CoCO2
Prototype system for a Copernicus CO_2 service

Requirements for data streams from additional tracers and new instrumentation

Alex Vermeulen
Ute Karstens
Dario Papale
Konstantinos Politakos
Leo Rivier
Elena Saltikoff
Mirosław Zimnoch

Co-ordinated by ECMWF

CoCO2

Prototype system for a Copernicus CO_2 service

D7.7 Requirements for data streams from additional tracers and new instrumentation

Dissemination Level:	Public
Author(s):	Alex Vermeulen (ICOS ERIC) Ute Karstens (ULUND) Dario Papale (CMCC) Konstantinos Politakos (FORTH) Leo Rivier (LSCE) Elena Saltikoff (ICOS ERIC) Mirosław Zimnoch (AGH)
Date:	23/02/2023
Version:	1.0
Contractual Delivery Date:	01/01/2023
Work Package/ Task:	WP7/ T7.4
Document Owner:	ICOS ERIC
Contributors:	ULUND, CMCC, FORTH, LSCE, AGH
Status:	Final

CoCO2: Prototype system for a Copernicus CO_2 service

Coordination and Support Action (CSA)
H2020-IBA-SPACE-CHE2-2019 Copernicus evolution – Research activities in support of a European operational monitoring support capacity for fossil CO2 emissions

Project Coordinator: Dr Richard Engelen (ECMWF)
Project Start Date: 01/01/2021
Project Duration: 36 months

Published by the CoCO2 Consortium

Contact:
ECMWF, Shinfield Park, Reading, RG2 9AX,
richard.engelen@ecmwf.int

The CoCO2 project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 958927.

Table of Contents

Figures

Tables

Glossary

ATC	Atmosphere Thematic Centre
CAL	Calibration Laboratory
CC0	Public domain data licence, meaning data is free to use, copy and distribute
CC4BYPID	Creative Commons Attribution 4.0 International data licence, meaning that data is free to use and distribute, but requires that attribution to the data provider is required, and that the licence should be passed on at redistribution or derived products
CP	Carbon Portal
CRDS	Cavity Ring Down Spectroscopy
Datacite	Provider of DOI registration for research data, extending the DOI metadata with research data specific additional metadata
DOI	Data Object Identifier, a special PID that requires a specific set of metadata to go with the data object, describing provenance and attribution
ECMWF	European Centre for Medium-range Weather Forecasts
Eddy flux	Eddy-covariance (also known as eddy correlation and eddy flux) is a key atmospheric measurement technique to measure and calculate vertical turbulent fluxes within atmospheric boundary layers
ETC	Ecosystem Thematic Centre
GUI	Graphical User Interface
Handle	Provider of 'bare-bone' PIDs with minimal metadata requirement, basis of the DOI and ICOS PIDs
ICOS	Integrated Carbon Observations System
JSON(-LD)	JavaScript Object Notation (for Linked Data)
Jupyter	Project to develop open-source software, open standards, and services for interactive computing across multiple programming language
ICOS Data Levels	
Level 0 (L0)	Raw data as delivered by the instrument, uncalibrated and not necessarily in useful units (e.g. mV, mA, µS)
Level 1 (L1)	Near real time data, automatically quality controlled, reduced (averaged) and calibrated
Level 2 (L2)	Final quality controlled data, including manual control and data reduction
Level 3 (L3)	Elaborated data, where additional information has been added to ICOS data, using a model and/or other observations, often gridded data
netCDF	Network Common Data Form, set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data
Ontology	Way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject. Used mainly to organize and represent knowledge in a machine-readable way.
OTC	Ocean Thematic Centre
PID	Persistent Identifier
QA	Quality assurance, measures taken to make sure the data has the proper quality

QC	Quality control, checks performed on the data to see whether the desired quality level has been reached
RDF	Resource Description Framework
RI	Research Infrastructure
Sparql	Query language for triple stores
SSL	Secure socket layer, basic protocol for safe encrypted one-to-one data transfer over the internet
TC	Thematic Center (ATC, OTC, ETC)
TCCON	Total Column CO2 Observing Network
Triple	A semantic triple (or RDF triple) is the atomic data entity in the RDF data model. A triple is a set of three entities that codifies a statement about semantic data in the form of subject–predicate–object expressions (e.g., "Bob is 35", or "Bob knows John")
Triple store	Database for storing and retrieval of triples
Turtle	Terse RDF Triple Language
WMO	World Meteorological Organisation
XML	eXtended Markup Language

1 Executive Summary

Current operational ICOS data streams have been well defined, including rich metadata and a process for clearly marking the level of quality control and timeliness as well as versioning, licensing and citing using DOIs.

New data streams such as those created in CoCO2 campaigns can follow the ICOS process with different degrees. In the platinum option, the full ICOS pipeline can be integrated. This is, however, only possible for instruments which provide raw data identical or very similar to existing ICOS instruments. In the gold option, the quality control normally undertaken by one of the Thematic Centres is replaced by a new partner. In the Silver option, the quality control and metadata for a cumulating timeseries are provided daily by the data producer with support of the ICOS Carbon Portal. In the bronze option, the ready processed data arrives in large batches and a DOI is minted for each batch. The fifth option, Flex ingest, is suitable for short-lived campaign datasets, and does not include any processing or QC at ICOS side. This means that raw data can be uploaded through a dedicated sftp account to the ICOS file storage and that the data will be read-only accessible through the ICOS Nextcloud instance for further processing up to L1 or L2 level and then be published using the manual upload process.

2 Introduction

The ICOS research infrastructure generates and processes many data streams that have a dedicated pipeline that involves one or more parties next to the ICOS Carbon Portal (ICOS-CP) central data repository, but all data is in the end published by the ICOS-CP together with rich metadata. Most of the data streams have their own specific requirements for processing and usage. In the ICOS data streams we differentiate between observational data streams and elaborated data streams, the latter are mostly based upon ICOS observational data and combine these with other observational or modelled data. Elaborated data are mostly contributed by other parties or the ICOS community.

New observational data streams can be incorporated, and this will require in most cases definition of new stations and instruments, new data types for raw and processed data and the associated descriptive metadata, setting up a processing pipeline to process the raw data and creating the needed account and right authorisations for upload of the data.

A provision needs to be made to develop and test the new data pipelines without disturbing the operational data provision and in the early development phase before regular data provision some temporary solution based on file sharing, e.g. though Nextcloud of sftp.

3 Principles of data ingestion and storage in ICOS

3.1 ICOS Data and its data level definition

ICOS data is data that has been generated by the ICOS infrastructure. Data is always associated with a defined project, ICOS is just one of the projects. At the ICOS-CP we differentiate between three major data levels for all data:

- L0: Level 0 or raw data. This is untouched data as received by the instrument or data logger. Data could be compressed by a loss-less algorithm like zip or gzip for the sake of efficiency. L0 data is in general not usable for normal users, as units are not always meaningful and information on what the values are representing is lacking from the data object itself.
- L1: Level 1 or near-real time data. Data produced from L0 data and that has been converted to meaningful units, has been automatically quality controlled and

calibrated, and in most cases reduced due to averaging and other data reduction techniques. L1 data is used extensively by station PIs and thematic centres for quality control and monitoring the instrumentation but can also be used by external users for analysis for specific applications with (near-)real time requirements. L1 is usually a time series.

- L2: Level 2 or final quality-controlled data. These are the main data product of ICOS and fulfil all the quality control requirements of ICOS, including the manual quality control and data flagging by the PI and thematic centre. L2 data is usually a time series.
- L3: Level 3 or elaborated data. This is data that has been generated in most cases using L1 or L2 data but then combined with other (model) data. Examples are spatially extrapolated data (SOCAT, FLUXCOM) or inverse model estimated priors or posterior flux data. L3 products are mostly contributed by the ICOS community or external parties and are usually spatiotemporal data, often formatted in netcdf.

3.2 FAIR and data object granularity

The ICOS data system is designed from the ground up to be fully FAIR, open and transparent. The data lifecycle is in detail explained in the ICOS Data Improved Lifecycle (2020, https://doi.org/10.18160/D2JV-KB6B).

Granularity of the data objects (binary blobs represented by files) is decided by the uploading user and in case of operational ICOS observational data it is mostly as daily files for L0+L1 and annual releases for L2 (see ICOS data types in chapter 3.2).

3.3 Data ingestion

Ingestion and storage of any observational data stream by the ICOS-CP follows the following principles:

- Data ingestion is a completely machine to machine operation that does not require any human intervention
- Data is accepted only for known and registered data types, each data type requires a specific authorisation for upload and for data levels>0 the variables that the data contains are defined in rich metadata
- Data is stored without any modification, including their original filename
- Data should be accompanied by a descriptive metadata package, that among others contains the SHA256 checksum, the data type and detailed provenance information, link to where applicable stations, instruments and persons, protocols and descriptive documents or related scientific publications
- Only data is registered that has been completely transferred and fulfils all requirements with regards to metadata, passed the validation checks (when applicable), and passed the checksum calculation
- The data object receives a unique Handle Persistent identifier (PID) that is linked to the ICOS-CP database and when resolved results into a machine and human readable landing page. The landing page provides access to the data and lists the metadata linked to the ICOS-CP ontology. PID resolving is completely dynamic and the landing pages are generated on the fly based on the current state of the metadata database. The ICOS PID contains a character encoded part of the data object checksum so it is uniquely linked to the data object.
- Data objects are stored permanently and can be always retrieved through their PID together with all the rich metadata.

- The data object is stored at the ICOS-CP and a copy is simultaneously stored at the EUDAT CDI B2SAFE services at CSC (Finland) with a duplicate at KFZ Jülich (Germany)

3.4 Data upload

Data ingestion is technically described at the ICOS-CP Github repository (https://github.com/ICOS-Carbon-Portal/meta#upload-instructions-scripting, https://github.com/ICOS-Carbon-Portal/data#instruction-for-uploading-icos-data-objects). There is also a user friendly GUI to assist in the manual upload (https://meta.icos-cp.eu/uploadgui/), or update metadata elements. All upload requires login and proper authorization for the respective data type to be uploaded.

3.5 Data license

All ICOS data is licensed under the Creative Commons Attribution 4.0 International (CCBY 4.0, https://creativecommons.org/licenses/by/4.0/). All metadata is licensed as CC0.

The ICOS-CP also supports other data licenses, most notably Public Domain (CC0, https://creativecommons.org/share-your-work/public-domain/cc0/) for some of the L3 data.

As often projects or scientific journals require that data publication is postponed until a certain release date, ICOS-CP supports that at ingestion an optional moratorium date is associated with the data object, which makes that the data object is only visible and accessible after that date.

3.6 Timeliness

All ingested data from L0 to L3 is available immediately at the ICOS-CP to all users without restriction, unless a moratorium has been specified in the ingestion metadata. Another exception is the L0 data from the atmosphere domain that is only available on request.

Observational data are the core of ICOS data and each of the domains provides in principle the raw data provided by the instruments as prompt as possible (within 24 hours) to ICOS-CP. This raw data can be sent to ICOS-CP directly by the instruments or data loggers, for later pick up and further processing, or after collection at one of the thematic centres. In case no internet connection is available, like in the case of ships of opportunity or buoys without satellite data connection in the ocean domain, data is transferred as soon as the ship enters the harbour, or when the station is visited for maintenance.

ICOS L1 data is provided usually within 24 hours after the generation of the underlying L0 data. L1 data always covers the period starting from the end of the latest L2 release to the current day in so called growing time series.

ICOS L2 data is generated at least once per year as major release in the first half of the new year. In case of the ecosystem a second annual release is provided at the end of the growing season of the northern hemisphere. L2 data always covers the complete period that the station provided ICOS data. From the ICOS-CP side there is nothing preventing from more frequent L2 releases.

3.7 Data versioning and collections

At upload, data objects can be marked in the metadata as new versions of previous data objects by linking to their PID. The older version will normally get a deprecated state in the database and by default won't be visible any more in the ICOS-CP data portal. However, the older version PID(s) are permanent and remain valid and landing pages will be generated when called. Both the landing page of the latest and the older versions of the data will point to the respectively previous and newer version of the data, forming a linked chain up to the earliest data version.

Next to L0 to L3 data objects it is possible to generate so-called collections of multiple data objects. This done by minting a Handle PID for a list of PIDs that all point to data objects. All landing pages of data objects in a collection will refer to the collection(s) to which they belong. Collections can also be versioned. Collections can be downloaded and then the system will on the fly compose a zip file that contains all the data objects in the collection, including an index of all included data objects and their PID.

3.8 Datacite DOI minting

Datacite DOIs (Data Object Identifier) can be minted for data objects or collections. Handle PIDs prescribe only a minimal amount of metadata to be registered. ICOS adds to this its own rich metadata model that varies with data type to the PID associated metadata. Datacite DOIs like ICOS use Handle PIDs but then use another simpler metadata model that mainly takes care of the attribution and some provenance of the data. The metadata schema currently is version 4.4 (https://schema.datacite.org/meta/kernel-4.4/).

Datacite DOI minting is a manually curated process that at the ICOS-CP is assisted by the upload GUI and the DOI GUI (https://doi.icos-cp.eu/). All major ICOS products such as L2 data collections and L3 products are minted DOIs. Data objects or collections with a DOI show in their landing page information the reference to their DOI and show the citation string connected to the DOI.

Users can pre-mint Datacite DOIs through the ICOS-CP DOI GUI, for example for a L3 product they want to publish through ICOS-CP. After curation by the ICOS-CP and ingestion of the data this DOI will then be published to and indexed by Datacite. These pre-minted DOIs are often required by publishers to include in a draft scientific paper to refer to supplementary data sets.

3.9 Data upload façade for daily packages of half-hourly files

An exception in the standard workflow is made for the ingestion of the raw data from ecosystem eddy covariance towers and phenocams, to accommodate for real time, half hourly transfer of data by instrumentation in the field that is not able to transfer data in the most up to date SSL protocol and does not have the computational power to calculate SHA256 checksums.

For this ICOS-CP provides the so called façade API for data upload (https://github.com/ICOS-Carbon-Portal/data#simplified-etc-specific-facade-api-for-data-uploads). Data are then merged into daily data files that are submitted by the façade on behalf of the station as daily L0 data through the standard ingestion routine of the ICOS-CP.

4 Data access and data processing integration

4.1 Data access through the ICOS Data Portal graphical user interface

At https://data.icos-cp.eu/portal a multi-faceted search function allows the user to drill down through the over 1 million data objects stored at the portal, to the data objects needed. Data objects can be selected for example by any combination of project(s), station(s), data level(s), variable name(s), domain(s), keyword(s), acquisition or submission interval, and/or through a spatial selection on the map. From the list of data objects the user can select all or some and download them directly,or add them to a virtual shopping cart for later download together with data from other queries. Also for most data types one or more data objects can be selected for preview as time series or spatiotemporal data on the map. The data object list of course also contains the PIDs as links that provide access to the data object landing pages (also machine readable by content negotiation and also encoded in the schema.org scheme) that contain the rich metadata. From each landing page the data can be downloaded directly or if available also previewed. Downloading requires the user to accept the ICOS data licence.

To present all the results the portal GUI is using the same linked open data technology that is used in the back-end. The actual (SPARQL) query resulting from the GUI settings in the faceted search can be viewed and copied for execution in an automated machine to machine operation. These queries and all data and metadata access uses standard and simple open web technology without requiring login or authentication (see section 4.2). It is possible to register an account and login, which will enable the user to bypass the licence acceptance page through a one-time acceptance of the data licence in the account profile. In the user profile also a record of previous downloads and all individual preview settings are stored for convenience.

4.2 Data access through standard internet protocols

Just like upload, also access to download of data and metadata through PID landing pages is done through basic http GET calls. Metadata exchange can take place in a variety of metadata profiles (ICOS JSON, rdf/xml, rdf/turtle, iso-19115 XML) through content negotiation. For download of ICOS data the download requires authentication through an electronic (?) token, available at the user profile page or using the procedure described at https://github.com/ICOS-Carbon-Portal/meta/#authentication. Authentication is required to assure that the ICOS data licence is accepted and to allow usage tracking and service abuse prevention.

4.3 Data access through python

ICOS provides a python library for easy access to the ICOS data and metadata. The library can be installed through pip. Detailed information is available at https://pypi.org/project/icoscp/ and https://icos-carbon-portal.github.io/pylib/. When using the ICOS Jupyterhub this library is already available, and all data will be available directly without need for download to local storage. In the Jupyterhub dedicated project groups can be defined that have access to a common shared storage area for exchanging data and code and that can also have access to dedicated associated group disks at the ICOS Fileshare, for example to access non-public measurement data from campaigns of data pipelines in the test phase and results from prototype data pipelines.

4.4 Data processing workflows

All three ICOS thematic centers (TCs) and the central labs have developed automated and standardized workflows that process raw data to daily NRT L1 data and regular final quality controlled L2 data releases. These workflows are following detailed specifications that are based on the community guidelines for high precision and long term consistent and highest quality data, following and in many cases even leading the WMO, FLUXNET recommendations. The general dataflow is shown in Figure 1

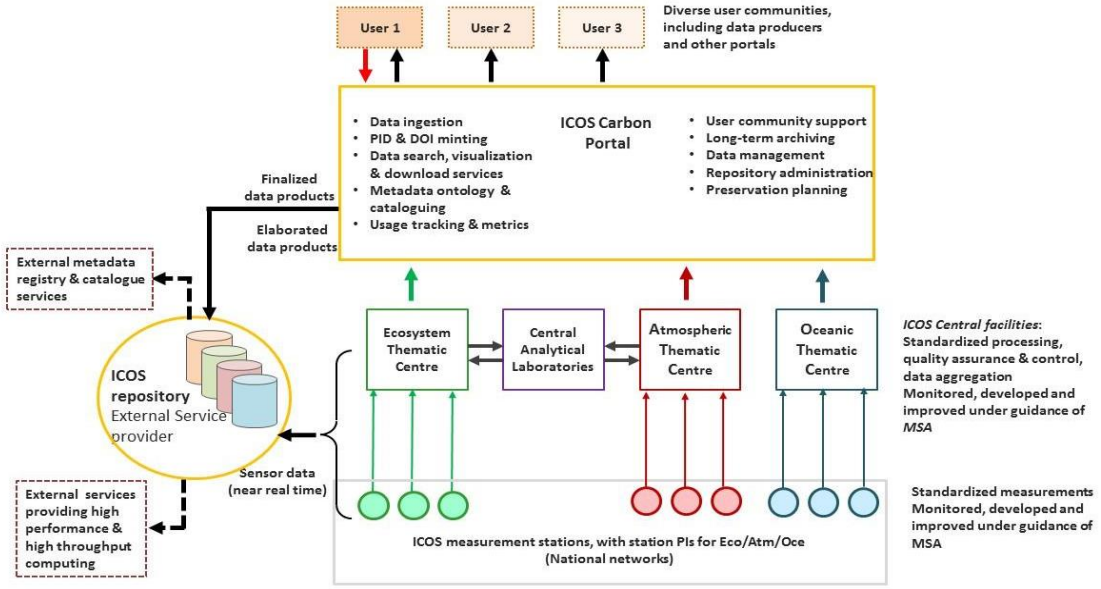

Figure 1 Overview of the ICOS RI data flow. Arrows indicate the transfer of data and metadata objects between or within sub-communities of the ICOS -RI. Black arrows indicate both data and describing metadata, red indicates only data and broken line arrows only

The Atmosphere TC (ATC) provides the station PIs with daily detailed overviews of the processed data, including statistics on uncertainty and calibration quality. Both the OTC and ATC provide GUI applications for the PIs to flag and asses the quality of the measurements and provide important metadata on the configuration and state of the measurements. In the ETC all metadata is exchanged with PIs using the BADM standard (https://ameriflux.lbl.gov/data/badm/). The daily metadata exchange between thematic centers is processed by the ICOS-CP into a unified and harmonized metadata model that describes the stations, instruments, measured quantities, methods, provenance, persons and organisations and their relations.

5 New data pipelines for CoCO2

5.1 Requirements for new data pipelines

New observations that cannot use any of the existing pipelines in ICOS, for example because the instruments are of a new type or because new variables not yet described are measured or through methods that do not follow the ICOS protocols, will in many cases require the formulation of a new data type. If the data will not be ICOS data one needs to define an associated project, which can be one of the existing projects, e.g. *Miscellaneous* or a newly defined project, for example named CoCO2. Only the ICOS-CP can create new projects in the ICOS metadata, upon request by email to info@icos-cp.eu.

Each data type is connected to a single project. For the project “Miscellaneous” there are at the moment three data types available that are available for new observational data. These are the data types “Fluxnet Product”, “Atmospheric Measurements Results Archive” and “NetCDF spatial data”. The disadvantage of these data types is that they are either very specific or very unspecific respectively. Data type can also connect to a value type and a variable specification. Again data types can only be added by the ICOS-CP and decisions on new data types will be made through consultation with the data provider during the first data curation steps.

Data objects of a certain data type can be associated with a station and sampling height, and one or more instruments. Stations must exist in the ICOS-CP database before they can be referenced as associated to the data object. The same applies to instruments. Stations

minimally need metadata with coordinates in longitude, latitude, and elevation (WGS84), but can be also associated with ICOS, WMO, organisations, personnel, and their roles. In principle this metadata is provided by the thematic centre of the respective domain of the station. But in case of no association with a Thematic Centre when new measurement data are added through a new data type, responsibility for the provision of metadata needs to be assigned to the data provider or an associate so that the ICOS-CP can create in the metadata database the associated stations, instruments, organisations, and persons. In the upload metadata these need to be referenced using the URL's that the ICOS-CP will provide. In case the data processing pipeline will be managed by an existing thematic centre this will be arranged between the TC and the ICOS-CP based on input from the measurement PI, otherwise by direct contact between the data processor and the ICOS-CP.

In case a new project is planned involving a central data processing unit that will handle the metadata and data pipeline, access can be given to a specific part of the metadata that as a start can be managed directly in the ICOS-CP metadata database and ontology through the manual admin interface, this is also the way that the OTC interfaces with the ICOS-CP. A relatively easy to use metadata editor can be made available to provide metadata, an example is the OTC entry editor at https://meta.icos-cp.eu/edit/otcentry/.

A decision flowchart schema is shown in 2.

5.1.1 Flex ingest

Data that will be generated for a short campaign or experiment (less than 1 year) (and that does not follow any of the ICOS protocols completely so that it can be processed using existing data pipelines) can be supported using the Flex ingest option. This means that raw data can be uploaded through a dedicated sftp account to the ICOS file storage and that the data will be read-only accessible through the ICOS Nextcloud instance (https://fileshare.icos-cp.eu). In there, data can be directed towards a separate group disk only accessible to assigned group members. Data pipelines can then be developed and applied by accessing the data using authenticated sftp or WEBDAV or by direct access using the ICOS Jupyter hub (https://jupyter.icos-cp.eu), using Python, Julia and/or R code. L1 and L2 files can then in an intermediate step be uploaded through the existing standard API and previewed through the Jupyter.

As soon as the routines are ready for operational use and the resulting data is of sufficient quality data providers could proceed to the next steps in the flow diagram of Figure 2, starting with the Gold option.

Final data resulting from this option can be published using other repositories like Zenodo/ Pangaea or after ingestion by ICOS using the Bronze option or after minting specific new data types through the silver option. If the Gold option is the goal, one can wait with publishing the data until the complete operational workflow has been established.

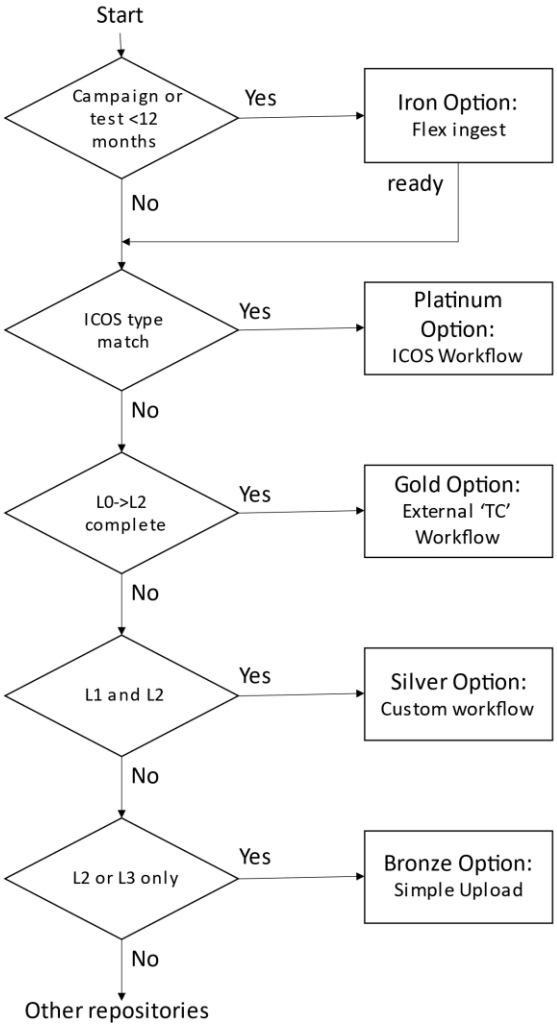

Figure 2 Decision flowchart for inclusion of new data streams to the ICOS data pipelines. Explanation of the options in section 5.1.2 to 5.1.5

5.1.2 Platinum option: full ICOS data pipeline

In case that the observations in principle follow ICOS protocols using instruments that are approved for ICOS and that provide raw data that already or with small adaptations can be or are planned to be ingested by a thematic centre, the integration is straightforward. The station, instrument and person data are added to the TC metadata and will be automatically exchanged with the data portal. As the data will not be marked as ICOS data this will require to use a possibly new data type that is connected to another (new) project. Data flow will furthermore follow the ICOS data pipelines. An example of this is provided in Section 5.2.1.1.

5.1.3 Gold option: External 'TC' workflow

For planned extensions of the ICOS network that do not match with any of the thematic centres and other thus far homeless data, that will have a sustained capacity for handling the data pipelines there is the option to define a new project with eventually a new (virtual) domain/theme and new or existing data types. Stations can be existing stations or other stations. Even instruments can be shared between projects. An example would be the tall towers and Picarro instruments used in both ICOS Cities and ICOS.

The new central data processing unit that will handle the metadata and data pipeline can be given access to a specific part of the metadata store that as a start can be managed directly in the ICOS-CP metadata database and ontology through the manual admin interface, this is

also the way that the OTC interfaces with the ICOS-CP. In this case one of the existing metadata integration options that have been developed for ATC and ETC can be used.

5.1.4 Silver option: L1 and L2 data provision only

In this case one or more data providers provide regular (daily) growing datasets of L1 and lower frequency L2 data from instruments using specific data types for the (new) project without further automated dynamic integration metadata. The data providers will register with help from the ICOS-CP all metadata on not already registered stations, persons and instruments and make sure that each data upload receives the correct metadata using the URIs provided by the data portal, using either the manual update GUI or the API. Documentation of the provenance can be uploaded through documents that will be associated with the data. L2 data products could if needed be minted a DOI after curation by the ICOS-CP.

Examples of these are datasets of APO or isotopes that are not part of standard ICOS operation and/or from non-ICOS stations that are gathered for a campaign and need to be published as part of a scientific publication, while also access needs to be given to users of the L1 data. The L1 data can also be useful to the researchers and technicians themselves to track the quality and status of their measurements. Disadvantage of the approach is that the raw data is not automatically stored and that reproducibility of the data processing is not ensured.

5.1.5 Bronze option: L2 and/or L3 data only

In this case the data provider plans to provide at larger intervals final quality-controlled measurement or model data, for example after batch wise data processing and following extraction of the processed data from the providers' relational database. For this a new project can be minted in case the project will be sustained for a long period and the visibility and findability is critical (like for the Global Carbon Project), otherwise the data might be associated with the project "Miscellaneous". Parsing at ingestion is optional, for example by providing the data as an arbitrary zip or xls file, although absence of this will exclude the option of providing previews of the data and data access through the python library. The data provider will have together with the ICOS-CP register the stations, instruments and persons that are not already available and use the URIs provided for these in the upload metadata. In principle the data can be uploaded using the upload GUI, especially when this concerns routine upload of many files over an extended time period, but in most cases this should be done and curated by ICOS-CP personnel, who can then also take care of creating collections and DOI minting, if needed.

Examples of these are model results or campaign wise measured datasets of APO or isotopes that are not part of standard ICOS operation and/or from non-ICOS stations that need to be published as part of a scientific publication.

5.2 Description of the new data pipelines in CoCO2

5.2.1 CoCO2 campaign data pipelines

Three partners deliver new campaign-wise observations in the CoCO2 project.

5.2.1.1 FORTH's Heraklion eddy flux towers for urban CO_2 fluxes

RSLab, FORTH provides, for two locations in Heraklion, Crete urban eddy covariance flux observations of CO_2. Detailed metadata on the flux sites can be found at http://rslab.gr/fluxtowers.html. HECKOR (since 2016) is located in the city centre and HECMAS (since 2021) in at a residential area, Mastambas south of the centre. Furthermore, FORTH has established an online interface for both flux towers displaying near-real-time data (http://rslab.gr/heraklion_eddy.html).

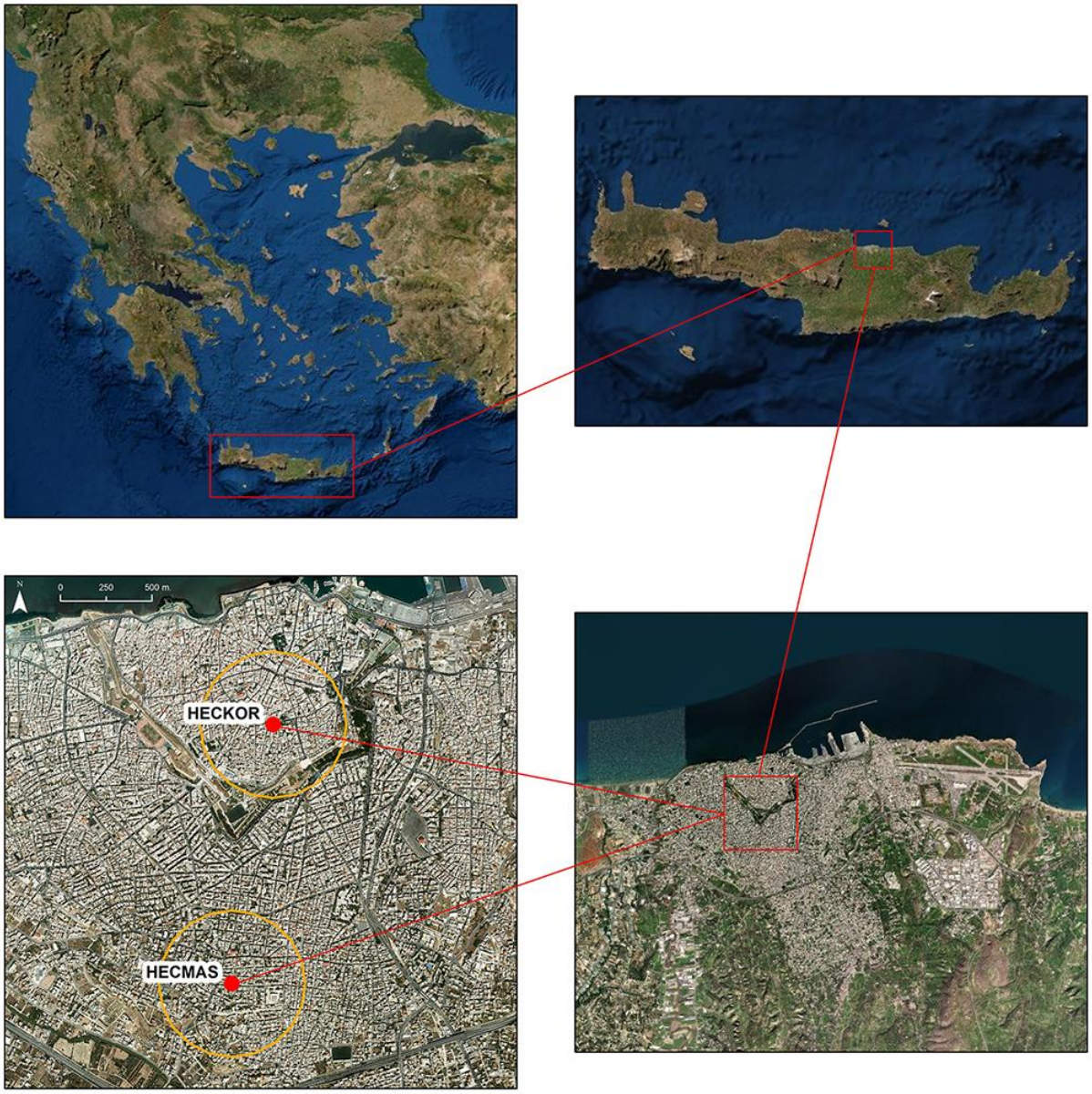

Figure 3 Location of the two FORTH eddy covariance flux towers in the town of Heraklion, Crete

Both towers deliver flux data using ICOS compliant hardware and their data streams can be integrated using the ETC data pipeline, using the platinum option, both towers delivering about 40 MB of raw data per day that can be ingested using the ICOS-CP ETC Façade interface (Section 3.7). From there the data can be picked up by the ETC for daily processing in the cloud, resulting in daily NRT-24h half-hourly automatically quality controlled data flux product and an annual Level 2 final quality controlled half-hourly data product. Published through the ICOS-CP together with the regular ICOS flux data products.

However, as the fluxes concern an urban area with no ecosystem, the station definition is different and also the standard gap-filling techniques used in the processing that take into account vegetation dynamics do not apply, so that some experimenting on choosing working approaches is requires, also considering the challenging conditions like sea breeze and high sensible heat fluxes.

HECKOR measurements have shown the impact to the CO_2 emissions due to an intervention started in January 2018 by the municipality of Heraklion, introducing traffic regulations and

rebuilding initiatives (road closures, pedestrianizations, minor planting), and further reductions during the COVID lock-downs in 2020 and 2021.

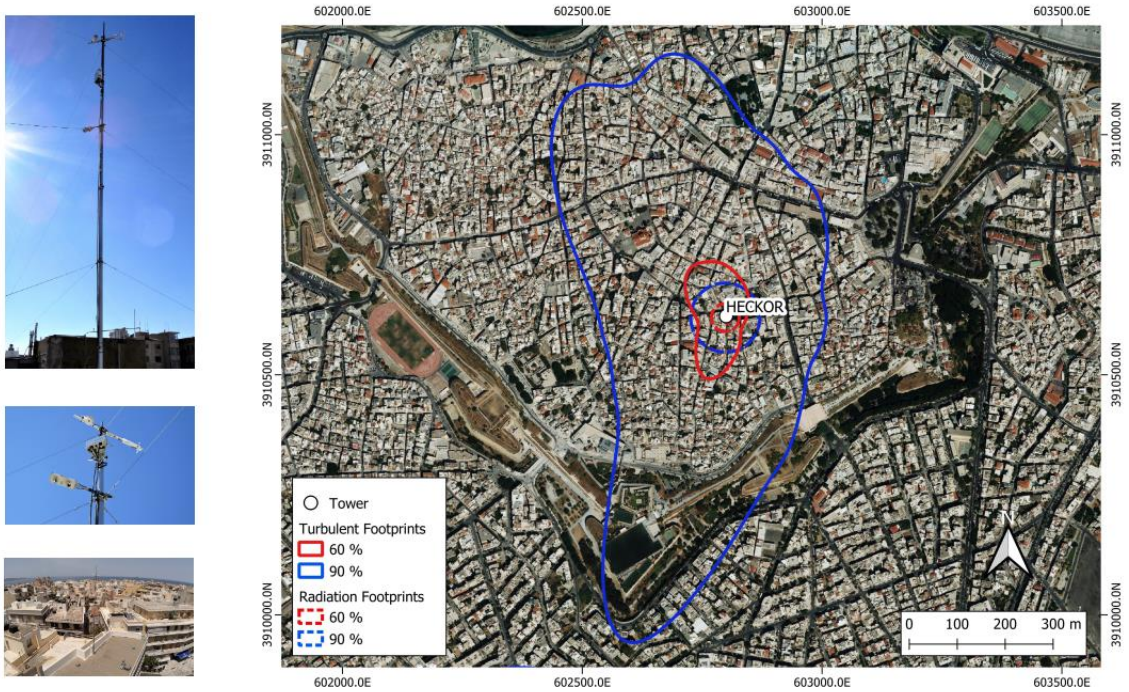

Figure 4 On the left, surroundings of HECKOR, on the right CO_2 eddy flux site and flux average (1-year) footprint projected on the map of Heraklion city center.

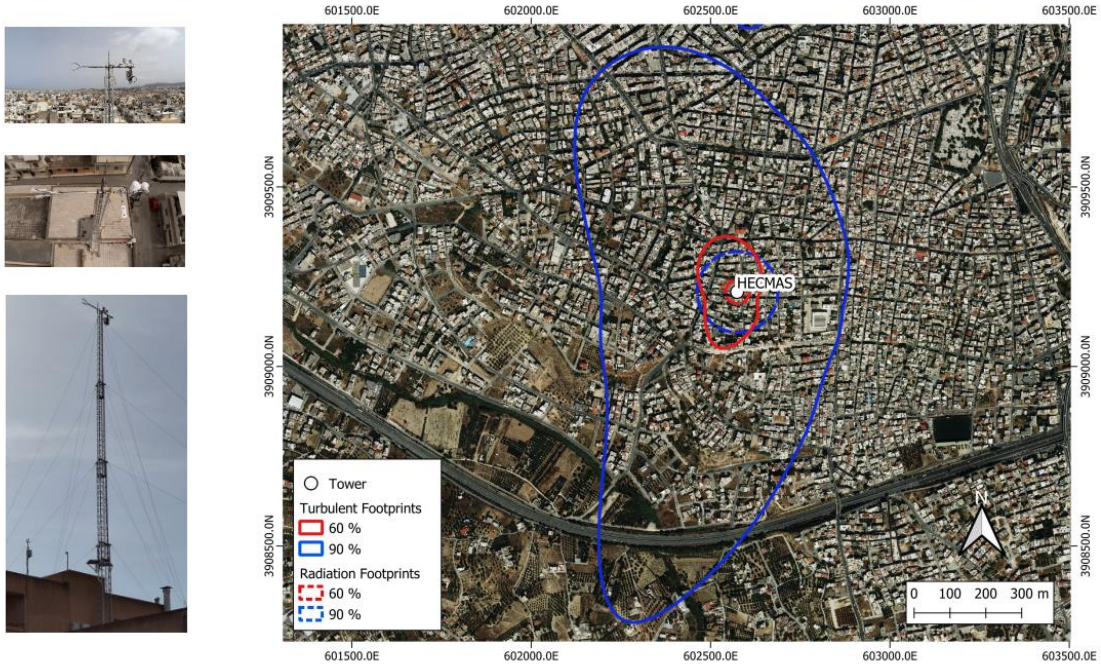

Figure 5 On the left, surroundings of HECMAS, on the right CO_2 eddy flux site and flux average (1-year) footprint projected on the map of Heraklion's residential area, Mastambas.

5.2.1.2 Krakow (Poland) eddy flux tower PL-Krk

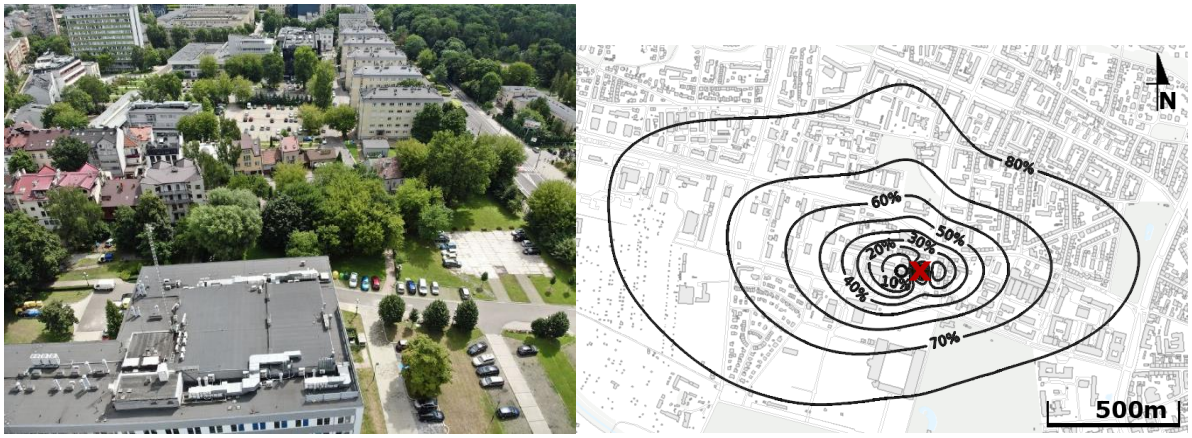

Figure 4 Surroundings of the Krakow PL-Krk CO_2 eddy flux site and flux average footprint projected on the map of Krakow. The measurement mast is visible on the rooftop of building at the bottom left corner.

The AGH University of Science and Technology (AGH UST) in Krakow provides data from a rooftop eddy covariance system set up at the rooftop of a university building in the city of Krakow, Poland. The CO_2 flux data are operational since February 2021 and is using an ICOS data pipeline compatible setup using a Smartflux station. This data stream can be integrated using the ETC data pipeline, delivering about 40 MB of raw data per day that can be ingested using the ICOS-CP ETC Façade interface and further analyses as described under 5.2.1.1 for the Heraklion towers, thus using the platinum option.

5.2.1.3 Krakow vertical gradients using balloon and drone lift

From March 2021 to February 2022, measurement campaigns were performed at Krakow, close to the eddy covariance site, each consisted of a number of flights. Air pressure, temperature, and humidity were measured as well as wind speed and CO_2 and CH_4 concentration during each flight. Each measurement campaign lasted at least from the afternoon until the next morning. At first, instruments travelled up and down inside the basket of an air balloon. This was later change by sampling from a line lifted through a drone. In total 269 vertical profiles have been obtained over the measurement period. Profiles consist of data on air temperature, pressure, humidity, wind speed, CO_2 and CH_4 mixing ratio.

Data will be provided according to the bronze option (Section 5.1.5) as L2 data files in the WMO time series CSV format, same as used by ICOS ATC, one file for each campaign. ICOS-CP will mint monthly collections and corresponding DOI for the complete dataset collection.

Relevant metadata has to be provided with each file in JSON as described at the ICOS-CP at https://github.com/ICOS-Carbon-Portal/meta#registering-the-metadata-package. The ICOS-CP will assist in this process of data upload.

Total data volume is about 87 MB. The data will have similar landing pages including preview of the data as ICOS L2 atmosphere data.

Figure 5 Photographs of balloon and drone launch for the greenhouse gas profile measurements in Krakow.

Required metadata at submission for each file:
Data affiliation: CoCO2
Data type: Atmospheric measurements results archive
Data level: 2
Licence: ICOS CCBY4 Data Licence
Format: CSV time series (WMO time series format)
Encoding: plain file
Acquisition:
- Station (name, station ID, lat/lon coordinates, sampling height)
- Start and end time (UTC)

Production:
- Produced by (author(s)): name. organisation, role
- (Contributors: Name, organisation, role)
- Production time (UTC)
- Comment

Submission
- Submitted by: name, organisation
- Publication time (UTC)
- Submission started (UTC)

5.2.1.4 Krakow $^{13}CO_2$ and $^{14}CO_2$ isotopic mixing ratios

In 12 campaigns lasting each 4 hours, flask grab samples have been collected at the AGH UST university campus in Krakow that are analysed for CO_2, $^{14}CO_2$ and $^{13}CO_2$ mixing ratios. Furthermore, monthly average samples were collected at the campus and at the Kasprowy Wierch mountain station, that were analysed for the same components.

Data will be provided according to the bronze option as L2 data files in the WMO time series CSV format, same as used by ICOS ATC, one file for campaigns and one for the monthly samples respectively. The ICOS-CP will mint a collection and corresponding DOI for the complete dataset collection.

Relevant metadata has to be provided with each file in json as described at the ICOS-CP at https://github.com/ICOS-Carbon-Portal/meta#registering-the-metadata-package. The ICOS-CP will assist in this process of data upload. The required metadata package for data ingestion has the same shape as the balloon vertical gradient data described in 5.2.1.3.

5.2.1.5 Sodankylä aircore vertical profiles by FMI

Since 02-2021 FMI has collected at the Sodankylä site more than 250 vertical profiles using aircores. Aircores are stainless steel tubes of a length of 100-200 m that are launched from balloons from heights up to 30-40 km. Before the ascent the aircore tube gets evacuated. During the descent it is opened and through the increasing air pressure the tube fills with air

and in that way stores a continuous record of air samples from top to ground. After retrieval of the aircore it is analysed by evacuating the tube and analysing this for CO_2 and CH_4 with a high precision analyser (usually Picarro CRDS). The aircore container also records pressure, air temperature and humidity during descent. Eventually N_2O analysis will be added to the measured mixing ratios at FMI.

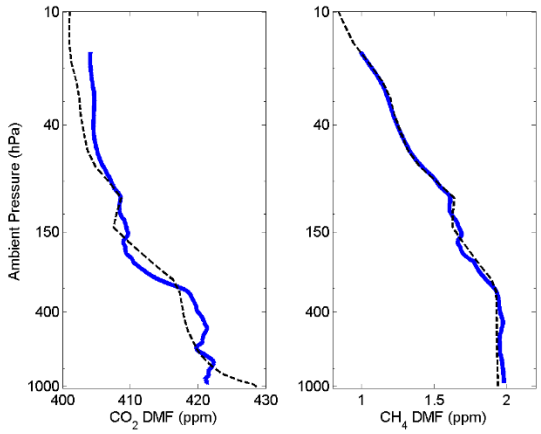

Figure 6 Example air core vertical profile at 26 April 2021 of CO2 and CH_4 mixing ratios at Sodankylä (blue lines) compared with simultaneous TCCON data

The final analysis results consist of vertical profile data on air temperature, pressure, humidity, CO_2 and CH_4 mixing ratio.

Data will be provided according to the bronze option as L2 data files in the WMO time series CSV format, same as used by ICOS ATC, one file for each air core launch. The ICOS-CP will mint one or more collections and corresponding DOI for the complete dataset collection.

Relevant metadata has to be provided with each file in json as described at the ICOS-CP at https://github.com/ICOS-Carbon-Portal/meta#registering-the-metadata-package. The ICOS-CP will assist in this process of data upload. The required metadata package for data ingestion has the same shape as the balloon vertical gradient data described in 5.2.1.3.

5.3 New data pipelines in ICOS Cities/PAUL

PAUL is short for Pilot Application in Urban Landscapes towards integrated city observatories for greenhouse gases. This project runs for 4 years, 2021-2025. PAUL was the name of the proposal, but externally it is now called ICOS Cities.

This project supports the European Green Deal by creating capabilities to observe and verify greenhouse gas emissions from densely populated urban areas across Europe. Cities are recognized as important anthropogenic greenhouse gas emission hotspots and therefore play a significant role in any emission reduction efforts. The PAUL project aims to increase our understanding of specific needs of greenhouse gas emission assessment in urban environments; it compares available and novel observational approaches and implements an integrated concept for a city observatory, providing unique data sets that feed diverse modelling approaches, scientific studies and will be the base of services towards the city administrations. Overarching goals of PAUL are to:

1. implement elements of a pilot city observatory in a large (Paris), a medium (Munich) and a small (Zurich) European city,
2. collaborate with city stakeholders and engage citizens in co-designing services that are required for GHG monitoring in order to validate the implementation of Paris Agreement, and

3. increase our understanding of specific needs of GHG assessment in urban environments and create a service portfolio for setting up an urban greenhouse gas observatory.

ICOS Cities data will be complementing the CO_2 MVS system by zooming in on the hot spots that cities are for the fossil fuel CO_2 emissions. In turn the analysis of these high resolution city data will require information from the Copernicus MVS to build its urban emission inversion systems on. For a large part ICOS Cities targets the same stakeholders as the Copernicus MVS and will build on the model and data developments from the CoCO2 project.

In Figure 9 an overview is given of the different measurements to be used in the ICOS Cities project.

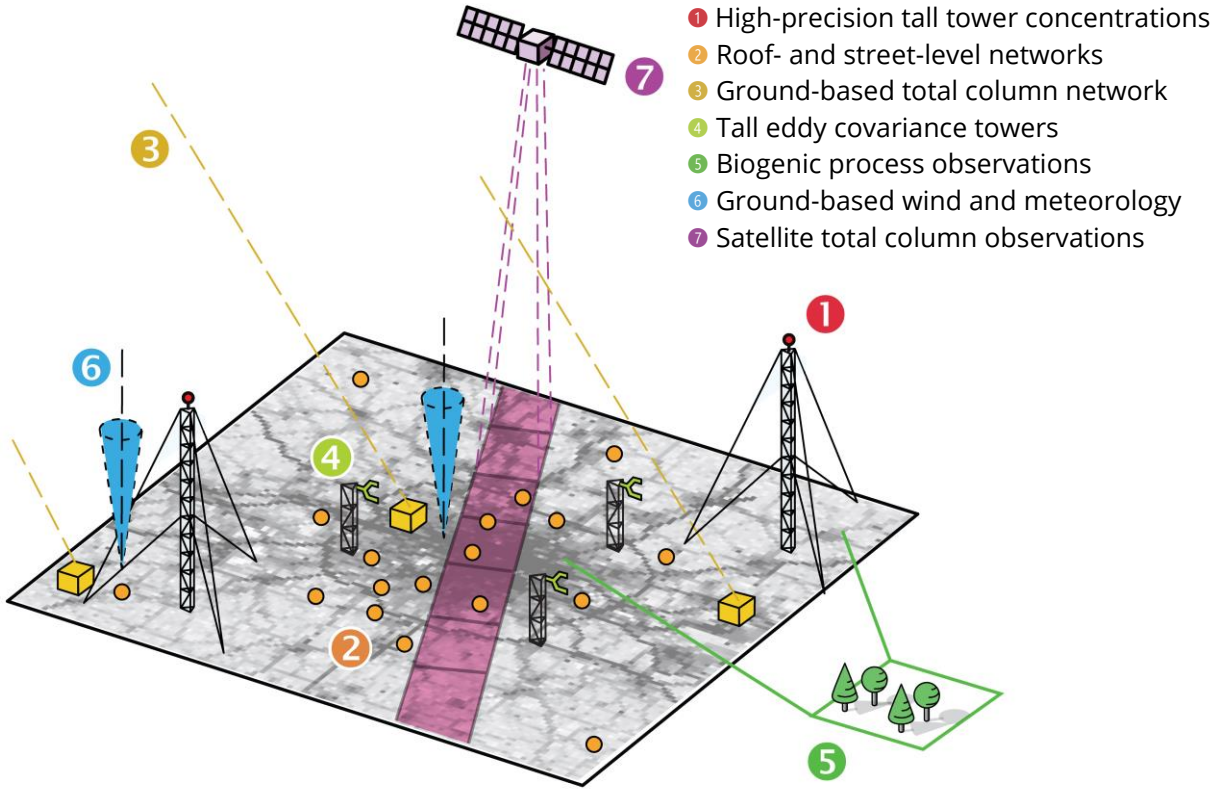

Figure 7 Schematic overview of measurements in ICOS Cities. Each city will have its own constellation of observations using one or more of the types shown here and of each deployed type one or more instances

In Figure 10 the time schedule for the implementation of the different measurements foreseen in ICOS Cities is shown. This includes many new measurements types and deployment of new instrumentation and existing measurements in different configurations. It is clear that together with rolling out the measurements also a lot of work is required for implementing new measurement data processing pipelines. The year 2023 is the year in which most of the instrumentation becomes operational and data will become available. 2023 will also be the year in which the new and updated data pipelines then should become available. This work is done in the framework of the ICOS Cities project but will be very relevant for the future CO_2 MVS.

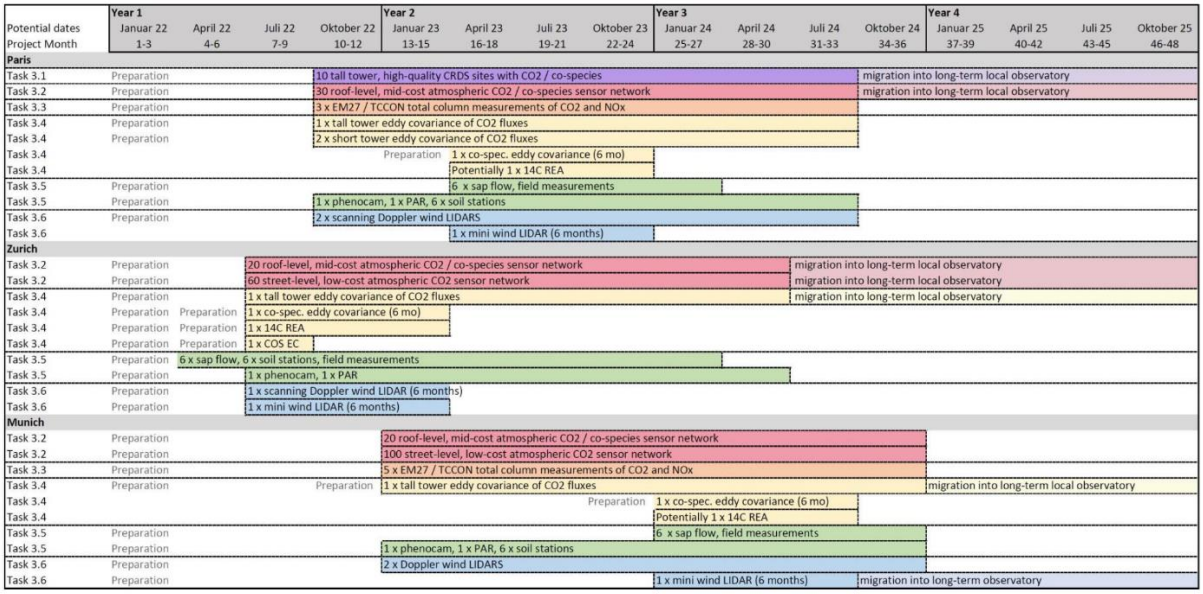

Figure 8 Time schedule for the implementation of the different measurement systems in the ICOS Cities project.

Data streams fitting to existing or updated ICOS pipelines (platinum, gold):

Tall tower sites (ATC, CAL)
- Picarro CRDS CO_2, CH_4, CO,
  - Paris: 9
  - Zurich: 9
- Flask sampler $^{14}CO_2$
  - Paris: 3

Ecosystem flux urban (ETC)
- Soil resp, LAI, Sap flow, TEROS, Phenocam, PAR:
  - Munchen, Zurich (UNIBAS; platinum): 10
- Eddy covariance (platinum)
  - Paris: 4
  - Zurich: 4
  - Munchen: 1

Synops (ATC)
- Paris: 10
- Zurich: 10
- Munchen:

ANSTO Rn (ATC)
- Paris: 4
- Zurich : 4

New observational data pipelines (silver unless mentioned otherwise):

TCCON
- Paris: 1
- Zurich: 1 (TTCON)

EM27 remote sensing (platinum: ATC)
- Paris: 2
- Zurich: 2
- Munchen: 5

MIRO Flux/REA/ICOS Flask sampler (gold): 1 (Zurich+Paris+Munchen respectively)

LIDAR Doppler

Paris: 2+1
Munchen: 2+1

Aethalometer
Paris: 2

Mid cost sensors CO_2
Paris: 10+16
Zurich: 14+20
Munchen: 20

Low cost sensors CO_2
Zurich: 60
Munchen: 60

CAPS N500 (NO, NO2)
Paris: 2

In addition to the observation pipelines there will be many data streams with either prior or ancillary data, like emission inventories and mapping information on land use and buildings. Most of this data will use the flex ingest (iron) option, but when the data is used to generate final model results and/or is ready for publication that data will be published through the bronze option and uploaded as elaborated data to fit with the documented modelling workflow pipelines.

At the final stages of the project (mid 2024 and further) final modelled inversion results will be uploaded and published as elaborated (level 3) data.

Almost all elaborated data will be spatio(temporal) that should be provided in cf-compliant (ATMODAT standard, https://www.atmodat.de/atmodat-standard) NetCDF format. A detailed protocol for ingestion at ICOS-CP of netcdf data that allows for support of discovery of variables, previews and subsetting is in preparation for summer 2023.

5.4 Progress in implementing the new data pipelines

Table 1 Overview of the implementation progress as of February 2023

Project	Location	Type	Provider	Data level(s)	Pipeline	Data Processor	Status	Data Start	Data End	Pipeline Completed
CoCO2	Heraklion	Urban CO_2 Eddy covariance (2x)	FORTH	L0, NRT, L2	platinum	ETC	○	1-2022	12-2022	6-2023
CoCO2	Krakow	Urban CO_2 Eddy covariance	AGH	L0, NRT, L2	platinum	ETC	○	2-2021		6-2023
CoCo2	Krakow	CO_2, CH_4 profiles	AGH	L2	bronze	AGH	○	3-2021	2-2022	9-3023
CoCO2	Krakow, KAS	$^{13}CO_2$, $^{14}CO_2$	AGH	L2	bronze	AGH	○	2022	2022	9-2023
CoCo2	Sodankylä	Aircore vert. profiles (CO_2, CH_4, (N2O))	FMI	L2	bronze	FMI	○	2021	2022	9-2023
ICOS Cities	Paris, Munchen, Zürich	Many different data types	Many	L0-L3	platinum& bronze	ATC, ETC,local	○	2021		12-2023

○ In progress, data flowing through Flex ingest ● Ready

6 Conclusion

All data streams that are not part of the current ICOS-CP process can be implemented to flow through ICOS-CP, but the management is different based on observation protocols, compliance of metadata standards, level of QA/QC and the requirements of timeliness, so the data will be ingested as project specific (non-ICOS) data, providing relatively limited metadata on provenance compared to full ICOS data.

Short campaigns (mainly useful for algorithm development purposes) can be saved as “flux ingest”, but if data is close enough to ICOS requirements, a NRT pipeline can be created (e.g. daily accumulating time series).

7 References

ICOS Data Improved Lifecycle (2020), https://doi.org/10.18160/D2JV-KB6B

Document History

Version	Author(s)	Date	Changes
0.1	Alex Vermeulen (ICOS ERIC)	16/01/2023	First version
0.2	Alex Vermeulen (ICOS ERIC)	12/02/2023	Update, add PAUL
0.3	Konstantinos Politakos	13/02/2023	Minor changes conc. FORTH
0.4	Miroslaw Zimnoch	13.02.2023	Minor changes conc. AGH
1.0	Alex Vermeulen (ICOS ERIC)	23/02/2023	Implemented comments and corrections of both reviewers

Internal Review History

Internal Reviewers	Date	Comments
Wouter Peters (WU)	20/02/2023	Good work!
Marko Scholze (ULUND)	21/02/2023	Heavy read. Please add glossary

Estimated Effort Contribution per Partner

Partner	Effort
ICOS ERIC	2 PM
ULUND	3 PM
FORTH	? PM
AGH	? PM
Total	**5 PM**