Greenhouse Gas Observation
& Climate-Smart Agriculture

THÜNEN

Report on the blueprint for an EU-Africa e-infrastructure

Ville Kasurinen, Wim Hugo, Emmanuel Salmon, Elena Saltikoff, Alex Vermeulen, Johannes Beck, Myléne Ndisi, Werner L Kutsch

European
Commission

Funded by the
European Commission
under the Horizon 2020
Work Programme

Project: 730995 - Supporting EU-African Cooperation on Research Infrastructures for Food Security and Greenhouse Gas Observations (SEACRIFOG)

Work package number: 5

Work package title:

Deliverable number:

Deliverable title:

Lead beneficiary: ICOS ERIC HO

Lead authors: Ville Kasurinen, Wim Hugo, Alex Vermeulen, Emmanuel Salmon, Elena Saltikoff, Werner Kutsch

Contributors: Myléne Ndisi, Johannes Beck

Submitted by: Veronika Jorch

Place, Date of submission: 28.2.2020

Outline

Executive Summary 1

List of Abbreviation 2

1. Introduction and background 5

2. SEACRIFOG Task 5.3: Roadmap for a common EU-African e-infrastructure 5

3. Overview of the landscape 6

3.1 SEACRIFOG project 6

3.2 Environmental and climate monitoring 7

3.3 Systematic observation – combining low and high cost sensors and data 7

3.3 Data Infrastructure 8

4. Blueprint of an EU-Africa e-infrastructure 9

4.1 Introducing principles for the blueprint data infrastructure 9

4.1 A simple solution 13

4.2 An advanced solution 14

4.3 Path from simple to advanced solution 14

4.3 Data related requirements of the infrastructure 18

4.4 Storage capacity requirements for in-situ measurements 18

4.5 Requirements for data transmission 20

4.5 Roadmap for the long-term sustainability 22

Literature 25

Annexure A: Concepts, Architecture, and Requirements 26

Annexure B: Actors, Systems, Services, and Stakeholders 28

Annexure C: Information Model 31

Annexure D: Use Cases 32

Annexure E: Specifications and Standards 35

Annexure F - Description of ICOS Carbon Portal 36

Executive Summary

This report is Deliverable 5.4 of the SEACRIFOG project funded by the European Union’s Horizon 2020 research and innovation program. The goal of the SEACRIFOG project is to support EU-Africa cooperation at several different levels including climate change, carbon cycle and greenhouse gas observations to support mitigation and adaptation under a changing climate. The overall purpose of SEACRIFOG is to promote the building of an integrative network for long-term and sustainable cooperation between African and European environmental research infrastructures.

Deliverable 5.4 “Report with the blueprint for an EU-African e-infrastructure” is summarizing elements introduced in previous deliverables from work packages 3, 4 and 5 with a focus on key elements required from the e-infrastructure. The blueprint e-infrastructure has been built by SAEON and the first version of the blueprint e-infrastructure is based on the SEACRIFOG Collaborative Inventory Tool that can be used to visualize available data products from Essential Climate Variables in Africa. The blueprint infrastructure is partly based on techniques and methodologies developed at the ICOS Carbon Portal and the current structure is open for the future development and needs.

While the deliverable 5.1 “Requirements and design considerations for an interoperable data portal” described FAIR principles related to available research data and currently used technical solutions, this report is describing the roadmap for an EU-African e-infrastructure. The blueprint infrastructure compiled by SAEON will merge and demonstrate efforts carried out in SEACRIFOG regarding existing data sets. The blueprint infrastructure will act as the first brokering registry for available data products related to environmental monitoring in Africa and can be extended to serve for example needs of the future African Research Infrastructure measuring carbon and GHG emissions.

The roadmap for an EU-African e-infrastructure describes the status and technical readiness level achieved during the SEACRIFOG project and documents the required steps from the state of the art to a simple and advanced solution. The e-infrastructure is designed to serve the EU-African research infrastructure including operational measurement infrastructure that would utilize services provided by the e-infrastructure. Deliverable 3.2 is reporting details regarding the technical requirements and estimated costs for the measurement infrastructure. The conclusion is that the costs of the e-infrastructure are small compared to investments and operational costs of measurement stations. However, a well designed and implemented e-infrastructure providing storage, data processing and analytical services will play an important role in the operation of the measurement infrastructure.

List of Abbreviation

Abbreviation	Explanation
AfriGEO	The African Group on Earth Observations
AOSP	African Open Science Platform
API	Application Programming Interface
AU	African Union
CLIVAR	Climate and Ocean - Variability, Predictability and Change
CMIP	Coupled Model Intercomparison Project
CP	Carbon Portal
CSIR	Council for Scientific and Industrial Research
DEM	Digital Elevation Model
EC	Eddy Covariance
ECMWF	The European Center for Medium-Range Weather Forecasts
ECV	Essential Climate Variable
eLTER	Long-Term Ecosystem Research in Europe
ENVRI RM	Environmental Research Infrastructure Reference Model
EOSC	European Open Science Cloud
EOV	Essential Ocean Variable
ERIC	European Reserearch Infrastructure Consortium
ESFRI	The European Strategy Forum on Research Infrastructures
ETC	Ecosystem Thematic Center
EU	European Union
EUROCOM	EUROpean Atmospheric Transport Inversion COMparison
FAIR	FAIR principles, Findable, Accessible, Interoperable, Reusable
FLUXCOM	An Iniative to upscale bioshpere-atmosphere fluxes from FLUXNET sites to continental and global scales
GAW	Global Atmosphere Watch
GCOS	Global Climate Observing System
GEO	Group on Earth Observations
GEOBON	Group On Earth Observations Biodiversity Observation Network
GEOSS	Global Earth Observation System of Systems
GHG	Green House Gases (CO2, NH4, N3, water vapor)

GIS	Geographical Information Systems
ICOS	Integrated Carbon Observation System
ICOS CP	Integrated Carbon Observation System Carbon Portal
ICOS ERIC	Integrated Carbon Observation System European Research Infrastructure Consortium.
ICOS ETC	Integrated Carbon Observation System Ecosystem Thematic Center
ICOS RI	Integrated Carbon Observation System Research Infrastructure
ICOS RO	ICOS Romania
ICOS TC	Integrated Carbon Observation System Thematic Center
ICSU-WDS	International Council for Science - World Data System
IG3S	An Integrated Global Greenhouse Gas Information System
IOC	Intergovernmental Oceanographic Commission
IPCC	Intergovernmental Panel on Climate Change
KII	Key Impact Indicators
KPI	Key Performance Indicator
LAI	Leaf Area Index
LC	Land Cover
LCCS	Land Cover Classification Scheme
LTER	Long-Term Ecosystem Research in Europe
LULUCF	Land Use, Land Use Change and Forestry
LW	Longwave
MISR	Multi-angular Imaging Spectral Raiometer
MSA	Monitoring Stations Assembly
MVS	Monitoring and Verification Support
NEE	Net Ecosystem Exchange
NEON	National Ecological Observatory Network
NIR	Near Infrared
NPP	Net Primary Production
OADC	Open Access Data Centre
OPD	Open Data Platform
OpenID	OpenID allows you to use an existing account to sign in to multiple websites, without needing to create new passwords
PI	Principal Investigator
PID	Persistent Identifiers
QA	Quality assurance

QC	Quality control
RDA	Research Data Alliance
RI	Research Infrastructure
RICOM	Research Infrastructure committee
RINGO	Readiness of ICOS for Necessities of Integrated Global Observations
ROI	Research Output Infrastructure
SADC	South African Development Community
SAEON	South African Environmental Observation Network
SAEOSS	South African Earth Observation System of Systems
SASDI	Sout African Spatial Data Infrastructure
SASSCAL	Southern African Science Service Centre for Climate Change and Adaptive Land Management
SEACRIFOG	Supporting EU-African Cooperation on Research Infrastructures for Food Security and GHG Observations
SOC	Soil Organic Carbon
TAHMO	Trans-African Hydrometeorological Observatory
TC	Thematic Center
TCCON	Total Carbon Column Observing Network
UN	United Nations
UNCBD	United Nations Convention on Biological Diversity
UNFCCC	United Nations Framework Convention on Climate Change
VERIFY	Verifying Greenhouse Gas Emissions
WMO	World Meteorological Organization
WP	Work Package

1. Introduction and background

The purpose of the project "Supporting EU-African Cooperation on Research Infrastructures for Food Security and Greenhouse Gas Observations" (SEACRIFOG) is to develop a continental network of joint EU-African research infrastructures (RIs) for monitoring GHG emissions and observing the climate system in Africa. This report constitutes Deliverable 5.4 "Roadmap for a common EU-African e-infrastructure" including a technical description regarding the blueprint e-infrastructure, which will be demonstrated by SAEON.

This report has considered previous work carried out in Work Package 3 "Developing a common research agenda to promote Carbon, GHG and aerosol observation in Africa to fill gaps in a global observation system" and Work Package 4 "Improving technical harmonization and data quality in environmental monitoring and experimentation". Previous reports from Work Package 5 "Interoperability of RIs, Access, Data Sharing" have described existing and theoretically available technical solutions and FAIR principles that can be used to facilitate platforms supporting scientific community. Deliverable 5.1 "Requirements and design considerations for an interoperable data portal" and Deliverable 5.2 "Promoting open access policies and licenses" has tried to create a rough overview from the landscape pointing out difference and similarities between different data infrastructures and policies that can be applied, when building a prototype of e-infrastructure.

As stated in the previous Deliverable 5.1, an interoperable data portal storing and providing easy and fast access to reliable, high-quality environmental data is a fundamental component for the future African Research Infrastructure. SEACRIFOG reports by Work Package 4 indicate that environmental observation data have several gaps over the African continent causing higher uncertainties to future predictions than well-monitored areas. WP4 has provided three reports compiling the most essential information regarding the existing data availability and relevant standards that can be used to design a GHG monitoring network. The SEACRIFOG Collaborative Inventory Tool will create the basis for a brokering registry of the available earth observation data products and sources and will be implemented in the blueprint e-infrastructure. During the blueprint infrastructure implementation phase, we will also investigate the possibility to implement automated a raw data processing pipe that has been previously developed in ICOS RI thematic centers for Atmospheric and Ecosystem stations. This roadmap report together with the blueprint e-infrastructure will try to provide the first glimpse of those benefits that a modern and well-designed monitoring infrastructure can provide for the scientific community locally and globally.

2. SEACRIFOG Task 5.3: Roadmap for a common EU-African e-infrastructure

This task aims to design a roadmap for a common EU-African e-infrastructure and is connected to the blueprint e-infrastructure for an EU-African Research Infrastructure.

Using the requirements gathered in task 5.1 we will identify the gaps and opportunities for a common e-infrastructure that supports a common research infrastructure. SAEON will develop for this a brokering registry, where the parties involved can offer standard-compliant services (clients - such as discovery, visualization, processing, analysis or query tools - and data or meta-data sources) within an infrastructure. SASSCAL can contribute with its experience of such an infrastructure in the region. In cases where there are deviations from the standard, mediation actions and mappings should also be accommodated. This brokering framework and its operational implementation directly contribute to the goal in respect of improved interoperability, but we will also be able to document our approach and contribute to the proposed roadmap.

3. Overview of the landscape

3.1 SEACRIFOG project

The overall aim of the SEACRIFOG project is designing an integrated Research Infrastructure (RI) including an e-infrastructure for the data measured and provided by this RI as well as external data (e.g. satellite data, see Figure 1). The e-infrastructure shall furthermore comprise modelling capacities and services for researchers to work scientifically on the data. WP1 in SEACRIFOG has summarized needs and gaps in terms of data, knowledge and RI suggesting recommendations for a joint EU-Africa Research Infrastructure and reflected stakeholder expectations and requirements for capacity building. Deliverable 1.1 points out several gaps based on stakeholder interviews and workshops that need to be filled to ensure that African knowledge will be integrated into the pan-African observational system of Earth Observations and GHGs (López-Ballesteros et al., 2018). Cooperation with European Research Infrastructures like ICOS and LTER and organizations like GEO and AfriGEO can significantly support capacity building and knowledge transfer. For example, existing management structures in established RIs and environmental measurement protocols developed by the ICOS thematic centers provide a starting point for fulfilling the needs of stakeholders and users of an African RI.

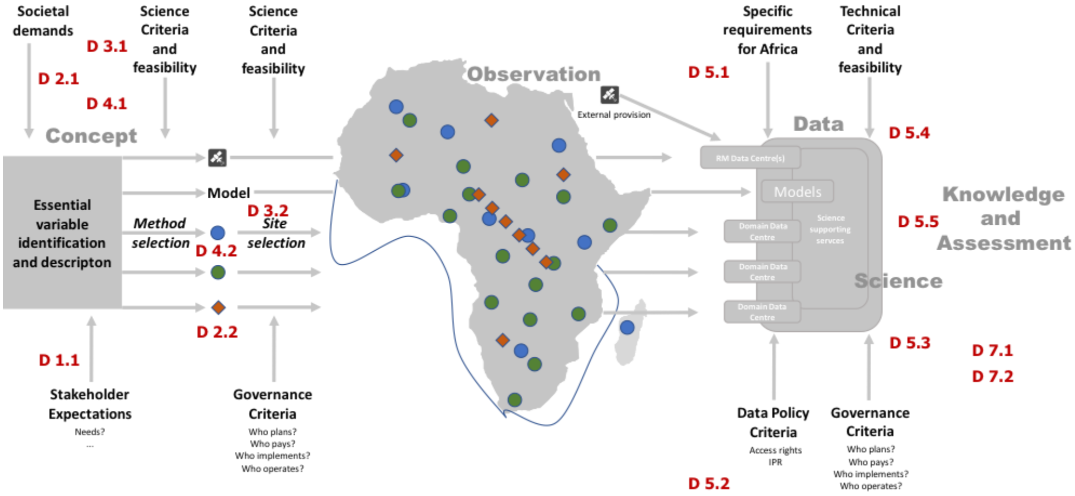


Figure 1: An overview of the general approach of the SEACRIFOG project.

The work carried out in WP3 has compiled information regarding the observational networks in Africa and documented the required variables for a comprehensive monitoring system. The requirements arising from Deliverables 3.1 and 3.2 have been considered when designing a generic solution for the blueprint e-infrastructure. For the e-infrastructure roadmap for African environmental and GHG monitoring the most advanced system is described in D3.2. In this advanced system the observational in situ measurement network will be connected to a data infrastructure in Africa. It will also be possible to calibrate remote sensing-based data products and to run models that require inputs from remotely-sensed variables and in situ measurements. However, such advanced system would require significant investments into an IT infrastructure in Africa and maybe a similar operator to the European Open Science Cloud (EOSC) that can facilitate, support and help build advanced solutions for evidence-based decision making. The African Open Science Platform (AOSP)(http://africanopenscience.org.za/) project managed by the Academy of Science of South Africa is one of the first openings that can lead to the development and fostering of cloud-based hubs serving scientific users.

3.2 Environmental and climate monitoring

Environmental monitoring as well as food production data are key to integrate greenhouse gas observations with impact and adaptation studies. For example, the Trans-African Hydrometeorological Observatory (TAHMO) aims to develop a network of weather stations across Africa using low-cost sensors, train the local people to operate the devices and provide data for farmers. Data collected in the TAHMO project is available freely for scientific use, but not for business solutions. The low-cost meteorological sensor network is a welcome addition to sparse environmental monitoring data on the African continent, but requires quality controlling and continuous maintenance activities.

African environmental monitoring data can also contain a tremendous potential for private companies, if these build a monitoring network carrying the risk of investments and then providing business-oriented solutions for data access. The European commission, European RIs, the African Union and international supporting organizations should actively work to ensure that the environmental monitoring data will be accessible for the people in Africa and for scientific use.

3.3 Systematic observation – combining low and high cost sensors and data

Several recent attempts trying to characterize the reliability of low-cost CO_2 sensors found in the literature are briefly summarized here. Case studies have varied from the wireless indoor air quality sensors measuring CO_2 concentrations to city scale studies (https://www.slideshare.net/ICOS_RI/a-lowcost-sensor-network-to-monitor-the-co2-emissions-of-the-city-of-zurich). The measurement accuracy has been reported to be dependent on ambient air temperature, relative humidity and calibration frequency. A network of 209 sensors can in some degree detect the variability of CO_2 emissions of the city of Zurich but required calibration with an intensive network of CO_2 measurements carried out with gas analyzers having a high detection accuracy. Therefore, low-cost CO_2 sensors cannot be utilized independently and without posterior drift corrections and quality controlling of data. From the scientific point of view, low-cost sensor devices can, if well maintained, produce additional information that can be useful for example for inverse modelling studies. For the estimation of carbon sources and sinks, the trace gas concentration alone cannot be used to estimate the ecosystem or regional carbon balance. Additional information regarding the meteorological variables and turbulent transport of trace gases will be required for such analysis.

The previous work carried out in WP4 has summarized Essential Climate Variables (ECV) and available data sources when building the SEACRIFOG Collaborative Inventory Tool and suitable measurement protocols that can be used in network and measurement station design. Deliverables from WP4 have also reported the most important gaps in the current observation network and data availability. WP4 has also suggested an optimal observational network design for atmospheric towers that can be used to detected content scale carbon sink and sources (Nickless et al 2019, submitted).

3.3 Data Infrastructure

The previous deliverables in WP 5 have described principles and practices that can be followed when building an interoperable data portal (D5.1) using FAIR data principles. Deliverable 5.2 has made suggestions regarding the data policy considerations related to data portal and data use. This report together with the blueprint e-infrastructure is trying to describe the benefits and technical solutions that can be used when building a data portal serving the operational needs of an RI. The existing e-infrastructure solutions available in SASSCAL and SAEON have been described through generic use cases in D5.1, as well as the structure and design of the ICOS Carbon Portal. ICOS CP has developed automated work flows for the post-processing of raw data derived from sensors using standardized methods.

The approach in ICOS RI has been to standardize GHG measurements including used sensors, post-processing and quality controlling of the data. In the African context, the existing measurement sites cannot immediately start to follow the measurement protocols suggested by ICOS if for example the technical requirements are not fulfilled. The African RI should agree on adapted measurement protocols. At least in the beginning there will be a need to handle data deriving from heterogeneous measurement setups. However, technical solutions developed at ICOS CP and ICOS ETC can be adapted and measurement stations can be connected to a centralized data base and post-processing pipelines.

These technical enhancements are likely to increase data quality and data quality control and allow the RI to document processes related to its data life cycle. According to previous findings associated to the work carried out in the RINGO project, the post-processing of eddy covariance raw data is one factor influencing data products and their uncertainty. Harmonization is difficult in general because post-processing is dependent on the specific characteristics of the site and the analyzer. For example, the applied spectral corrections may have a large influence on calculated fluxes of CO_2, water vapor and methane (Mammarella et al 2016, Fratini & Mauder 2014).

An interoperable data portal integrating measurements from different observational networks and providing a Jupyter hub in the cloud is essential for the scientific community. Scientific analyses and use cases on a virtual platform will allow users to analyze and visualize data in a cloud environment without downloading all data products. The virtual platform for scientific analyses is especially useful for such users who do not have access to a high-speed telecommunication network. An interoperable data portal, which has been built using FAIR principles introduced in D5.1, should also be connected to the governance structure and the high-level dialogue platform (WP7) in order to ensure that the e-infrastructure or the RI can fulfill the requirements of the stakeholders.

In situ monitoring networks for ecosystems (ICOS, AmeriFlux, Fluxnet, AsiaFlux) and stations monitoring trends in the Earth's atmosphere (GAW, GCOS, TCCON) are very important components when estimating the global GHG balance. High-quality measurements are also crucial for achieving the goals of the Paris Agreement and for monitoring and verification support (MVS) mechanisms. They are also playing a significant role in other H2020 projects. For example, VERIFY develops a system to estimate greenhouse gas emissions to support countries in their emission reporting to the UN Climate Change Convention. In VERIFY, the emissions are estimated based on land, ocean and atmospheric observations using several methodologies and with a focus on carbon dioxide, methane and nitrous oxide.

4. Blueprint of an EU-Africa e-infrastructure

4.1 Introducing principles for the blueprint data infrastructure

This section describes the design for an e-infrastructure aggregating multiple layers of data processing, storage, archiving, and processing in support of the SEACRIFOG project, with a prioritization for development towards an initial basic implementation of such a system. The design is based on the considerations outlined in SEACRIFOG Deliverable 5.1 (Kutsch et al, 2019).

The designed system needs to meet three overarching requirements. It needs to be:

- **Interoperable:** data sources, processing services, components (for example for visualization), and supporting vocabularies are increasingly open, and a system needs to be developed in compliance with community-agreed standards and specifications for interoperability.
- **Configurable:** data services, visualizations, and processing services are often not exactly suited to the end use - and need to be configurable to suit the context of the end user (largely via semantic mapping and transformation).
- **Federated:** data services, vocabularies, components, and processes need not be centralized, and any architecture we adopt needs to support this distributed, federated model for the construction of web-based resources.

As a practical example, these requirements have been implemented and fulfilled in the data infrastructure developed in ICOS RI as follows:

Interoperable: the ICOS Carbon Portal system is based on a triple store for linked data and a SPARQL endpoint to access all metadata and data objects. The data portal is interoperable and openly accessible.
Configurable: many of the aspects mentioned above are implemented and available in the ICOS Carbon Portal. For example, the preview of time series is available as iframe and can be embedded directly by end-users. For the time being, there is no instance of a third party semantic available, mainly because there was no need so far. The near future development and collaboration with the NEON network in the USA may require third party integration.
Federated: ICOS is by essence a distributed, non-centralized system. The thematic centers for Atmosphere, Ecosystem and Ocean have custom-made and specific locations, systems and methods. The Carbon Portal acts as a collector and distributor. The lesson learned is that it is difficult to bring together different heterogenous data sources to provide a simple and homogenous interface for the end-user.

The designed system in this report and the blueprint infrastructure have to contribute to the following objectives:

- Provide an overview and a single access point for understanding and exploring the scope of carbon-related observation in Africa;
- Express this scope of observation in terms of space and time, as well as essential or standard variables being observed, instruments and sensors employed for observation, and its deployment within institutions, networks, projects, and initiatives;
- Provide background information on the protocols used for observation, to maximize the potential for collation of data across the continent;
- Provide links, whenever possible, to contributing infrastructures for the purpose of data access, preferably via standardized data services;
- Serve as an interface and a channel of site-level and dataset-level metadata to international and regional initiatives, such as GEOSS and ILTER;
- Implement the data and dissemination policies adopted in SEACRIFOG as required.

Such a system has to contribute to the following objectives:

SEACRIFOG requirement	**ICOS Carbon Portal as an example**
Provide an overview and a single access point for understanding and exploring the scope of carbon-related observation in Africa	Single point of access, one stop shop, one website
Express this scope of observation in terms of space and time, as well as essential or standard variables being observed, instruments and sensors employed for observation, and its deployment within institutions, networks, projects, and initiatives	Data portal allows to select different projects as well as search and filter data objects. Data objects are linked to stations, instruments, projects, etc. and an automatic history of change is stored for provenance and reproducibility
Provide background information on the protocols used for observation, to maximize the potential for collation of data across the continent	Each thematic center has created a set of protocols on how to perform measurements and collect data. For calibration and harmonization purposes, central facilities are available (instrument calibration for example)
Provide links, whenever possible, to contributing infrastructures for the purpose of data access, preferably via standardized data services	Persistent Identifiers (PID) for each digital object. This allows to create automatically citation strings and back links to the originator
Serve as a channel of site-level and dataset-level metadata to international and regional initiatives, such as GEOSS and ILTER	Support or participation for EUROCOM, IG^{3}IS, TRANSCOM, FluxCOM
Implement the data and dissemination policies adopted in SEACRIFOG as required	Published data policy and open data portal

The above largely addresses requirements for the horizontal integration of systems across different institutions, consortium members, and networks. In addition, the blueprint needs to include example implementations of representative value chains (vertical integration) that will commonly be encountered by contributing organizations to serve as models. These are aligned with the need for support of generic data families (Mirtl et al, 2018, Hugo et al, 2017). In short, one needs to demonstrate usefulness for the following:

1. Traditional spatially referenced datasets, based on either remote or in situ observation (or for socio-economic data, surveys and government statistics). These datasets are typically explored online, linked or downloaded for use in desktop GIS environments, and linked into online atlases, indicators, and other decision support tools;
2. Multidimensional model or remotely sensed data cubes, typically explored online, and then downloaded or subsetted and downloaded for inclusion into models, analyses, and scientific workflows;
3. Time series observations, with or without raw processing of real-time or near-real time data, for which quality assurance and automated publication of data is an important aspect;
4. Other digital objects, which can include small, schematically diverse datasets, reports, and other resources required by the community;
5. Demonstrate how evidence can be applied to context-specific decision, planning, and policy support.

These typical applications need to be assessed and a synthesis, leading to illustrative use cases, has to be developed.

To achieve these goals, the SEACRIFOG implementation needs to do the following:

SEACRIFOG implementation	ICOS Carbon Portal as an example
Confirm a systems architecture that can address the requirements and use cases expressed for the system	The existing data portal (https://www.icos-cp.eu) is a running implementation.
Define the main systems components that will be required to support the objectives of the system, as well as the major APIs required to communicate between them	We have our own servers to provide the frontend. High availability data center for backup SPARQL endpoint to access all of the metadata and data
Define the actors and user roles that will be using the system	We have many different roles. But in a nutshell: Station Principal Investigator (PI) responsible for: a. Data collection b. Adhere to the ICOS protocols c. Data submission to the Thematic center d. Final quality control of data Thematic center a. Automatic data quality assurance and quality control (QA/QC) b. Authority of Metadata collection c. Data and metadata submission to the Carbon Portal Carbon Portal a. One stop shop b. Long-term data storage c. High availability of Data services d. User friendly interface e. Adhere to FAIR principles, open access
Define APIs for contributing infrastructures that provide background information on variables, protocols, sensors and instrumentation, stations and deployment, and potentially the management and organizational structure of the observation network. Many of these service APIs will need to align with existing specifications or services, and several may make use of existing vocabularies, registries, or metadata catalogues	Bespoke interface for Stations and Thematic center (upload) and a SPARQL endpoint as API implementation for end-users

Define APIs and service infrastructure for the harvesting of metadata associated with the data generated by the observation network. In addition, there is a need to define the standard data service APIs that will be used to access, subset, download, and visualize or explore data. Such data will generally be stored in suitable infrastructure supporting one or more of the standard data families defined earlier for SEACRIFOG (Kutsch et al, 2019)	All data objects are available through the SPARQL endpoint. Data slicing is not commonly available for the moment (internal discussions ongoing).
Create a conceptual model for the system of systems data layer, with proper definition of the main entities, relationships between them, and cardinality of such relationships.	
Define the use cases and functionality required from system components, and support the requirements of the different actors and user roles.	Internal roles are defined and active: Principal Investigator → Thematic Centre → Carbon Portal Definition of third party / end-user use cases is in progress.

4.1 A simple solution

The simple solution should provide a basic value chain to researchers, decision-makers, and policy-makers: understanding the scope of published evidence in respect of carbon observation, being able to explore and download or access the datasets and products linked to observed variables, and being able to include standardized data services into simple societal benefit applications (for example, thematic atlases based on datasets and products that are federated).

A simple solution for an initial implementation should be based on the following broad considerations:

- Making maximum use of existing infrastructure from SEACRIFOG consortium members, taking the requirements and needs identified in the project into account.
- Allow for extensions and scalability, since the consortium may grow, and more functionality will be added in the future.
- Largely be sustainable on the basis that funded contributions from consortium members will form the bulk of the available resources, with as little funding as possible required for governance, coordination, and technical integration.

4.2 An advanced solution

Advanced solutions need to broaden the scope of the simple solution in a number of ways:

Governance aspects:

- Establish some form of coordination, or secretariat, possibly utilizing AfriGEO or GEO as a platform;
- Agree on a consortium framework, constitution, and bylaws that define participation, governance structures, membership, policies, and financial or in-kind contributions (if any). Bodies such as GEO-BON and ILTER can serve as models of initiative building.

Technical aspects:

- Provide ways for more data providers to submit site and data product metadata to the SEACRIFOG infrastructure;
- Extend the value chain in both directions - adding pre-processing and raw data pipelines/ resources where required, and adding research and societal benefits (for example helping researchers include data services into workflows, or contributing data to more complex decision and policy support applications).
- ICOS Carbon Portal as an example of advanced solution

4.3 Path from simple to advanced solution

The following table summarizes the actions to be taken to move from a simple to a more advanced solution, together with estimates of time frames and responsibility or proposed responsibility for each action.

Aspect	Element	State of the Art	Simple Solution	Advanced solution
Governance	Framework	SEACRIFOG Collaborative Inventory Tool	The roadmap (D5.4) defines a proposed framework, Feb 2020 - SEACRIFOG	Inception meeting of stakeholders, facilitated development of a consortium agreement, framework, constitution, bylaws, and data policy. Possible support organizations AfriGEO/GEO, ICOS, ILTER Target: 2021
	Constitution	SEACRIFOG project	The roadmap defines a broad outline Feb 2020 - SEACRIFOG	
	By-Laws	No actions	No development	
	Data Policy	No actions	Deliverable 5.2 provides an outline and proposals for a data policy	

Aspect	Element	State of the Art	Simple Solution	Advanced solution
			Feb 2020 - SEACRIFOG	
	Finances	SEACRIFOG project funding	Financed from SEACRIFOG project until Feb 2020	Maintained by SAEON as a public platform but without specific support, no extensions or additions. Prepare funding proposals in collaboration with willing consortium partners. Target: 2021
Standards and Specifications	Protocols	Protocols associated to data listed in collaborative tool	Review of appropriate protocols and observation standards - Deliverable 4.3. Feb 2020 - SEACRIFOG	Adoption by a General Assembly or designated committee of an initiative, with periodic review. Promote standardization while recognizing diversity. Target: 2021 - Managed participation in standards-setting or consensus-seeking organizations and initiatives, such as RDA, OGC, TDWG, and GEO. Target: 2022
	Metadata Standards	Metadata associated to listed data	Review of appropriate content and service standards - Deliverable 5.1. Feb 2020 - SEACRIFOG	
	Data Standards	Data accepted as it is	Review of appropriate content and service standards - Deliverable 5.1. Feb 2020 - SEACRIFOG	

Aspect	Element	State of the Art	Simple Solution	Advanced solution
Systems and Infrastructure	Architecture	Collaborative Inventory Tool hosted by SAEON pointing to data sources	Develop a simplified architecture that is extensible and scalable. Indicate how architecture can be extended to accommodate more elaborate cases. Deliverable 5.4. Feb 2020 - SEACRIFOG	Advanced architecture documented and implemented, dependent on the nature of contributions expected from consortium members. Target: 2021
	Pre-processing	no preprocessing	Implement a carbon flux processing pipeline based on ICOS software stack. Adjust for non-standard sensors in use at an African site. Document the changes and process. Deliverable 5.4. February 2020 - SEACRIFOG	Create a publicly maintained fork of the ICOS software stack for non-standard sensors, and promote its use in non-ICOS observation infrastructures. Target: 2021
	Metadata Management: Sites	No actions	Develop a prototype application capable of API exchanges with contributing systems and with aggregators. Deliverable 5.5. February 2020 - SEACRIFOG	Obtain funding - financial or in-kind - for operational maintenance of the application. Target: 2021

Aspect	Element	State of the Art	Simple Solution	Advanced solution
	Data Management and Access	No actions	Develop example use cases demonstrating access to standardized data via metadata records for datasets and data products. Deliverable 5.5. February 2020 - SEACRIFOG	Extend use cases to supplementary functi ons and requirements. Target: 2021-2024
	Value Addition and Societal Benefit	No actions	Develop example use cases demonstrating access to standardized data via metadata records for datasets and data products. Deliverable 5.5. Feb 2020 - SEACRIFOG	Extend value chains to all appropriate data families and variables. Target: 2021-2024
	Physical Infrastructure	Hosted by SASSCAL & SAEON	Host on project consortium infrastructure	Find cloud-based infrastructure, preferably in Africa, funded from consortium or consortium-arranged funding. Target: 2021-2024
	Connectivity	No actions	Use existing research and commercial networks. Feb 2020 - SEACRIFOG	Find alignment and collaboration with initiatives such as UbuntuNet. Target: 2021-2024

4.3 Data related requirements of the infrastructure

Requirements for building an in-situ measurement network to observe GHG emissions and uptake in Africa has been described in Deliverable 3.2. including estimated costs for the measurement infrastructure. In this report we assume that such in situ measurement station network would be the customer of the designed data infrastructure and they would together constitute the Research Infrastructure. The assumptions of this report are limited to cover the data infrastructure that is required for a ground observational network. The estimated cost required to operate a larger data infrastructure including remote sensing-based data products and models is reported in Deliverable 3.2.

At the time of writing this report, we do not have information regarding existing data centres specialized in scientific computing in Africa, that could host physically servers and computational processes required by the RI. Investments in new servers are expected to take place every second or third year, when the amount of data collected by the infrastructure increases. Commercial service providers can be a good solution in the early phase of the infrastructure and do not require investments in devices. In case long-term funding

As an example, in Europe, the Copernicus Data and Information Access Services (DIAS) is currently providing five cloud-based online platforms that can be used to process, visualize and store datasets related to earth and environmental observations. However, these platforms are not meant for long-term storage of raw data. Cloud-based environments can typically provide several different technical options for the implementation and they are scalable in terms of computational capacity, size of the repository and temporal time scale. Service agreements can be made for periods varying from a few days to several years. For example, the DIAS WEkEO price list provides several different options for computational resources and their prices vary from 66€ to 24 000€ per month (https://www.wekeo.eu/web/guest/price-list). Estimated costs and more detailed description of the data and monitoring infrastructure including a centre concentrated on modelling and remote sensing-based data are given in Deliverable 3.2. In this report. In this report, we have concentrated to estimate the minimum requirements for the data storage capacity excluding person costs and data transmission costs.

4.4 Storage capacity requirements for in-situ measurements

One ground station using the eddy covariance method in the ecosystem or atmosphere domain produces roughly 20 GB of raw data per year. The final post-processed files averaged on half-hourly values are significantly smaller and can vary from 4 to 100 MB depending on the number of sensors used (Flux, meteorological variables and soil water and temperature profiles) and gap filling strategies. Final post-processed and gap-filled data files can be either downloaded using slower connections or visualized in cloud services. The transfer of raw data files to a centralized storage will require the transfer of 48 half-hourly raw data files every day (i.e. approximately 1-2 MB per file or 50 to 100 MB per day).

One possible solution for the storage is to store centrally measurements and raw data for the whole Research infrastructure. The other possibility is to distribute the data storage to domain specific centres or to different regions, based on the connectivity and bandwidth capacity. The collection and co-location of raw data files is crucial for post-processing activities and the system must be designed in such way that all measurement stations can store raw data in a centralized or distributed infrastructure. Raw data files should also be secured using B2SAFE replica (https://www.eudat.eu/b2safe) or similar techniques. Storage capacity can be arranged using commercial providers like Google and Amazon or from governmental or intergovernmental operators like EOSC (or the planned African Open Science Cloud, http://africanopenscience.org.za).

Based on experiences of ICOS RI, 2 TB storage is required to produce services for 50 measurement stations for one year. Obviously, the second measurement year doubles the storage requirements and during the first ten years storage capacity will increase to 20 TB for 50 stations (Table 1). It should be noted that all data does not need to be available all the time. Raw data sets are typically archived after they have been post-processed to Level 1 and Level 2 products.

Table 1. Data storage capacity requirements

Stations	**1st year (TB)**	**10th year**	**20th year**	**30th year**
50	2	20	40	60
100	4	40	80	120
150	6	60	120	180

According to prices derived from commercial service providers like Google and Amazon, the storage cost per € and B2B contracts can reduce the price approximately by 30% (Table 2). Long-term storage that does not need to be directly available is also less expensive and the cost is approximately half if data is infrequently accessed and in long-life archive. Depending on how the post-processing of raw data files is arranged, a temporary disc space extension might be required if compressed raw data files must be extracted for analyses. Uncompressed files require approximately 10 times more storage capacity than compressed ones. However, there should not be a long-term need to store uncompressed files for longer periods of time.

Table 2. Variation range of data storage costs based on commercial service providers (€).

	10th year			**20th year**			**30th year**		
Stations	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
50	4000	5400	6240	8000	10800	12480	12000	16200	18720
100	8000	10800	12480	16000	21600	24960	24000	32400	37440
150	12000	16200	18720	24000	32400	37440	36000	48600	56160

The estimated storage costs per measurement station during the first operative year is roughly 11€ per station and 540€ for a network of 50 stations (Table 2 & Table 3). However, it should be noticed that the storage cost is not the only factor that has an influence on operational costs. Larger storage capacity, computational power and human resources will be required to operate a data centre that has a focus on modelling and remote sensing-based data products.

Table 3. Data storage requirements and costs per year. The price has been calculated using the mean price from Table 2.

	Raw data (GB)	**Level 1 data (GB)**	**Level 2 data (GB)**	**Storage capacity total**	**Total price (€)**	**Price for 50 stations (€)**
Ecosystem station	40	0.1	0.1	40.2	11	550
Atmosphere station	40	0.1	0.1	40.2	11	550
Ocean station	10	0.1	0.1	10.2	3	140
Ship line	24	0.2	0.2	25	7	350

Although initial storage space requirements and costs for 50 measurement stations seems to be rather low, human resources are required for technical support, maintenance and data curation like quality control and quality assurance. Ground station measurements using the eddy covariance technique require post-processing or raw data, which is computationally demanding. The post-processing chain can be in some degree automated and the workflow for 50 measurement stations requires a server with a storage capacity of 100 TB and 20 cores. In the case the computational processes are outsourced to a service provider, the cost of 200 000-400 000 core hours can be expected to be from 10,000€ to 20,000€ per year. Compared to the human costs of doing software development and curation, for the volume and processing requirements of eddy covariance and atmosphere observations, the investments and running costs are small (See D3.2).

4.5 Requirements for data transmission

The estimated costs of data transfer are dependent on the local circumstances and access to data transmission networks in Africa. Based on the available information until 2021, African subsea cables on the west coast of Africa are connected to South America (Brazil) and Europe (Great Britain, Portugal, France) while east coast cables are connected to Mediterranean data transmission networks.

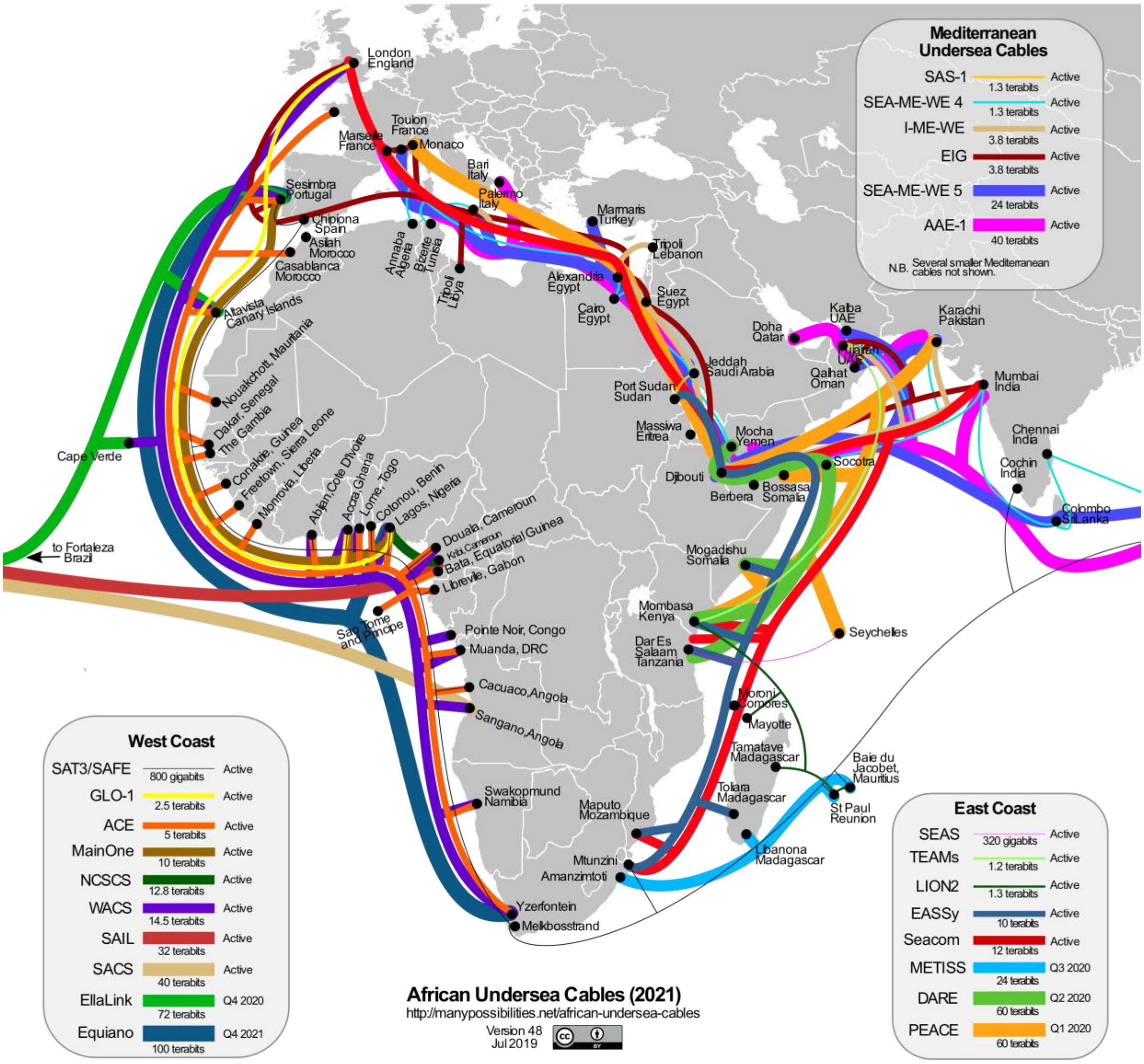


Figure 2. African undersea cables updated on July 2019. (https://manypossibilities.net/african-undersea-cables/)

A map of African undersea and terrestrial fibre optic cables (https://afterfibre.nsrc.org) provide an updated view of the existing networks and those currently under construction. It should be noticed that while some parts of the continent, like South Africa and countries on the eastern and western coasts, have established fibre optic cable networks, in several countries in central Africa fibre networks are still under construction. Measurement stations that are already connected to data transmission networks can easily use a centralized data portal or a distributed data infrastructure. However, when establishing new measurement stations, the building of a data transmission network cannot be included in the initial investment costs unless this is agreed. For measurement stations in remote areas it might be necessary to use microwave transmission links between the remote measurement point and a local research station that can provide access to a fibre cable network. Microwave technology can be used for data transfer up to 80 km distances. In case the distance to the remote measurement station is larger, data transmission through a satellite connection can be the only available option. The costs of microwave transmission and satellite transmission are not estimated in this report nor in Deliverable 3.2 because they are highly dependent on local existing infrastructures and circumstances. For example, the availability of technical stuff able to build and maintain such systems is crucial for the implementation and long-term sustainability of local solutions.

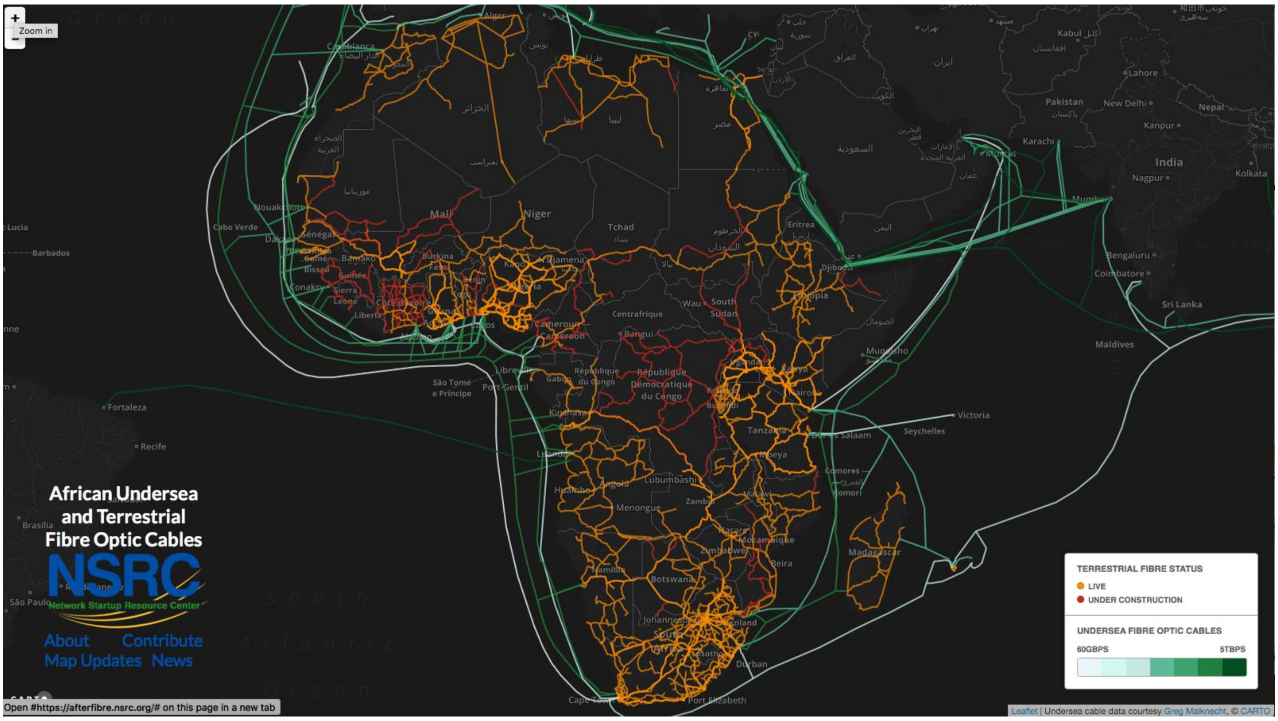

Zoom in
African Undersea and Terrestrial Fibre Optic Cables
NSRC
Network Startup Resource Center
About
Contribute
Map Updates
News
TERRESTRIAL FIBRE STATUS
LIVE
UNDER CONSTRUCTION
UNDERSEA FIBRE OPTIC CABLES
60GBPS
5TBPS
Open #https://afterfibre.nsrc.org/# on this page in a new tab
Leaflet | Undersea cable data courtesy Greg Malknecht, © CARTO
Mali
Niger
Tchad
India
Madagascar
Botswana
Durban
Port Elizabeth
Fortaleza
Recife
Seychelles
Victoria
Maldives
Sri Lanka

Figure 3. African undersea and terrestrial fibre optic cables according to the Network Startup Resource Center (https://afterfibre.nsrc.org).

4.5 Roadmap for the long-term sustainability

To get the commitment of partners, it is imperative to convince them of the long-term sustainability of the infrastructure. A typical time perspective of a research or development project, 3–5 years, is far too short to justify the starting investment to build a useful infrastructure. The World Meteorological Organization (WMO) is using 30-year periods to calculate the arithmetic average of a climate variable like temperature and precipitation.

Building a Research Infrastructure observing GHG emissions in Africa is therefore a long-term commitment and requires a preparatory phase project. For example, the ICOS Stakeholders' Interim Council (ISIC) was established in April 2010 as a high-level council for country representatives. ISIC corresponded in the preparatory phase to the current General Assembly of ICOS ERIC and decided upon strategic issues such as legal, governance and financial implementations or locations of facilities. ICOS achieved the European Research Infrastructure Consortium (ERIC) status in 2015. The initial expected lifetime of ICOS ERIC is twenty years, that is until 2035.

The following goals and objectives have been identified:

1. Partnerships
2. Funding Stability
3. Regular system evaluation and upgrades
4. Effective communications

4.5.1. Partnerships

Before a Research Infrastructure can be built, a formal decision is required including a strategy related to participation and resources. African countries should find a solution how an infrastructure consortium will be set up with the support of international and governmental organizations like the African Union, the European Union, the United Nations and WMO. For example, ICOS has at the moment 12-member countries. ICOS ERIC is open for new countries to join. The previous experience based on research infrastructures in Europe can be used when a similar environmental monitoring infrastructure is established in Africa.

The infrastructure must have in its bylaws clear roles of the partners at the level of the organization and at national level. Initial contributions are often based on enthusiasm of individual researchers, but for long-term sustainability it is important to have procedures that do not depend on individuals. In Europe, infrastructures often have a General Assembly with national representation as the highest decision-making body. Building collaboration with end-users from early stages is also fruitful, and leads often to better motivation of the funding agencies as well as to structures and services which are better suited for purpose. An advisory board consisting of users is recommended.

4.5.2. Funding stability

Data infrastructure as a part of an established environmental monitoring infrastructure require commitments from participants and a strategy for long-term funding requirements including initial investments, as well as operational and maintenance costs covering the estimated lifetime of the infrastructure. A reasonable lifetime for an infrastructure monitoring environmental conditions and climate related phenomena should be at least 30 years. The bylaws should also define winding-up processes, as well as rules and responsibilities related to remaining payments and debts of the RI among the members of the consortium. A plan defining how data collected by the RI will be maintained., stored and kept open after winding-up process is already necessary.

An adequate and consistent funding base through a variety of sources, both national and international, is needed. Note that some institutions such as the Word Bank only support investments, not the maintenance of ongoing services. It is not enough to buy servers once, they get old and must be updated regularly. At the time of writing this report, we could not find partners that would have been able to host physically servers in scientific computing centers in Africa. Such partners might still exist and develop projects like the African Open Science Platform which can be useful for data infrastructure development in Africa.

4.5.3. Regular system evaluation and upgrades

Computer technology - both hardware and software - is developing fast. Even though a large infrastructure cannot jump to the latest fashionable approach, it is advisable to monitor developments in e.g. database structures at a system level. A data infrastructure must also adapt to the development of the measurement infrastructure. For example, new LIDAR-based techniques can be used for landscape and forest structure mapping and analyses and data products are storage and computationally intensive.

The continuous development in measurement techniques requires also coordination in the measurement networks and the ability of the data infrastructure to add new instruments to the network. One possibility to coordinate protocols and standards used in measurements is to organize domain-specific thematic centers like it is done in ICOS or to agree on a tight cooperation with institutes that can assist in domain-specific questions. It is a fair assumption that measurement techniques and devices will not remain the same during the period varying from 10 to 30 years. During few decades measurement precession of various sensors will probably increase significantly and require adaption also from the post-processing and quality controlling of the data.

4.5.4 Effective communication

Environmental Research Infrastructure in Africa measuring carbon and other GHG dynamics in several ecosystem types and atmosphere would significantly improve our understanding how African continent contributes to global greenhouse gas and climate dynamics. The in-situ measurement network would be the key component producing timely and real-near-time information from various African ecosystems and atmospheric transport of trace gasses and would therefore be interesting also for wider scientific community.

Research Infrastructure need to define co-operation and agree on processes where essential information is exchanged and communicated. This includes communication between in-situ measurement network, funding agencies, research institutes involved in operations and national and international level stakeholders. Communication strategy and management plan will be required for internal and external communication needs. For example, in ICOS annual report is published every year, which reports the core activities carried out in the RI for stakeholder groups.

EU-African Research Infrastructure can also play a role in capacity building in Africa and promote the growth of the scientific community. Funding strategy described in detailed in SEACRIFOG Deliverable 3.2 is describing a vision for the next three decades. EU-African RI should also be in active interaction with other African science related iniatives like African Open Science Platform (AOSP) and seek potential ways to integrate RI to evidence based decision-making processes. This work will be continued in the last SEACRIFOG deliverable “D7.1. Integrated strategy for a sustainable EU-Africa research cooperation on food security and GHG observations” and “D7.2. Co-financing concept for the establishment and long-term maintenance of adapted observation systems on food security and GHGs”.

Literature

Arzoumanian, E., Vogel, F. R., Bastos, A., Gaynullin, B., Laurent, O., Ramonet, M., & Ciais, P. (2019). Characterization of a commercial lower-cost medium-precision non-dispersive infrared sensor for atmospheric CO_2 monitoring in urban areas. *Atmospheric Measurement Techniques*, *12*(5), 2665-2677.

Mueller, E., Graf, P., Meyer,J., Pentina,A., Dominik, B., Perez.Cruz, F., Hueglin, C., & Emmenegger, L. (2019). Integration and calibration of NDIR CO2 low-cost sensors, and their operation in a sensor network covering Switcherland. *Atmospheric Measurement Techniques Discussions*, https://doi.org/10.5194/amt-2019-408, in review.

Fratini, G., & Mauder, M. (2014). Towards a consistent eddy-covariance processing: an intercomparison of EddyPro and TK3. *Atmospheric Measurement Techniques*, *7*(7), 2273-2281.

Hugo, W., Hobern, D., Kõljalg, U., Tuama, É. Ó., & Saarenmaa, H. (2017). Global infrastructures for biodiversity data and services. In *The GEO handbook on biodiversity observation networks* (pp. 259-291). Springer, Cham.

Kutsch et al. (2017) Data interoperabilty between European Environmental Research Infrastructures and their contribution to global data networks, AGU Abstracts

Kutsch et al (2019). Requirements and design considerations for an interoperable data portal. SEACRIFOG Deliverable 5.1.

Mammarella, I., Peltola, O., Nordbo, A., Järvi, L., & Rannik, Ü. (2016). Quantifying the uncertainty of eddy covariance fluxes due to the use of different software packages and combinations of processing steps in two contrasting ecosystems. *Atmospheric Measurement Techniques*, *9*(10), 4915-4933.

Mirtl, M., Borer, E. T., Djukic, I., Forsius, M., Haubold, H., Hugo, W., ... & Orenstein, D. E. (2018). Genesis, goals and achievements of long-term ecological research at the global scale: a critical review of ILTER and future directions. *Science of the total Environment*, *626*, 1439-1462.

Ndisi, M., Nickless, A., Beck, J., López Ballesteros, A., Kasurinen, V., Merbold, L., Scholes, R.J., Salmon, E., Vermeulen, A., Kutsch, W.L. Concept paper for an adapted observation system for Africa including special and sectoral observational requirements to integrate multiple Grand Challenges in Africa. SEACRIFOG Deliverable 3.2.

Annexure A: Concepts, Architecture, and Requirements

A.1 Main Entities

1. The main entity in the SEACRIFOG conceptual model is the concept of an Initiative. An initiative has different levels of granularity, and can be a global, regional, or local network, a program or project with a start or end, a research infrastructure, or a specific institution.
2. The typology of Initiatives are described in a controlled vocabulary (Initiative Typology).
3. An Initiative can obtain data from, be linked to, or operate one or more Sites. A Site is synonymous with a Platform, and although it can be stationary (as are many terrestrial sites), it need not be - vessels and floats in the oceans are also observation platforms, as are satellites, drones, and BRUVs.
4. Sites have a Site Typology as a controlled vocabulary, distinguishing whether observations are made in situ or remotely, and potentially distinguishing subtypes of in situ and remote observation.
5. The 'Observed Feature' or Station is fixed in three-dimensional space, but not all sites or observed features are predetermined, and the station or feature of observation need not be in the same location as the site or platform.
6. Some stations are opportunistic (citizen science observations or voyages of opportunity), some are virtual (pixels in a remotely sensed scene), and some are partially or wholly random (ocean floats).
7. Stations have a Station Typology that takes account of the nature of the station.
8. Sites and Stations can be combined in a two-level hierarchy, called SiteStation, because the nature of the relationship between stations, sites, initiatives, and sensors are different depending on the typologies involved[1].
9. SiteStation has a defined locality in two-dimensional space, together with an elevation
10. The relationship between Stations, Sites, and Sensors (implying the Variables being measured) are captured in a construct referred to as a SensorDeployment. SensorDeployment takes care of the sensor deployment at a given station or site, both in terms of temporal coverage, and in terms of offset from a StationSite elevation. This accommodates the common scenario where multiple Sensors of the same type are deployed in the same Station at different elevations, or where a sensor measures multiple elevations from the same position[2].
11. Each SensorDeployment references a Sensor and a SiteStation, but also needs to define a Protocol, since the same Sensor can be used in different SiteStations with different Protocols, or different Sensors can be used with the same Protocol.

[1] As an example, for remotely sensed satellite photography there is no need to describe the sensors for each of the virtual stations covered by a scene, since it is impractical and the sensors remain exactly the same within a given temporal coverage of that sensor.
[2] For example ADCP

1. SensorDeployment can also make provision for recording the instance of a sensor, for example by recording the serial number or similar unique identifier for an individual sensor.

This allows sensor calibration, and tracing of individual sensors that may move between stations over time - for example after being repaired.

2. Each SensorDeployment can also reference a Calibration, and a Calibration can be used for more than one SensorDeployment.
3. Sensors and Instruments are to some extent interchangeably used by the community, and in some cases an Instrument can be composed of more than one sensor. This is accounted for in the arrangement for a hierarchy of Sensors, and the Sensor hierarchy could potentially have a typology to describe levels of detail.
4. Protocols and Sensors are both linked to Essential or Standard Variables. The same Variable can be measured with different Sensors and/ or Protocols in any valid combination. Variables preferably have authoritative descriptions in Ontologies or in external metadata defining the variable.
5. Some Variables are primary measurements (for example the voltage from a thermocouple) that is transformed to a temperature value using a Protocol.
6. Some Variables are phenomena (primary observation data), while others are called offerings (derived observation data, usually from statistical aggregation or filtering)[1]. The relationship between primary and derived variables can be stored in the same Variables structure with a parent-child relationship.

A.2 Special Case: Remotely Sensed Datasets

Remotely sensed datasets, whether from satellite, aerial platforms, or any other vehicle, share some characteristics that result in a slightly different conceptual model. The main differences are:

1. Satellite observation, in general, is accomplished within a Mission, and the platform (satellite) used for observation may contain one or more Instruments. The Mission corresponds to an Initiative in the sense of other networks, but the metadata and instrument inventories associated with a mission is generally maintained externally, in services operated by OSCAR and CEOS.
2. The Platform or Site recording the observations are distinct from the observed Feature or Station, and these stations are virtual, as explained above: usually corresponding to the individual pixels associated with the satellite scene or mosaic.

A.3 Vocabularies

[1] Many data loggers can record a mixture of these. It is common to find derived values such as maximum and minimum temperature or averages in data logger outputs in addition to the near-continuous temperature recorded.

In some instances, lookup values and lists need to be obtained from Vocabularies and Registries. These services can be classified generically as follows:

1. Formal ontologies and vocabulary services, maintained by the community or a reputable authority. Examples include services such as EnvThes and BODC.

2. Registry and catalogue values, such as DataCite DOIs for datasets, code, and protocols, re3data registry for data centres and repositories, and similar. These services provide a persistent identifier (PID) for each concept or thing that it references.
3. Informal vocabularies, temporarily hosted by SAEON, that are either
    a. Official lists for which an authoritative service does not exist (for example a list of local authorities in South Africa or Namibia)
    b. Or lists that are context-specific, but for which no formal vocabulary exists (for example a list of instruments and sensors deployed by a consortium member).
    c. Either of the above may at some point be standardised, and need to be configurable in respect of its source and API specification.

One of the major challenges faced by systems developers in respect of vocabulary and registry services is the large range of API schema and syntax employed by services. It will be a goal (but not a formal deliverable) to investigate measures whereby the number of supported API schema can be reduced via simple brokering operations, and potentially establish such a brokering registry.

Annexure B: Actors, Systems, Services, and Stakeholders

B.1 User Roles

1. Normal (casual) user (no registration or login): These users will typically attempt to find data associated with a given variable, protocol, and/ or network, and often require data for a specific temporal and spatial coverage.
2. Registered users
    a. Default - Such users are registered solely for the purpose of unique identification, and as such can store preferences, provide feedback, and configure views of the system tailored to a specific end use.
    b. Accredited - Users for which more is known (for example associated with an accredited network or institution), and who can be relied on to assist with review or endorsement of new content.
3. Administrator for a network or initiative: Able to confirm accredited user accounts, and configure the (automated) links to for contribution of data, metadata, and services.
4. Administrator/ Curator for the SEACRIFOG RDI - all privileges, systems deployment and configuration.

B.2 Stakeholders

The following stakeholder groups need to be served by the SEACRIFOG Data Infrastructure, each having slightly different objectives for their participation:

1. Researchers: typically interested in understanding scope of observation, and optionally obtaining metadata or data.
2. Research Institutions and Infrastructures (Research Councils, Universities, Institutes, …): interested in contributing and disseminating data and services that are in the public domain, or is made available within the consortium.
3. Government Infrastructures : contributing and finding data.
4. Research Infrastructure Management (SASSCAL, SAEON, ...): Collaboration and synergy from shared data and resources.
5. Private Sector Institutions (Insurance, Investors, …): Provide more detailed evidence as a measure to reduction in hazardous event and investment risks.
6. Policy and Decision-makers, Planners: Societal benefit from value added to evidence.
7. General Public: largely unpredictable, but assumed to be largely as a source of scientific data.

B.3 Systems

The following types of systems and contributions have to be integrated into a comprehensive SEACRIFOG data infrastructure. There are precedents for such an infrastructure, as shown in Figure B.3.1 - based on Kutsch et al. (2018[1]).

- Individual data centres and institutions contribute to an initiative or network. Such arrangements are shown in the lower part of Figure B.3.1, and accommodates two situations: a tightly constrained network, such as ICOS, in which the protocols, sensors, and variables under observation is specified, and a loosely managed network (such as SAEON) where this is not the case .
- These networks may perform analysis, data processing and quality assurance, and any number of additional functions prior to offering data to end users. These datasets should be ‘published’, and be provided with a citable PID.
- The networks may also be offering data and/ or metadata to higher-level aggregating networks, such as FluxNet. These, in turn, also offer datasets and services to end users in an easily citable and persistently identifiable way.

It is assumed that use of data and services will be tracked so that credit (citation and usage statistics) can be correctly attributed.
Reference to be confirmed

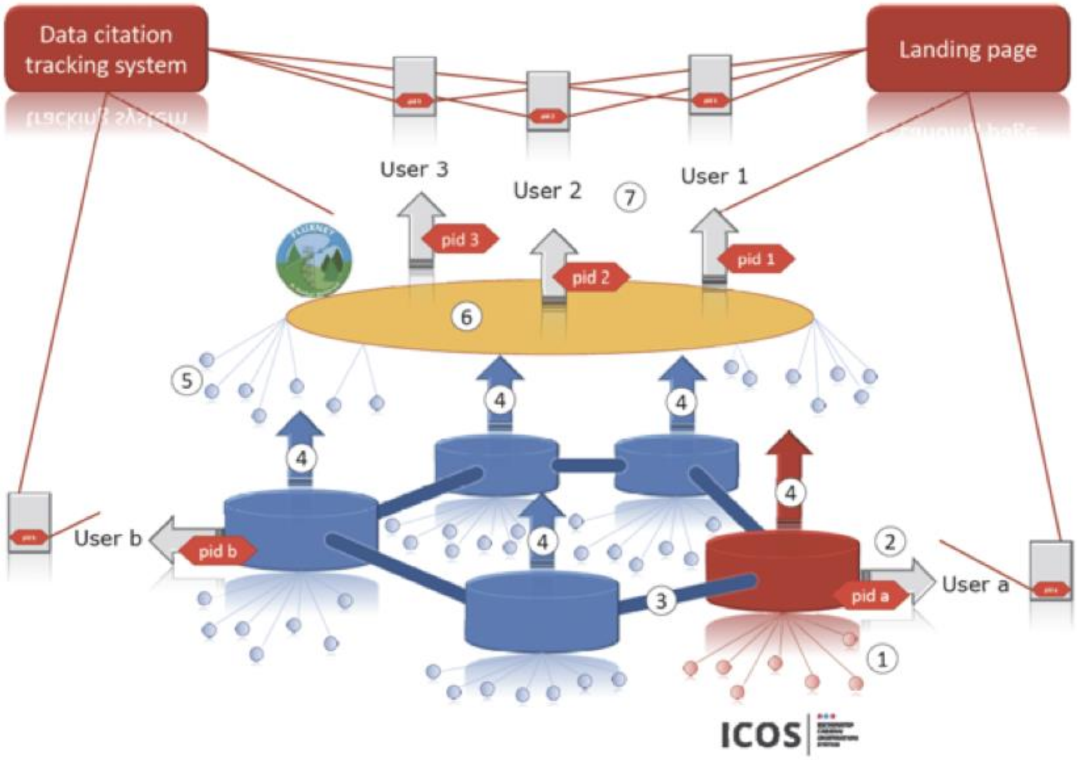


Figure B.3.1. Fluxnet as example for the structure of a typical global data infrastructure.

It should be clear that initiatives such as GEO and ILTER more or less mirror this 'system of systems' approach.

The table indicates the typical data and hosting services associated with each group, as well as the typical method of metadata aggregation.

#	Systems Description	Data Hosting and Services	Metadata Aggregation
1	Individual SItes	Internal	Internal
2	Standalone Data Infrastructure (ICOS, SASSCAL)	Internal	Internal
3, 4	Distributed/ Integrated RDI (SAEON, …)	Yes, Links	Internal and External
6	Regional Thematic Network of Networks (SEACRIFOG)	Links	External Only
6	Aggregators (GEOSS, Fluxnet, GBIF) (Systems of Systems, Networks of Networks, …)	Links, Aggregation	External Only

B.4 Services

-

1. Vocabulary Services (EOL, EnvThes, …): These are required to properly reference and tag metadata, and to properly describe content of datasets.
2. Brokering Services (GEOSS DAB, …): Mechanisms for ensuring crosswalks between standardised data and metadata services, and correction of non-standard datasets and metadata services.
3. Data Services: standardised data services for specific data families (Mirtl et al., 2019).
4. Discovery/ Harvesting Services: Catalogues of metadata harvesting endpoints, used by the SEACRIFOG data infrastructure to automate aggregation of metadata to the maximum possible extent.
5. Visualisation/ Exploration Services: Providing data exploration and visualisation tools to end users, enabling them to understand the nature and scope of data to be sourced and/ or downloaded.

Workflow and Processing Services: Likely future extensions, but not in the immediate future, except for specific use cases (carbon flux pre-processing as an example).

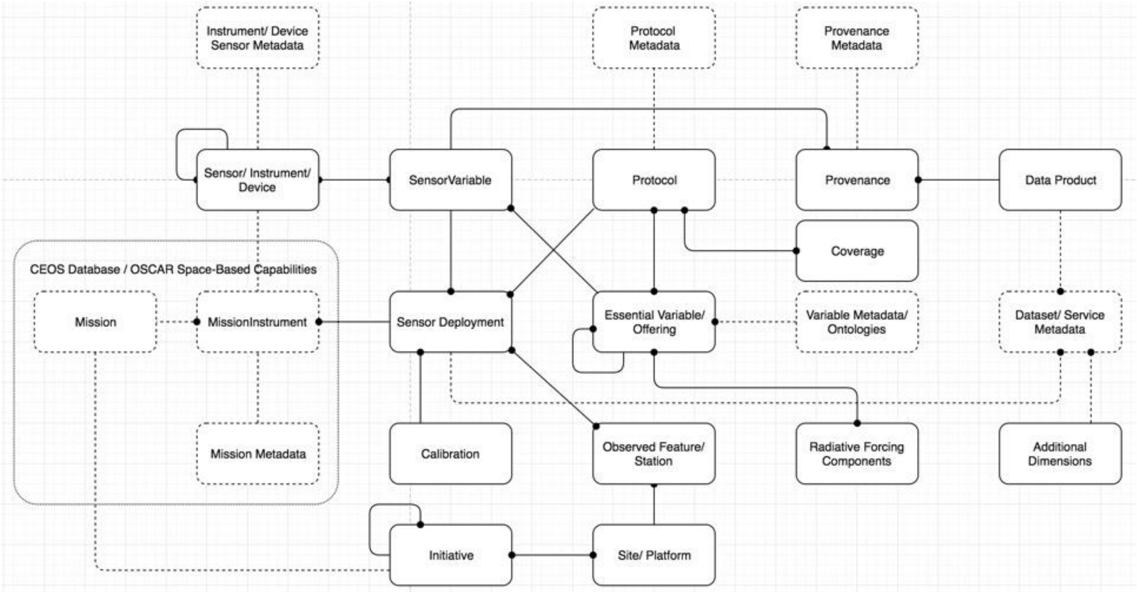
Annexure C: Information Model

Based on the information presented in Annexure A, we propose an information model that will be used as the basis for database and API service design. The model comprises three parts:

1. Entity-Relationship Model
2. Data Flow Pipelines
3. External Resources, Linked Open Data, Vocabularies and Name Services

C1: Entity-Relationship Model

Figure C.1.1 shows the entities and relationships between entities, based on the narrative description in Annexure A.



Instrument/ Device Sensor Metadata
Protocol Metadata
Provenance Metadata
Sensor/ Instrument/ Device
SensorVariable
Protocol
Provenance
Data Product
Coverage
CEOS Database / OSCAR Space-Based Capabilities
Mission
MissionInstrument
Sensor Deployment
Essential Variable/ Offering
Variable Metadata/ Ontologies
Dataset/ Service Metadata
Mission Metadata
Calibration
Observed Feature/ Station
Radiative Forcing Components
Additional Dimensions
Initiative
Site/ Platform

Figure C.1.1 - Entity-Relationship Model. Dots denote 'many' sides in a relationship. Broken lines: attributes are external to the system, and preferably denoted by metadata found through a PID.

Annexure D: Use Cases

D.1 The Research Output Life Cycle and Value Chain

Scientific data or evidence progresses through a value chain that is associated with a life cycle. In each of the life cycle stages, or value chain stages, the requirements for metadata are slightly different (but additive). Our implementation needs to take research output state into account, and make provision for use cases associated with these.

Relative Value Chain State	**Digital/ Physical**	**Type of Metadata Required**	**Typical Purpose**	**Part of Core Metadata?**
Identification/ Archiving	Yes/ Yes	PID only, fixity	Reliable reference, preservation	Yes
Citation	Yes/ Yes	Ownership, Authorship, ...	Reliable attribution	Yes
Publication Ready	Yes/ ?	Data dictionary, Methodology, Protocols, License, ...	Reproducibility and Re-usability, Discoverability	Yes, but references external protocols, ...
Analysis Ready	Yes/ ?	Contextual metadata, external references and multiple end-uses, provenance graphs, interoperable services	Context-specific application and semantics,	Partly, but references one or more format or discipline specific external metadata records
Indicator Ready	Yes/ ?	End user annotations, fitness for use assessments	Bindings to end uses and workflows, role definitions	External to metadata, references PIDs

D.2 Generic Use Case: Published Resources

Research outputs

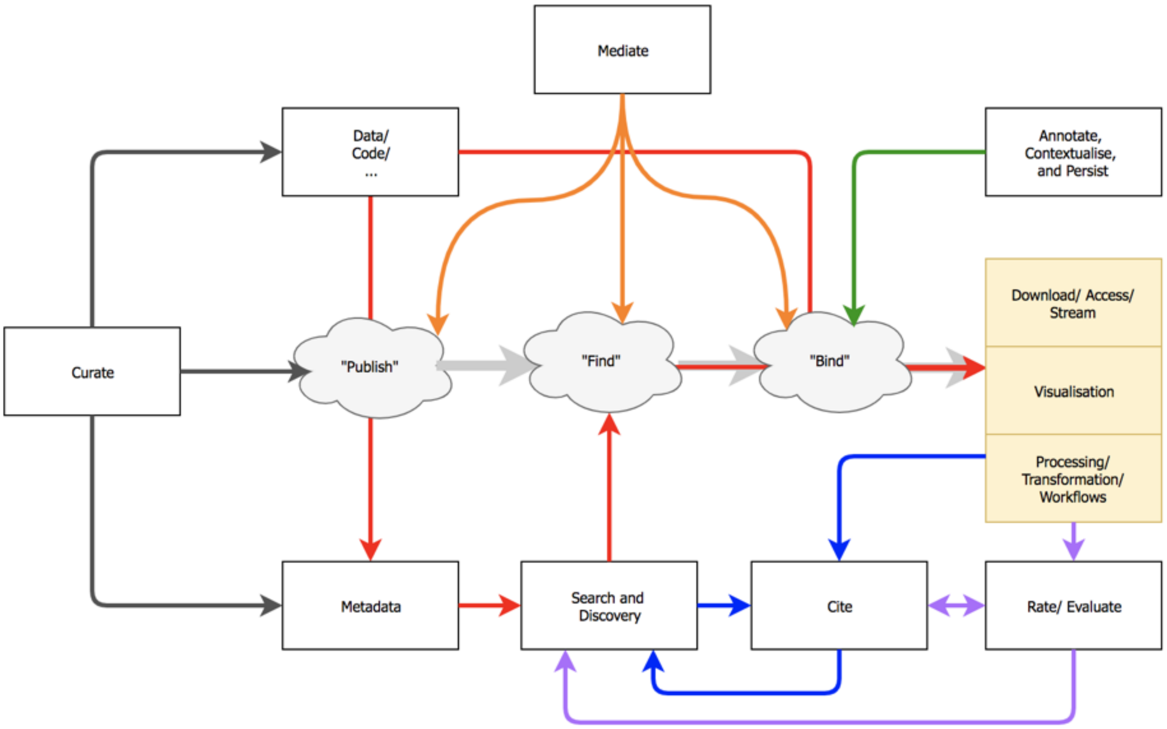


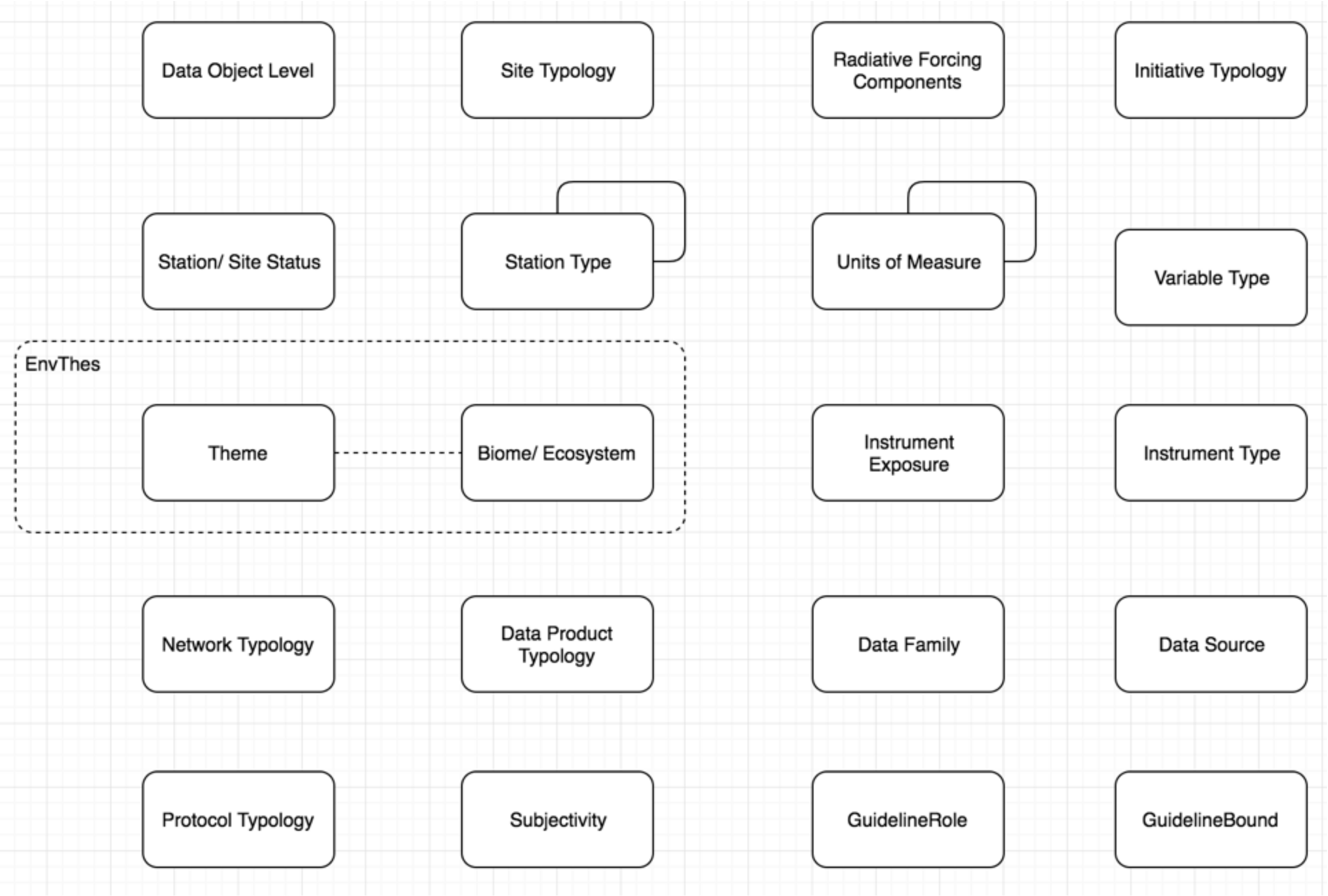
Figure D.2.1 - Scope of the Generic Research Output Publication Use Case

D.3 Contextual Metadata and Use Cases

#	Use Case	Description	Reference/ Examples
SSI	Site/Station Inventory	Standardised Filters Standardised Categories Standardised Views: • Maps, Lists, Charts, Timelines, Detail View Standardised Export of Data and Synthesis Reports	WMO OSCAR SASSCAL Data Portal Tableau DEIMS SAEON ObsDB
IPN	Initiatives/ Projects/ Networks	Standardised Filters Standardised Categories Standardised Views: • Maps, Lists, Charts, Timelines, Detail View Standardised Export of Data and Synthesis Reports	WMO OSCAR SASSCAL Data Portal Tableau DEIMS
DAT	Data Inventory	Standardised Filters Standardised Categories Standardised Views: • Maps, Lists, Charts, Timelines, Detail View Standardised Export of Data and Synthesis Reports	GEOSS Portal SASSCAL Data Portal SAEON Discovery
VAR	Browsing by Variable	Standardised Filters Standardised Views: • Maps, Lists, Charts, Timelines, Detail View Standardised Export of Data and Synthesis Reports	SEACRIFOG-Tool
CUD	Editing	Edit, add, and delete entity instances Edit, add, delete relationships between entities	SEACRIFOG-Tool DEIMS

Annexure E: Specifications and Standards

E.1 Vocabularies and Registries



Annexure F - Description of ICOS Carbon Portal

The ICOS Carbon Portal (CP), https://www.icos-cp.eu, has been developed on the basis of the Carbon Portal white paper written in 2012. The white paper document is available on ICOS Alfresco. CP was developed as the one stop shop for all ICOS data products, i.e.: “a virtual data centre where ICOS data can be discovered, accessed and visualized, and where users can also deposit data products based on ICOS data”. The data system is integrated with a metadata system that describes the data and its provenance. Furthermore, CP provides capabilities for advanced web based services that provide researchers, general public and decision makers with useful higher level products based on ICOS data.

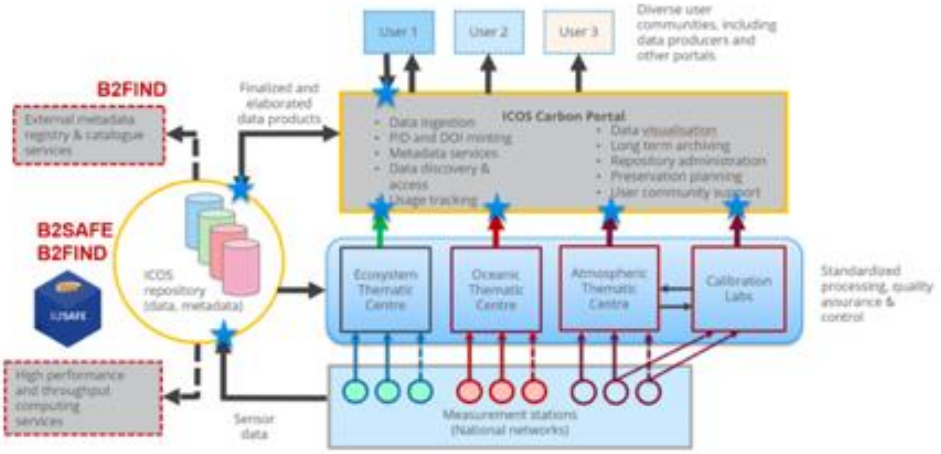

User 1
User 2
User 3
Diverse user communities, including data producers and other portals
B2FIND
External metadata registry & catalogue services
Finalized and elaborated data products
ICOS Carbon Portal
Data ingestion
PID and DOI minting
Metadata services
Data discovery & access
Usage tracking
Data visualisation
Long term archiving
Repository administration
Preservation planning
User community support
B2SAFE
B2FIND
B2SAFE
ICOS repository (data, metadata)
Ecosystem Thematic Centre
Oceanic Thematic Centre
Atmospheric Thematic Centre
Calibration Labs
Standardized processing, quality assurance & control
High performance and throughput computing services
Sensor data
Measurement stations (National networks)

Figure 5.3.1 Simplified data flow within ICOS Research Infrastructure

The basic principles of the CP are data security, long term archiving through a trusted repository, enforcing the data policy and user-friendly operation. As a service to the data providers, CP will keep track of the use of the data and its citation. By default, the CP supports machine to machine access to data and metadata. For human users CP adds user friendly web services on top for data discovery and access.

All ICOS data is open data, licensed under a Creative Commons International 4.0 Attribution (CC4BY).

All developments by CP are open source and are based on open source libraries and tools. The sources are licensed under GPL and are available from https://github.com/ICOS-Carbon-Portal. The backend skeleton Portal metadata and data services are generic and fully customizable and can be adapted to any project or look and feel. Landing pages can be stylized to an identity that is coupled to the data object type, so can depend on theme or data provider. All services in backend and frontend are dockerized and are fully scalable.

Carbon Portal Data Ingestion

The philosophy of CP is to treat all data objects equal and preserve the complete integrity of all data objects, so the actual data is never touched or changed up to the bit level. This goes for all data levels, i.e. from raw data, NRT data, final quality controlled data up to elaborated data products. CP strives for the maximum granularity of Data Objects.

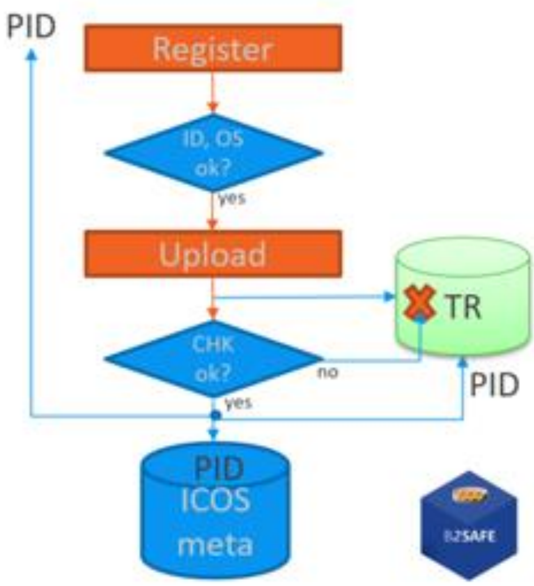


Figure 5.3.1 A simplified schematic of the ICOS central data ingest that enables robust, persistent identification and transparent and secure data ingest directly into the trusted repository (see also https://github.com/ICOS-Carbon-Portal/meta)

Before ingestion CP requires the uploader to calculate the SHA256 checksum of the data object. All ingestion data transport uses standard http(s) put and get methods, and can be invoked by for example using the curl program. In the first stage of ingestion the uploader informs through a small metadata packet in JSON format of the object specification and the checksum of the data object together with some minimal provenance metadata that informs on the uploader, the spatial and/or temporal coverage that the data relates to for as far as applicable and depending of the object specification also on other important information like station, measurement level and instrument ID. Only objects with a known and registered Object Specification type are accepted. After successfully registering in this first step the user can start uploading the data object. While the uploader streams the data to CP, the data is forked and streamed at the same time to the B2SAFE trusted repository.

When the object specification defines the data format of the file, a check is performed after the complete upload, to check the compliance to the data format and even possibly the validity of the data columns and spatial and temporal coverage as contained in the data file. Any deviation from the definition or prescribed metadata results in refusal of the file and abortion of the ingestion. The successful parsing of the data for text files also results also in the generation of binary CP-internal representations of the data that are used for the visualization of time series in the data preview.

After upload completion, the checksum of the upload is compared with the registered checksum and when ok, a handle PID is minted for the data object and returned to the user. The metadata from the metadata packet is then added to the metadata repository and enriched with information on the PID, the checksum and other Object Specification dependent metadata. The suffix of the data object PID consists of the first 18 characters of the checksum of the data object and is thus unique for the data object. Later the PID suffix can at any time be compared with the SHA256 checksum of the data object to ensure that the data is up to the bit and exact copy of the original data object.

The CP metadata system

The metadata that accompanies the data objects is maintained in a versioned so called RDF triple store, following the Web 3.0, linked open data approach. The database can be queried using an open SparQL endpoint at https://meta.icos-cp.eu/sparql. The metadata store fully supports date versioning and data collections. It is machine actionable through standard http(s) protocol. The metadata store is fully described by the underlying ontology, that again itself is defined in RDF through the OWL language.

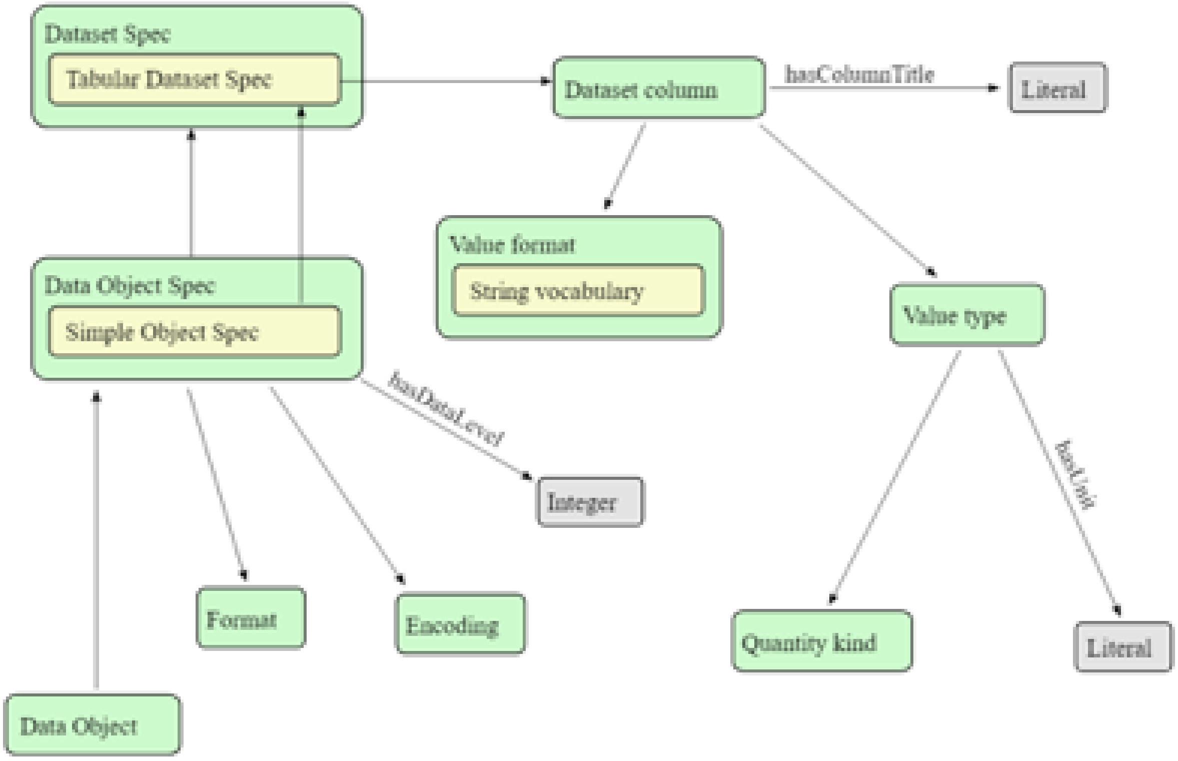


Figure 5.3.2 The simplest ICOS data object model for time series (see also https://github.com/ICOS-Carbon-Portal/meta)

The design of the metadata system is fully configurable to act with a single or multiple portal front ends using a single or multiple metadata stores. This means that for example multiple infrastructures can have their own differently styled data portal and use one single metadata store, or that one infrastructure has one portal that uses several external metadata stores, or that several infrastructures use one common portal that relies on a set of federated metadata stores, one per infrastructure. All completely transparent to the outside user.

The ICOS CP metadata store is for example shared with the Swedish SITES national infrastructure that has its own dedicated and styled portal, while ICOS Sweden is just using the metadata store backend and data is served through the ICOS CP portal.

The metadata system supports versioning of data, dynamically growing data and collections.

CP data discovery

The main entry point for data discovery for humans is https://data.icos-cp.eu. Here a set of filters can be easily set to filter to the data sets that the user actually is looking for. The list of data objects that fulfils the set of filters is display dynamically. Changing the filters also dynamically updates the remaining options for the other filters that comply with the other filter settings. Filters can and will be added, removed and applied incrementally. From the results page the user can view the most relevant information on the data object and/or drill down to the data object landing page for all relevant metadata. Most data objects can be previewed, see data visualization. Most data objects can also be added to the user's data cart for easy download, see data access.

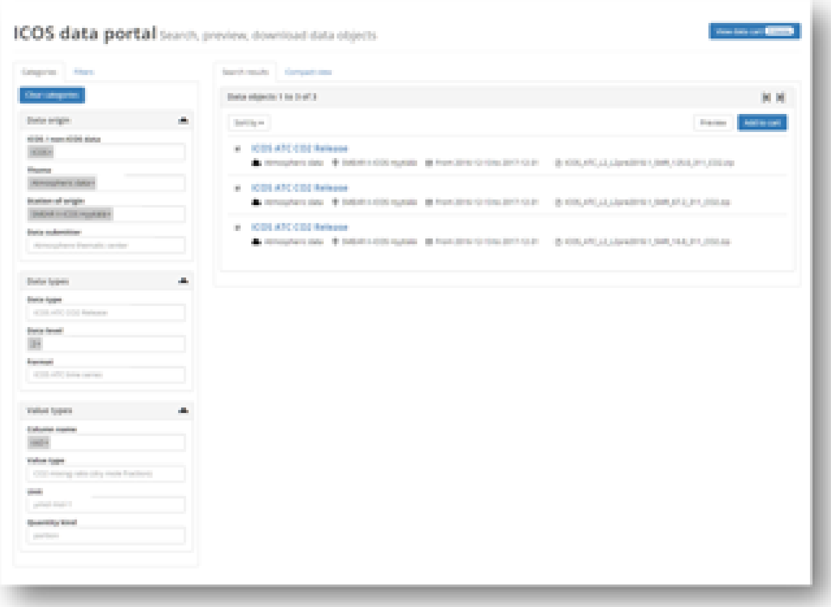


Figure 5.3.3 Example of the ICOS Data Portal Search Results

CP data access

Data access is provided through the PID (or DOI) of the data objects. Resolving this PID through the handle or doi system leads normally to a landing page that contains a link to the data object(s). In case of non-ICOS data objects this link can point to another data portal due to data license restrictions. Raw data objects are currently also not directly downloadable but require contact with the relevant thematic center.

The data discovery tool allows to add selected data objects to the user's data cart from where the collected objects can be downloaded in one batch into a single zip archive.

All data downloads are logged and ICOS data has a data license check implemented before the download to inform the user of the ICOS CC4BY license and its implications. Users can easily track the number of downloads per dataset, country, station, contributor and/or theme, categorized by time and country of the download.

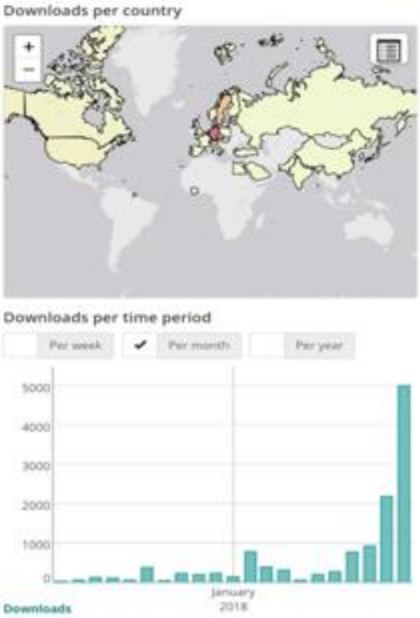


Figure 5.3.4 Example of Download Metrics

CP data visualization

Time series and spatial data sets can be previewed directly from the data portal in the search results and in the data cart for a quick check just before download. The visualization supports the overlaying to append time series for a single column and the overlaying of overlapping time series of from different stations, instruments and/or measurement heights. A fully interactive map or chart is shown that can be reproduced in any web portal or page in an iframe by using the provided link.

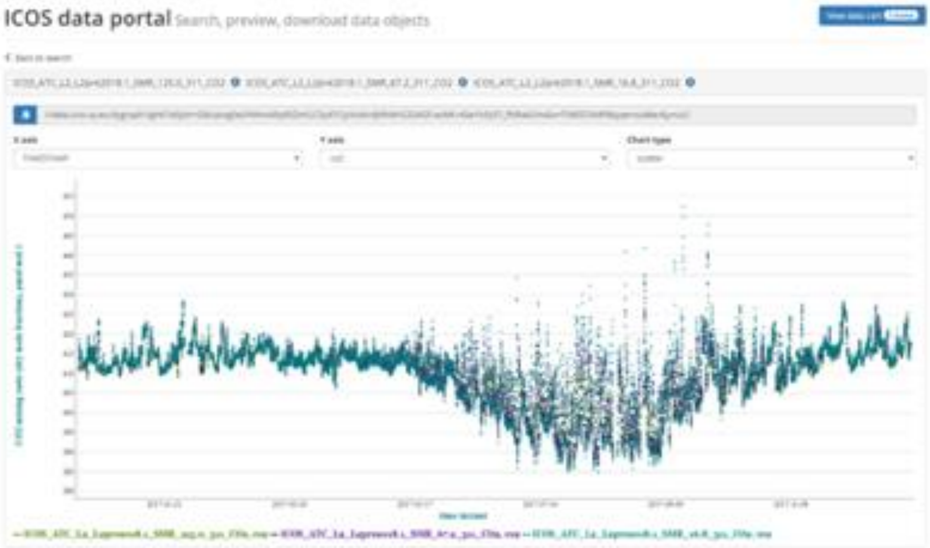


Figure 5.3.5 Example of Data Visualisation

An interactive tool allows to link atmospheric footprint data with modelled and measured time series at https://stilt.icos-cp.eu/viewer.

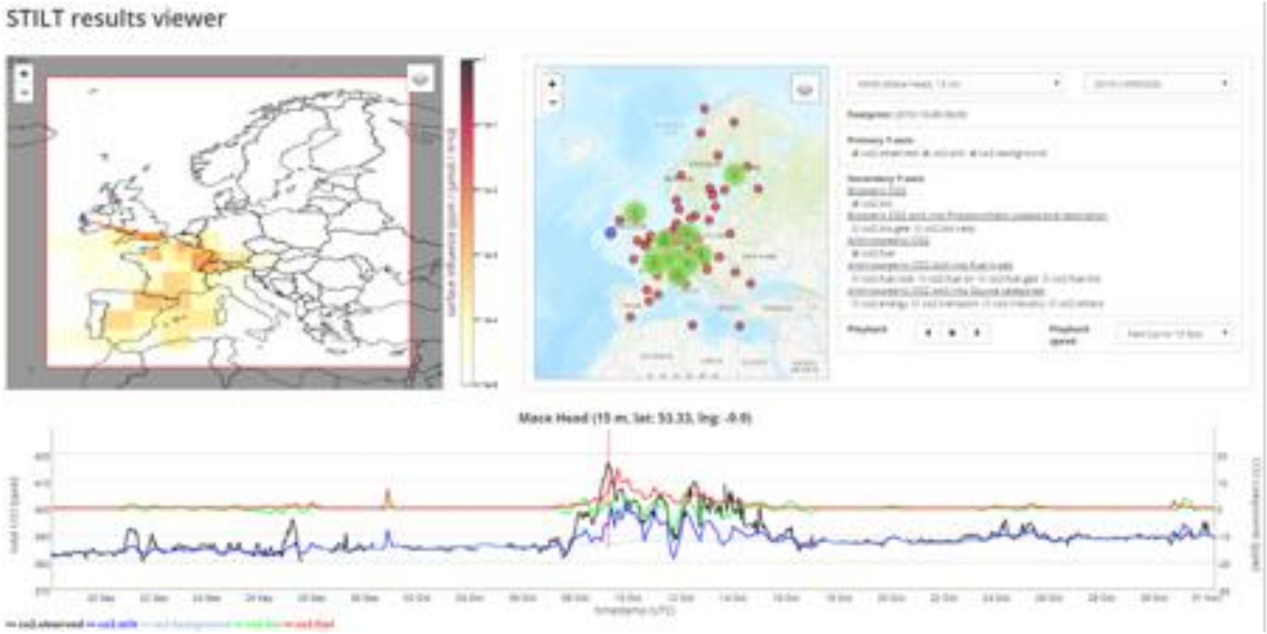
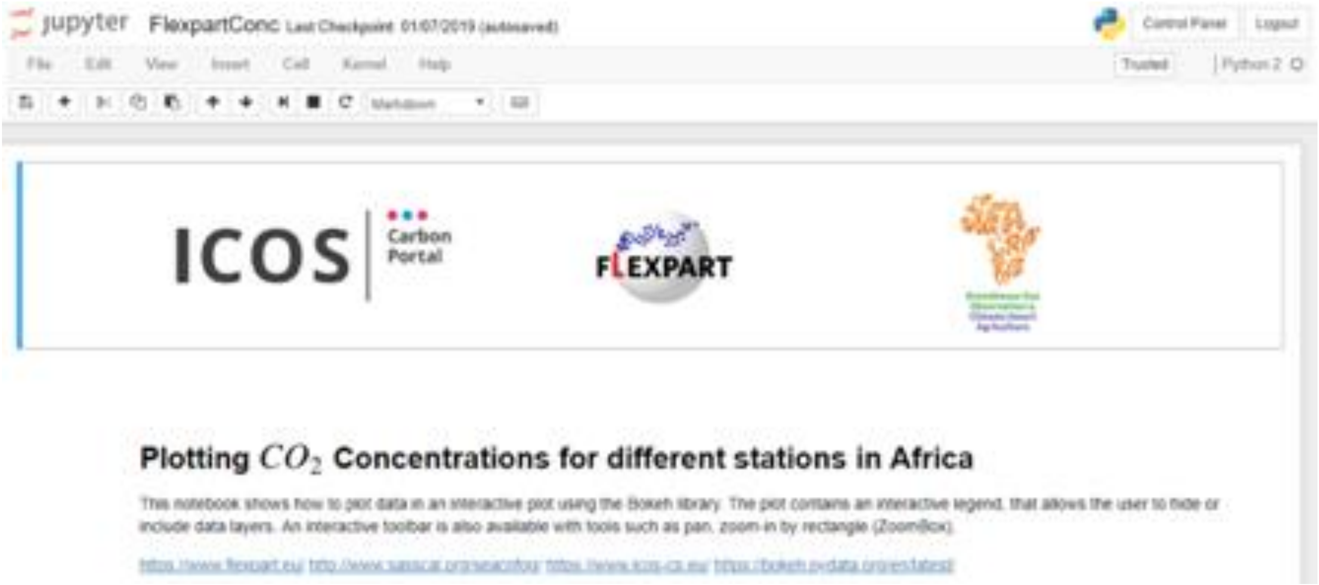


Figure 5.3.6 STILT Results Viewer

Virtual Research Environments

Carbon Portal offers scientists access to Jupyter notebooks that give direct access to the data objects stored at the ICOS CP through either an API or a library of python functions. These notebooks are either run directly on the CP servers or are operated as virtual machines in the cloud, making use of a changeable amount of memory and cores for even the most demanding analyses. Notebooks can be shared among colleagues for collaborative analysis of for example model ensembles, sharing common input and output data and modelling resources.

For less advanced users that would have difficulties with programming, CP plans to provide interactive tools that give access to powerful models and data analysis tools. One example is the Stilt footprint calculator that allows users to perform footprint calculations using the Stilt Lagrangian footprint model for any point in Europe and period within the provided range in space and time. The results are immediately after calculation available in the Stilt results viewer and for download, together with the forward prognosis of CO_2 concentrations at the chosen receptor point.



Plotting CO_2 Concentrations for different stations in Africa

This notebook shows how to plot data in an interactive plot using the Bokeh library. The plot contains an interactive legend, that allows the user to hide or include data layers. An interactive toolbar is also available with tools such as pan, zoom-in by rectangle (ZoomBox).

https://www.flexpart.eu/ http://www.sasscal.org/seacrifog/ https://www.icos-cp.eu/ https://bokeh.pydata.org/en/latest/

Figure 5.3.7 Incorporating ICOS Data into Jupyter Notebook