D3.12 - In Situ Data Hub II

WP3 – Large Scale Demonstrators

Authors: Stavros Tekes, Yanis Nasiopoulos

Date: 28.07.20

Full Title	**Promoting the international competitiveness of European Remote Sensing companies through cross-cluster collaboration**		
Grant Agreement No	824478	**Acronym**	PARSEC
Start date	1st May 2019	**Duration**	30 months
EU Project Officer	Milena Stoyanova		
Project Coordinator	Emmanuel Pajot (EARSC)		
Date of Delivery	**Contractual** 31.07.2020	**Actual**	28.07.2020
Nature	Other	**Dissemination Level**	Public
Lead Beneficiary	DRAXIS		
Lead Author	Stavros Tekes	**Email**	stavros@draxis.gr
Other authors	Giannis Nasiopoulos		
Reviewer(s)	Dimitar Misev (RASDAMAN)		
Keywords	API Service, EO, dockers, containers, database, architecture, hardware		

Document History

Version	Issue date	Stage	Changes	Contributor
1.0	15.07.2020	Draft	First draft	DRAXIS
1.1	27.07.2020	Final	Content Corrections	DRAXIS
---	---	---	---	---

Table of Contents

List of Acronyms

API	Application Programming Interface
RESTful API	A web based HTTP Application Programming Interface
B-tree	Data structure that maintains sorted data and allows fast searches
C3S	Copernicus Climate Change Service
CAMS	Copernicus Atmosphere Monitoring Service
CSV	Comma Separated Values
Datacube	Multi-dimensional Array of values
EO	Earth Observation
FTP	File Transfer Protocol
GDAL	Geospatial Data Abstraction Library (software library for handling geospatial data)
GIST	Generalized Search Tree (allows fast searches)
GIN	Generalized Inverted Index (allows fast searches)
GIS	Geographic Information System
Hash	Data structure that maintains sorted data and allows fast searches
HTTP	Hyper Text Transfer Protocol
IoT	Internet of Things
JSON	JavaScript Object Notation
Kong	Software to securely manage communication between clients and microservices
MongoDB	Cross-platform Document Oriented Database
NGINX	Software that Accelerates Content & Application Delivery
NetCDF	Network Common Data Form
PHP	General-purpose Programming Language
PostgreSQL	Advanced Open-source Database
PostGIS	Software program that adds support for geographic objects to the PostgreSQL
RDBMS	Relational Database Management System
VM	Virtual Machine (multiple PC-like machines running under one computer)
WRF	Weather Research and Forecasting Model
XML	Extensible Markup Language

Executive Summary

This report is an updated version of "D3.4-In Situ Data Hub I", and it includes the same content as in D3.4 with additional information on the new data sources included in the second stage of enriching the In-Situ Data Hub. Overall, this report "D3.12-In Situ Data Hub II", has general improvements in almost all chapters, but the essential chapters updated are summarized below.

- *Chapter 2-refined content presenting the data hub and its functionality*
- *Chapter 3- additional data sources briefly presented and new screenshots of the hub provided*
- *Chapter 4-updated infrastructure contains additional information and figures on weather and climate services (subchapter 4.4)*

The report documents the PARSEC In Situ Data Hub which acts as a repository hosting geo-referenced field observations coming from sensor networks, independent IoT devices, citizen observatories and other online structured and unstructured data sources available.

The In Situ Data Hub provides access to real and past time data through pre-populated measurements derived from institutional in situ datasets hosted by major stakeholders responsible for monitoring of climate, environment and different sectors of the economy (e.g. food security). Additionally the hub collects open sensor measurements available on the internet either in structured formats (e.g. XML or JSON services, databases, csv or excel files) or as unstructured content such as sensor measurements in web page tables in a structured way.

Initially the overall platform design and architecture of the data hub is presented, followed by the hardware and software requirements for the deployment of the hub.

1 Introduction

The In Situ Data Hub is a multi-source repository with automated discovery, retrieval, harmonisation and transformation services. Geospatial data from a wide range of sources is acquired and transformed into time series datacubes (multi-dimensional array of values), which are made available for use in a combined, uniform, analytical environment.

The acquired datasets are made publicly available for access and download. The retrieved information can be used for various purposes, such as

i) the production of value-added products, particularly in combination with satellite data
ii) validation of parameters extracted with EO techniques and algorithms
iii) training neural networks
iv) calibration or visualisation purposes,
v) discovering insights based on big data analysis that can combine observations, time and geolocation information, etc
vi) services of operational monitoring, warning and reporting to public institutions or businesses.

Currently the In Situ Data Hub provides air quality data for approximately 10 pollutants originating from:

- Official ground-based monitoring stations, with approximately 10651 daily measurements for 12000 stations and provide information as of August 2019 and onwards.
- Low cost sensors, with approximately 27394 measurements per minute, per hour and daily for 30000 sensors and provide information as of September 2019 and onwards.
- CAMS (Copernicus Atmosphere Monitoring Service) with a 3 days forecast and 10km resolution providing continuous past and present data and information on atmospheric composition.

Additionally the hub delivers high-resolution weather and climatology information, covering the wider area of Europe, including region of North Africa and the Middle East.

2 In Situ Data Hub

Currently millions of sophisticated embedded measuring devices are being networked together providing an extensive monitoring system for the earth. Remote sensing observations and in situ measurements are being assimilated offering systematic capabilities to develop geophysical and atmospheric information products and services for use in models at relevant scales. Intelligent sensor-webs, based on converging technologies of micro-sensors, computers, and wireless telecommunications facilitate and support critical activities such as the monitoring of remote environments, risk assessment & hazard mapping, and renewable resource information management, providing an integrated earth sensing framework.

Figure 1- Cloud of in-situ measuring devices

For the most part, the available sensor networks are streaming the produced measurements making the information publicly available for individuals, companies and organization to consume and integrate in different services and products. Nonetheless, despite the vast availability of information, one of the major problems of the in situ data in general, is that the information is decentralized. Practically every new project or initiative creates a small information hub to serve its needs. This clustering, leads to data fragmentation, and mostly inaccessible and unattainable information sources. Attempting to portray the available data sources on a data map, would produce a chaotic network with multiple interconnected hubs.

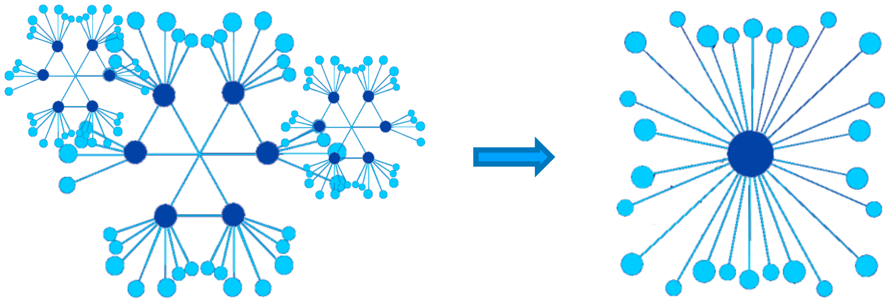

Figure 2- From clustered data hubs to centralized information structures

The challenge of the In Situ Data Hub is to collect and centralize the available data, and make this information readily available to the user through a convenient and friendly user environment. In order to achieve this, the first step is to unify the available data sources such that to create a single access data point, so the data will cease being decentralized into small data hubs and will fall under the big In Situ Data hub umbrella.

Technically the In Situ Data Hub is an automated data acquisition and pre-processing system, acting as a multi-source repository of field measurements, and data derived from instruments located directly at the point of interest and in contact with the subject of interest. The hub is offering automated discovery, retrieval, harmonisation and transformation services, collecting geospatial data from a wide range of sources acquired from sensor networks, independent IoT devices, citizen observatories, and many others. The acquired information is transformed into time series datacubes, and where necessary, ingested into the hub's own database, or directly made available for use.

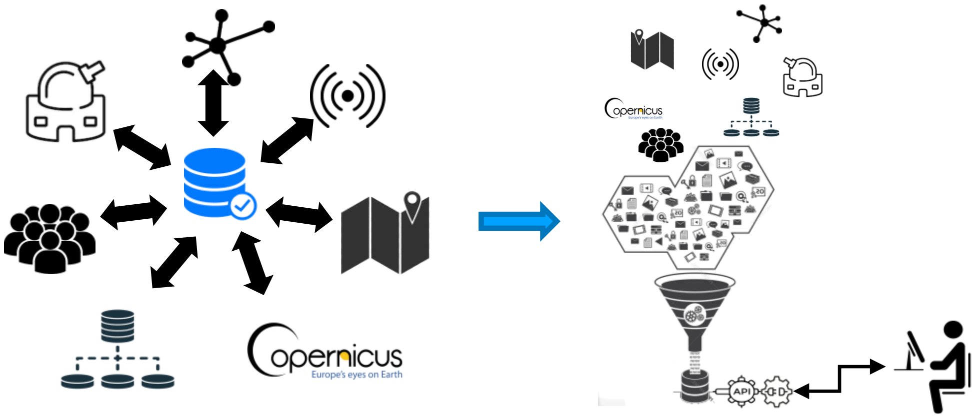

Figure 3 - Acquired information is stored and made readily available through a uniform and user friendly environment

All the acquired information is provided effortlessly to the user through a combined, uniform, analytical environment over the exposed API of the hub, while the documented APIs allow any in situ requester to view and, in some occasions, to create or modify the data of the hub. The main objective of the hub is to offer SMEs, a centralised, expedite access to valuable resources that is often key to realising value-added EO (Earth Observation) services.

3 Main Features

The In Situ Data Hub is an automated data acquisition and pre-processing system, which exposes API services for data provision. Geospatial data from a wide range of sources are acquired (sensor networks, independent IoT devices, citizen observatories, etc.), transformed into time series, and ingested into the database, ready for use in a combined, uniform, analytical environment.

Although the in situ data hub has been designed around the needs of the manager, in reality it is oriented for programmers. More specifically the hub is offering to the manager a user-friendly environment, with intuitive search and filtering options, with small descriptions of each information source available, with the ability of dataset-preview and details such as latest dataset update, data format and many more. At the same time virtually any programmer possessing basic programming skills, can read and interpret the fully documented APIs, the in situ data hub exposes. Through the environment the programmer can acquire all the necessary documentation on the methods to access all the available data sources. All information is thematically grouped offering examples of dataset requests, filters and several other options to facilitate the programmer.

Currently the hub integrates in-situ air quality (AQ) data, from official monitoring stations, low-cost sensors (IoT-Internet of Things networks) and air quality forecasts from the Copernicus Atmosphere Monitoring Service (CAMS).

Additionally the hub delivers high-resolution weather forecasts based on the Weather Research and Forecasting model, and climatology information, both covering the wider area of Europe, including region of North Africa and the Middle East.

The In Situ Data Hub can be accessed at https://insitu-datahub.draxis.gr/
An indicative screenshot of the interface is presented below.

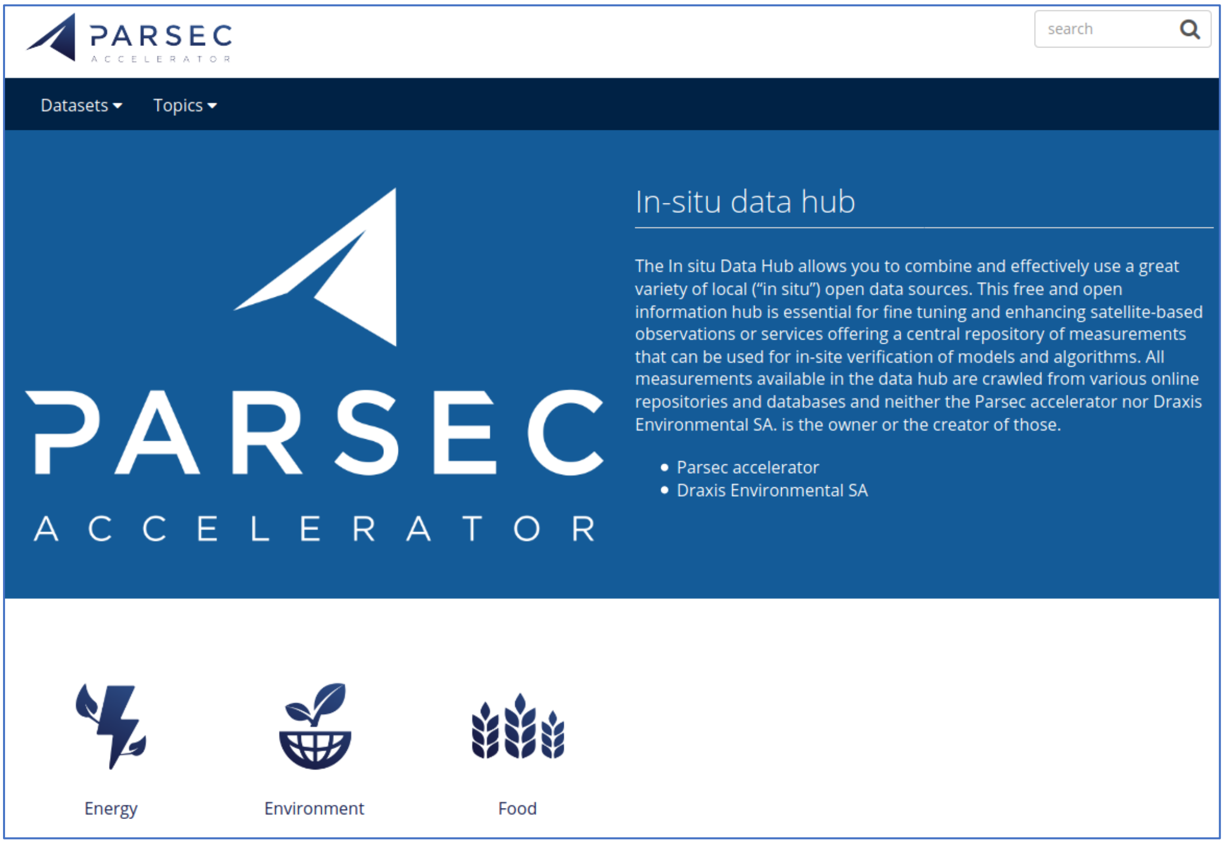

Figure 4: The In Situ Data Hub manager's user interface (home page)

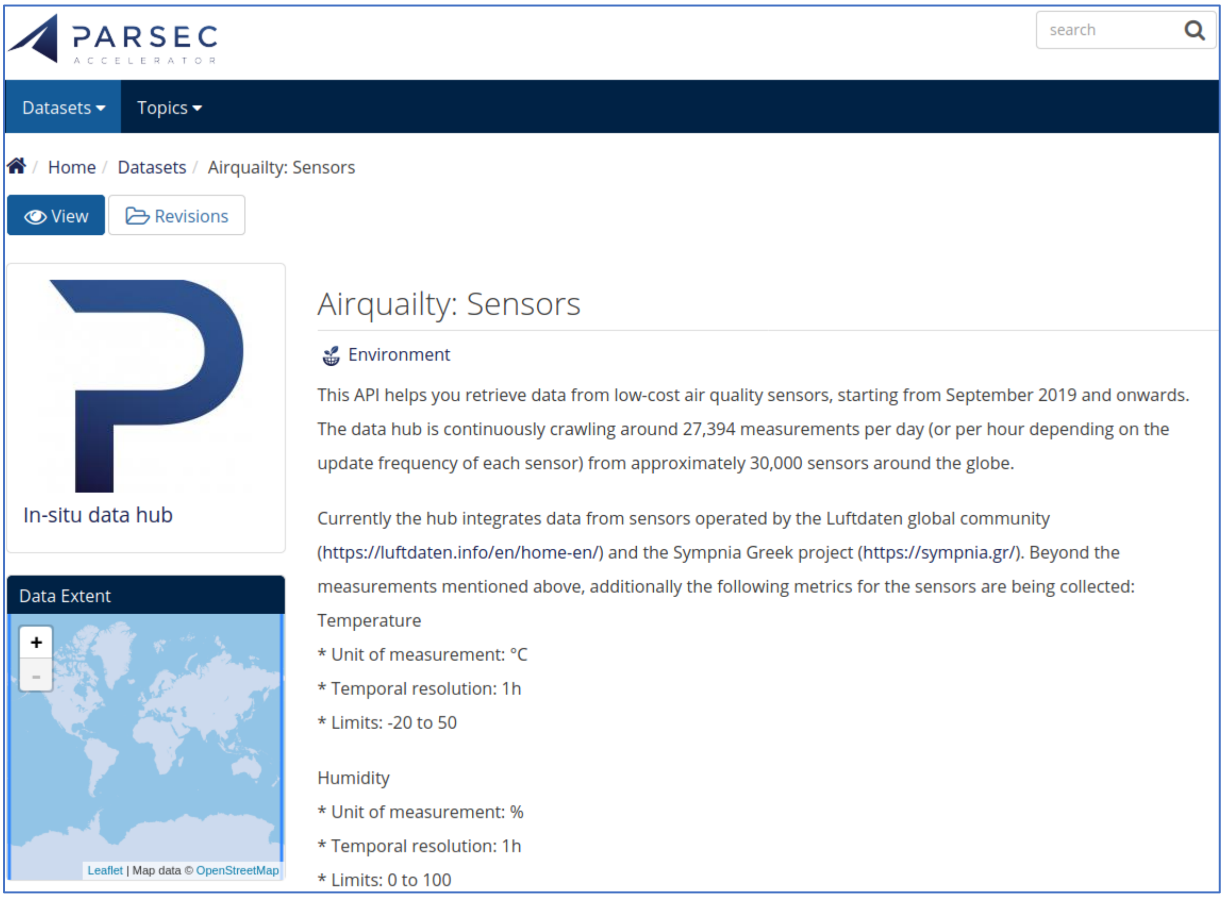

Figure 5: The In Situ Data Hub manager's user interface (description of data source)

Field	Value
Publisher	In-situ data hub
Modified	2020-04-07
Release Date	2020-03-20
Homepage URL	https://insitu-datahub-api.draxis.gr/documentation
Identifier	07515a61-1dab-430d-ac40-68b2cf7d4b36
Spatial / Geographical Coverage Area	POLYGON ((-180 -90, -180 90, 180 90, 180 -90, -180 -90))
Language	English
License	Creative Commons Attribution
Author	DRAXIS ENVIRONMENTAL SA
Public Access Level	Public

Figure 6: The In Situ Data Hub manager's user interface (data source information table)

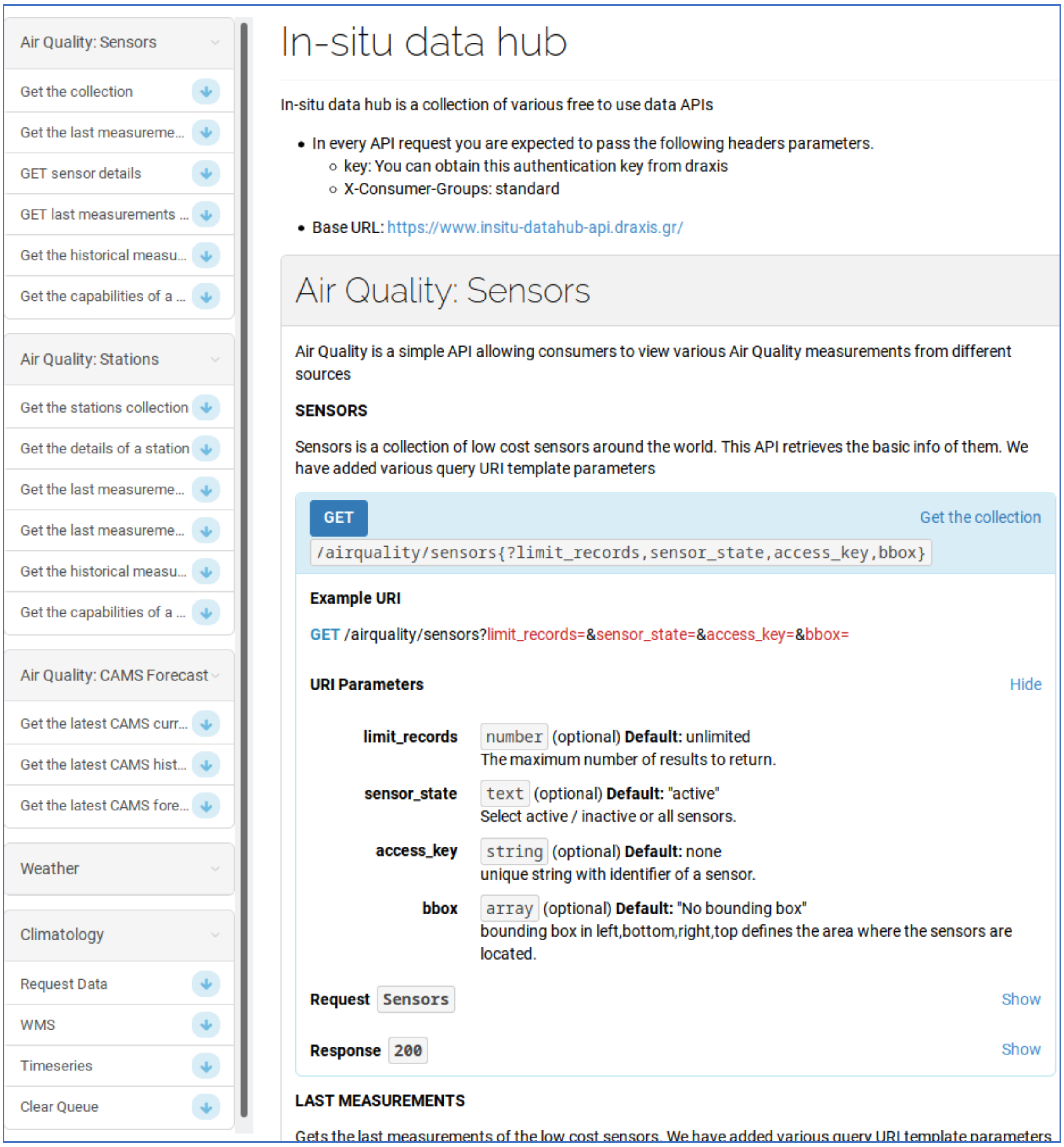

Figure 7: The In Situ Data Hub developer's user interface (API documentation)

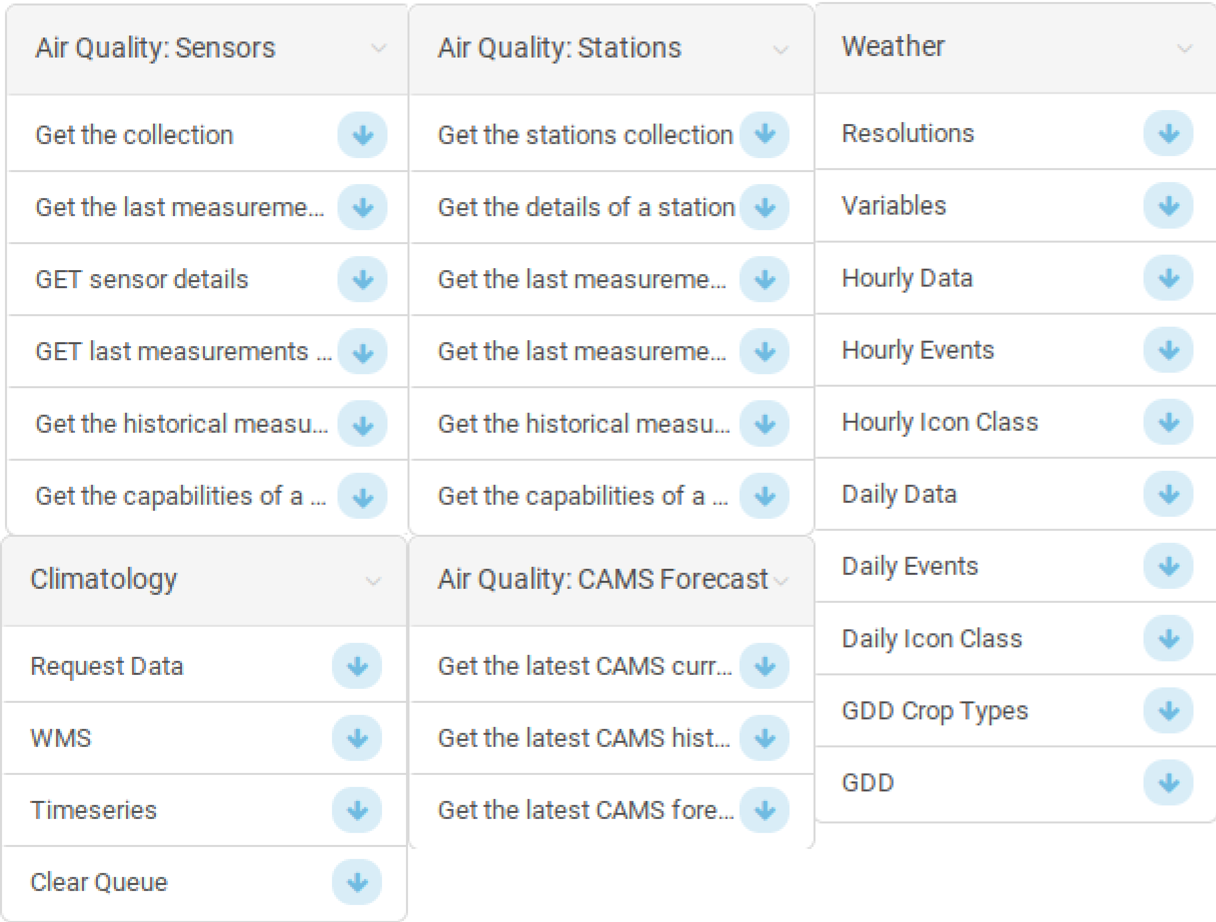

Figure 8: The In Situ Data Hub developer's user interface (API groups of datasets)

4 Architecture

The In Situ Data Hub adopts an agile system architecture ensuring the seamless integration of various technological components, enabling optimal use of resources, and driving EO data exploitation and EO service provision in the future. As seen in figure 2, further below, the high-level architecture of the hub will consist of three interconnected service tiers.

- Backend API Service
- In Situ data storage Service
- CAMS Service

An illustrative example of how this agile architecture helps deliver EO-based value to end-users, in key emerging industries, is presented below.

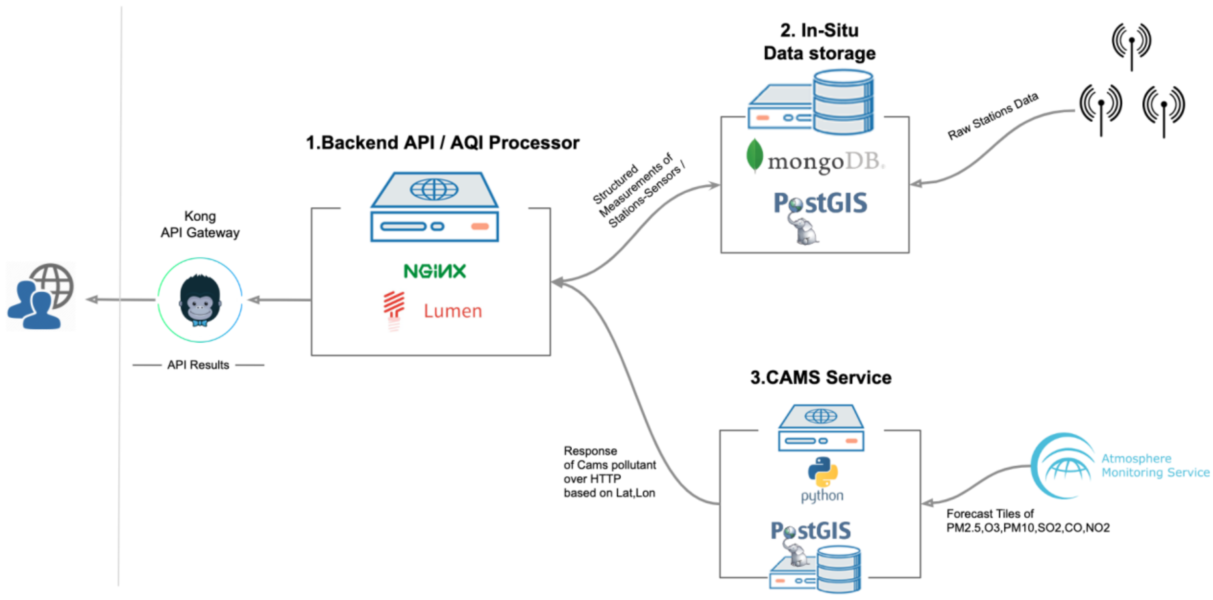

Figure 9: In-situ data hub architecture.

4.1 Backend API Service

A RESTful API (web based HTTP Application Programming Interface), based on the PHP framework Lumen (a popular general-purpose programming language), is responsible to apply air-quality intelligent algorithms on the retrieved raw pollutants. Lumen is an integrated framework that facilitates application development through built-in programming tools, allowing for a fast and ease deployment compared to other frameworks of the PHP programming language.

The design of the In Situ Data Hub source code follows the MVC development model, Model - View - Controller, which ensures the best distinction between the logic of the application and data models and the presentation of the retrieved results. It enables the development process to be separate and thus achieves improved performance in every aspect of the application's development lifecycle (Growth Speed, Debugging, Upgrades).

Additionally the NGINX Component (software that accelerates content and application delivery), improves security and acts as the web server of the API's stack. It is an open-source software for web serving, reverse proxying, caching, load balancing, media streaming, and other services.

API Gateway

The Kong gateway (software to securely manage communication between clients and microservices) is the single entry point into the In Situ Data Hub infrastructure. It handles a request by invoking multiple rules such as authentication mechanisms, throttling and aggregating the results. It can also translate between web protocols such as HTTP and WebSocket and web-unfriendly protocols that are used internally.

4.2 In Situ Data Storage Service

Postgres database is the in situ data storage, powered by the geospatial component PostGIS required for georeferenced storage of measurements coming from sensors, stations and raster data of the Copernicus Atmosphere Monitoring Services.

PostgreSQL is an open source RDBMS (Relational Database Management System) which is not developed by a single company but by a global community of users, companies, and institutions. PostgreSQL supports functions such as B-tree, hash, GiST and GiN, automations, which enable high speed search and data retrieval from the database, and additionally a wide range of predefined and user-defined data types and objects.

MongoDB
The In Situ Data Hub infrastructure generates a large number of events (i.e. logging,) which contains useful information about their operation including errors, warnings, and user's behaviour. Due to its non-sql architecture MongoDB (document oriented database) is suitable for storing log data from the multiple subsystems and provides an easy way to monitor the overall architecture.

4.3 CAMS Service

The CAMS Service acts as a standalone micro service where requests for data are submitted from the core, Backend API service, of the In Situ Data Hub. Due to the nature of data provided by CAMS (raster raw data), particular programming expertise has been adopted to achieve image processing in combination with GIS technologies deployed by the core in situ service.

The services offered by CAMS have been carefully studied, to decide the most suitable way to integrate the plethora of variables (3-days hourly forecast datasets) provided daily, and combine those datasets with the in situ data. Namely special focus has been given to the following variables PM25, O3, PM10, SO2, CO.

CAMS supports data provision through API service and FTP server (File Transfer Protocol services). As CAMS API service is not for operational use, suffering from delays and throttling (as stated in Copernicus site), the FTP server has been chosen, which offers NetCDF (or grib) raster data with the most updated forecast cycle.

Additionally historical data stored in the PostGIS database, of the in situ data storage service, has been combined to produce structured GIS queries for statistical reasons.

The CAMS service components consist of the following.

1 A standalone service which downloads NetCDF layers
A Python service which makes a connection to the FTP endpoint of CAMS and downloads the latest 3-day forecast for the parameters of our interest.

2 A post processing service which converts downloaded raw data, ready to be stored into the relational PostGIS database of the In Situ Data Hub storage.

A python service retrieves and posts processed raw data using GDAL and Numpy libraries (software libraries for handling geospatial data) in order to extract data from satellite bands and store them into the PostGIS database.

3 An API which serves Hourly forecasts for PM25, O3, PM10 , SO2 , CO
A lumen based API has been built as the interface of the micro service, capable of retrieving Lat, Lon coordinates and responding with parameters of our interest in a json readable format (open standard data and file, human-readable interchange format)

4.4 Weather Prediction Service

To meet the needs of the PARSEC beneficiaries the in-situ data hub has integrated high-resolution meteorological forecasts based on the WRF numerical model (Weather Research and Forecasting Model). The WRF numerical model equations are integrated in four nested domains with a spatial resolution of 18 Km, 6 Km and 2 Km, for the d01, the nested domains d02 and d03, and the d04, respectively depicted in figure 3 below. Namely the domain d01 covers the wider area of Europe, including North Africa and the Middle East. The nested domain d02 covers the Italian and Balkan Peninsulas, the nested domain d03 covers the Iberian Peninsula, while the nested domain d04 covers Greece, part of Albania, Skopje, Bulgaria and Turkey. Figure 10 illustrates the integration regions of the WRF model.

.

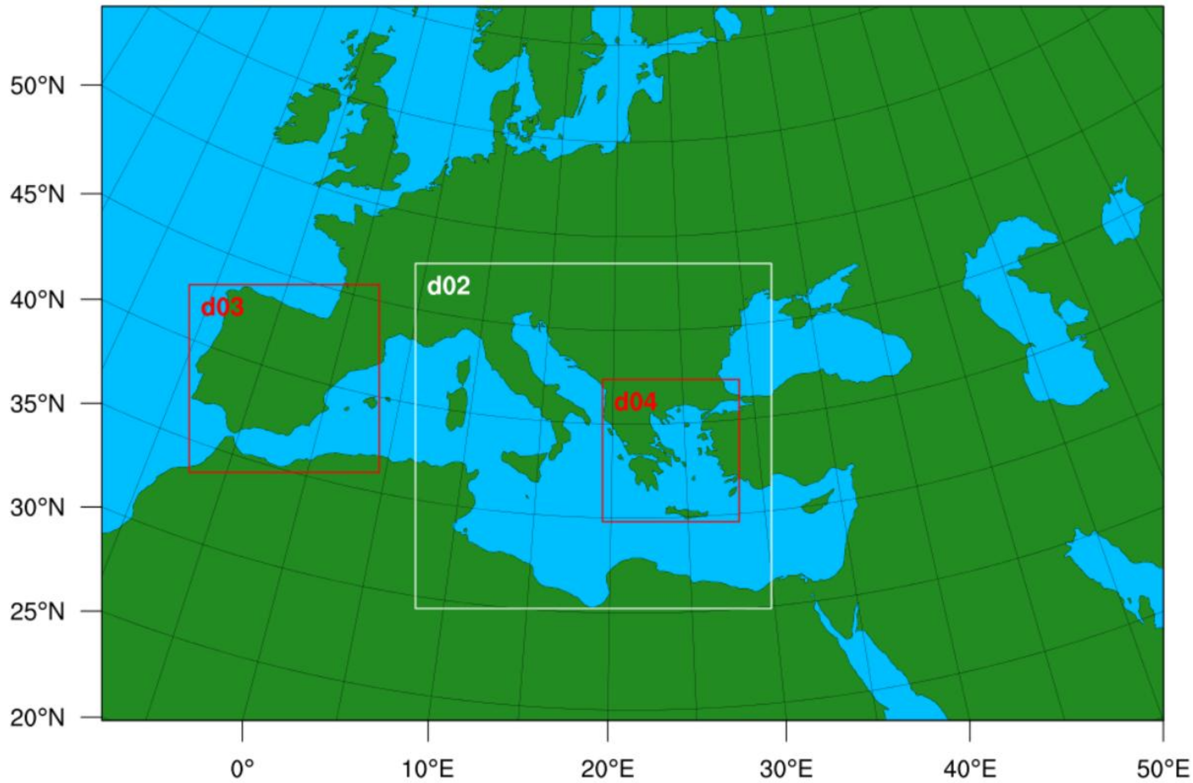

Figure 10: Weather prediction map.

The provided weather model delivers a plethora of operationally generated atmospheric fields including but not limited to

- Air Temperature at 2m Height
- Relative Humidity at 2m Height
- Wind Speed and Direction at 10m Height
- Precipitation and Precipitation Rate
- Reference Evapotranspiration
- Leaf Wetness
- Precipitation Types: Hail, Graupel, Ice Pellets and Snow
- Snow Height, density and cover
- Solar Radiation

Except from the weather prediction service, the In Situ Data Hub is providing also a suite of pre-computed climate indices for the area of Europe (including North Africa and the Middle East)- Figure 10, based on the reanalysis data and seasonal climate forecasts. The operational climate data provision is based on the available reanalysis data and seasonal climate forecasts of Copernicus Climate Change Service and particularly the ERA-5 Reanalysis data and the C3S multi-system seasonal forecast service. In the following list all available indices provided by the hub are presented.

- Standardized Precipitation Evapotranspiration Index
- Palmer Hydrological Drought Index
- Palmer Drought Severity Index
- Maximum Length of Dry Spell
- Maximum Length of Wet Spell
- Annual Count of Days with Precipitation > 20mm
- Number of Frost Days
- Number of Icing Days
- Cold Spell Duration Index
- Warm Spell Duration Index
- Total Totals Index
- K-Index
- Growing Degree Days
- Growing Season Length
- Soil Water Balance
- Cooling Degree Days
- Heating Degree Days
- Leaf Wetness Index
- Temperature Humidity Index

5 Hardware requirements

The infrastructure of the entire system is based on a single server running under Ubuntu Linux operating system and consisting of 9 Docker Containers. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application. The Dockers allow the application to run anywhere, regardless of the operating system, the existence or not of a cloud infrastructure, or whether it is a physical or virtual environment. This technology enables the In Situ Data Hub to be easily migrated and deployed under any operating system and hardware infrastructure.

Through packing ("shipping") any application into a container, acting as a separate Virtual Machine (VM), it enables the seamless integration of all the necessary dependencies to operate the service it represents, anywhere and under any operating system in exactly the same way.

The table below depicts the hardware needs of the system and the software running.

Characteristics	Specifications
Operating System	Linux Ubuntu 18.04 (or higher)
Required Installed Software	Docker
CPU	4 Cores
RAM	8GB
HDD	300GB

Table 1: Characteristics and Specifications of the infrastructure

6 Software requirements

Container based software architecture

The In Situ Data Hub follows a container based software architecture where docker images, containers, networks, volumes and other components, "ships" each subsystem making up the overall architecture acting as a unique service. Containers, working just like small-scale virtual machines (VMs), but in a far more specific and granular way, isolate each subsystem and its dependencies—all of the external software libraries the app requires to run—both from the underlying operating system and from other containers.

Some of the major advantages of Dockers and containers are listed below:

- Enables more efficient use of system resources
- Allows for faster software delivery cycles
- Enables application portability
- Reduces dependency issues

The below table along with the server-docker schema provided, depicts the different containers of the In Situ Data Hub infrastructure.

Container	Subsystems
Nginx-proxy	Proxy server filters and forwards requests to related containers
Data-api-kong	Container of API gateway (Kong)
Data-api-kong-db	Container of API gateway(Kong) Database
Data-API-environment	Codebase of Core/AQ API
Cams-API	Codebase of CAMS microservice
Mongo-logs	Container of MongoDB logging database
AQ-DB	Postgres/PostGIS Database container of In-situ data
Cams-DB	Postgres/PostGIS database container for Copernicus netcdf
Letsencrypt	Container that issues SSL certificates automatically
Docker-gen	Auto deployment container

Table 2: In Situ Data Hub container based software architecture.

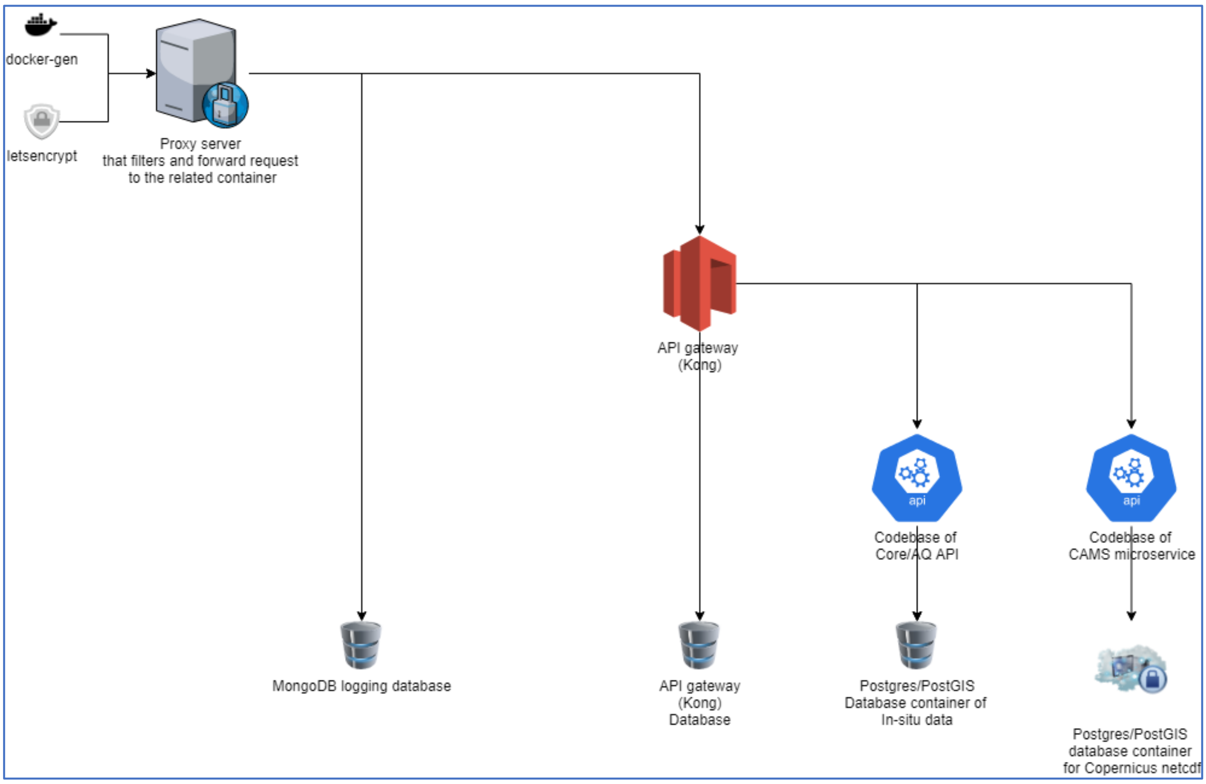

Figure 11: Server-docker schema

Scheduled data retrieval
The In Situ Data Hub contains crawlers which parse structured contents, extracting measurements with relevant information, based on predefined guide rules. Periodically scheduled cron-jobs (software used for scheduling tasks to run on the server) will trigger the crawling mechanism enabling any new data to be crawled or polled by the In Situ Data Hub. In case new data exists, it will be grabbed and ingested in the system and then forwarded to the content parser which will extract the information and store it in the database.

7 Hub Expandability

The In Situ Data Hub architecture is developed such that to allow future integration and expandability of data provision services. The software infrastructure of the hub is able to accommodate any datasets which will facilitate beneficiaries in deploying their business cases.

If further integration of data sources is needed a careful evaluation of the possible sources of information will be conducted, and once agreed with the PARSEC beneficiaries, it could be incorporated and offered through the hub. Throughout this process datasets of major stakeholders, open sensor measurements available on the internet, and information from web page tables will be retrieved, parsed and ingested into the In Situ Data Hub in a structured way.

All recordsets of the In Situ Data Hub categories can be expanded and enriched with new sources of information providing access to data on different domains, from air-quality, to energy, and economy.

PARSEC ACCELERATOR

Our Partners

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824478.