General requirements for ICOS

Context of general requirements in ICOS

Complete report on general requirements for ICOS available at: https://envriplus.manageprojects.com/projects/requirements/notebooks/470/pages/60

The detailed RI response to this part of the requirements questionnaire is attached: ICOS - 0 - General questions 2016-01-27.docx

Summary of ICOS general requirements

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.

Detailed requirements

Identification & citation

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.

Curation

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.

Cataloguing

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.

Processing

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.

Provenance

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.

Optimization

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

Community support

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.

Formalities (who & when)

Go-between	Alex Vermeulen
RI representative	Maggie Hellström
Period of requirements collection	September 2016 - December 2015
Status	Finished