The ICOS OTC Data Lifecycle Plan

Camilla Stegen Landa, University of Bergen and Bjerknes Centre for Climate Research (Norway), Camilla.Landa@uib.no
Steve Jones, University of Exeter (UK), s.d.jones@exeter.ac.uk
Benjamin Pfeil, University of Bergen and Bjerknes Centre for Climate Research (Norway), Benjamin.Pfeil@uib.no
Truls Johannessen, University of Bergen and Bjerknes Centre for Climate Research (Norway), Truls.Johannessen@uib.no

Introduction

The Integrated Carbon Observation System (ICOS) is a Pan-European research infrastructure aiming to provide the long-term observations required to understand the present state and predict future behavior of the global carbon cycle and greenhouse gas emissions. To facilitate the research that will lead to this understanding, ICOS will offer a high precision, long-term observing network of stations across Europe and adjacent regions, which measures greenhouse gas fluxes from ecosystems and oceans, and greenhouse gas concentrations in the atmosphere. The specific tasks of collecting and processing data measured at national network stations are divided among the Ocean Thematic Centre (OTC), the Ecosystem Thematic Centre (ETC) and the Atmosphere Thematic Centre (ATC). All quality controlled data will be made available through the Carbon Portal (CP).

The distributed, interdisciplinary and complex nature of this infrastructure enforces that a data lifecycle plan is needed in order to be able to make the data accessible for current and future climate change research. The data lifecycle plan is an essential part of data management plans for European Infrastructure projects, and becomes increasingly important for projects dealing with large data volumes and quality control performed by various entities. The data lifecycle plan documents in detail all changes applied to a dataset from its collection at the sensor until its final publication, and allows a transparent tracking of all quality control procedures applied. Responsibilities should also be assigned in the plan.

Data definitions, transfers and responsibilities have been discussed within ICOS for several years and the data lifecycle plan is aligned with recommendations of the *Common Operations of Environmental Research Infrastructures Reference Model (ENVRI RM)*. ENVRI RM is a common ontological framework and standard for the description and characterization of environmental research infrastructures. The ENVRI RM identifies a set of common functionalities in research infrastructures and provides solutions to common problems. The concepts and frameworks of the ENVRI RM were used to examine the requirements and optimize the design of the ICOS data infrastructure.

Method

Our focus is on the data lifecycle plan for the ICOS OTC. The OTC is hosted by Norway and the United Kingdom and is responsible for coordinating the marine network of ICOS. The marine network consists of instrumented Ships of Opportunity/Voluntary Observation Ships (SOOP/VOS), fixed ocean time series stations and repeat hydrography (Fig. 1). SOOP/VOS and fixed time series stations are equipped with a suite of automated instrumentation to measure atmospheric and surface ocean pCO2, sea surface temperature, salinity and related variables. During repeat hydrographic cruises discrete samples of total alkalinity, total carbon, pH, nutrients and related variables are taken. Approximately 50 different lines and stations operated from 12 countries are planned for the ocean network.

The data lifecycle within OTC is of complex nature with many steps and transmissions of data and metadata between the OTC, the CP and the principal investigators. However, the data lifecycle can be boiled down to data acquisition, data curation and data access. Within these basic pillars are many steps which include different levels of quality checks and assurance, version control, data archiving, assigning of persistent digital identifiers, and data publications. The OTC will aim to follow international procedures and best practices for the above steps.

An automated system for data submission and quality control will be established for the automated data streams from the marine network of ICOS. This system will be based upon the automated data ingestion system as developed within the marine biogeochemistry community for the Surface Ocean CO2 Atlas (SOCAT) project (www.socat.info), which contains amongst others user friendly quality control tools, automated range checkers and visualization options. These features will be optimized to OTC's needs. The datasets submitted to the automated ingestion system for SOCAT are quality controlled by the principal investigators, and in many cases made public, prior to submission. OTC plan to extend and optimize this infrastructure and speed up the process of data release. All tool-kits developed will be made available to the international marine community with the aim that raw data is streamed directly from the instruments into automation systems.

All changes applied to a dataset will be logged and archived using a version control system so that older versions can be restored if needed. In addition, different data levels have been defined, starting from raw data and ending with the published dataset (or data product) which is the final data level. The data levels in between differ in the amount of processing and quality control applied. Users can get access to data from different levels depending on their needs, which are often a counterbalance between the data quality needed and the freshness of data. It will be possible for users to get access to near real-time data, final quality controlled ICOS datasets, and data products where ICOS data are included. Persistent digital identifiers (DOIs) will be assigned for data citation to ensure that data sources, providers and funding sources are being acknowledged.

Standardized vocabularies from the *Natural Environmental Research Council* (NERC) vocabulary server will be implemented for the OTC data. This vocabulary server was developed by the NERC DataGrid program as part of an effort to support uniform data and metadata discovery and access. Standardized vocabularies will ease the use of the data by the modeling community and exchange of data with different communities, and ensure interoperability with existing systems (e.g. SeaDataNet and EmodNet).

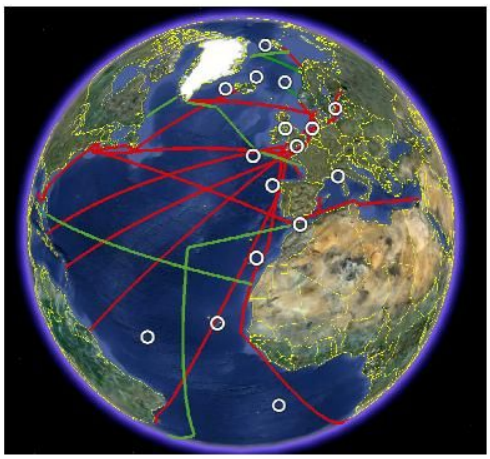

Fig. 1: The suggested network of stations for the ocean network: Fixed Ocean Stations (cycles), Ships of Opportunity/Voluntary Observation Ships (red lines) and Repeat Sections (green lines).