D3.13 – Manual
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)	(EVERSIS)		
Keywords	API Service, EO, datasets, IoT devices, sensors, air-quality, json, datacubes		

Document History

Version	Issue date	Stage	Changes	Contributor
1.0	17.07.2020	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
AQ	Air Quality
C3S	Copernicus Climate Change Service
CAMS	Copernicus Atmosphere Monitoring Service
CSV	Comma Separated Values
Datacube	Multi-dimensional Array of values
EO	Earth Observation
HTTP	Hyper Text Transfer Protocol
IoT	Internet of Things
JSON	JavaScript Object Notation
Kong	Software to securely manage communication between clients and microservices
PHP	General-purpose Programming Language
SME	Small Medium Enterprise
UTC	Universal Time Zone
WRF	Weather Research and Forecasting Model
XML	Extensible Markup Language

Executive Summary

This report is an updated version of "D3.5-In situ Data Hub Manual I", and it includes the same content as in D3.5 with additional information on the new data sources included in the second stage of enriching the In-Situ Data Hub. Overall, this report "D3.13-In situ Data Hub Manual 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 presented and new screenshots provided*

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 the 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 use and purpose of the data hub are presented, followed by detailed guidelines on the actual requests needed to retrieve the information.

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, 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-day 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 broader area of Europe, including the region of North Africa and the Middle East.

2 Hub Use and Purpose

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 facilities, support critical activities, such as the monitoring of remote environments, risk assessment and 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 organizations to consume and

integrate into 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.

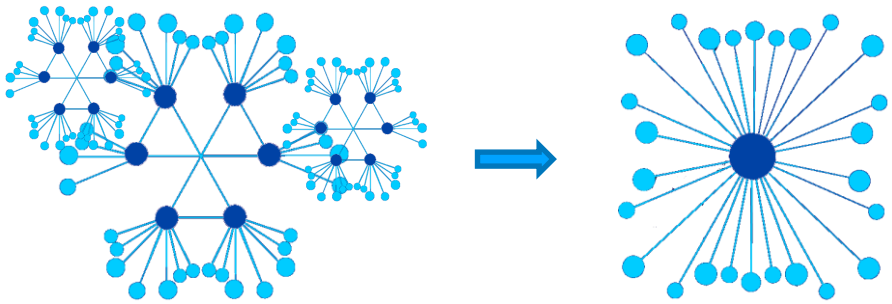

Figure 2 – From clustered data hubs to centralized information structures

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. To achieve this, the first step is to unify the available data sources 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.

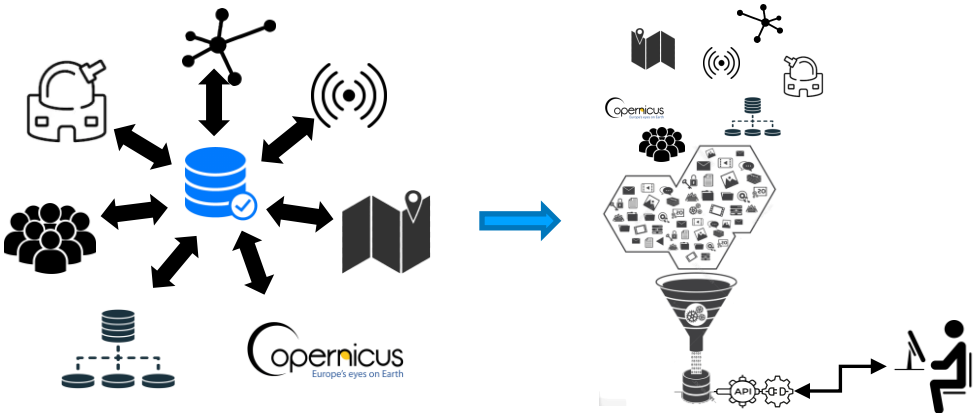

Figure 3 – Acquired information is stored and made readily available through a uniform and user friendly environment

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 requesters to view and, on 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 are often key to realising the value-added EO (Earth Observation) services.

3 Types of Datasets Provided

Currently, the hub integrates the 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).

Official ground-based monitoring stations

Specifically, data from official stations as of August 2019 and onwards are daily crawled and stored into the database of the hub. Approximately **10,651 daily air-quality measurements from 12,000 stations around the world** are retrieved, parsed, harmonized, and ingested into the storage of the hub.

The following table presents an indicative list of the data sources already integrated and provided to the PARSEC beneficiaries:

- https://openaq.org/ (Global) Measurements are retrieved hourly
- www.ypeka.gr (Attica, Greece) Measurements are retrieved daily at 16:00
- www.pkm.gov.gr/ (Central Macedonia, Greece)
- http://www.envdimosthes.gr/deltioPop.php (Thessaloniki, Greece) Measurements are retrieved daily at 12:30

An indicative list of the cities for which we already provide AQ information is presented below.

London	Los Angeles (USA)	Milwaukee	Cleveland	Copenhagen	Long Beach
Paris	San Francisco (USA)	Miami	Baltimore	Luxembourg	Kansas City
Istanbul	Colorado Springs (USA)	Oklahoma	Mexico city	Madrid	Atlanta
Munich	Boston (USA)	Minneapolis	New Delhi	Barcelona	Tulsa
New York	Las Vegas (USA)	Philadelphia	Sydney	Visalia, USA	New Orleans
Ho Chi Minh City	Chicago	Eugene (USA)	Phoenix	Hamilton, Canada	Bridgeport
Bangkok	Denver	Indianapolis	Albuquerque	Memphis	Sheffield
Washington DC	Oakland	Virginia Beach	Sacramento	Nashville-Davidson	Bergen
Saint Luis, USA	Columbus	Houston	Roseville	Lisbon	Jacksonville
Philadelphia	Charlotte	New Haven	Shreveport	Berlin	Hartford
San Antonio	Detroit	Omaha	Springfield	Geneva	Richmond
San Diego	El Paso	Orlando	Tacoma	Oslo	Lyon
Dallas	Bridgeport	Pittsburgh	Worcester	Amsterdam	Hong Kong
New Orleans	Cincinnati	Providence	Thessaloniki	Utrecht	Bangkok
Austin, USA	Eugene	Raleigh	Athens	Zurich	Reykjavik

Table 1 – Collection of cities the In Situ Data Hub retrieves AQ information

All measurements are stored according to the universal time zone – UTC.

For each source, the temporal resolution is defined, ranging from 1h to 8h to 24h.

For some sources, the temporal resolution may not fall under any of these categories (e.g., in Thessaloniki, the measurements are taken hourly from 00:00 pm until 11:00 am (11 measurements) and the max value from all collected is provided). In these cases, the hub makes an assumption that the temporal resolution is for 1h. Also, there are many cases where the temporal resolution is unknown (e.g., openAQ stations worldwide). In these cases, the hub also makes an assumption that the temporal resolution is for 1h and accepts all measurements into the database even if a pollutant's

default temporal resolution is different than 1h. Additionally, supportive information such as max/average/latest values are also registered.

For each station, the following data is stored:

- City, County
- Coordinates of the station
- Latest measurement (Date, Time). Depending on the source, but usually the hourly max values of the day or the daily average is used.
- Pollutants' concentrations:
    - PM10 (usually provided as 24h or 1h average), PM2.5 (24h or 1h average) -> $\mu g/m^3$
    - SO2 (24h or 1h average), CO (8h or 1h average), O3 (8h or 1h average), NO2 (1h average) -> ppm

Low-cost sensors

With respect to the low-cost sensors, stored measurements can be retrieved from the hub starting from September 2019 and onwards. The data hub is continuously crawling **around 27,394 measurements per day** (or per hour, depending on the update frequency of each sensor) from approximately **30,000 sensors around the globe**.

Currently, the hub integrates data from sensors operated by the Luftdaten global community (https://luftdaten.info/en/home-en/) and the Sympnia Greek project (https://sympnia.gr/). Beyond the measurements mentioned above, additionally, the following metrics for the sensors are being collected:

- Temperature
    - Unit of measurement: °C
    - Temporal resolution: 1h
    - Limits: -20 to 50
- Humidity
    - Unit of measurement: %
    - Temporal resolution: 1h
    - Limits: 0 to 100

CAMS forecasts

The services offered by CAMS have been carefully studied to decide the most suitable way to integrate the plethora of variables (3-day hourly forecast datasets) provided daily, and combine those datasets with the in situ data.

- Spatial resolution: 0.1° - around 10km for the European region
- Time resolution: 1h
- Forecast horizon: 3 days
- Type of level: Surface
- Species: PM2.5, PM10, O3, CO, SO2, NO2
- Domain: Europe

The CAMS forecast is based on seven state-of-the-art numerical air quality models developed in Europe from all seven models. They are combined via an ensemble approach consisting in calculating the median value of the individual outputs. Usually, better estimates of air pollutant concentrations are generated by CAMS regional Ensemble.

The Ensemble forecast is the median of the forecasts from 7 different state-of-the-art atmospheric modelling systems from institutions across Europe, which all use the same emissions data, the same

weather forecast data, and the same boundary conditions (from the CAMS Global forecast system) and combine these with Earth observation data to provide the air quality forecasts.
`

Currently, the partner models involved in Ensemble production are:

- CHIMERE from INERIS – A multi-scale chemistry-transport model for atmospheric composition analysis and forecast provided by the Institut national de l'environnement industriel et des risques (France).
- EMEP from MET Norway – The co-operative programme for monitoring and evaluation of the long-range transmission of air pollutants in Europe (European Monitoring and Evaluation Programme) provided by the Norwegian Meteorological Institute.
- EURAD-IM from University of Cologne – The EURopean Air pollution Dispersion – Inverse Model provided by the University of Cologne.
- LOTOS-EUROS from KNMI and TNO – A 3D chemistry transport model aimed to simulate air pollution in the lower troposphere provided by the Royal Netherlands Meteorological Institute and the Netherlands Organisation for Applied Scientific Research.
- MATCH from SMHI – The Multi-scale Atmospheric Transport and Chemistry a three-dimensional, Eulerian model provided by the Swedish Meteorological and Hydrological Institute.
- MOCAGE from METEO-FRANCE – The Modèle de Chimie Atmosphérique de Grande Echelle a Chemistry – Transport Model provided by Meteo-France.
- SILAM from FMI – The System for Integrated modeLling of Atmospheric coMposition provided by the Finnish Meteorological Institute.

European Air Quality Index

For each of the above-mentioned data sources, the European Air Quality Index, recently created by the European Environment Agency (https://www.eea.europa.eu/themes/air/air-quality-index/index) is being estimated and provided. For NO2, O3, and SO2, hourly concentrations of NO2, O3, and SO2 are used for the calculation of the index. For PM10 and PM2.5, the 24-hour running average is considered for the calculation of the index. All concentrations are in µg/m3. A sub-index for all the available pollutants is being requested and estimated in the temporal resolution (hourly concentrations for NO2, O3, and 24-hour running means for PM10/PM2.5).

Weather Prediction Services

The in-situ data hub integrates high-resolution meteorological forecasts based on the WRF numerical model (Weather Research and Forecasting Model). The WRF numerical model equations are integrated into 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 broder 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 4 illustrates the integration regions of the WRF model.

.

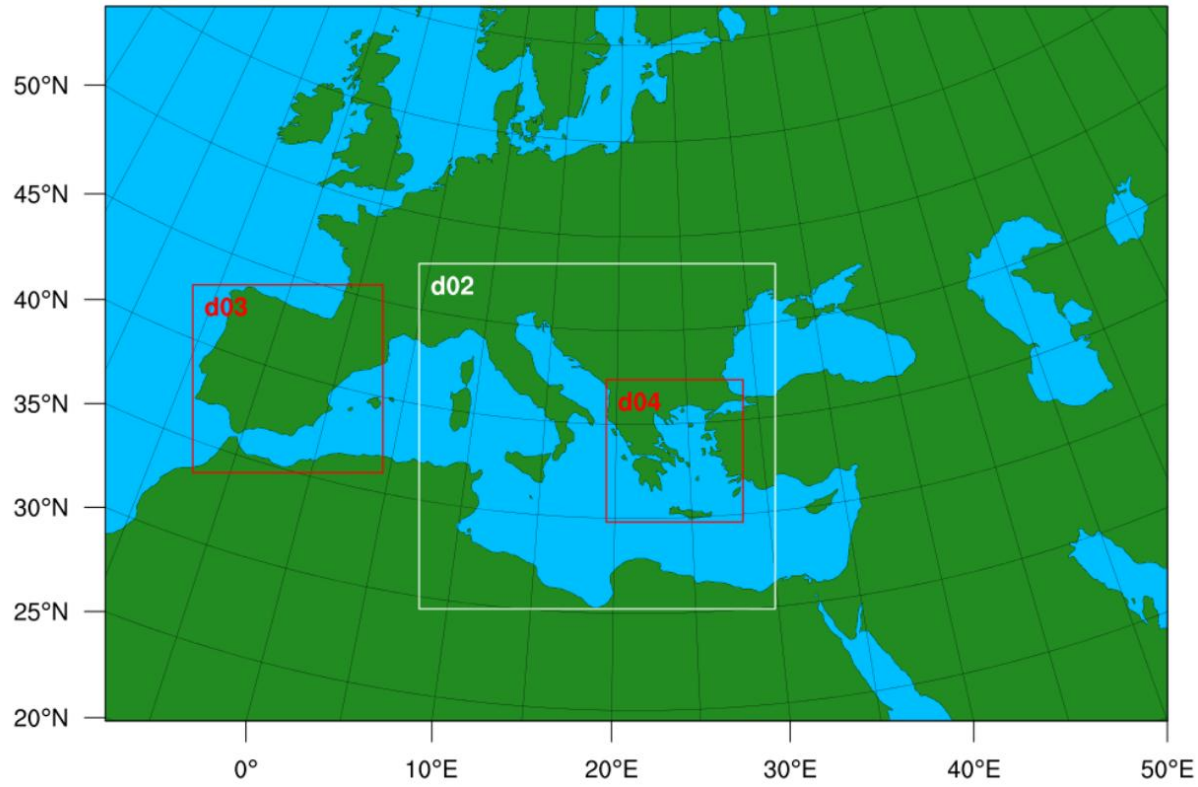

Figure 4 – Weather prediction map

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

Climate Services

Except for the weather forecasting service, the In Situ Data Hub is also providing a suite of pre-computed climate indices for the area of Europe (including North Africa and the Middle East) – Figure 4, 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, there are presented all available indices provided by the hub.

- 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

4 Information Retrieval

A RESTful API (web-based HTTP Application Programming Interface), based on the PHP framework Lumen (a popular general-purpose programming language), is responsible for applying the air-quality intelligent algorithms on the retrieved raw pollutants from the integrated resources, and serve the results to the requester. Additionally, a Kong gateway (software to securely manage communication between clients and microservices) acts as the single entry point into the In situ Data Hub handling requests by invoking multiple rules such as Authentication mechanisms, throttling, and aggregating the results. It also translates between web protocols such as HTTP and WebSocket and web-unfriendly protocols that are used internally.

Detailed documentation of the In situ Data Hub APIs is provided below. The documented APIs allow the requester to view and, on some occasions, to create or modify the data of the hub. The documentation can be accessed through the following url:
https://insitu-datahub-api.draxis.gr/documentation.

The use and retrieval of information are straightforward, for instance, if the user requests to collect information from air-quality sensors, the corresponding API-set is selected from the available list (to the left). Once selected, the respective documentation and examples provide the guidance to the user on how to retrieve the datasets of the low-cost air quality sensors from around the world.

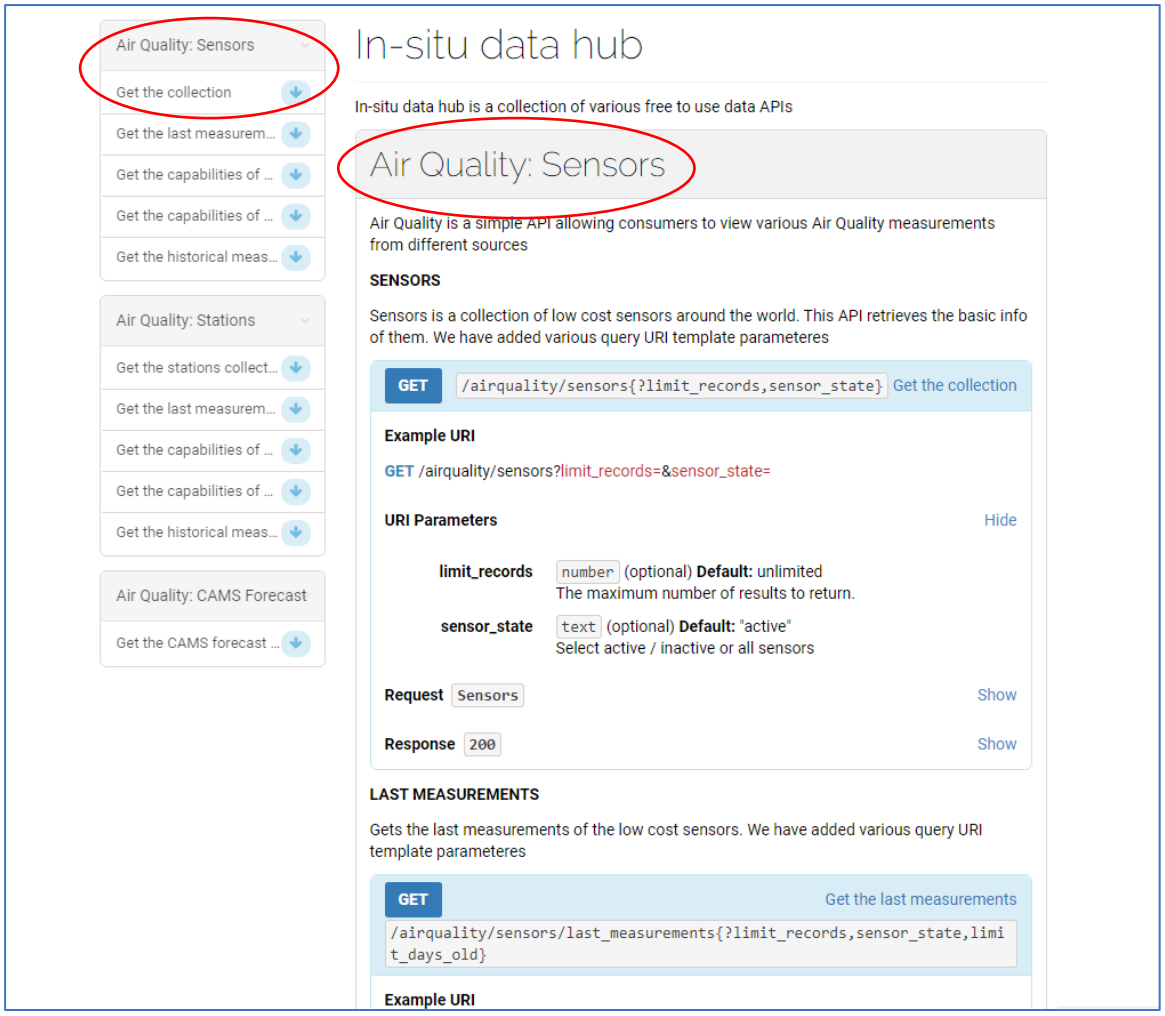

Figure 5 – In Situ Data Hub User Interface

For every API, the user can view a small description of the data provided, the API entry point, an actual GET request, as an example for sample data to be retrieved. Additionally, the user is instructed on the different parameters available in order to narrow down (limit) the requested information along with a short description.

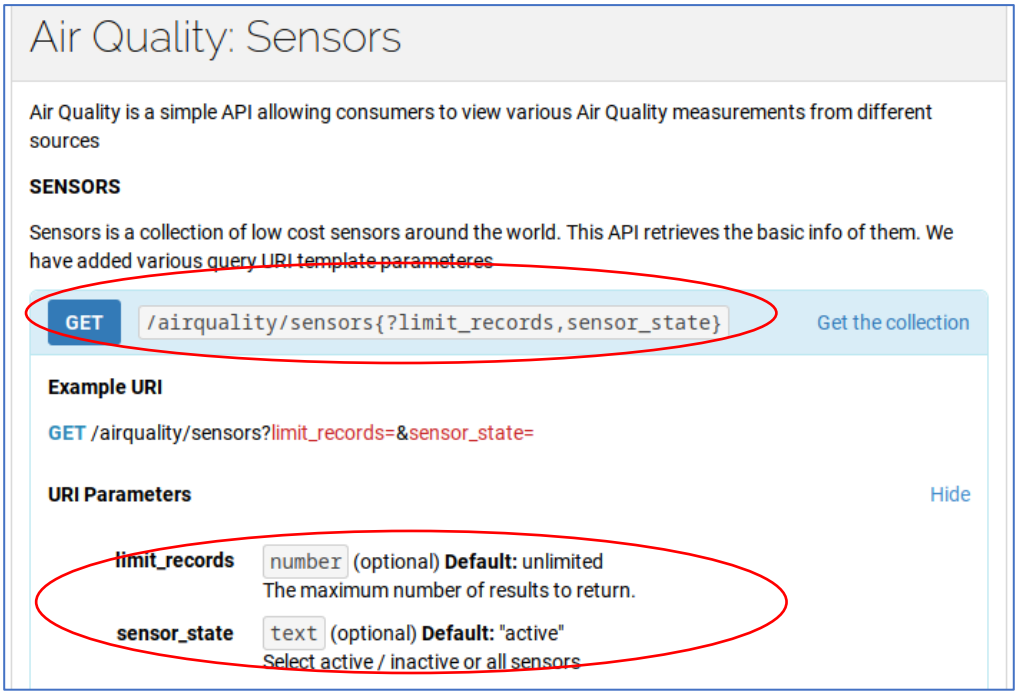

Figure 6 – In Situ Data Hub GET Request & Parameters

Once the request is obtained, the user is introduced with a sample response. Pressing on the available "show" buttons displays that the response headers are presented along with the "body" of the response, which is a json file with a response status "success" plus the requested data in the "data" field.

Figure 7 – In Situ Data Hub Response

```
{
  "status": "success",
  "message": "",
  "data": [
    {
      "id": 2,
      "access_key": "3133443bea0d7f71171706ceb7541a97",
      "name": "DriveIt Sensor 1",
      "project": "driveit-sensor",
      "created_at": "2019-07-26 13:10:57",
      "updated_at": "2019-07-26 13:10:57",
      "deleted_at": null,
      "lng": "22.894333",
      "lat": "40.655102"
    },
    {
      "id": 10,
      "access_key": "test-1002",
      "name": "Test sensor",
      "project": "sympnia-sensor",
      "created_at": "2019-08-30 11:32:22",
      "updated_at": "2019-08-30 11:35:48",
      "deleted_at": "2019-08-30 11:35:48",
      "lng": "23.727539",
      "lat": "37.98381"
    },
    {
      "id": 13,
      "access_key": "83ee72848e5be03a9c78cdfa91b5aa9e",
      "name": "Aek",
      "project": "sympnia-sensor",
      "created_at": "2019-08-30 14:09:37",
      "updated_at": "2019-08-30 14:09:54",
      "deleted_at": "2019-08-30 14:09:54",
      "lng": "18.798452489848",
      "lat": "25.968551630399"
    }
  ]
}
```


Figure 8 – In Situ Data Hub Body of response in JSON

5 Scheduled data retrieval

All data of the hub are continuously renewed and enriched. New sources of information are being added, providing access to information on different domains, from air-quality to energy and economy. Following the above-mentioned guidelines, the user can retrieve any dataset needed on a daily bases, choosing any of the available domains.

In order to retrieve the latest datasets available, the In situ Data Hub contains crawlers that parse structured contents, extracting measurements with relevant information, based on predefined guide rules. Periodically triggered, hourly, or daily, crawling mechanisms search for newly available datasets for retrieval and ingestion into the hub. In case new data exist, data 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. Through this process, the user is provided the latest datasets available for the domain of interest.

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.