Near Real Time Data Processing In ICOS RI

Margareta Hellström, Alex Vermeulen and Oleg Mirzov
ICOS Carbon Portal
Lund University
Lund, Sweden

Simone Sabbatini, Domenico Vitale and Dario Papale
ICOS Ecosystem Thematic Centre
University of Tuscia
Viterbo, Italy

Jérôme Tarniewicz, Lynn Hazan and Leonard Rivier
ICOS Atmosphere Thematic Centre
Laboratoire des Sciences du Climat et de l'Environnement
Gif-sur-Yvette, France

Steve D. Jones, Benjamin Pfeil and Truls Johannessen
ICOS Oceanic Thematic Centre
University of Bergen
Bergen, Norway

***Abstract*— This paper describes the implementation of (near) real-time (NRT) data processing in the recently launched European environmental research infrastructure ICOS. NRT applications include handling of raw sensor data (including safe storage and quality control), processing and evaluation of greenhouse gas mixing ratios and exchange fluxes, and the provision of data to the RI's user communities.**

Keywords*— *near real-time; data processing; research infrastructure; greenhouse gases; environmental monitoring

I. INTRODUCTION

A. What is ICOS?

ICOS (Integrated Carbon Observation System; see https://icos-ri.eu/) 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. After careful quality control, processing and aggregation, these ICOS data will contribute to research aiming to describe and understand the present state of the global carbon cycle in Europe and adjacent regions, as well as help predict future behaviour of GHG emissions in the face of the on-going global climate change [1].

Historically, the processing and subsequent dissemination of environmental observation data of the types covered by ICOS tended to be performed on a seasonal or annual time cycle - mainly due to delays in collecting and synchronising all related measurements, ancillary information and metadata. During the last decade, an increased demand for more frequent data releases - for example to allow rapid assessment of the impact on ecosystems from extreme weather events, and for contributing to global air quality forecasting services - has prompted the development of more automated data handling and processing work flows. At the same time, demands have increased for more sophisticated data management practices, including storage also of raw data in trusted repositories, implementation of rich metadata schemes, and the assignment of persistent identifiers to support citation and provenance tracking.

With its concerted efforts to both achieve a high degree of standardisation and measurement quality across a dense network of measurement stations, and to provide fast access to data on greenhouse gas concentration and flux estimates, ICOS represents the new generation of European environmental observation initiatives.

B. Overview of the ICOS organization

At the time of writing (September 2016), ICOS RI has 12 member countries: Belgium, the Czech Republic, Denmark, Finland, France, Germany, Italy, the Netherlands, Norway, Sweden, Switzerland and the United Kingdom. The ICOS organization is distributed, with the Head Office located in Finland, the Carbon Portal data centre in Sweden, and a number of central facilities (thematic centres and central analytical laboratories) hosted by Belgium, France, Germany, Italy, Norway and the United Kingdom. Fig. 1 shows a schematic of the data flow in ICOS.

The measurement station networks of ICOS span three domains - atmosphere, ecosystem and ocean. Together, these provide information on greenhouse gas concentrations and exchange, meteorological and other environmental variables. Measurements are carried out at ecosystem sites, in tall towers and on oceanic platforms and vessels. The stations are operated by the ICOS member countries. An up-to-date interactive listing of all ICOS stations is available at https://www.icos-cp.eu/node/83.

C. Data flow in ICOS

The sensors deployed at the ICOS stations mainly fall into three categories: i) infra-red absorption gas analysers (dye lasers) with readout frequencies on the order of seconds (dependent on gas species); ii) sonic anemometers for two- or three-dimensional wind velocity, with readout frequencies of ca 10 Hz; iii) devices measuring meteorological and environmental variables by monitoring resistance, voltages or

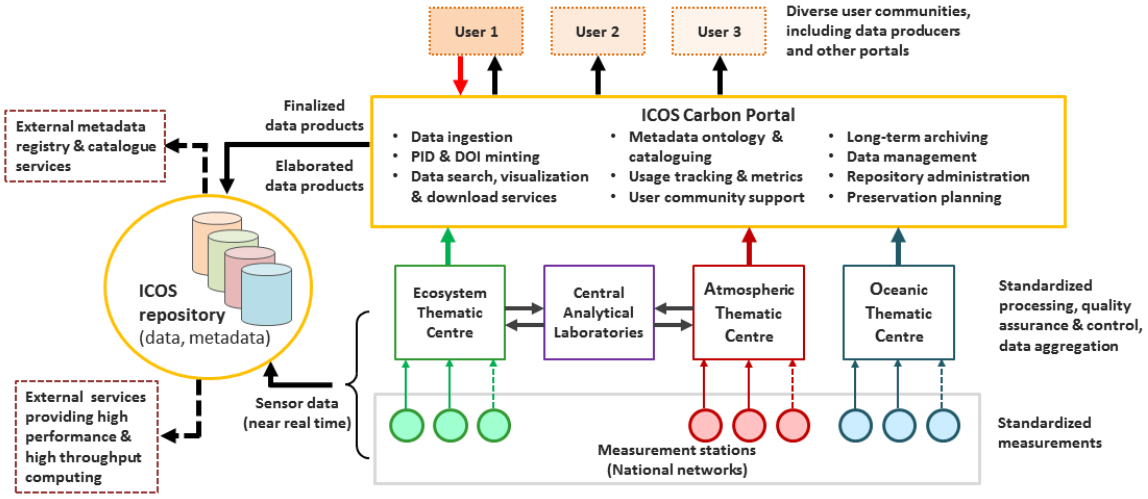

Fig. 1. Schematic of the data flow in ICOS. Sensor data, collected at ICOS observation stations, are archived in a central, trusted repository. Standardized data processing is then performed by experts at one of three Thematic Centres. Near-real time data, as well as quality controlled & aggregated ICOS data products are published via Carbon Portal, which also disseminates user-derived "elaborated" data products. A centralized ICOS metadata catalog supports data discovery & usage tracking. ICOS data can be staged to high throughput (HTC) and high performance (HPC)computation facilities.

currents at readout frequencies ranging from seconds to minutes. Sensor output is typically collected and formatted either inside the instruments themselves, or by using data logger systems. All readout values must be provided with timing information, to allow parallel data streams to be synchronized during the subsequent processing.

After measurements are performed at stations, the sensor data from all stations are transferred to the central ICOS data repository for safe storage. From there, the raw data are staged to the common thematic centres, one for each main branch (atmospheric, ecosystem, ocean), for processing. The thematic centres each operate local computing facilities, where data are processed in a standardized way according to well-defined protocols. In addition, the centres offer expert advice and support in technical matters to both ICOS observation station personnel and end users of ICOS data products.

Quality-assured and -controlled data products from the thematic centres are transferred to the Carbon Portal (https://icos-cp.eu/), 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.

A general characteristic of ICOS is diversity, not only among data products but also data producers and data users. In the following, we briefly describe these three aspects, and the challenges they bring to ICOS, especially in terms of standardisation and comparability necessary in such kind of networks.

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 centres 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 "elaborated data 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 centres' work flow. The challenge is to bring together all outcomes under a common data curation scheme.

Data products: These consist mainly of three kinds: i) raw sensor data collected at the measurement stations associated with ICOS RI (known as Level 0 data); ii) aggregated and quality-controlled observational data that are produced by ICOS expert centres based on the sensor data (Level 1 and Level 2); and iii) so-called "elaborated" data produced by researchers external to ICOS, but based (in part) on ICOS observational data (Level 3). The latter are typically results from calculations modelling global or regional greenhouse gas budgets. A special, fourth case comprises near real-time data sets (of e.g. greenhouse gas concentrations and/or fluxes) that are provided to users after applying basic quality controls.

The ICOS 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 catalogue and as a knowledge repository with respect to e.g. data types. Importantly, all data produced and distributed by ICOS will be openly available to everyone wishing to use

them, under a Creative Commons 4 Attribution-ShareAlike license.

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.

II. Processing Of ICOS Data

A. *Why is NRT data processing important for ICOS?*

Each of the three ICOS domains has its own definitions and time scales for what is considered to be NRT data. These differences are mainly grounded in the characteristics of the analysis and quality control techniques and algorithms that are used. Thus, typical time delays between sensor data collection and NRT data dissemination range from 36-48 hours for atmospheric data, 10-30 days for ecosystem data, to 1-2 months for ocean data. (These domain-specific aspects are discussed more in detail in Sections II.B-D below.)

But regardless of the actual time scale involved, near real-time data processing occurs at several levels of ICOS, as discussed in the following:

Station level: Because ICOS measurements capture characteristics of the natural world the observations are time and location-unique, and therefore not reproducible. This implies that the collected sensor data needs to be both safely stored and quality controlled as soon as possible after acquisition. The first step depends on implementing fast, preferably automated transmission of data and metadata from all ICOS measurement stations to a common trusted repository operated under strict data security principles. The second step requires the data can undergo at least a basic quality control (and in some cases also specific processing & evaluation), preferably within hours of collection, in order to quickly detect any issues, e.g., with the instrumentation. In ICOS, this near real-time quality checks can be performed both (semi-) manually by the staff at the measurement stations (not covered in this article), and by automated processing at the TCs. If bad or missing sensor data are detected, rapid action can be taken to remedy the problems. (This should not be confused with the much more rigorous quality assessment and flagging that is applied during the production of the final Level 2 data products.)

Thematic Centre level: Data processing at the TCs should be as automated as possible, but it is of course also highly reliant on intervention by skilled experts. The first stage of the data quality control is focused on flagging instrument malfunction, e.g. via applying threshold filtering and statistical detection of outliers. Information on detected problems is directly communicated back to the observation stations. While some aspects of the data analysis do require data collected over a longer period to be treated concurrently, much of the processing is performed as soon as the data are available - i.e. in near real-time.

Global monitoring system level: Typical users of ICOS near real-time data include projects aiming to produce short-term predictions for greenhouse gas concentrations and air quality on a global scale. For this purpose, measured data on e.g. CO_2 concentrations should be available with as little delay as possible in order to evaluate trends. The related data assimilation systems are configured to work with non-finalized input data with relatively high levels of uncertainties - i.e. the output of the automated first-step quality control.

Extreme event detection: One of the missions of ICOS (see https://www.icos-ri.eu/about-us/our-mission) is to be able to quickly provide information to stakeholders and policy makers on effects of "weather and climate extreme events", such as droughts, heat waves or intense cold spells, on the overall greenhouse gas budget of Europe. This requires that preliminary values on e.g. carbon dioxide and methane source and sink strength are being continually evaluated at close to real-time scales (days or weeks after the start of an event).

Outreach activities: As part of the standard measurement program, ICOS stations also monitor meteorological variables, including air temperature and pressure, precipitation and wind speed & direction. By providing this information in near real-time to the general public living in the vicinity of the stations can be a useful complement to data from the respective national weather services.

In the following, the data processing performed at ICOS three Thematic Centres (for atmosphere, ecosystems and oceanic data) is described in more detail. We also outline the data and metadata ingestion service implemented by the ICOS Carbon Portal data centre, which focuses on providing support for automated near-real time processing at the Thematic Centres.

B. *The Atmospheric Thematic Centre (ATC)*

The ICOS Atmospheric Thematic Centre (ATC; see http://www.icos-atc.eu) has main facilities at LSCE in Paris, France. ATC also includes a mobile laboratory for quality control at stations operated by FMI, Finland. The ATC automatically processes atmospheric greenhouse gases mole fractions of data coming from sites of the ICOS network. The centralized data processing aims to reduce inter-laboratory differences and facilitate the production of a coherent dataset in near-real time i.e. on a daily basis for atmospheric measurement. Currently (September 2016), there are 11 stations being processed by ATC on a daily basis, and this will increase as the atmospheric station network is fully developed during 2017.

The instrumental raw data (Level 0 or L0) are transferred once a day from the monitoring sites to the ATC using the Secure File Transfer Protocol (SFTP). The files are first archived, and then automatically processed. The processing is based on shell scripting, Java and MySQL for the database. Traceability of the data downloads and long-term archival for raw data, are being implemented in collaboration with the

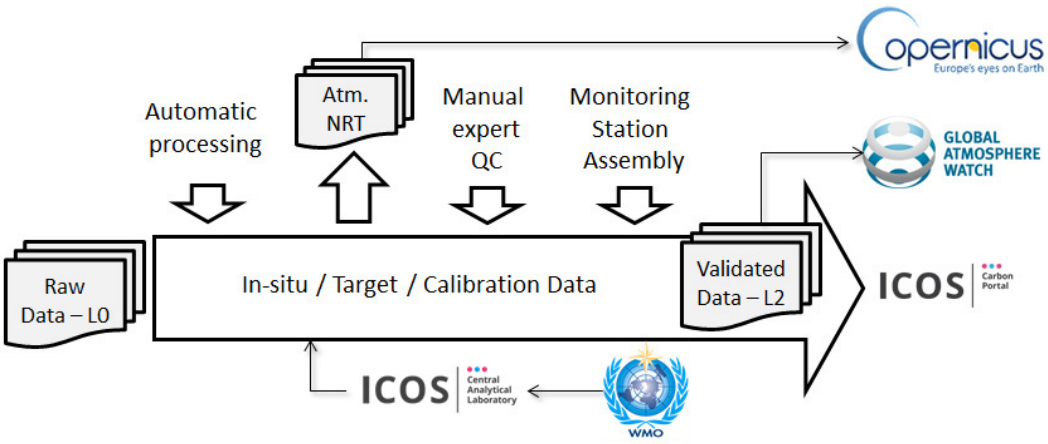

Fig. 2. The Near-Real-Time data processing flow implemented at the ICOS Atmosphere Thematic Centre, from L0 to L2 data. Atmospheric NRT data is a intermediate product of the whole processing chain at ICOS ATC. The Monitoring Station Assemblies (MSAs) for atmosphere, ecosystem and marine networks consists of scientific and technical experts from the stations of the ICOS National Networks of the ICOS ERIC member countries. The MSAs monitor, develop and improve the scientific and technical basis of the ICOS RI and work closely with ICOS Central Facilities.

Carbon Portal of ICOS. First step of the processing is the automatic flagging out of raw data using instrumental parameters (e.g., temperature, pressure, flowrate) which are checked against a valid interval or threshold.

Then CO_2 and CH_4 measurements have to be corrected for the effects of water vapour dilution and pressure broadening. Using calibration gases provided by ICOS Central Analytical Laboratories (CAL; see https://www.icos-cal.eu/), CO_2 and CH_4 mole fractions are corrected in a final step, by application a transfer function which is instrument, location, species and type of gas dependent. This step allows CO_2 and CH_4 measurements to be traceable to the WMO mole fraction scale. The automatic processing of atmospheric CO_2 and CH_4 mole fractions at the ICOS ATC is described in detail in [2].

The challenge of traceability is also handled using flagging of the data. Three types of flags are defined during the processing. The 'internal' flag is associated with the different steps of the processing and ensures the traceability of the process. Then a 'user' flag allows separating the valid and invalid data. A third type of flag named 'descriptive flag', allows the Principal Investigator (PI) of a station to provide codified reasons for invalidating data or useful information for validating data. ICOS ATC quality control and assurance procedure is a mix of automatic control and manual PI data screening.

The amount of data uploaded from GHG instruments to ATC is on the order of 6.5 MB/day per CO_2/CH_4 in situ analyser (not taking into account the complete high-resolution absorption spectra), corresponding to a total ~71 MB a day for 11 stations. Time synchronization is negligible with respect to the time transit from the inlet to the air analyser (~100m for tall towers). The synchronization between stations is achieved by requiring that all the data timestamps (or the beginning of the time intervals, if averaging) must be given in UTC and that all the used time servers are synchronized [3]. (The ATC system takes this for granted and does not perform checks on this.) L0 Data are ASCII space-separated values compressed files. Uncompressed version of data files is ~50 MB, corresponding to ~550 MB of L0 data being daily processed by ATC, for one GHG instrument.

Once data are handled and automatically processed at ATC, daily increment in ICOS ATC database represents ~275 MB a day for 11 stations. For end users, the default CO_2/CH_4 file size (hourly GHG concentrations with a basic metadata subset), directly extracted from the database is 3 KB/day per in situ analyser. More verbose version of NRT output files (same as previous output but with added ancillary data such as target concentration and analyser parameter) are bigger, with a total amount of 100 KB/day for 11 stations.

Two major end-user communities for NRT atmospheric measurements are identified to date. The first community is composed of the station PIs and station operators. NRT datasets facilitate station management and allows getting a fast feedback on the data; it improves reactivity in case of disruption in the data flow and thus limits data gaps. NRT atmospheric data allow for early-warning monitoring systems, for example, in the case of extreme GHG events (e.g., linked to droughts, high-pollution events). It allows, for example, to adjust the campaign setup and observation plan or to place more emphasis on a specific phenomenon.

A second type of end-user community are operational operators in the field of earth monitoring, e.g., the Atmosphere Monitoring Service CAMS of the Copernicus program (see for example https://atmosphere.copernicus.eu/services/air-quality-atmospheric-composition), and the European Centre for Medium-range Weather Forecasting (ECMWF) that computes GHG forecast-only and global GHG reanalysis products, with a temporal resolution of at least 6 hours (see e.g., http://www.ecmwf.int/en/forecasts). NRT atmospheric data sets can also be used to as input to operational systems for data assimilation. These assimilation model calculations also tend to improve operational meteorological forecasting. All the processing of NRT data at ATC, including interactions with international organizations and ICOS facilities (Central Analytical Laboratory and Carbon Portal) are summarized in fig. 2.

C. *The Ecosystem Thematic Centre (ETC)*

The ICOS Ecosystem Thematic Centre (ETC; see http://www.icos-etc.eu) has facilities at CMCC and UNITUS in Viterbo, Italy, at PLECO in Antwerp, Belgium and at INRA in Bordeaux, France. The ETC has implemented an Eddy Covariance (EC) raw data processing scheme that is designed to provide the scientific community with standardized half-hourly datasets of net ecosystem exchange (NEE) of CO_2, sensible (H) and latent (LE) heat fluxes, momentum flux (T) and several ecological and bio-meteorological time series collected at the ICOS ecosystem stations distributed all over the Europe. The final products will be both long term time series (Level 1 and 2) and Near Real Time (NRT) data that will include fluxes and variables from the previous 10 days.

The EC raw data processing scheme was developed following the most recent recommendations on flux calculation (see Chapters 3-4 of [4]). Quality assurance and quality control (QA/QC) procedures and uncertainty estimation are based on performing several combinations of processing options, with the aims to i) check and correct EC data in order to ensure that products are of the highest quality; and ii) provide information about both random [5] and systematic uncertainty. The implemented processing steps are outlined in Fig. 3. Both L2 and NRT products will be handled following the same steps. However, mainly because of the computational time constraints described below, some of the procedures will be adapted to the characteristics of the NRT data,

The raw sensor data, which consist of both high-frequency (acquisition frequency of 10 Hz) and low-frequency (acquisition frequency below 0.05 Hz) components, are collected either through data loggers or dedicated personal computers at the ecosystem stations, following the instruction protocols prescribed by the ETC. Data are transmitted every day as sets of half-hourly files (high frequency data) and daily files (meteorological data) to the Carbon Portal (CP), using the scheme outlined in Section II.E. All data files (in either ASCII or binary format) are stored in the ICOS repository, and assigned a persistent identifier that may be used to record provenance in the data processing system. The measurement metadata, together with other ancillary information needed to correctly process the data, are submitted directly from the stations to the ETC using the international Biological, Ancillary, Disturbance and Metadata (BADM) standard protocol (see http://ameriflux.lbl.gov/data/aboutdata/badm-data-product/).

The amount of raw data volumes for each day (dominated in size by the 48 half-hourly files from the EC sensors) will range from few to hundred(s) MB. These data will be staged on-demand from the ICOS repository storage, and processed by the ETC daily as to produce NRT datasets. Every 6-12 months, L1 and L2 datasets will be produced, by a process that requires re-analysis of all raw data from the corresponding period. The ETC processing scheme uses the gCube software system (https://www.gcube-system.org/) (supported by the D4Science organization; http://www.d4science.org), and is an implementation of the Hybrid Data Infrastructure (HDI) approach. The HDI is a recently developed effective solution for managing the new types of scientific datasets, and

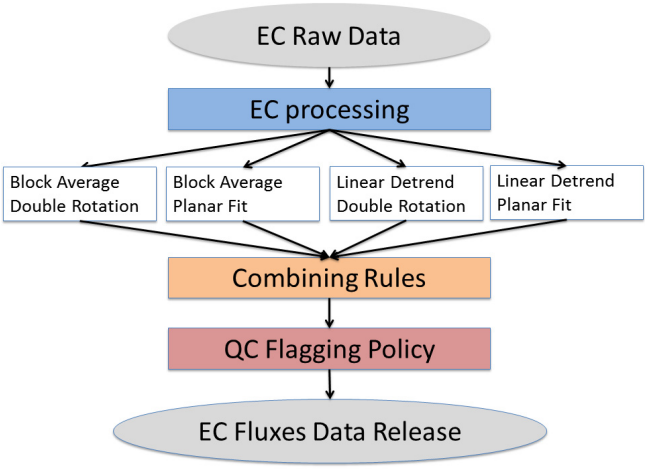

Fig. 3. The Near-Real-Time data processing flow implemented at the ICOS Ecosystem Thematic Centre. The same processing steps are used for the production of both NRT and Level 2 aggregated data products.

integrates several technologies, such as Grid and private and public Cloud, to provide flexible access to and usage of data and data-management capabilities [6]. With HDI it will be possible to easily provide the high-throughput computational power that is required to quickly process the large EC datasets from the network of over 60 ecosystem stations. The data processing algorithms are based on a combination of R scripts and the EddyPro® software package (LI-COR Biosciences, https://www.licor.com/env/products/eddy_covariance/eddypro.html).

GHG fluxes will be calculated according to four processing schemes combining two options for the anemometer tilt correction (Double Rotation and Planar Fit, see Chapter 3 of [4]), and two options to correct for trends (Block Average and Linear Detrending, see Chapter 7 of [4]). In the case of NRT processing, parameters for time lag optimization, spectral correction and planar fit method will be calculated periodically from a preliminary analysis of longer datasets (few months). The fluxes obtained from the previous steps will be combined according to rules described in the ICOS Processing Protocol (see [7], pp. 69-70), in order to allow the quantification of the total uncertainty arising from a combination of random and systematic (i.e. those associated to different processing schemes) sources.

The quality of the calculated half-hourly fluxes are subsequently flagged according to a set of quality tests performed during the different moments of the processing chain: at the high-frequency level a flagging strategy adapted from Vickers and Mahrt [8] will be applied, while the quality flagging policy from Mauder and Foken [9], involving the steady state and well-developed turbulence tests, will be used for half-hourly data. In addition, fluxes identified as spikes and NEE calculated below a given u* friction velocity threshold (estimated as in [10]) will also be flagged. (In contrast to the standard approach implemented for long time series of data, the u* threshold for NRT will be estimated through a preliminary analysis of the previous year of EC data from each

station.) A 1-year half-hourly file and a 10-day half-hourly NRT file will be about 30 MB and 0.5-1.0 MB, respectively. The calculated half-hourly fluxes files, together with aggregated meteorological variables and BADM metadata files are finally uploaded to the CP for publishing and dissemination.

D. The Ocean Thematic Centre (OTC)

The ICOS Ocean Thematic Centre (OTC; see http://www.icos-otc.eu/) has facilities at the University of Bergen, Norway and at the University of Exeter, UK. The OTC will receive continuous measurement data from instruments deployed on fixed buoys and voluntary observing ships (VOS), which are typically commercial ships that donate space for scientific work. As of writing (September 2016), 18 ocean stations have registered to participate in the ICOS project, all measuring CO_2 at the ocean surface.

The instruments in use are deployed in hostile environments: on buoys they are exposed to the elements, and VOS instruments are typically installed in the ship's engine room, with extreme temperatures and vibration. The combination of specialist application and small research budgets prevents the market from producing off-the-shelf products with the desired resilience. Indeed, many of these instruments are hand-built by the scientists themselves from basic components. Maintenance windows are typically small, with buoys being visited at sea, and ship access limited to a few hours during port visits. Failures are commonplace.

While satellite transmission of data is possible, budget constraints mean that this is rarely used. Data retrieval is therefore limited to routine maintenance visits every few weeks. After this, the data must be processed to convert the raw measurements into the required CO_2 values (calibration, temperature and moisture adjustments, etc.). Extensive quality control must be performed to ensure that all sensors are working correctly (the small signals being measured makes this is non-trivial). Much of the code for this work was developed by individual scientists leading to inconsistencies in processing methods. While expert manual quality control is still required, many simple checks (range checking, outliers etc.) could be automated to reduce the workload.

The lack of perceived need for fast access to this data historically meant that many months could pass before it was made public. Recent demands from a variety of stakeholders have brought all these shortcomings into focus, and created the need to expedite the entire process. The SOCAT project is the major central database for these measurements, and only in 2016 has this started to be published annually. This in turn feeds into the annual Global Carbon Budgets, an annual assessment of the state of the world's carbon balance. Although oceans tend to change slowly, there are sub-annual and inter-annual dynamics that are poorly understood, and the slow release of data means that these are often not studied until years after the fact.

The OTC is responding to these demands by developing tools to expedite the process of turning raw data into fully processed, quality controlled ocean surface CO_2 data. While much of the delay between measurement and publication cannot be removed, it should be possible to reduce the time required from months to weeks, while improving quality and consistency across the whole observation network. The core of this effort will be QuinCe, a web-based tool that will automate the processing of measurements and the basic quality control. Raw data (Level 0) from the instruments will be uploaded and converted to a common format. This will be processed according to the methods described by Pierrot et al. [11], ensuring consistency across all stations in the network and reducing the likelihood of unnoticed bugs.

Next, a selection of automatic quality control routines (GPS validation, range checking, spike detection etc.) will identify and flag possible problems. These routines will be based on those used in the SOCAT project [12] to maintain consistency with the wider global surface ocean carbon community. These reduced and automatically quality controlled measurements will constitute Level 1 data. The original scientist will review the Level 1 data and the assigned quality control flags directly

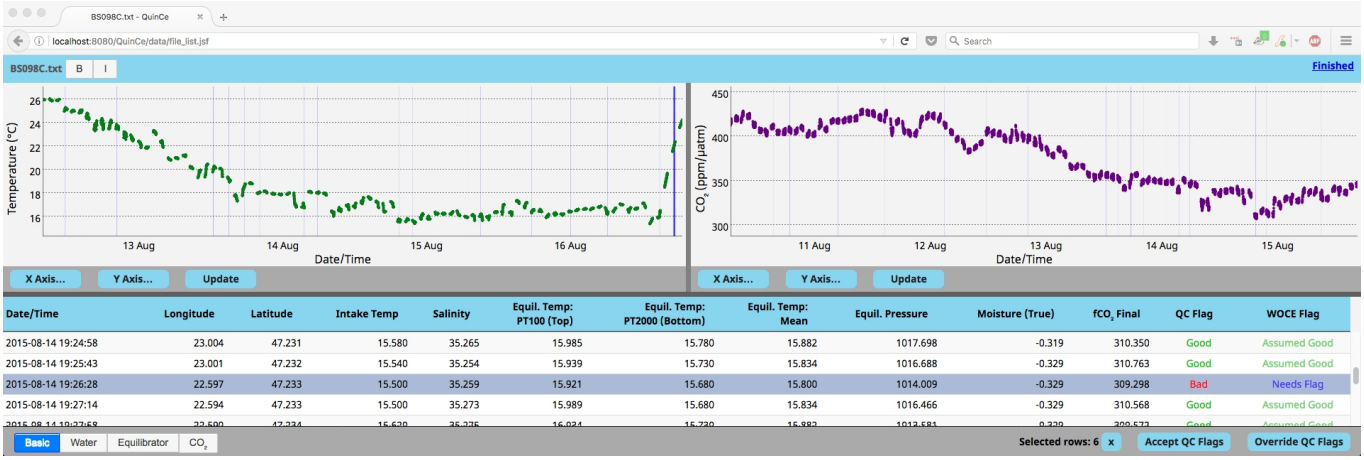

Fig. 4. Screenshot of the QuinCe QC system for use by ICOS ocean station Principal Investigators (PIs). Custom plots and maps can be shown for any variable, with the numeric data displayed beneath. Flags from the automatic QC routines are shown, and the PI can either accept them or override them. Interactive graphical-interface-based tools like QuinCe offer PIs an efficient and easy-to-use platform to quickly assess the quality of the data from their stations.

in QuinCe (Fig. 4) and either agree with them or override them if they decide that the automatic routines were mistaken. QuinCe will provide the ability to view the data using customizable plots and maps to assist with this process. Once all the flags have been checked, the data will be given a final check by the OTC, who will then publish the data set as Level 2 data via the ICOS Carbon Portal.

Further improvements in getting the data to potential users will be made by granting limited public access to the Level 1 data online. Since this is produced automatically, it can be ready in a matter of minutes from the time of upload. This data cannot officially be used since it has not had complete quality control and is therefore not deemed publicly available. Instead, previews of the data will be published in the form of visualisations, which can be used to compare with existing observational data. If users notice items of potential interest, communication channels will be available via the OTC to allow discussion with the originating scientists to identify potentially fruitful areas of further investigation.

E. Ingestion of data via the Carbon Portal

The ICOS Carbon Portal (CP; see https://icos-cp.eu/) is hosted by Lund University, Sweden. The CP is tasked with designing and implementing the set of services that form the backbone of the data management of the whole RI. The list includes services supporting i) data ingestion & storage, including minting of persistent identifiers (described in more detail below); ii) staging data from the repository to high-performance and high-throughput computing resources; iii) easy-access cataloguing on top of an ontology-based metadata database (RDF triple store accessible via a SPARQL endpoint); iv) single-sign on AAI (authorization, authentication and identification) for ICOS services; and v) a virtual research environment (VRE) platform for user-initiated data processing (based on Jupyter Notebook running on virtual machine instances). As far as possible, the CP bases all its data management and computing services on Open Source technology.

Of interest here, the CP is currently implementing a first version of a system for automatically ingesting data objects into a trusted repository - based on the EUDAT B2SAFE service (see https://eudat.eu/services/b2safe) - for archiving. This ingestion system, with its RESTful HTTP API, will ultimately be used for all kinds of ICOS data, from raw sensor data to user-generated elaborated products. ICOS will be using the services of two providers of persistent identifiers (PIDs), namely ePIC (the European Persistent Identifier Consortium; http://www.pidconsortium.eu/) for Level 0, NRT and Level 1 data, and DataCite (https://www.datacite.org/) for Level 2 and Level 3 data products.

Fig. 5 illustrates the main steps of this data ingestion process, which is explained in more detail in the following (numbering refers to the figure).

1) The "Uploader" prepares a Data Object (DO) and an associated Metadata Object (MO), containing information about owner, origin, file title, checksum and data type. The MO is uploaded to the "Meta" service of the Carbon Portal. "Meta" checks consistency & contents of the MO and stores

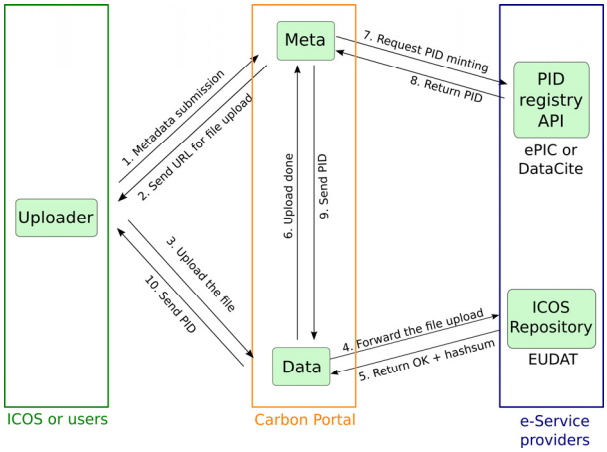

Fig. 5. The Carbon Portal "Meta" and "Data" services facilitate the ingestion of data from an Uploader into the trusted ICOS Repository (see also Fig. 1). All ICOS data objects are automatically assigned a persistent and unique digital identifier at the time of ingestion.

the information in the ICOS metadata store. If there is a problem, "Meta" informs "Uploader" and requests a re-submit of the MO.

2) If all is OK, "Meta" creates a unique hexadecimal string, based on the checksum. This string will be used as suffix for the persistent identifier that will be created for the DO, and will also be contained in the URL of the landing page of the object. Finally, "Meta" generates a specific URL, to be used for the DO upload, and returns this to "Uploader".

3) "Uploader" initiates the transfer of the DO to the Carbon Portal "Data" service.

4) The DO is transferred as a bit stream to the "Repository". During streaming, the checksum of the DO is calculated on the fly. Depending on the specified data type, a local copy of the DO may be created by "Data", but the default is to only store the DO at the "Repository".

5) After the transfer is completed, "Repository" returns a receipt containing local file information (path, local identifier etc.) and the checksum. If a checksum discrepancy is found, "Data" returns an error code to "Uploader", who can perform a re-submit.

6) "Data" sends a "storage completed" message to "Meta", including the relevant repository metadata.

7) "Meta" sends a request to the "PID registry" to register a PID for the DO. The request includes the ICOS-generated suffix and all relevant metadata, e.g. hash sum and other DO properties, as well as the resource locator of landing page.

8) The "PID registry" registers a PID handle for the DO. If something is wrong, e.g. the suggested suffix already exists, an error message is returned to "Meta"; else, the newly minted PID is returned.

9) "Meta" stores the confirmation of the PID minting, and passes the PID to "Data". "Meta" also updates the DO's record in the ICOS Metadata Store to indicate that the uploading is completed. From this point, the dynamically created landing page for the DO will list all associated metadata, including the

link pointing to the Repository file system. (By default, only authorized users are allowed to connect to the "Repository" file system.)

10) "Data" returns the URL of the landing page to "Uploader" as a receipt of the successful conclusion of the ingestion process.

III. Summary and Outlook

ICOS RI is characterised by a network of monitoring stations collecting data in three main domains – atmospheric, ecosystem, oceanic – with the aims to capture and describe the current GHG budget of Europe, and to understand patterns and issues related to global climate change. ICOS data products are high-quality data that are collected following strict standards, and then openly and freely distributed as long-term and near real time datasets to different users around the world.

Through its data products, ICOS is anticipated to significantly contribute to the work of many researchers across a wide range of research fields, including environmental and atmospheric sciences. In addition, both directly through its own expertise and indirectly via the output of its user base, ICOS will be able to support national and international policy makers in their efforts related to climate change mitigation and adaptation, especially in the context of the recent COP-21 Paris agreement [13].

The increasing availability of computational power and the high standardization level agreed inside ICOS allow the infrastructure to process data in Near Real Time, opening to a number of new applications. The implementation of early alert systems for weather extremes, built on fast and streamlined sensor data analysis, is quickly becoming a key aspect in facing climate change and the ICOS NRT data processing will give a crucial contribution from the "ground observation" point of view that together with satellite remote sensing and NRT modelling can constitute a strong real time observatory.

Acknowledgment

This paper has been produced within the framework of the ENVRIplus project (see http://www.envri.eu/envriplus) with support from the European Union's Horizon 2020 research and innovation program under grant agreement No 654182.

References

[1] IPCC (2014). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland, 151 pp.

[2] Hazan, L., Tarniewicz, J., Ramonet, M., Laurent, O., and Abbaris, A. (2016). Automatic processing of atmospheric CO_2 and CH_4 mole fractions at the ICOS Atmospheric Thematic Centre, Atmos. Meas. Tech. Discuss., doi:10.5194/amt-2016-53.

[3] Laurent, O. et al. (2016). ICOS Atmospheric Station Specifications, available on-line at https://icos-atc.lsce.ipsl.fr/doc_public

[4] Aubinet M, Vesala T, Papale D (2012). Eddy Covariance: A Practical Guide to Measurement and Data Analysis, Springer Atmos. Sciences

[5] Finkelstein PL and Sims PF (2001). Sampling error in eddy correlation flux measurements. J. Geophys. Res. 106, 3503–3509.

[6] Candela, L, Castelli D, and Pagano P. "Managing big data through hybrid data infrastructures." ERCIM News 89 (2012): 37-38.

[7] Sabbatini S, Mammarella I, Arriga N, Graf A, Hortnagl L, Ibrom A, et al (2016). Protocol on Processing of Eddy Covariance Raw Data. ICOS.

[8] Vickers, D and Mahrt L (1997). Quality control and flux sampling problems for tower and aircraft data. Journal of Atmospheric and Oceanic Technology 14(3): 512-526.

[9] Mauder M and Foken T (2004). Documentation and instruction manual of the eddy covariance software package TK2. Arbeitsergebn, Univ Bayreuth, Abt Mikrometeorol (ISSN 1614-8916), 26-42.

[10] Papale D, Reichstein M, Aubinet M, Canfora E, Bernhofer C, Kutsch W, et al (2006). Towards a standardized processing of net ecosystem exchange measured with eddy covariance technique: algorithms and uncertainty estimation. Biogeosciences, 3: 571-583.

[11] Pierrot, D. et al. (2009). Recommendations for autonomous underway pCO_2 measuring systems and data-reduction routines. Deep Sea Research Part II: Topical Studies in Oceanography, 56, 512–522. doi:10.1016/j.dsr2.2008.12.005

[12] Bakker, D. C. E. et al. (2016) A multi-decade record of high quality fCO_2 in version 3 of the Surface Ocean CO_2 Atlas (SOCAT). Earth System Science Data, 8, 383-413 doi:10.5194/essd-2016-15.

[13] United Nations (2015). Paris agreement. (English version.) Available at http://unfccc.int/files/essential_background/convention/application/pdf/english_paris_agreement.pdf

About the Authors

Carbon Portal: Alex Vermeulen is the CP's director, Oleg Mirzov is the portal's system architect, and Margareta Hellström is the ICOS liaison with EUDAT and RDA Europe. Margareta is also the corresponding author of this paper (e-mail margareta.hellstrom@nateko.lu.se).

Atmospheric Thematic Centre: Leonard Rivier is the ATC's director, Lynn Hazan is in charge of the ATC data bases, and Jérôme Tarniewicz is responsible for data distribution.

Ecosystem Thematic Centre: Dario Papale is the ETC's director, Simone Sabbatini is responsible for protocol development, and Domenico Vitale is responsible for data analysis.

Ocean Thematic Centre: Truls Johannessen is a co-director of the OTC, Benjamin Pfeil works with international ocean data coordination, and Steve Jones is the OTC data management specialist.