

D3.5 – Manual In Situ Data Hub I WP3 – Large Scale Demonstrators

Authors: Stavros Tekes, Yanis Nasiopoulos Date: 12.03.20

PARSEC

Full Title	Promoting the Sensing companie	international con es through cross-cl	npetitiveness uster collabor	of European Remote
Grant Agreement No	824478	Acronym	1	PARSEC
Start date	1 st May 2019	Duration	l	30 months
EU Project Officer	Milena Stoyanova			
Project Coordinator	Emmanuel Pajot (EARSC)		
Date of Delivery	Contractual	29.02.2020	Actual	06.03.2020
Nature	Other	Dissemin	ation Level	Public
Lead Beneficiary	DRAXIS			
Lead Author	Stavros Tekes	Email		<u>stavros@draxis.gr</u>
Other authors	Giannis Nasiopoul	los, Alex Economou	l	
Reviewer(s)	(EVERSIS)			
Keywords	API Service, EO, da	atasets, IoT devices	s, sensors, air-	quality, json, datacubes

Document History				
Version	Issue date	Stage	Changes	Contributor
1.0	20.02.2020	Draft	First draft	DRAXIS
1.1	27.02.2020	Draft	Merge review comments	DRAXIS
1.2	06.03.2020	Draft	Review comments	DRAXIS
1.3	12.03.2020	Final	Formatting	DRAXIS

Disclaimer

Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains

Copyright message

© PARSEC consortium, 2020

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgment of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Table of Contents

e of Contents	3
of Acronyms	3
cutive Summary	4
Introduction	5
Hub Use and Purpose	5
Types of Datasets Provided	5
Information Retrieval	8
Scheduled data retrieval	12
	e of Contents

List of Acronyms

API	Application Programming Interface
RESTful API	A web based HTTP Application Programming Interface
AQ	Air Quality
CAMS	Copernicus Atmosphere Monitoring Service
CSV	Comma Separated Values
Datacube	Multi-dimensional Array of values
EO	Earth Observation
НТТР	Hyper Text Transfer Protocol
IoT	Internet of Things
JSON	JavaScript Object Notation
Kong	Software to securely manage communication between clients and microservices
РНР	General-purpose Programming Language
SME	Small Medium Enterprise
UTC	Universal Time Zone
XML	Extensible Markup Language

Executive Summary

This report documents the PARSEC In situ Data Hub which will act as a repository hosting georeferenced field observations coming from sensor networks, independent IoT devices, citizen observatories or 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 use and purpose of the data hub is 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 days forecast and 10km resolution providing continuous past and present data and information on atmospheric composition.

2 Hub Use and Purpose

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.

The main objective of the hub will be to offer SMEs, a centralised, expedite access to valuable resources that is often key to realising value-added EO(Earth Observation) services.

3 Types of Datasets Provided

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).

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.

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

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

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 one of these categories (e.g. in Thessaloniki the measurements are taken hourly from 00:00pm until 11:00am (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:
 - \circ PM10 (usually provided as 24h or 1h average), PM2.5 (24h or 1h average) -> μ g/m³
 - 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 (<u>https://luftdaten.info/en/home-en/</u>) and the Sympnia Greek project (<u>https://sympnia.gr/</u>). 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
 - o 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-days 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

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 air quality forecasts.

Currently 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.
- <u>EURAD-IM</u> 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 threedimensional, 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 (<u>https://www.eea.europa.eu/themes/air/air-quality-index/index</u>) 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 μ g/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).

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 to apply 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.

A detailed documentation of the In situ Data Hub APIs is provided below. The documented APIs allow the requester to view and, in 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 is 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 guidance to the user on how to retrieve the datasets of low cost air quality sensors from around the world.

PARSEC

Air Quality: Sensors	In-situ data hub	
Get the collection	In-situ data hub is a collection of various free to use data APIs	
Get the last measurem 👽		
Get the capabilities of 🔸	Air Quality: Sensors	
Get the capabilities of 🔸	Air Quality is a simple API allowing consumers to view various Air Quality measurements	
Get the historical meas 👽	from different sources	
	SENSORS	
Air Quality: Stations \sim	Sensors is a collection of low cost sensors around the world. This API retrieves the basic inf of them. We have added various query URI template parameteres	0
Get the stations collect 👽	GET /airquality/sensors{?limit records.sensor state} Get the collection	
Get the last measurem 👽		
Get the capabilities of 🔸	Example URI	
Get the capabilities of 🔸	GET /airquaitty/sensors?limit_records=&sensor_state=	
Get the historical meas 👽	URI Parameters Hide	
	limit_records number (optional) Default: unlimited	
Air Quality: CAMS Forecast	The maximum number of results to return.	
Get the CAMS forecast 👽	sensor_state [text] (optional) Default: "active" Select active / inactive or all sensors	
	Request Sensors Show	
	Response 200 Snow	
	LAST MEASUREMENTS	
	Gets the last measurements of the low cost sensors. We have added various query URI template parameteres	
	GET Get the last measurements	
	<pre>/airquality/sensors/last_measurements{?limit_records,sensor_state,limi t_days_old}</pre>	
	Evample IIDI	
	Example on	

Figure 1: 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 a 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.

Air Quality: Sensors
Air Quality is a simple API allowing consumers to view various Air Quality measurements from different sources
SENSORS
Sensors is a collection of low cost sensors around the world. This API retrieves the basic info of them. We have added various <u>query URI template parameteres</u>
GET /airquality/sensors{?limit_records,sensor_state} Get the collection
Example URI
GET /airquality/sensors?limit_records=&sensor_state=
URI Parameters Hide
limit_records number (optional) Default: unlimited The maximum number of results to return. Image: content of the second
sensor_state text (optional) Default: "active" Select active / inactive or all sensors

Figure 2: 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 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.

sensor_state	The maximum number of results to return. text (optional) Default: "active" Select active / inactive or all sensors
Request Sensors	Hide
Headers	
Accept: applicati	ion/json
Response 200	Hide
Headers	
Content-Type: app X-My-Message-Head	olication/json der: 42

Figure 3: 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"
   }
 1
}
```

Figure 4: 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 latest datasets available, the In situ Data Hub contains crawlers which 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 a 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. Through this process the user is provided the latest datasets available for the domain of interest.



Our Partners















_Dbw**con**





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