DATA INTEGRATION, HARMONIZATION AND PROVISION TOOLKIT FOR WATER RESOURCE MANAGEMENT AND PREDICTION SUPPORT

GEORGIOS VOSINAKIS*, EVANGELOS MALTEZOS, MARIA KROMMYDA,
ELEFTHERIOS OUZOUNOGLOU & ANGELOS AMDITIS
Institute of Communication and Computer Systems (ICCS), Greece

ABSTRACT

Timely and reliable information is critical to organizations managing water resources. Drinking water is one main source of risk when its safety and security is not ensured. Early prediction and mitigation of such risks relies on prediction models that depend on live and historical data. Such data are quite heterogenous in nature, including sensor measurements, satellite imagery and radar readings, unmanned aerial vehicle (UAV) images and videos as well as results of prediction algorithms (flood risk, oil spills etc). AQUA3S is an EU funded project which combines novel technologies in water safety and security, aiming to standardize existing sensor technologies complemented by state-of-the-art detection mechanisms. Sensor networks are deployed in water supply networks and sources, supported by complex sensors for enhanced detection. Sensor measurements are supported by videos from UAVs, satellite images and social media observations from the citizens that report low-quality water in their area also creating social awareness and an interactive knowledge transfer. Semantic representation and data fusion provides intelligent decision support system (DSS) alerts and messages to the public through first responders' mediums. This study presents the data ingestion, integration and harmonization platform that was developed to support the systems of the project, consisting of the necessary APIs, to ingest data, a harmonization layer and a data store layer The data is harmonized and indexed using the NGSI-LD model to make sure information can be indexed and served both is real time through a live context broker, as well as in the form of historical time series through a dedicated historical data service. The data store layer includes provisions for the storage of annotated binary files (images, videos, etc.) as well as georeferenced map layers following OGC protocols such as web feature service (WFS), web map service (WMS), and web coverage service (WCS).

Keywords: water safety, water management, WFS, WMS, WCS, decision support, NGSI, linked data, water resource management, digital twins.

1 INTRODUCTION

The management of water networks and water reservoirs is a critical aspect for any society. Water management combines a variety of fields from resource management to predicting and defusing crisis situations. Timely information and readily available, well indexed data are critical to informed decisions during a quickly evolving crisis, such as water contamination incidents or flooding of inhabited areas due to adverse weather conditions. At the same time recording and monitoring of land use in areas around critical water sources is vital to the provision of adequate clean water. An increasing number of algorithms and prediction systems have been suggested that use a variety of diverse data sources. In this context several approaches have been suggested utilising sensor measurements [1], [2] as well as satellite imagery [3], [4] and even citizen information available through social media [5], [6].

AQUA3S [7] is an EU funded project that aims to combine novel technologies in water safety and security and standardize existing sensor technologies complemented by state-of-the-art detection mechanisms, providing water suppliers with timely information and

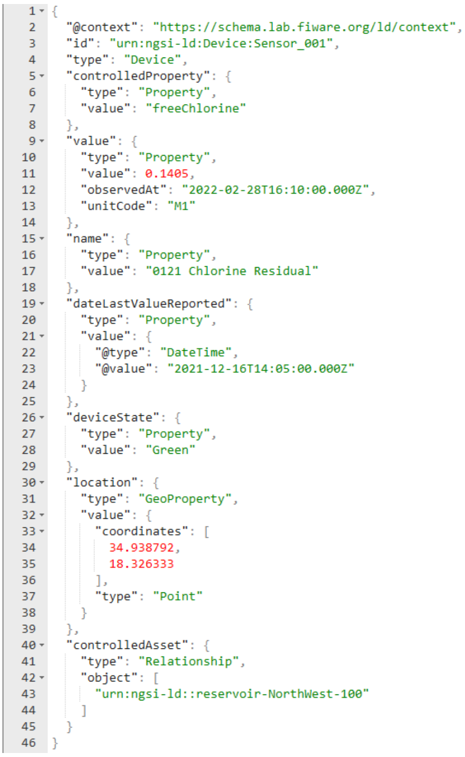
* *ORCID: http://orcid.org/0000-0003-2029-8409*

WIT Transactions on The Built Environment, Vol 208, © 2022 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/FRIAR220071

assessments of evolving situations. In this context the harmonization and availability of data in a ready to be consumed format is integral to the success of the project and the operation of the platform. Based on an analysis of the available data types, as well as user and technical requirements gathered during the initial phase of the project, a data ingestion and harmonization layer was designed and developed, together with a data store layer and the necessary API services that allow for the indexed storage and request of data by analytical modules, data visualization modules, as well as authorized data consumers directly.

The available data belongs to two main categories, primary data, available through a variety of sources and legacy sensor systems and secondary data, produced by the analytical modules of the project platform. After ingestion, data are stored and indexed to facilitate availability. This is achieved through a harmonization layer. Data were harmonized using Fiware compatible smart data models [8] and following the FIWARE NGSI-LD standard [9]. Several new data models, used in indexing the data types of the project were developed by the consortium partners (e.g. for satellite imagery, social media data, risk management). The main focus was given to providing a solution that would be easy to extend and adapt to new data types, as well as configure to new requirements.

Fig. 1 shows an example of an NGSI device entity in JSON-LD format. The entity includes a unique URN ID for the entity (line 3). The linked data [10] format of these data models allows (through their URN ID to establish connections between entities. In this case a “relationship” attribute (lines 40–44) connects this sensor to a specific real-world asset (water reservoir).



```
 1 {
 2   "@context": "https://schema.lab.fiware.org/ld/context",
 3   "id": "urn:ngsi-ld:Device:Sensor_001",
 4   "type": "Device",
 5   "controlledProperty": {
 6     "type": "Property",
 7     "value": "freeChlorine"
 8   },
 9   "value": {
10     "type": "Property",
11     "value": 0.1405,
12     "observedAt": "2022-02-28T16:10:00.000Z",
13     "unitCode": "M1"
14   },
15   "name": {
16     "type": "Property",
17     "value": "0121 Chlorine Residual"
18   },
19   "dateLastValueReported": {
20     "type": "Property",
21     "value": {
22       "@type": "DateTime",
23       "@value": "2021-12-16T14:05:00.000Z"
24     }
25   },
26   "deviceState": {
27     "type": "Property",
28     "value": "Green"
29   },
30   "location": {
31     "type": "GeoProperty",
32     "value": {
33       "coordinates": [
34         34.938792,
35         18.326333
36       ],
37       "type": "Point"
38     }
39   },
40   "controlledAsset": {
41     "type": "Relationship",
42     "object": [
43       "urn:ngsi-ld::reservoir-NorthWest-100"
44     ]
45   }
46 }
```

Figure 1: Example of an NGSI-LD device entity, used to map sensors (in this case a residual chlorine sensor).

A central context broker allows for the storage and retrieval of data in NGSI-LD compatible format. With the harmonization of available data to the NGSI-LD standard and the indexing of binary data (e.g. satellite images) using the data models developed, the integration platform can be connected to multiple services and expanded in the future to include more data sources, or new data types and can support the data requirements of any client module (analytical, visualization, etc.) that might be developed in the future based on the NGSI-LD standard.

The platform has been designed to accommodate the needs of the AQUA3S project but, by offering basic data harmonization and storage services it can be considered "data type agnostic" and can be easily adapted to other fields, beyond water quality and management.

2 DATA TYPES CONSIDERED

The data ingestion architecture developed has been designed based on the data types available to the project. The aim of this layer is to collect data from multiple heterogenous sources that will be harmonized, processed and indexed. Primary data, available from various sources comprise:

- Live sensor data;
- Sensor measurements from legacy SCADA systems;
- Satellite imagery, with accompanying metadata;
- Social media data, with accompanying metadata;
- Call complaint data from water suppliers' call centres;
- Water Network data, available as EPANET files;
- CCTV and Drone images and footage, with accompanying metadata.

While secondary data, produced by analytical modules within the AQUA3S platform comprise:

- Results of social media data analysis;
- Results of satellite image analysis;
- Geo-referenced map layers with accompanying metadata;
- Crisis classification analysis results.

The aim of our ingestion platform is to collect data from multiple sensors and other sources and has a central role in the integration of heterogeneous sources of data. It essentially consists of a logical bus implementing a publish–subscribe mechanism aiming to advertise the field devices data and metadata as resources as well as the relevant adapters for translation of the proprietary sensor data to a harmonized data format implemented by a middleware layer.

3 SYSTEM ARCHITECTURE

The architecture of our system consists of two distinct layers:

- A data ingestion and harmonization layer, responsible for ingesting data into the system, as well as indexing and aligning data received with the required data models.
- A data store layer, responsible for storing indexed data and the accompanying APIs that allow for the effective retrieval of data in a structured manner.

A general overview of the system architecture is shown in Fig. 2.

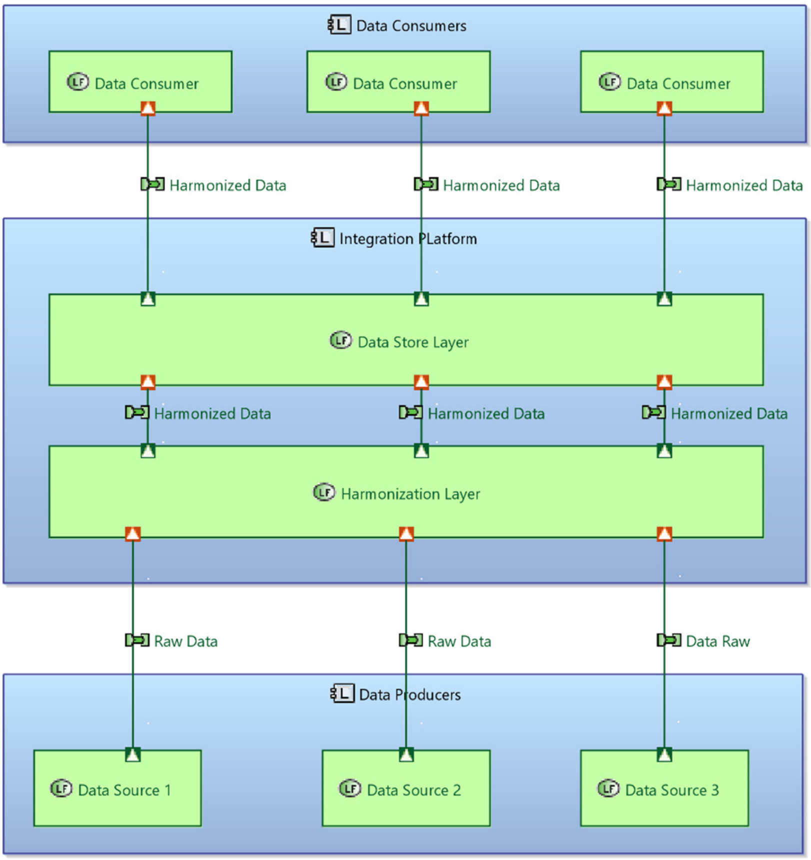


Figure 2: The general block diagram of the system architecture, showing the Integration platform with its distinct layers.

3.1 Data store layer

The data store layer consists of a context broker, holding information in JSON-LD format on sensor measurement and analysis results. Binary files are stored in a series of binary storage modules while a corresponding entity is created within the context broker, recording important information on the binary file, as well as its location in the binary storage. In this context, the broker is the central module of the structure holding the current, up-to-date image of the entire system. It provides, directly or indirectly through links, all the information ingested by data consumers and analytical modules.

The core of the data store layer is based on four main, open-source tools:

- An Orion-LD context broker [11] – It stores and serves the current state of the system (latest values) using the JSONLD format and adhering to Fiware smart data models.
- Cygnus-LD historical enabler [12] – Paired with a historical service API, developed within the project to serve queries to the historical database, it stores and serves historical information and time series of historical values.
- A GeoServer [13] – It stores and serves georeferenced images and data.

- A WebDAV server [14] – It stores and serves large binary files (e.g. images, videos and documents) and allows the distribution and collaboration on files.

Orion-LD allows for the creation of entities (e.g. corresponding to specific sensors) in JSON-LD format. Entities are indexed based on a unique entity ID and categorized based on entity type. Entities can be queried through HTTP requests based on their type or unique entity ID. This allows data consumers to receive the current status of a specific entity or the current status of all entities of a specific type (e.g. all sensors). Entities can be updated with new values. Orion-LD will hold and provide, when queried, the current, latest status of each entity monitored. When an entity is updated with a new value (e.g. a new measurement) Orion-LD will replace the previous with the latest. Orion-LD does not store previous values of entities, this is achieved through the Historical Service.

The Historical Service is responsible for storing and providing historical information of all entities held in Orion-LD. It consists of two components, Cygnus-LD, which is responsible for storing the historical information and the Historical Service API which provides the historical data to the data consumer modules through HTTPS requests. For every entity within Orion-LD a subscription is being created to the specific entity, establishing Cygnus-LD as the endpoint. Whenever the entity is updated in Orion-LD, the broker sends an automatic notification, including the previous status of the entity to Cygnus-LD. Cygnus-LD stores the status to a Postgres Database, indexed by its timestamp. Whenever a consumer module needs a timeseries of historical values for any entity, a request can be sent to the Historical Service API that will return a list, containing the historical values. Requests can be made using the specific entity ID and can either specify a time interval (starting date/time, ending date/time) or get values from a specific time/date to the present. Cygnus-LD is an open-source solution, but while it records data to a historical database, it does not provide an API to retrieve this data. This is the reason the Historical API was developed within the project to provide a Fiware compatible API service for the querying and provision of historical time series. Fig. 3 shows a general diagram of the historical service's architecture.

3.2 Integration and harmonization layer

A series of adaptors, developed in the Python 3.6 programming language [15] have been developed to facilitate the ingestion and harmonization of available data to the data models used within the project.

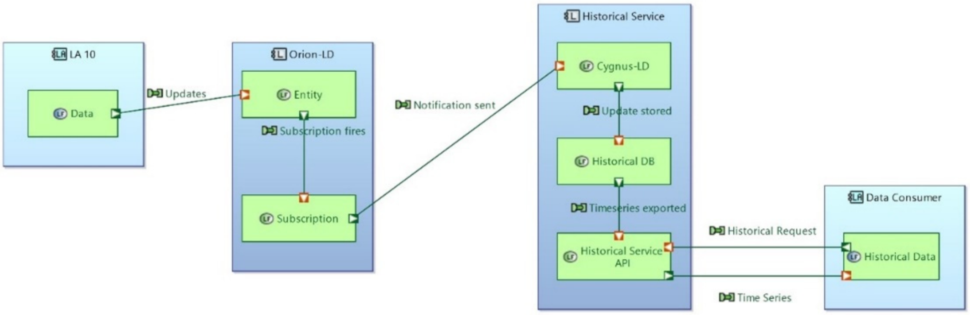


Figure 3: A block diagram of the historical service, comprising the Cygnus-LD connector (open source) and the Historical Service API developed within the AQUA3S project.

The harmonization process is performed in two ways, depending on the specific requirements of each data source:

- Client scripts – Installed as embedded service within the data producing modules and transmitting data in a decentralized manner.
- Services that query available APIs or servers, and centrally collect information from available sources.

Clients are developed as lightweight scripts. They are meant to be installed in either AI units or single board computers used by independent sensors without overloading their operation. In the case of clients, the harmonization layer, which aligns data with the NGSI-LD standard, is included with the client so that data transmitted by the device is NGSI-LD compatible thus making the device an independent, integrated data transmitter. The client is installed on the device and the script is run either at regular intervals to receive updated measurements (in the case of sensors) or when new data is available for transmission (in the case of analytical services).

Centralized ingestion services incorporate various modules depending on the required communication protocols used to query data. All such services are available as Docker containers that run centrally on the AQUA3S platform. Depending on the availability and requirements for data, they run either periodically (in the case of sensors and legacy SCADA systems) or on demand when an analytical service needs new data to consume. Fig. 4 shows a diagram of the two harmonization methods.

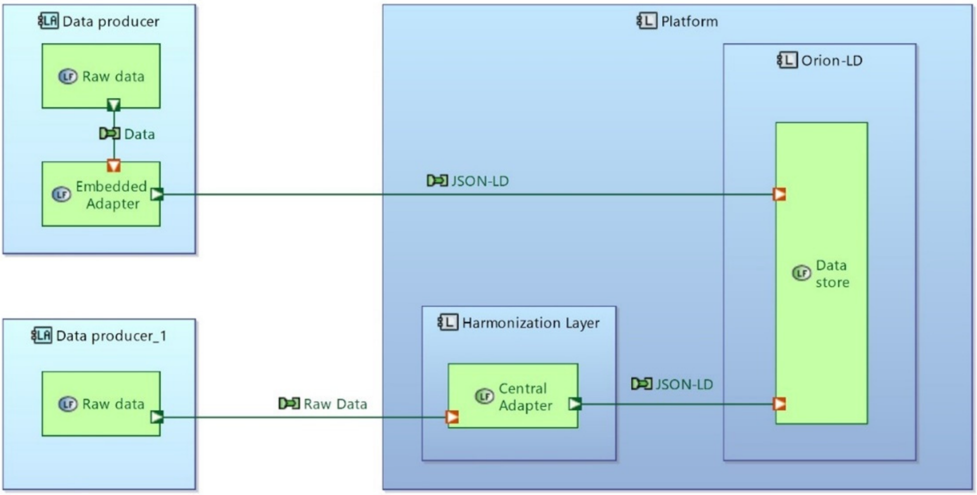


Figure 4: Block diagram of the two harmonization methods, through decentralized embedded adapters (clients) and a central harmonization service.

To guarantee the reusability and adaptability of the adaptors used within the project to the various heterogenous data types ingested two Python 3 libraries were developed:

- Harmonization library: Implements a class of JSON parsers that include methods to ingest raw data and parse them into the correct JSON-LD structure, according to the corresponding data model.

- Connection library: Implements a class of connectors that handle connections to Orion-LD, providing methods to create, update or delete entities in Orion-LD, perform diagnostic checks and handle subscriptions and notifications.

4 DATA HARMONIZATION, STORING AND INDEXING

Data considered can be divided, based on their availability and format, into four basic categories: live data feeds, recorded data, georeferenced map layers and images and binary files. To keep our architecture "data type agnostic" in order to guarantee expandability to different data sources we developed a generalized strategy in handling each distinct data type.

4.1 Live data feeds with an embedded client

Live data feeds originate either from sources external to the project platform (like sensors that provide a live feed of readings) or internal sources (like analytical tools that process available data to produce analysis results).

Sensors provide periodical measurements at medium or high frequencies (1 Hz to 1/60 Hz). These measurements are either stored locally in a (usually unstructured) JSON file or made available through a socker or API. Sensors are installed at the end users' locations, where they need to connect to a wi-fi or mobile network to transmit data. Analytical tools can either operate automatically, to provide results in a periodical manner similar to sensors, or run on demand, to produce results on an ad hoc basis.

In both cases an on-board Python service was developed that will be installed on the sensor or analytical tool as an embedded client. In the case of a sensor, the service polls the measurements provided by the sensor at specific internals, harmonizes them to the JSON-LD data model and transmits them to the Orion-LD context broker. In the case of an analytical tool, the client is accessed by the tool itself when data is ready to be transmitted. Analytical tools that have been integrated using this method include:

- Satellite imagery analysis modules, processing satellite images and analyzing water resource related risks and crisis.
- Social media modules, including a crawler that collects social media posts relevant to water supply issues and an analytical tool that uses the posts as input.
- A crisis analysis module, combining several data sources and producing risk estimations on water supply and flooding.

Communication with the platform is through a wi-fi or mobile connections. Fig. 5 shows a diagram of this structure.

This decentralized solution was selected for independent sensors, based on common IoT practice, to make sure the ingestion service is expandable and adaptable. A configuration file provided with the adapter installed on the sensor allows end user to successfully configure the adapter to specific sensors.

When a new sensor needs to be added all the end user has to do is install the adapter on the sensor's hardware and update the configuration file with the new sensor's information. When the service connects to a network it will transmit measurements to the configured Orion-LD context broker.

The available configurations allow the users to select what happens in case of connection loss. The sensor can store measurements locally and transmit them when connection is established (to avoid data loss), or it can send only the latest measurement to ensure speed of service.

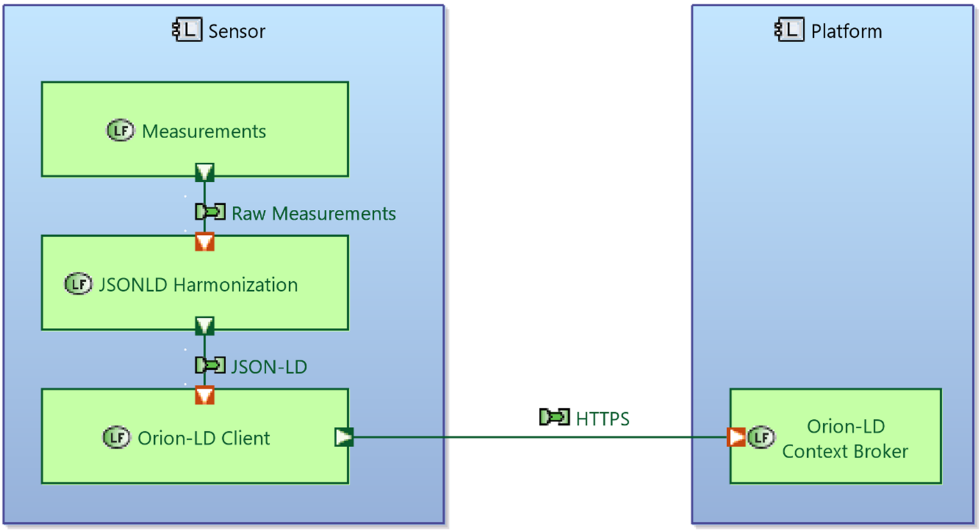


Figure 5: Block diagram of the embedded ingestion and harmonization client, using a sensor as an example.

4.2 Near real time and other recorded data

This category comprises a variety of data types that are available irregularly in the form of CSV or TXT files. These files, in the case of the AQUA3S project include bulk measurements provided by legacy SCADA systems, as well as call complaint records from water suppliers' call centres and EPANET [16] files that map water networks in a text format.

Data are collected from SCADA systems periodically and are made available to the project in batches, at specific time intervals. As a result, such data are near-real-time. Every time a batch is available a CSV file with the measurements is exported and uploaded to a dedicated sFTP server. CSV files generated in this manner have a filename starting with the unique sensor ID and include a column of measurements and a column with the corresponding timestamps. In the case of call complaints data are uploaded periodically by end users to a sFTP server in the form of csv files. The files list anonymized information on each call and its subject.

For sensor measurements, the ingestion service consists of a Python 3 application that acts as a server-side component that has the role of connecting to multiple sFTP sites, reading a list of asset files and reading a number of different measurement csv files. This functionality is based on periodic tasks running which connect and retrieve csv files. The service manages a list of queues on a per sensor/device basis. Queues contain the retrieved sensor measurements from the sFTP sites. The previous periodic tasks feed these queues.

It also handles a periodic task that reads one by one sensor measurements from the previous queues, performs harmonisation to the NGSI-LD data model and, via a rest client, uploads them to the Orion-LD context broker. To guarantee the expandability of the ingestion service each partner maintains within the corresponding sFTP server a configuration CSV file that lists the sensors providing measurements and their basic information (e.g. ID, name, location). Fig. 6 provides a general block diagram of this process.

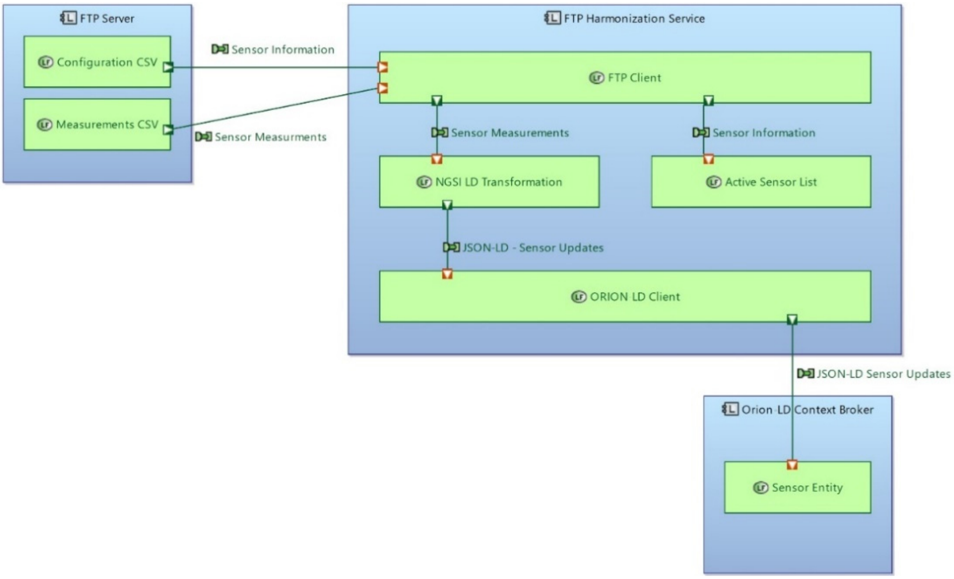


Figure 6: Ingestion and harmonization service for legacy SCADA systems.

When a new sensor needs to be added all the end user must do is update the configuration file with the new sensor's information. The service periodically ingests the configuration file and then automatically retrieves CSV files corresponding to those sensors. If a sensor is removed from the configuration file, the service will stop ingesting files from the corresponding sensor.

Call complaint records are integrated in a similar manner. An sFTP client periodically polls an sFTP server for new uploaded files. If new files are found, they are sent to the harmonization service which harmonizes data to the data model and uploads them to Orion-LD.

The analytical tool receives call complaints from the context broker and, after analysing them, sends the analysis results to its own harmonization service that publishes analysis data to the Orion-LD context broker.

4.3 Georeferenced map layers

This category includes satellite data comprising georeferenced satellite photographs that are acquired from online, open-source satellite data hubs (e.g. The Copernicus Hub), as well as the products of analyses produced by analytical modules developed within the project. The available data include georeferenced photographs and JSON files that include information of the performed analysis, as well as metadata on the photographs and their source.

After processing in the analytical service of the platform, photographs are uploaded to a GeoServer instance loaded on the platform. Information on the analysis performed, as well as metadata, including the photograph's location in the GeoServer are then sent to the harmonization service, which is responsible for harmonizing the data to the Orion-LD, Fiware compatible data model. The harmonization service includes a JSON-LD parser and an Orion-LD client and is attached to the analytical service. It is called ad hoc by the analytical service when new data is available for uploading (Fig. 7).

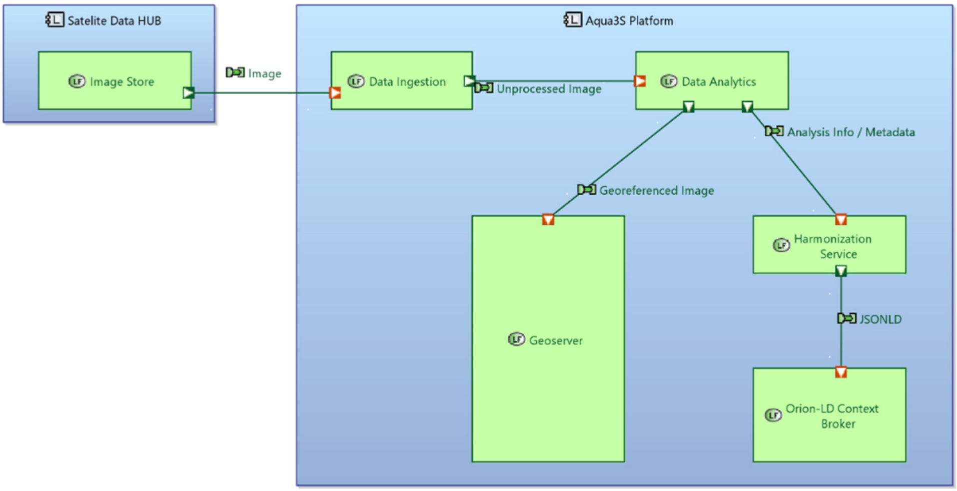


Figure 7: Ingestion and harmonization of satellite images.

4.4 Binary files

Large binary files include photographs and videos from drones or CCTV cameras operating at the water supplier's facilities or at important bodies of water. Such files are saved in the WebDAV server provided, while any connected metadata are harmonized and saved in the Orion-LD context broker, along with the location of the relevant file in the WebDAV server. If an end user, or analytical/visualization tools need to access the file they can query the broker with the relevant information and receive a link to the file in WebDAV. Fig. 8 shows a diagram of this architecture.

5 CONCLUSION

The ingestion platform is overall a lightweight and easy to install application based on open-source tools and unique integration solutions developed in Python 3. It is focused on being easily adaptable and expandable to different frameworks, providing libraries that can be reused in different applications as well as targeted configuration files that can be used to adjust the platform's operations in an intuitive manner.

The harmonization layer consists of a library used to parse Fiware compatible JSON-LD payloads, incorporating methods that can be used to expand the library to new data models if the need arises. It also incorporates an Orion-LD client that manages entity creation, updates, deletes and queries, as well as the handling of multitenancy and subscriptions.

Based on the heterogenous data that were available to the project in many different formats, the focus was given to providing an easily configurable and adaptable solution. With the harmonization of available data to the NGSI-LD standard and the indexing of binary data (e.g. satellite images) using the data models developed, the integration platform can be connected to multiple services and expanded in the future to include more data sources, or new data types and can support the data requirements of any client module (analytical, visualization etc) that might be developed in the future based on the NGSI-LD standard. The platform can be considered "data type agnostic" and can be easily adapted to other fields, beyond water quality and management.

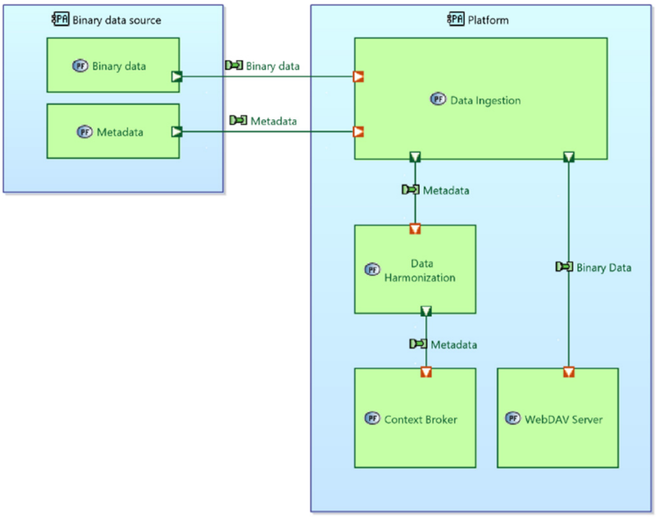


Figure 8: Schematic of binary data ingestion process.

Several steps have already been taken and will continue to be taken to allow for the fast, effective, and easy deployment of the platform and facilitate its connection with new data providers and consumers. Such steps include the standardization of development libraries that support data creation and queries as well as generalizations on the queries necessary for data retrieval by data consumers or analytical tools.

ACKNOWLEDGEMENT

This work is part of the AQUA3S project. AQUA3S has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.832876. Content reflects only the authors' view.

REFERENCES

[1] Garcia, D., Puig, V. & Quevedo, J., 2020. Prognosis of water quality sensors using advanced data analytics: Application to the Barcelona Drinking Water Network. *Sensors*, **20**(5), p. 1342, 2020. DOI: 10.3390/s20051342.

[2] Grbčić, L., Lučin, I., Kranjčević, L. & Družeta, S., A machine learning-based algorithm for water network contamination source localization. *Sensors*, **20**(9), p. 2613, 2020. DOI: 10.3390/s20092613.

[3] Jiang, W., He, G., Long, T., Ni, Y., Liu, H., Peng, Y., Lv, K. & Guizhou Wang, G., Multilayer perceptron neural network for surface water extraction in Landsat 8 OLI satellite images. *Remote Sensing*, **10**(5), p. 755, 2018.

[4] Altenau, E.H., Pavelsky, T.M., Durand, M.T., Yang, X., Prata de Moraes Frasson, R. & Bendezu, L., The Surface Water and Ocean Topography (SWOT) Mission River Database (SWORD): A global river network for satellite data products. *Water Resources Research*, **57**(7), e2021WR030054, 2021.

[5] Johns, R., Community change: Water management through the use of social media, the case of Australia's Murray–Darling Basin. *Public Relations Review*, **40**(5), pp. 865–867, 2014.

[6] Harou, J.J., Garrone, P.A.O.L.A., Rizzoli, A.E., Maziotis, A., Castelletti, A., Fraternali, P., Novak, J., Wissmann-Alves, R. & Ceschi, P.A., Smart metering, water pricing and social media to stimulate residential water efficiency: Opportunities for the smarth2o project. *Procedia Engineering*, **89**, pp. 1037–1043, 2014.

[7] AQUA3S. https://aqua3s.eu/. Accessed on: 28 Feb. 2022.

[8] Smart Data Models. https://smartdatamodels.org/. Accessed on: 28 Feb. 2022.

[9] The ETSI standard. https://www.etsi.org/deliver/etsi_gs/CIM/001_099/009/01.01.01_60/gs_cim009v010101p.pdf. Accessed on: 28 Feb. 2022.

[10] Heath, T. & Bizer, C., Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, **1**(1), pp. 1–136, 2011.

[11] Orion-LD repository. https://github.com/FIWARE/context.Orion-LD. Accessed on: 28 Feb. 2022.

[12] Cygnus-LD repository. https://github.com/telefonicaid/fiware-cygnus. Accessed on: 28 Feb. 2022.

[13] Geoserver website. http://geoserver.org/. Accessed on: 28 Feb. 2022.

[14] WebDAV website. http://www.webdav.org/. Accessed on: 28 Feb. 2022.

[15] Python version 3.6 website. https://www.python.org/downloads/release/python-360/. Accessed on: 28 Feb. 2022.

[16] EPANET website. https://www.epa.gov/water-research/epanet. Accessed on 28 Feb. 2022.