS-internal work flows and to end users of ICOS data products. Since much of ICOS data will be dynamic in nature, the RI is interested in implementing a system for data citation that allows for versioning, as outlined by the Research Data Alliance WG on Data Citation.

Read more at the "Identification and citation in ICOS" page!

Curation

Curation in ICOS
ICOS intends to set up and operate its own community data repository, with a design based on the Open Archival Information System (OAIS) reference model. The intention is to apply for Data Seal of Approval (DSA) status for the ICOS repository. In OAIS terms, the repository functionality will include most of the main functions of a data archive: Ingestion, Management and Access, as well as relevant parts of the Administration, Preservation Planning and Management layers. Long-term archival storage is foreseen to take place at an external trusted data center operating the EUDAT B2SAFE service.

Read more at the "Curation in ICOS" page!

Cataloguing

Cataloguing in ICOS
Central to all ICOS operations will be a comprehensive and continuously updated *metadata database that describes all ICOS data objects* - including sensor data, aggregated data products, observation station information and measurement protocols. This database will be the backbone of the ICOS cataloguing service, serving the data discovery functionalities of the Carbon Portal, and supporting the long-term repository archiving. The data object metadata database (DOMDB) design must be flexible in order to both handle (merge) the various ICOS-internal metadata schemas, as well as allowing efficient interfacing with other data portals and cataloguing services.

Read more at the "Cataloguing in ICOS" page!

Processing

Processing in ICOS
ICOS handles both observational data (collected by its own networks of measurement stations) and data associated with atmospheric and ecosystem modeling (mainly performed by external research groups). The requirements differ somewhat: the observations are mainly sensor time series data in tabular form, which have to be calibrated, quality controlled, and gap filled. The evaluation of greenhouse gas and energy fluxes involves quite complex calculations. Modelling, on the other hand, involves both, preparation of input data (combination of many different data sources), various calculation steps, and processing of the output data. Model runs typically require access to high-throughput and high-performance computing facilities.

Read more at the "Processing in ICOS" page!

Optimisation

Optimisation in ICOS
ICOS has chosen not to address this issue at the moment, as the RI is just starting its operational phase.

Read more at the "Optimisation in ICOS" page!

Provenance

Provenance in ICOS
ICOS observational data products comprise data collected at 100+ individual stations, each equipped with hundreds of sensors, so it is very important to record exactly where data values come from. Also the entire chain of processing steps - including calibrations, quality control, and gap filling - must be traceable. Some variables are calculated and/or aggregated from high-frequency observations using complex data processing, and to ensure reproducibility, information about which software versions and parameter sets were applied is crucial. Similar considerations are also applicable to ICOS “elaborated” data products.

Read more at the "Provenance in ICOS" page!

Community support

Community support in ICOS
ICOS must be prepared to support its user communities on several levels, including user-friendly web-based interfaces for discovering and accessing ICOS data products, providing documentation and support information (e.g., as a wiki and a FAQ), operating a “help desk”, and offering training (tutorials, workshops). At the same time, ICOS also recognizes the need to provide training on ITC- and e-Infrastructure-related topics to its own personnel.

Read more at the "Community support in ICOS" page!

Summary and conclusions