D3.1.2 Linked Water Dataspace

Project Acronym:	**Waternomics**
Project Title:	**ICT for Water Resource Management**
Project Number:	**619660**
Instrument:	**Collaborative project**
Thematic Priority:	**FP7-ICT-2013.11**

D3.1.2

Work Package:	*WP3*	
Due Date:		31/01/2016
Submission Date:		
Start Date of Project:		31/01/2016
Duration of Project:		36 Months
Organisation Responsible of Deliverable:		NUIG
Version:		1.8
Status:		Draft
Author name(s):	Edward Curry	NUIG
	Wassim Derguech	NUIG
	Souleiman Hasan	NUIG
	Umair Ul Hassan	NUIG
Reviewer(s):	Christos Kouroupetroglou	Ultra4
	Schalk-Jan van Andel	UNESCO-IHE
Nature:	☒ R – Report ☐ P – Prototype ☐ D – Demonstrator ☐ O – Other	
Dissemination level:	☒ PU – Public ☐ CO – Confidential, only for members of the consortium (including the Commission) ☐ RE – Restricted to a group specified by the consortium (including the Commission Services)	
Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013)		

Revision history			
Version	**Date**	**Modified by**	**Comments**
D3.1.2 1.5	Dec 2015	Umair Ul Hassan	Update based on advancement in implementation since D3.1.1
1.6	Dec 2016	Souleiman Hasan	Update based on advancement in implementation since D3.1.1
1.7	Jan 2016	Souleiman Hasan	Restructure based on internal review and update section 4
1.8	Jan 2016	Edward Curry	Updated section 1
1.9	Jan 2016	Wassim Derguech	Updated sub-sections 4.1.1, 4.2.1, and 4.3.1
2.0	Jan 2016	Umair Ul Hassan	Updated sub-sections 4.1.2, 4.2.2, and 4.3.2
2.1	Jan 2016	Souleiman Hasan	Updated sub-section 4.2.3
2.2	Jan 2016	Souleiman Hasan	Updated summary
2.3	Jan 2016	Souleiman Hasan	Updated formatting and layout

Executive Summary

This deliverable reports on the final design and implementation of the Linked Water Dataspace for the pilots. A dataspace is an emerging information management approach used to tackle heterogeneous data sources that support requirements, such as standardization, enrichment, and linking of data in an incremental manner.

Designing and implementing a dataspace is the core of the Waternomics information platform and is one of the major objectives for the Waternomics project. In the Waternomics vision collecting, standardizing, enriching and linking water usage data coming from sensors together with other relevant contextual data sources is a key step needed for effective analytics to drive decision making: e.g., planning, adjustments and predictions and to raise user awareness of water consumption.

In Deliverable 3.1.1 we have reported on the initial design and plans for the dataspace including background information and state-of-the-art analysis. In this report, we detail the design and implementation of the dataspace. The primary contribution of this report is a customized architecture for implementing a Linked Water Dataspace and how each pilot joins the dataspace and its data gets used from it. The primary lesson learned is that dataspaces can play a key role in distributed data production, processing, and consumption scenarios where a prior full control over the data flow is not guaranteed. The designed dataspace shows how a data source can join the space in a lightweight manner with the support of various services that facilitate its co-existence with other sources and potential consumers.

The reported dataspace builds on the lambda architecture which is aimed at Big Data processing using three layers: speed, batch, and serving. This architecture allows the processing of various types of data coming from water sensors or residing in building management systems. It allows the services and applications to transparently query data from these sources for further water domain-specific use cases. In this work we incorporated the Lambda architecture as a part of a bigger dataspace model. That means that additional dataspace support services can help supporting the lambda architecture. For instance, the WKAN catalogue can guide the serving layer of data by using meta-data and entities.

Table of Contents

List of Figures

List of Tables

List of Listings

1 Introduction

The goal of Waternomics is to explore how ICT can help households, businesses and municipalities with reducing their consumption and losses of water. A key component of the Waternomics information platform aims at collecting water consumption and contextual information from different sources to be used for effective data analytics to drive decision making: e.g., planning, adjustments and predictions and to raise user awareness of water consumption.

A key outcome of the work carried out in Work Package 3 consists of designing a Linked Water Dataspace. A dataspace is an emerging information management approach used to tackle heterogeneous data sources and that supports requirements, such as standardization, enrichment, and linking of data in an incremental manner. The primary contribution of this report is a customized first version of architecture for implementing such a Linked Water Dataspace. This architecture has been refined depending on the other tasks and work packages requirements.

1.1 Work Package 3 Objectives

The objective of Work Package 3 (WP3) is to develop the project software and user environment that delivers water information services to various targeted stakeholders. It allows linking sensors, data management systems and water meters. More specifically, the objectives of WP3 are as follows:

- To develop a linked dataspace able to capture and store data from various sources
- To develop a set of services based on the web as a platform allowing applications to interact and use the data stored in the linked dataspace
- To provide a set of applications consisting of various customisable and personalised components
- To deploy and customize the developed applications in appropriate sites for validation and evaluation

WP3 receives as input the high-level system architecture, usage and exploitation cases, and KPIs for reporting from WP1. WP3 provides as output the water information services platform for the pilots and future exploitation as a primary project outcome/result. In particular, WP3 develops and provides user friendly content, applications, and platforms that enable all stakeholders (utilities, commercial users and domestic users) to take decisions that result in reduced water consumption and losses, and increase overall awareness of drinking water supply issues.

1.2 Purpose and Target Group of the Deliverable

The objective of this deliverable report on the completion of Task 3.1 (Data sources adapters customization / development) and 3.2 (Linked water data infrastructure design). Both tasks contribute to the management of a dataspace containing water data that is used by support services for dedicated applications. We call this dataspace the **Linked Water Dataspace**. This data space has been refined with respect to the project evolution.

A **Linked Water Dataspace** is a central technological concept of the Waternomics information platform. Its role consists of providing an interoperability space with common access

mechanisms to the data. Important components of the dataspace are the data sources adapters that take as input sensors and other sources of data and provide a linked data-cloud rich with knowledge and semantics about the water consumption. These adapters require a predefined semantic model for describing both sensor data and meta-data.

The **semantic model for sensor data** is a formal ontology developed within the Waternomics project. This model has been designed with respect to the data models used by existing management systems.

Ontologies and adapters of the Waternomics project are proposed in order to:

- Create a linked data cloud,
- Facilitate data integration across multiple platforms,
- Facilitate data access and interoperability.

The main target groups for this deliverable are designers of water management systems as well as ontology designers and technicians. The minimal vocabulary shown in this document is also of interest to domain experts in water management.

1.3 Relations to other Activities in the Project

Figure 1 illustrates the relations of this deliverable to other activities in the Waternomics project. These relations are represented as links numbered from 1 to 4 and are described as follows:

Link 1: The design of the Linked Water Dataspace follows the set of dataspace requirements identified in Section 2.4.2 of the Deliverable D 1.3. These requirements are further refined and explained in Section 2.

Link 2 and 3: This deliverable reports on the design of the Linked Water Dataspace. This feeds into D3.2 (Support Services APIs and Components Libraries) and D3.3 (Waternomics Apps) as both require the technical details of the dataspace for the developments of support services and applications.

Link 4: Pilot planning in WP5 also uses the initial version of this deliverable which was communicated in deliverable D3.1.1 as it constituted a detailed background analysis and technical design work towards the Linked Water Dataspace.

Link 5: A previous deliverable (D3.1.1) has reported on the initial design in M12.

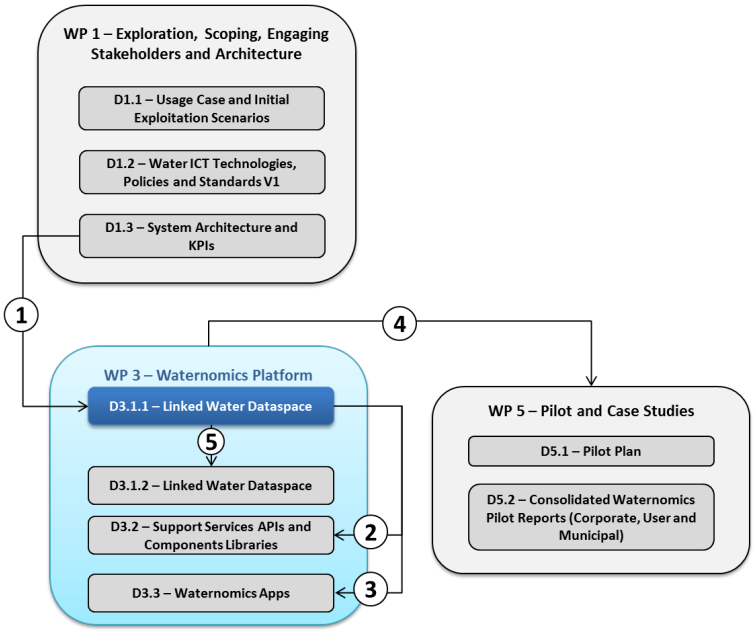

Figure 1: Relationships between D3.1.1 and other activities in Waternomics

1.4 Document Outline

The remainder of this document is organised as follows:

- Section 2: Functional Dataspace Requirements – defines the set of requirements for building the Linked Water Dataspace.
- Section 3: A Realtime Linked Water Dataspace- defines the architecture for managing water data and exposing it into the Linked Water Dataspace.
- Section 4: Concrete Pilot Cases- presents concrete cases for datasets joining the dataspace in various pilots: Galway's NUIG engineering building and CnaC school, Thermi, and Linate airport.
- Section 5: Summary- concludes the deliverable.

1.5 About Waternomics

Climate change, increased urbanization and increased world population are several of the factors driving global challenges for water management. In fact, the World Economic Forum has cited "The Water Supply Crises" as a major risk to global economic growth and environmental policies in the next 10 years. In parallel, the United Nations has called for intensified international collaboration. To help reduce water shortages, Waternomics explores the technologies and methodologies needed to successfully reduce water consumption and losses from households, companies and municipalities. Waternomics is a three year EU-funded project that started in February 2014 that develops and introduces ICT as an enabling technology to manage water as a resource, increase end-user conservation awareness and affect behavioural changes, and to avoid waste through leak detection. In saving water, energy is also conserved (treatment and pumping) as is the CO_2 associated with energy production. Unique aspects of WATERNOMICS include personalized feedback about end-user water consumption, the development of a methodology for the design and implementation of systematic and standards-based water resource management systems, new sensor hardware developments to make water

metering more economic and easier to install, and the introduction of forecasting and fault detection diagnosis to the analysis of water consumption data.

WATERNOMICS is demonstrated in three high impact pilots that target three different end users/stakeholders:

- Domestic users in Greece implemented by a water utility
- Corporate operator in Italy provided by a major EU airport
- Public and Mixed-use based demonstration in Ireland

Through these contributions, WATERNOMICS will pioneer a new dialogue between water stakeholders. It will enable the introduction of Demand Response principles and open business models through an innovative human centric approach that uses personalized water data, water availability based pricing, and gamification of water usage statistics. To maximize impact, the project highlights business development, exploitation planning, and outcome oriented dissemination.

1.6 Updates on Deliverable D 3.1.1

This deliverable is an updated version over the deliverable D3.1.1. The main work tackled since the submission of D3.1.1 has been to put the initial designed architecture reported in D3.1.1 into practice with implementation across pilot sites. Additional aspects reported in D3.1.1 such as support services have been reported in D3.2-Support services API's and Components' libraries. The main updates committed in this deliverable are:

- The CATALOG service implementation updates, the use of CKAN and its adaptation into WKAN, how data sources and entities are managed. This is done in Section 3.2.
- The QUERY service architecture updates the actual implementation and the realization of the Lambda architecture in Section 3.3.
- The steps towards joining the dataspace and updates as implemented. This is shown in Section 3.4.
- The concrete use case implementation in all pilots: Galway's NUIG engineering building and CnaC school, Thermi, and Linate airport, using Kafka and Druid as shown in Section 4.

The primary lesson learned is that the dataspace can play a key role in distributed data production, processing, and consumption scenarios where a prior full control over the data flow is not guaranteed. The reported dataspace builds on the lambda architecture which is aimed at Big Data processing using three layers: speed, batch, and serving. In this work we incorporated the Lambda architecture as a part of a bigger dataspace model. That means that additional dataspace support services can help supporting the lambda architecture. For instance, the WKAN catalogue can guide the serving layer of data by using meta-data and entities.

Background information on Linked Data and the state of the art of dataspaces as well as support information for the design of the dataspace have been communicated in Deliverable D3.1.1.

2 Functional Dataspace Requirements

Deliverable D1.3- Section 2 identified the functional requirements as well as the set of generic architecture functions in that the Linked Water Dataspace should respect when dealing with input data collected from sensors and other data sources. These requirements also specify how the output data is made available to support services and applications. In this section, we carry out a further analysis of these requirements and propose the following list of functional requirements for the dataspace:

- **Standardisation:** The system serves as a common dataspace for different stakeholders. Consequently, the data exchanged and published by the system should be standardised. The same data and support services are available to all applications. With respect to this requirement we use RDF for describing the various entities managed within the system. An analysis of existing RDF data adapters is reported in D3.1.1.

- **Consuming Open Data:** By definition Open Data *"can be freely used, modified, and shared by anyone for any purpose*" [1]. In our context, the system should be able to make use of relevant open data assets for proper analytics. Possible scenarios for consuming open data include the prediction of water consumption using open weather data. Consuming open data requires a proper selection and evaluation of data source in order to select the most suitable one for proper decision support [2].An initial survey of Open Data sources for water management are presented in D3.1.1.

- **Publishing Linked Data:** The data produced by adapters or support services should be published in the dataspace with respect to linked data principles[1]: available on the web, structured, not using a proprietary format, using Uniform Resource Identifiers (URIs) to denote entities and linked to other data sets [3]. Details about these principles are discussed in D3.1.1 and Section 3.3.

- **Data Linking:** When publishing water data to the dataspace, it has to be linked to other data sets. This linking is very useful for ensuring an optimal data management and integration. It helps enhancing their (re)use and discovering new knowledge from water data put into a wider context. It is important to assess and determine what data sets are relevant to be linked with water data. D3.1.1 provides further details about this requirements and how it is covered by the dataspace.

- **Real-time data / events:** The system will be handling continuous streams of data coming from multiple sensors and data sources. The system should be able to manage large quantities of data in real-time. Real-time processing of data requires the development of algorithms and tools for parallel processing of simple and complex events see D3.1.1 and D3.2.

- **Real-time Analytics:** Data analytics needs to be continuously made in order for consumers' applications to make timely decisions i.e. the analytics should make use of the latest data available. Speed layer from the Lambda architecture is covering this requirement. Details are given in Section 3.

[1] http://5stardata.info/

- **Data integration:** The system analytics should integrate both real-time and historical data for effective decision support. Therefore, the system should be capable of seamlessly analysing water consumption and provide integrated view of the data being analysed. The use of Lambda Architecture covers this requirement as in essence; it facilitates the integration of real-time (speed layer) and historical data (batch layer). For a discussion of existing implementations of the Lambda architecture, we refer to D3.1.1.

- **Heterogeneity of Sensor Data Events:** The system handles data from a wide variety of sensors and consequently a wide variety of data formats. The dataspace needs to be able to handle applications' queries across data formats with respect to their semantic similarity. Additionally, the dataspace should manage data produced by developed services from other work packages such as leakage detection data from WP4.

- **Enrichment of Sensor Data Events with Open Data:** Raw sensor data reports mainly on the observed values of a particular property. This data requires additional contextual information, such as the location of the sensor, in order to deliver accurate decision analytics. The dataspace needs to enrich sensor data with relevant information that are required by support services and applications.

In order to cover these requirements, we propose to use Linked Data as a set of best practices for publishing, sharing and interlinking structured data on the web. In such context, if all the data of our infrastructure is open and linked to other open data sets from the web, it would be easier to create a water information system combining various distributed data repositories. Thus this would enable access and sharing of water data without barriers for building effective services and application.

3 A Realtime Linked Water Dataspace

Deliverable D1.3-Section 3 discusses an overall system architecture that contains three main layers: the hardware, the data, and the software, as shown in Figure 2.

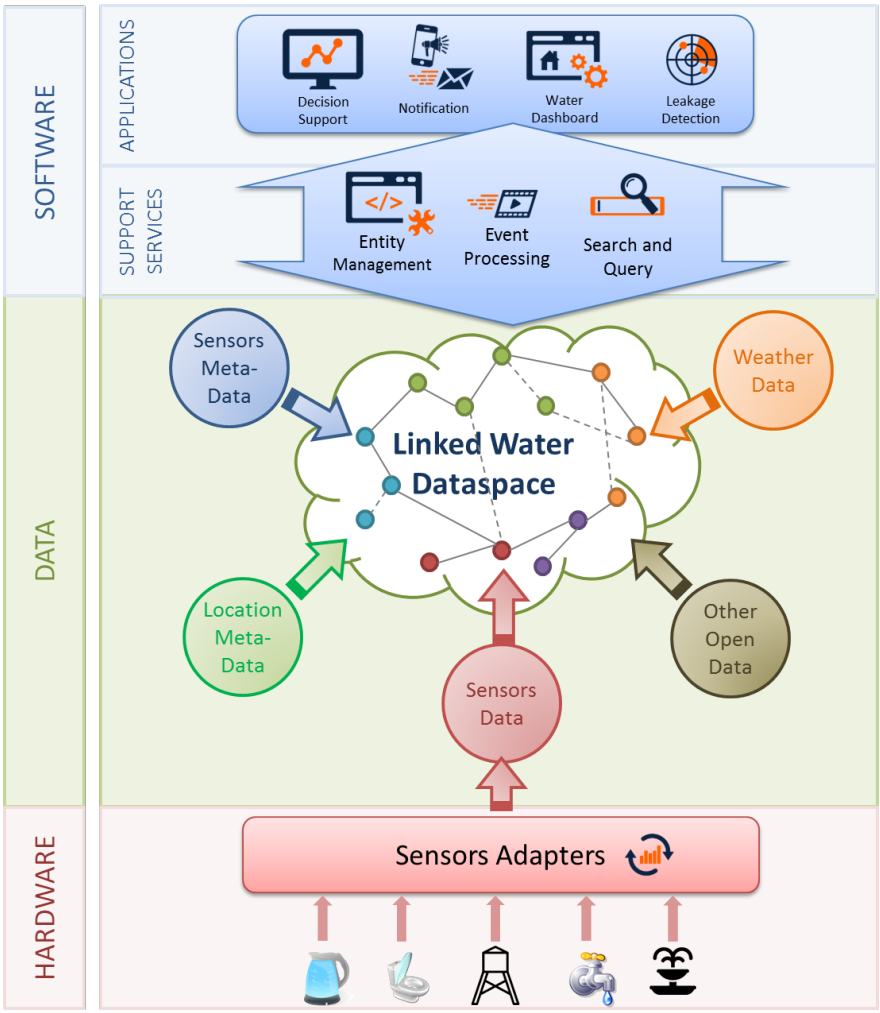

Figure 2: System Architecture.

In this section, we dig deeper into the data layer that is represented by the Linked Water Dataspace. By reviewing the current state of the art, we adopt the following approaches for handling the requirements in the Linked Dataspace as shown in Table 1.

Table 1: Used Approaches for the Dataspace

Requirement	Approach
Standardisation	Standard compliant ontologies such as SSN are used for describing sensors and their data. An entity management service that standardizes critical entities and maps their identifiers in different data sources. We extend CKAN data catalog for this purpose.
Consuming Open	One of the support services that are included in the dataspace consists of

Data	crawling weather data and integrating it in prediction services. Various open source data can have various types/formats, and an app might not know which open dataset is relevant to a particular task. Users register each source in CKAN data catalog as an explicit indication of joining the dataspace. We provide custom adaptors for each data source to join the dataspace.
Publishing Linked Data	The produced data by every support service is published in RDF using linked data principles. We register each RDF source in CKAN data catalog as an explicit indication of joining the dataspace.
Data Linking	Relevant open data sources are integrated in the dataspace through explicit links between entities. We maintain high level mappings in CKAN data catalog and low level mappings maintained by each source.
Real-time data / events	The dataspace is designed with respect to the lambda architecture that covers real-time event processing. We use an event processing engine for managing live sensor events, and deploy a scalable message oriented middleware for passing data between real-time sources and applications. We use middleware based on published/subscribe pattern, such as Apache Kafka, and Apache Spark Streaming for real-time aggregation jobs.
Real-time Analytics	Real-time analytics of data is considered in the platform as part of the speed layer of the lambda architecture. They seamlessly serve real-time and historical data in aggregated form to applications. We implement the serving layer of Lambda architecture using Metamarkets DRUID cluster.
Data integration	Lambda architecture in essence was used to deal with both historical and real-time data. This data integration is insured by relevant support services and follows a pay-as-you-go paradigm to ingest, integrate, and aggregate real-time data. We implement the speed and batch layers of Lambda architecture by employing Apache Spark for ingestion and aggregation of both real-time and historical data.
Heterogeneity of Sensor Data Events	The platform is designed to handle a variety of sensor types. Consequently for each sensor type, a dedicated collector is designed for collecting data and converts it into RDF for further processing. An event engine uses an approximate semantic matching model [4]–[11] to process sensor events.
Enrichment of Sensor Data Events with Open Data	Sensor readings are further enriched by dedicated support services using open data.

3.1 Dataspace Overview

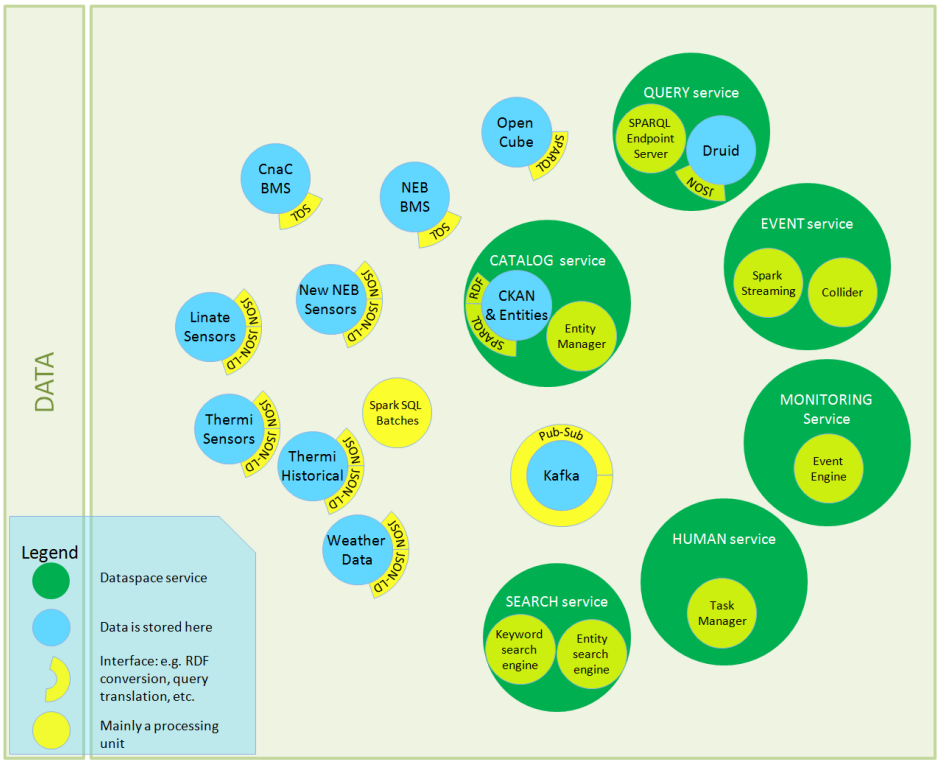

Figure 3: Linked Water Dataspace Overview- Components View (Componenets interact as shown in Figure 4 for instance)

The Linked Water Dataspace shown in Figure 3 is a collection of water datasets along with a set of services that supports the dataspace. The dataspace is designed to be an incremental view of how water datasets join the computational space targeted by applications. In contrast to the classical one-time integration of datasets that causes a significant overhead, the Linked Water Dataspace adopts a pay-as-you-go paradigm. Water datasets join the space in an incremental manner: the more interfaces they expose, the more links they provide, and the more linked dataspace services they support, the more integrated into the dataspace they become.

The diagram in Figure 3 illustrates components view of the Linked Water Dataspace. We can recognize three main concepts:

1. **Datasets** such as weather data, water sensor data, building management system data, etc. Those form the actual content of the water dataspace. They are the basics for all the insights that can be drawn from the dataspace. Datasets can join and leave the dataspace. In fact, that can very dynamic such as in the case of dynamic sensor environments. Joining the dataspace requires some "cost" to be paid. This cost takes various forms: the registration into a catalog so the dataset becomes visible by others, the conformance to a schemata or the mapping to other schema, the exposure of data into a set of formats such as the RDF serialization JSON-LD, etc. We adopt in the Linked Water Dataspace a pattern in which the publisher of the data is mainly responsible for paying this cost. That is a very pragmatic feature as it allows the dataspace to grow and enhance gradually. It is a quite scalable concept, and is followed in large scale environments such as the Web.
2. **Adapters** (or Interfaces) are the technical facades of the datasets that other members of the dataspace can talk to. Such facade are a way to quantify the degree of involvement of the

dataset in the dataspace. For example, a dataset that provides a JSON-LD interface allows structured queries to be executed, and thus it is superior and more integrated into the dataspace than a document that is only exposed by keyword search.

3. **Linked Dataspace Services** which form the backbone of the dataspace. While data is crucial for the space, support services are the platform that allows datasets to be visible, query-able, integratable, searchable, monitorable, and curatable. For instance, for a sensor to become visible to the dataspace, it must register itself in a catalog along with some information on how to get its data, how often and precise it is, etc. Such a service is provided by the CATALOG service in Figure 3.

Besides those three concepts, there is also the concept of a **Relationship** between two datasets, or between a dataset and a service. Relationships are omitted from Figure 3. for clarity, but an example would be a "replica" relationship between Druid and Data Cube for some datasets as will be discussed later on in this Section.

Applications surround the dataspace and make use of its support services to interact with the datasets. For instance, a Realtime dashboard checks out the real-time sources from the CATALOG service, and then consume data from the EVENT service. A data analytics app discover data from the CATALOG service and runs analytics algorithm over dimensional data queried through the QUERY service.

The Linked Water Dataspace emphasizes six main services; two of them are core and the focus of this document, while the others are support services and the focus of Deliverable D3.2. Services align with the requirements in Section 2 as shown in Table 2

Table 2: Linked Water Dataspace Support Services and Requirements

	3.1.1.1 Dataspace Service	3.1.1.2 Requirements									3.1.1.3 Enabling Technologies
		3.1.1.4 Standardisation	3.1.1.5 Consuming Open Data	3.1.1.6 Publishing Linked Data	3.1.1.7 Data Linking	3.1.1.8 Real-time data / events	3.1.1.9 Real-time Analytics	3.1.1.10 Data integration	3.1.1.11 Heterogeneity of Sensor Data Events	3.1.1.12 Enrichment of Sensor Data Events	
	CATALOG	+	+	+	+			+			CKAN
	QUERY	+		+			+	+			DRUID, SPARQL
	EVENT		+			+	+		+	+	COLLIDER, SPARK Streaming

	MONITORING		+			+					Event Processing Engine
	HUMAN				+			+			IMIRT Task Manager
	SEARCH				+						Entity and keyword search engines

The CATALOG and QUERY services are core to the Linked Water Dataspace as are usually emphasized in dataspaces literature [12], [13]. We developed the concepts and implementations of the Linked Water Dataspace based on these two services as described in this document. Further development and addressing on the remaining services is the subject of the future Deliverable D3.2 as discussed in Section 3.5.

3.2 The Dataspace CATALOG Service

The catalog is the central registry of the Linked Water Dataspace. Within the catalog all datasets and data sources are declared along with several meta-data about them. The meta-data can include (i) details about type of data, (ii) ownership of data, and (ii) the entities such as sensors or locations. The catalog might also contain open data sources that are relevant to water management such as weather observation stations or forecast services. More specifically and as illustrated in, the Waternomics catalog service provides and point of discovery for all the data contained in various sources:

- Historical Sensor data from a BMS. For example, the historical data for NEB (Galway pilot) stored in Amazon Cloud Storage.
- Real-time VTEC sensors installed in pilot sites
- Open Data from web sources such as the Weather Data.
- Real-time data from other pilot sites. Such data is sent to the dataspace through a RESTfull API[1] to the Kafka middleware.

The catalog service for the Linked Water Dataspace is built upon the CKAN portal. Therefore, it is appropriately named WKAN. In essence, WKAN serves a critical role of creating a data portal that is open and available online for developers and users of Waternomics project. WKAN extends the original CKAN portal with functionality of entity-centric view of data sources in Linked Water Dataspace. Registering at the catalog is the first step for a dataset or data source that joins the dataspace, as described in Section 3.4. The registration process entails detailing of meta-data which would help users of the catalog in locating and using the data source. Besides the core functionality of cataloguing the datasets and data sources, the WKAN also adds further information items to help organize Linked Water Dataspace.

[1] Example can be seen here: http://linkeddataspace.waternomics.eu:8003/message?topic=vtec.eindhoven.json

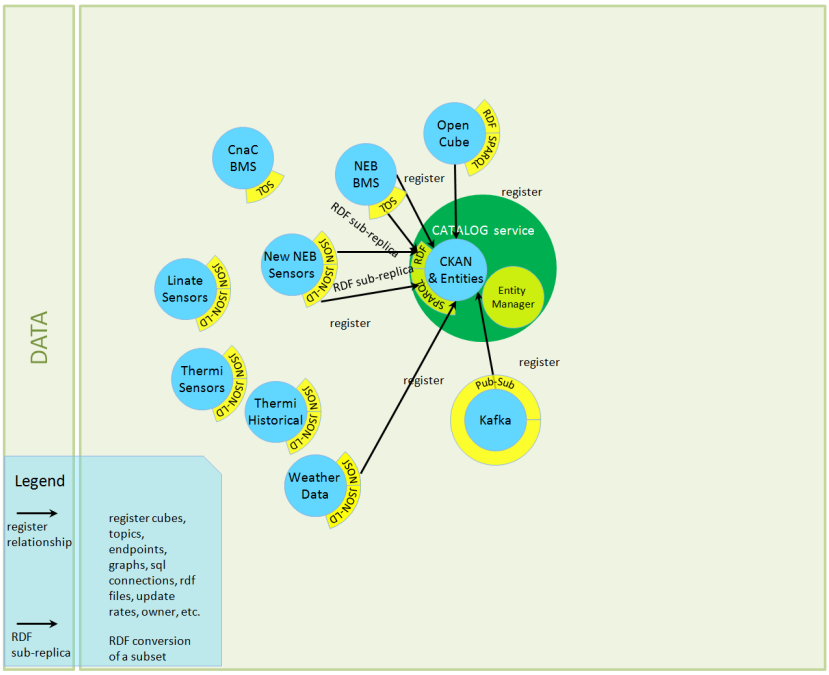

Figure 4: Linked Water Dataspace- The CATALOG Service View

Figure 4 shows the four types of information items stored in WKAN, as described below:

- **Datasets:** These information items describe either a static dataset or a dynamic data source. A dataset might contain contextual information about a building at pilot site. A data source might contain dynamic data generated by sensors installed in the building. Meta-data about a dataset includes source of dataset, state of dataset, timestamp of last update, data format, etc.

- **Organizations:** A pilot site in Waternomics projects is defined as an organization that groups a set of datasets together. For instance, all the datasets and data sources for Galway pilot are grouped together into organization named "NUI Galway". This grouping also enables separation between open and closed datasets. Datasets which are part of an organization can be declared private to hide them from general public access.

- **Entities:** An entity defines a concrete instance of a concept with in the Linked Water Dataspace. For instance, a sensor or a water outlet. WKAN catalogues critical entities in the Linked Water Dataspace and links those entities with the datasets which contain further information about entities. Meta-data about an entity includes identifier, entity type, associated datasets, etc.

- **Applications:** Applications are the descriptions of software and hardware that utilizes datasets from the Linked Water Dataspace. For example mobile applications, public displays, blog posts, web applications, dashboards, etc.

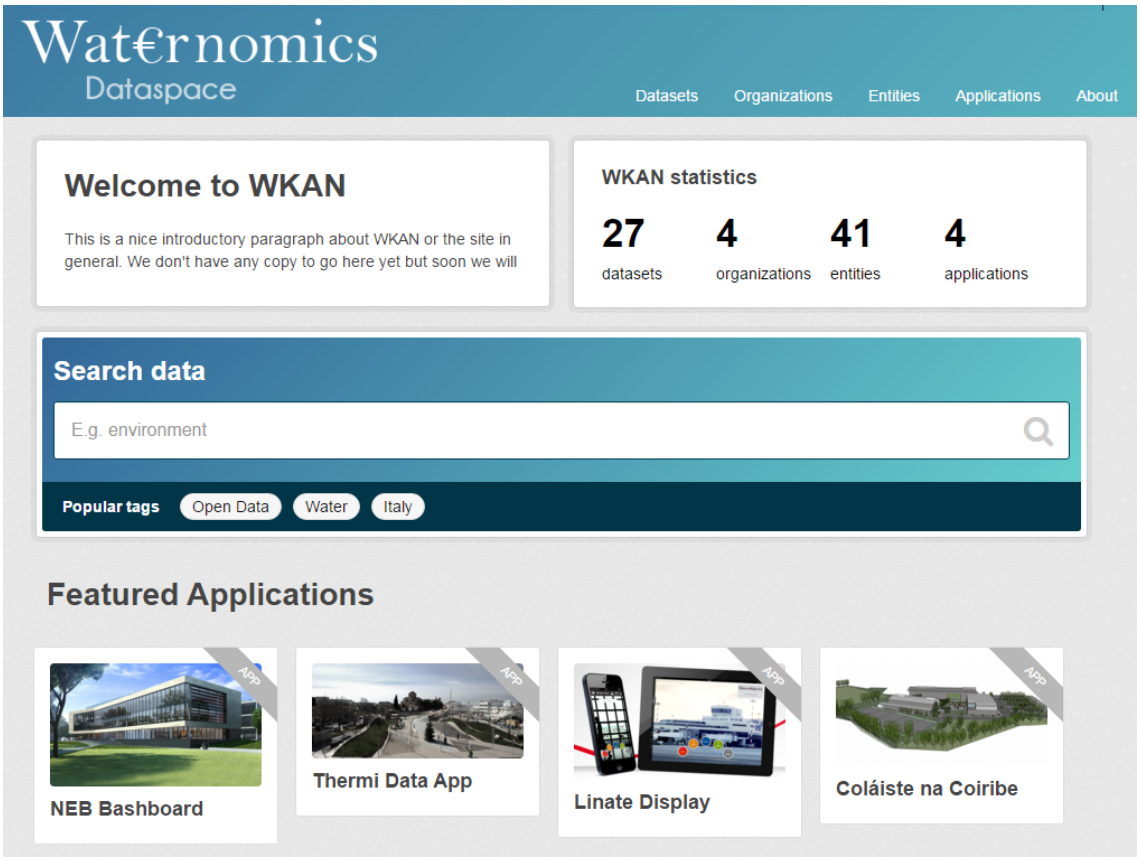

Figure 5: Screenshot of the WKAN catalog service

3.2.1 Datasets and Data Sources

The information about water management and its related entities is generally spread across various systems in an organization. The LWD is no exception to this observation, but a majority of data about water management can be found with the facilities management department or a similar unit. Building management systems (BMS) are a prime example of such information source. Modern BMS include a variety of sensors and actuators to control the behaviour of a building depending on the environment and water usage. Understandably, the BMS is considered the main source of information about sensors, outlets, and sites. However, this assumption might not be correct is various cases. It is not common to have BMS installed for large and old buildings. Similarly, airport and cities also lack the infrastructure required to install a BMS. In such cases, it is common to have individual databases for sensors, outlets, and sites. Each of those databases might have different data formats and management processes. It is essential to have such information integrated and linked through common semantics.

The Linked Water Dataspace consists of information related to the water consumption as well as the information that supports the decision-making process. In this regard, information from other sources is brought within the dataspace. Such sources generally exist outside an organization's control and require source specific data collection processes. For instance, it is now common for the weather websites and meteorological to provide interfaces for exporting their data. The interface can be in the form of RESTful API or comma separated text files. Essentially, these data sources provide information which might not be critical to dataspace but provides useful contextual information. For example, an undergraduate open day at the university might indicate

the higher than usual water consumption. A BMS does not have such contextual information in its database. In short, applications might require data from external and open sources as well.

So far we have discussed the information stored in typical databases. However, the information contained in databases might become inaccurate with the passage of time. Most importantly, the databases might not contain the complete information. To counteract such issues, it has been used to include the users of the dataspace in catalog management process. However, this is not a trivial task which might need careful consideration. Nonetheless, the inclusion of the users in the information curation process not only helps in maintaining accurate data but might also help in building user trust and ownership of the system. An appropriate process is required to source data from users. Such a process is the focus of the HUMAN service for LWD, which is further discussed in the Deliverable D3.2.

As soon as a dataset is defined in WKAN catalog, it becomes part of Linked Water Dataspace. The WKAN provides search and browse functionality for its user to navigate through datasets. shows a screen capture of the search and browse interface of WKAN. A search request with the term “Galway” reveals 4 data sets. We describe here only the first 3 data sets as they appear on . The first one concerns the NEB (Galway pilot) water sensors. The data format for this dataset is SQL, indeed, this describes the access to the NEB SQL database for collecting water sensors data. The second data set relates to the Galway city current weather observations taken from openweathermap.org. Access to the data, as described in the portal, is done through HTTP access and the returned data is in JSON format. The third data set is the weather forecast for Galway from met.no. It describes and provides the API link to the weather forecast data to be collected by relevant open data services.

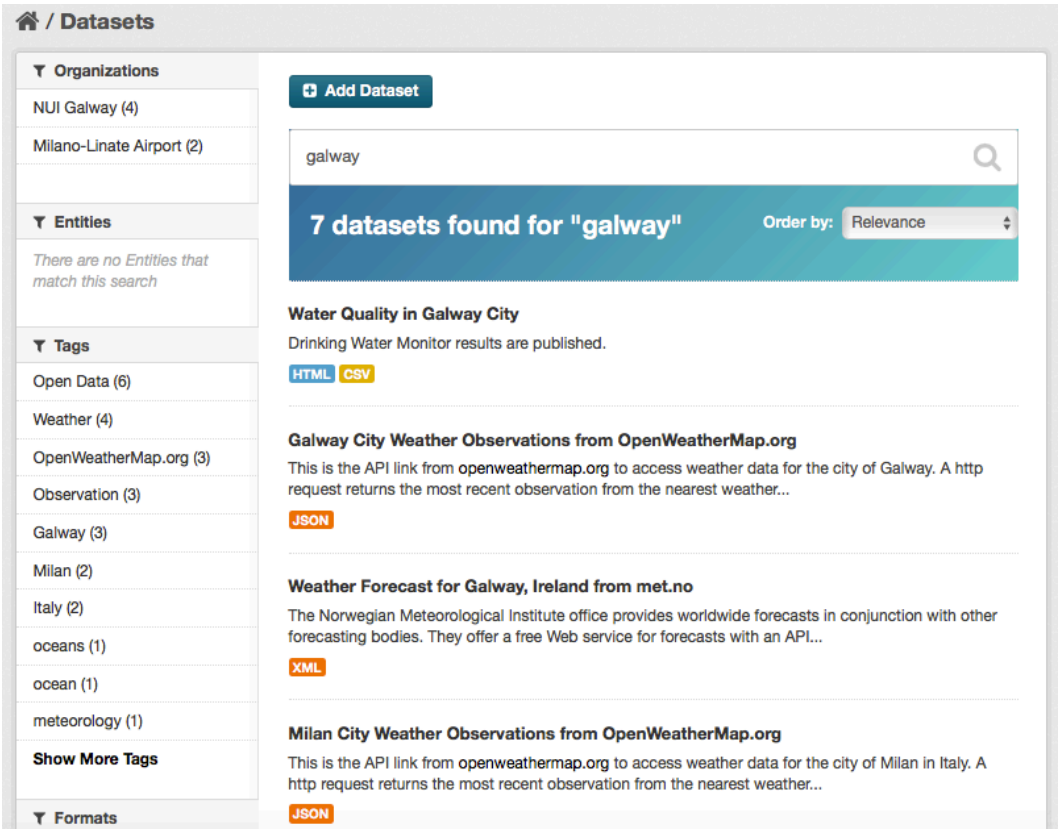

Figure 6: Search and browse in WKAN – data sets related to “Galway”

Further details about a dataset are viewable by selecting an individual dataset. For example, Figure 6 shows the set of rules defines for observing the retention time period for drinking water fountains in NUIG pilot. The meta-data reveals the URL to access location of the dataset, as well as additional information such as the timestamp when dataset joined the dataspace, format to be used while accessing data, etc.

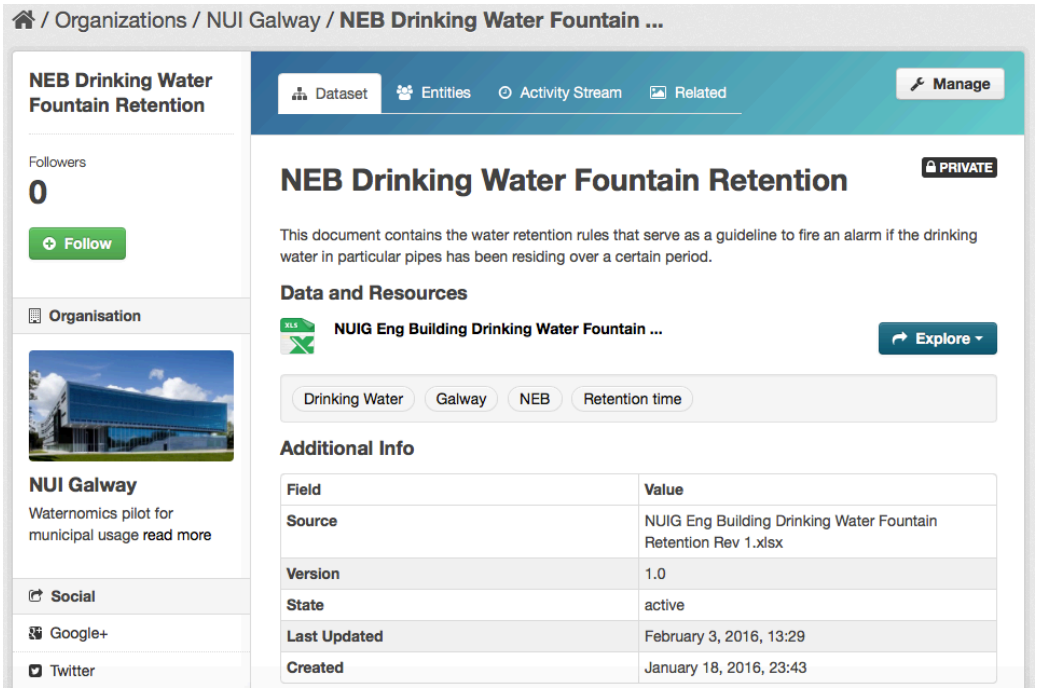

Figure 7: NEB Historical Logs in Waternomics dataspace

3.2.2 Managed Entities

An important aspect of WKAN catalog is the treatment of entities as the first-class citizen of Linked Water Dataspace. Managing information about the critical entities is one of the primary requirements in any information system. This is due to the fact that all decision-making applications rely on accurate entity information for delivering their functionality. The real-time Linked Water Dataspace (LWD) is no exception to this requirement. The entity management process is concerned with the maintenance of information about entities critical to the water data management and analysis. The expected outcome of this process is a database that severs as the canonical source of meta-data of entities for water management. In the case of Linked Water Dataspace, the primary set of entities includes sensors and their locations.

Beside these entities, the dataspace applications might also require information about the water users, groups, buildings, and water outlets. In short, all of the information that can help in understanding the water consumption, through association with real-world objects, is included in the entity management process. However, the level of quality control might be differentiated depending on the criticality of an entity for dataspace applications. This highlights the need for appropriate entity management process both in terms of software and operational guidelines. Figure 8 shows an example of an entity defined in WKAN, along with its critical meta-data. In this case, the sensor “M3” is defined and a resource identifier is assigned to the sensor.

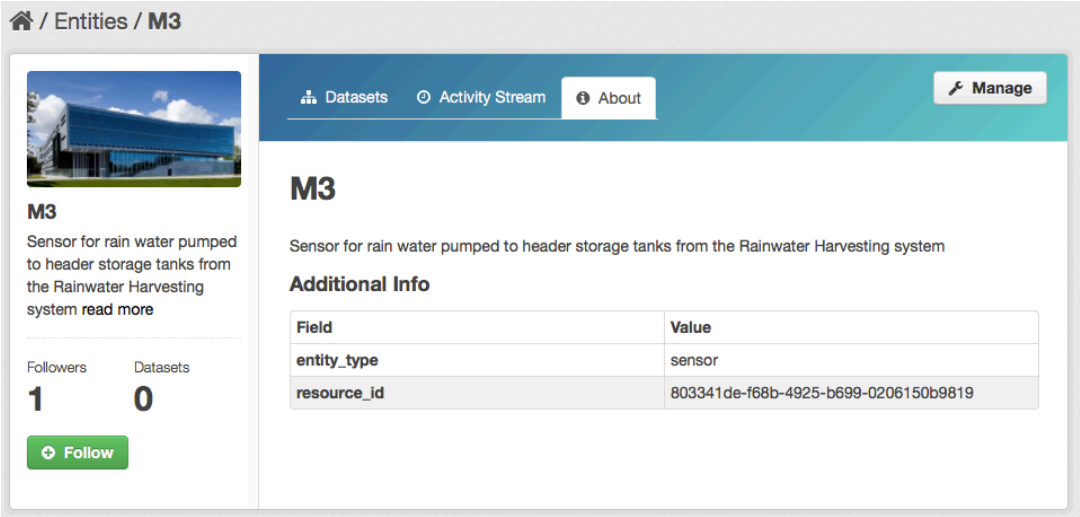

Figure 8: Example of a sensor entity in WKAN catalog

The resource identifier is used to access detailed data about M3. As shown, each dataset is associated with datasets that contain information related to sensor M3. For example, the BMS in Galway pilot contains daily readings generated by the sensor M3. Therefore, the BMS data source is associated with sensor M3 in a catalog. Such mappings between entities and datasets are extremely useful for dataspace users; whether its end users or application developers.

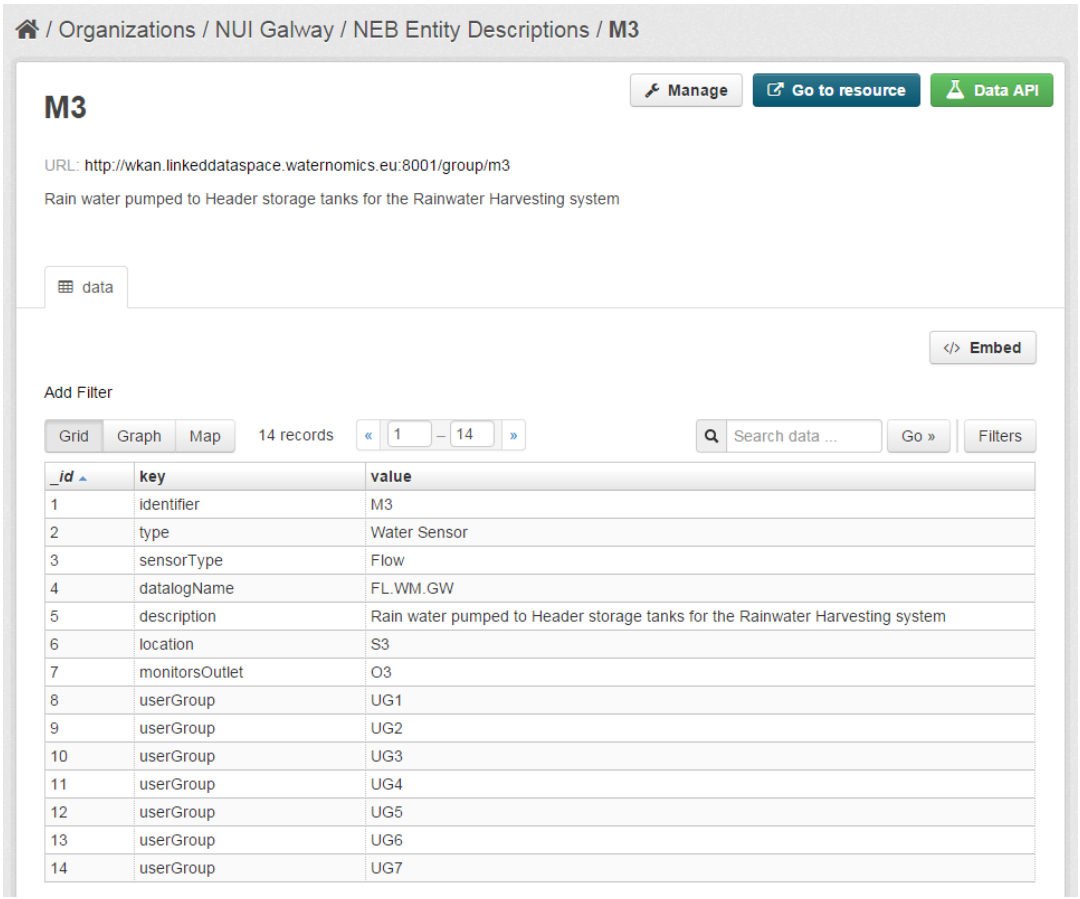

Figure 9: Detailed meta-data of a sensor entity in WKAN catalog

Figure 9 shows the detailed meta-data about sensor entity M3. The meta-data is uploaded to the WKAN as a JSON resources file during the entity definition process. The resource identifier for the JSON file is associated with entity, as discussed earlier. As can be seen, the detail meta-data lists association of the entity M3 with other entities such as water outlet (O3), location (S3), etc. These associations provide further contextual information for applications using Linked Water Dataspace.

In the following we describe example entities and their minimal set of attributes using standard RDF vocabularies.

- **Sensors:** The sensors that measure the flow of water generate the streams of data at different intervals. This data is used to calculate the water consumption levels of the area covered by a sensor. Generally, different types and forms of sensors are installed for metering water consumptions. Therefore, it is necessary to exactly describe each sensor, its capabilities, and its coverage. The sensor description may include the identifiers and labels. The sensor capabilities can include information like the units of measurements and rate of measurement. In Table 3 we describe some example attributes of a sensor entity. The attributes are stated in the RDF using the Dublin Core Metadata Initiative Terms (dct:) vocabulary.

Table 3: Required attributes for the entity Sensor

RDF property	Example Value
dct:type	Sensor
dct:identifier	M12
dct:title	Sensor 12
dct:description	Sensor 12 on Second Floor
dct:coverage	S2

- **Water Outlets:** Besides sensors, information about the actual physical water outlets is also important for analysis and decision making. It is possible that a single sensor might be installed for a set of outlets. In such cases, a cumulative assessment for water consumption is needed Table 4 shows the minimum set of attributes required for each entity that represents an outlet.

Table 4: Required attributes for the entity Outlet

RDF property	Example Value
dct:type	Outlet

dct:identifier	O2
dct:title	CW Tap - Floor 2
dct:description	Cold water tap on Floor 2

- **Location:** Information about sensors and outlets is not useful without the information about their associated spatial locations. For instance, it is difficult to perform meaningful analysis when only the water consumption of a pipe or a tap is known. Water flow of a pipe along with the locations serviced by the pipe constitutes more actionable information. Therefore, we consider sites as the third critical set of entities in the LWD. For simplicity, we assume that each sensor covers one or more sites and each outlet is installed at single site. Table 5 lists the minimum set of attributes required for each site entity.

Table 5: Required attributes for the entity Site

RDF property	Example Value
dct:type	Site
dct:identifier	S2
dct:title	Floor 2
dct:description	The Second Floor in NEB

3.3 The Dataspace QUERY Service and the Realtime Aspect

The architecture for the Linked Water Dataspace shown in Figure 3 is designed to meet the requirements for integration of multiple sources for water management, with various levels of data updates, coming from legacy data or from sensors. The QUERY service as defined here covers various aspects and processes of the dataspace that lead to the data being structuredly queried by the applications. The QUERY service also addresses low latency and fault-tolerant data analysis. The used architecture to realize The QUERY service follows the Lambda Architecture recommendations [45].

The Lambda architecture [45] is a concept which is getting acceptance by Big Data researchers and practitioners. Herein, the Lambda architecture realises the need for integrating water data within a data warehouse that is processed by batch processes and views that are pre-computed for fast access by applications. The Linked Data principles serve as a mediator to improve the integration within the batch layer for relatively not fresh data.

Besides, the architecture realises the need for a real-time aspect of the water data. Such an aspect could be crucial to support decisions relevant to fast detection of situations of interest such as leakage, and react accordingly. That is achieved by the speed layer that effectively works over streams of linked water data. Streams are not actually stored but rather processed in flow to guarantee a low latency view of the data that can complement the older views achieved by the batch layer.

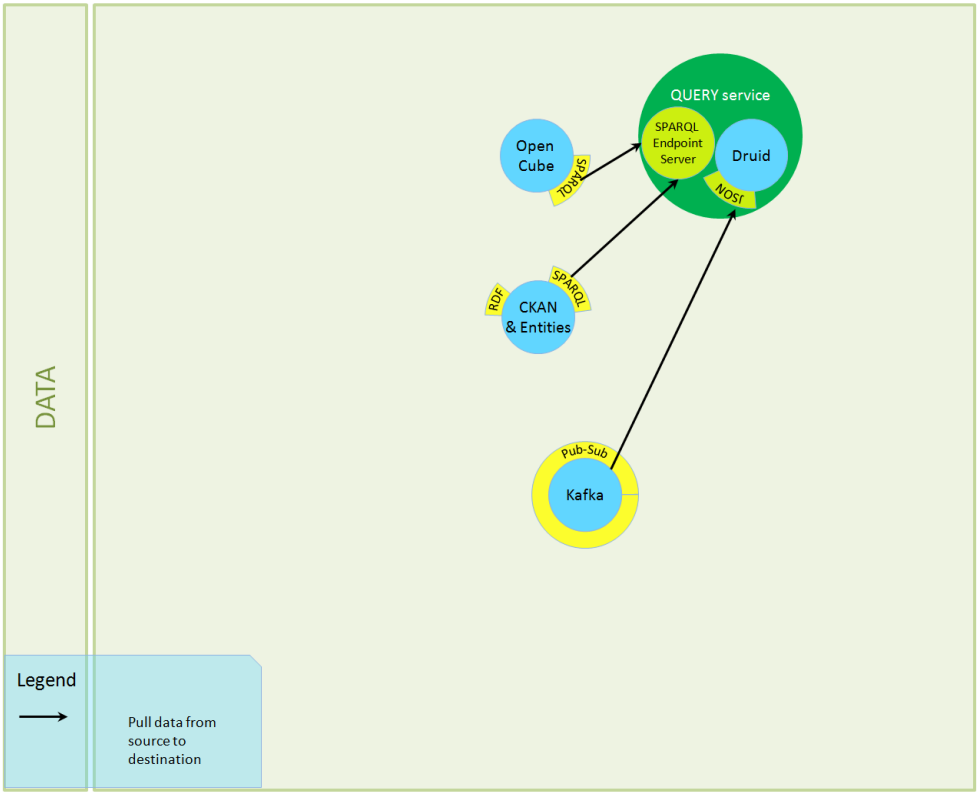

Figure 10: Linked Water Dataspace Lambda Serving Layer-- The QUERY Service View

To provide applications with a single interface for data access, a serving layer is provided. This forms the query interface for the applications and is the visible facet of the dataspace QUERY service as shown in Figure 10. This layer splits queries to the batch and speed layers to combine pre-computed results from the batch layer, with near-time fresh and maybe approximate results from the speed layer.

The query results are combined and transparently returned to the user. The following components are the defining keys of the Water Lambda Architecture and they allow data from a Pilot site to engage in the dataspace as shown in Figure 11:

- **Water Datasets**: those include the building management systems, legacy data such as water consumption logs or bills, as well as sensors installed on the pipes for instance.

- **Adapters**: those serve as the first step to normalise the water data into the Linked Data realm. They respect the semantic model and convert data items such as sensor events into their corresponding canonical linked data form. Details about the first version of the vocabulary used in this project are discussed in D3.1.1.

- **Management layer**: it is responsible for providing the functionalities and services essential for managing metadata such as ontologies as well as entity data such as people,

places, etc. This layer is handled within the CATALOG service as discussed in Section 3.2.

- **Batch layer**: it provides batch-based processing of the Linked Dataspace for accurate, but with some delay, data views, such as aggregation-based water data views, and basic analytics. This layer is exemplified in the processing of water consumption data from the building management system as discussed in Section 3.

- **Speed layer**: it provides real-time processing for water data with low latency processing such as approximate event matching, data enrichment and complex event processing. Details about real-time sensor data processing are provided in Section 3.3.

- **Serving layer**: it provides a transparent query of data from batch and speed layers, along with mid-level services such as activity detection and prediction services. This serving layer is realized via Druid and described in high level in Section 4, while it is detailed as part of deliverable D3.2.

- **Applications**: applications interact with the architecture via the serving layer and gain access to multiple water data views so their processing can take place starting from there, either of simply presenting data in user interfaces, or providing further processing capabilities and control mechanisms. Applications are covered in the deliverable D3.3.

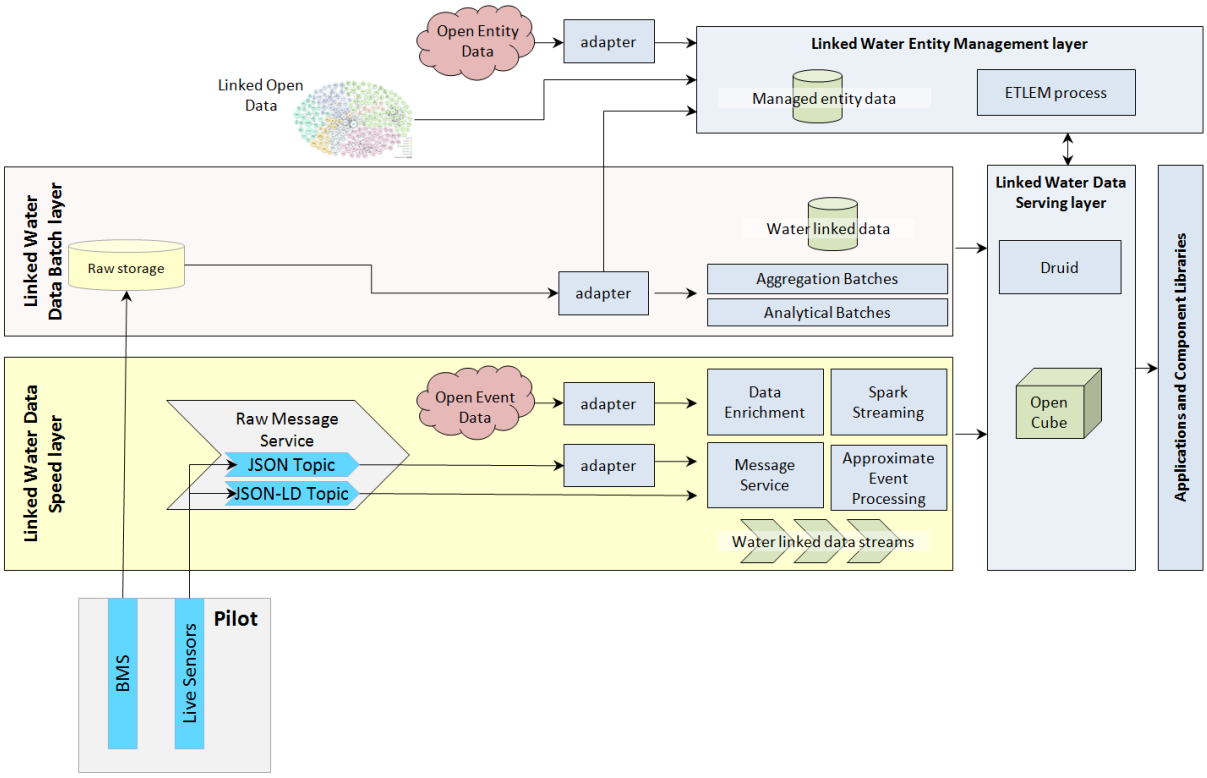

Figure 11: Linked Water Dataspace and the Lambda Architecture

An important aspect of the QUERY dataspace service is making data available in the statistical form through Linked Open Data principles. This is similar to availing a data warehouse query end to applications. Open Knowledge Foundation[1] defines Open Data as: "*data that can be freely*

[1] http://okfn.org/

used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike".

It means that particular data should be available free of charge to everyone for the personal as well as the commercial usage without any restrictions. This idea is not new and is similar to the "open" movements such as Open Source or Open Hardware.

The growth of the World Wide Web, Internet and easier access to the web allowed the Open Data movement to gaining the momentum. Moreover, some of the governments are willing to make their data available to the public in open formats. This ensures the government transparency and increases the trust. Furthermore it has an impact on the development of services and applications addressing public demands. The best-known examples of government open initiatives are data.gov and data.gov.uk.

The Waternomics platform should be able to make use of relevant Open Data assets for proper analytics. Possible scenarios for consuming Open Data include the prediction of water consumption using open weather data. Data consuming requires a proper selection and evaluation of data source in order to select the most suitable one for a proper decision support. The more "linked" datasets are available, the more accurate the predictions and analysis can be.

After analysing the existing and well-established Open Source projects that supports the operations on the Linked Data, in particular the Linked Open Statistical Data in D3.1.1, we notice that the described platforms may be used as an element of Waternomics Batch Layer, which provides batch-based processing of the Linked Dataspace for accurate data views, such as aggregation-based water data views, and basic analytics, with some delay.

The RDF store should support batch ingest capability with speed. The number of entities is expected to be large, so the Batch Layer should be fault free and relatively simple.

The contribution made here is that the data processing architecture is made a part of a bigger dataspace model. That allows the use of existing dataspace support services to support the data processing architecture. For instance, the WKAN catalogue can support the serving layer by enriching data with entities that can be useful at the consumption side.

3.4 Joining the Dataspace and Adapters

The Linked Water Dataspace is composed of multiple data sources, as defined in WKAN catalog. The types of data sources can include but is not limited to real-time sensor streams, historical databases, large text files, and spreadsheets. For a data source to become part of the Linked Water Dataspace, it must be discoverable and should conform to the 5 stars scheme of Linked data. The critical entities in the data source should be exposed as Linked data. We follow a seven step approach for including a data source into the LWD, according to the principles of Linked Data. The approach allows the conversion and publishing of data in standard formats for Linked data such as RDF or JSON-LD. It further facilitates the open availability of data in support of various applications.

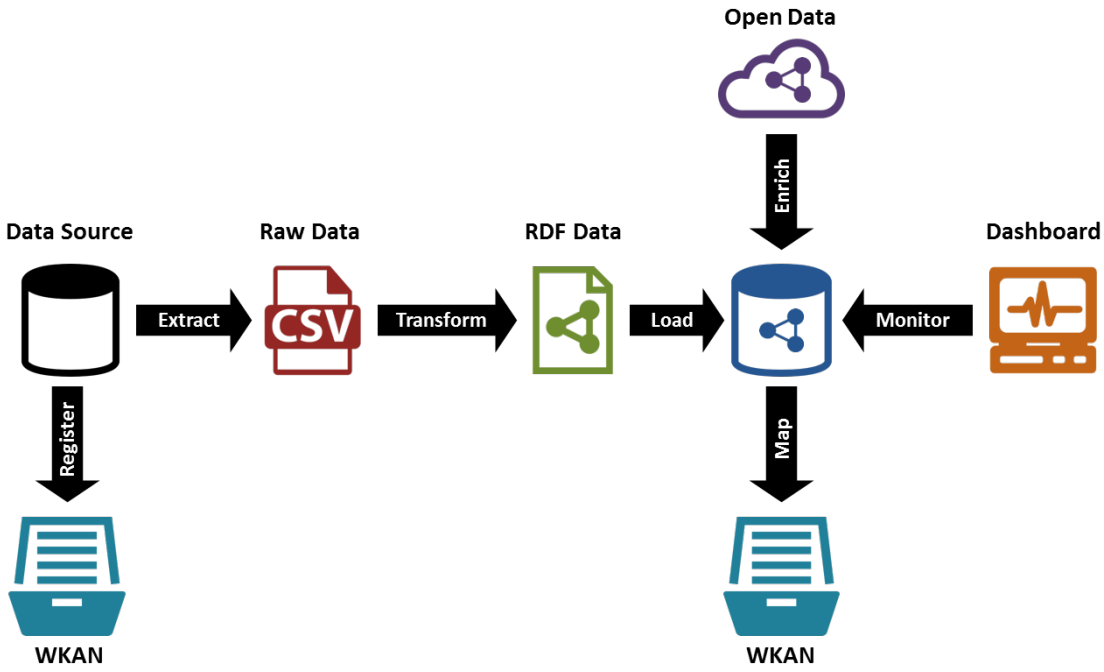

Figure 12: RETLEMM process for exposing entities as Linked data

Figure 12 provides an overview of the joining process for a data source.

- **Register:** A new data source joining the dataspace would require it to be registered in the dataspace catalog i.e. WKAN. The registration means that the catalog contains an entry describing the data source at minimum in terms of type, access, and format. Further information about the data source can include the physical address of files or a query endpoint. After this information a data source is considered part of the dataspace, since it can be accessed and used by application. Figure 13 shows the user interface for registering a new data source (or dataset) in WKAN.

- **Extract:** The second step of the process is to extract the data from the data source. For the sake of simplicity, it is assumed that such information are extracted in the form of CSV[1] files. Besides the data extraction from databases, this step can also include data collection from community of users. Essentially, the purpose of this step is to source relevant information from either digital or human sources.

- **Transform:** Given the CSV representation, the next step is to convert the data in appropriate format for publishing. In this regards, we adapt a simple semi-automated process for transforming the CSV files to RDF files using appropriate tools such as Microsoft Excel and Google Refine. This process is supported by the schema mapping document that define the correspondence between the attributes in sources with the RDF properties in the target Linked Data representation of entities. It is recommended to map each source attribute with a property in WATER or DCMI vocabularies that are defined later in this report.

- **Load:** Once the data has been converted and represented in RDF format, the next step is to store it in an appropriate data store. For this purpose, any general purpose RDF store

[1]Comma Separated Values

may serve the purpose. However, it is necessary for the RDF store to have necessary publishing, querying, and search functionalities to support applications. Additionally, the RDF store should support batch ingest capability. Since, the number of entities can be large the batch ingest should be fault free and simplistic.

- **Enrich:** Generally above steps are sufficient to support analytical and decision support applications. Nevertheless, it is desirable to enhance the meta-data with additional information such as links to external datasets and additional attributes extracted from non-critical data sources. Therefore, the purpose of this optional step is to add contextual information so that the overall vision of the Linked dataspace supported. Essentially, this step should be driven by the needs of individual applications to support their functionality.

- **Map:** Similar to enrich step the schema and entities of a data source may be mapped with other data sources and entities in the catalog. This facilitates integration and de-duplications of classes and entities. Additionally, it allows automated process of data collected from multiple dataset using advanced reasoning and schema agnostic query tools.

- **Monitor:** As discussed earlier, the metadata is considered to be immutable and append only. However, it is not unusual for the organizations to change or update the definitions and attributes of their core data sources. We handle this challenge through an appropriate monitoring process. As this point in time, we keep the monitor process to simple scheduled checks for changes in data sources. We expect this to change as the project matures and more insights are brought in terms of the data management process.

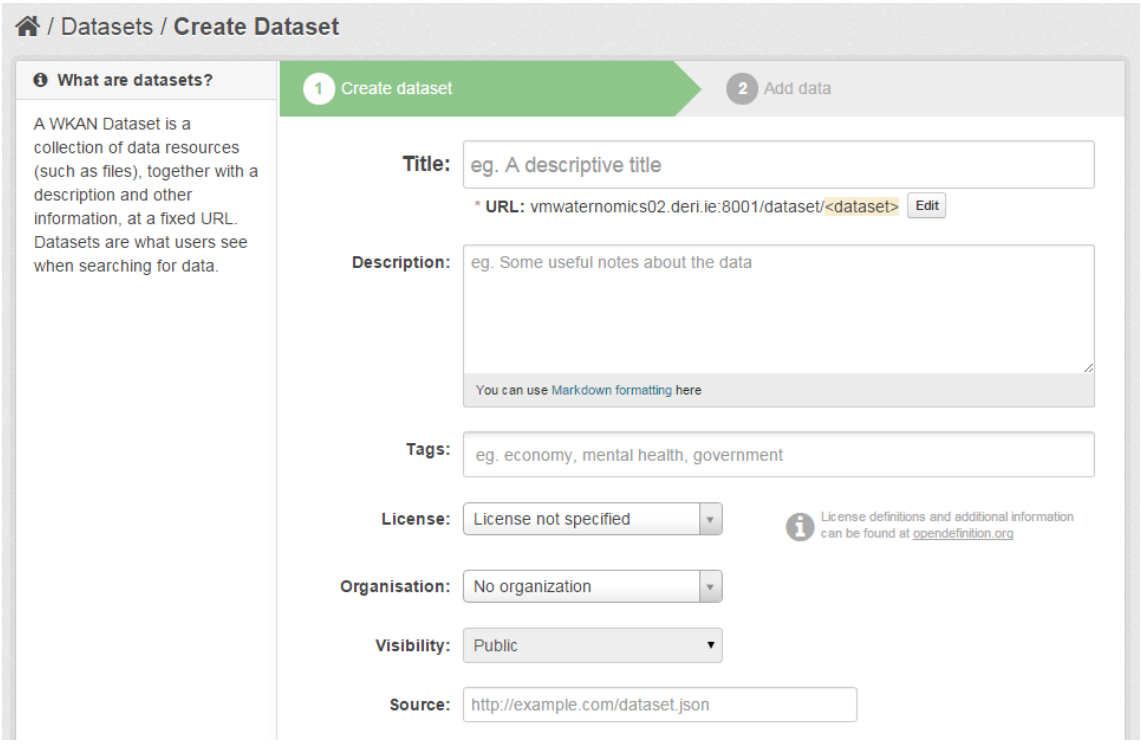

Figure 13: User interface for registering a new data source (or dataset)

At this end, the process applications developers should be able to find and understand the data source. Generally when a data source joins, above process can be performed manually by the data source owners with the help of dataspace administrators. However, the ETL step can be automated to speed-up the process. Automation is specifically desirable for large scale historical data and real-time metering data. In the following, we discuss the two alternatives for automation:

- **Adapters:** Adapters can be considered non-materialized view of a data source. They encode the ETL process in the form of mappings between source data format and the target data format. In case of a historical database, the data resides in the source and the ETL is performed on-the-fly every time as queries are posted on non-materialized view. This form of ETL process can be considered part of the serving layer, as discussed earlier. In case of a real-time data streams, the ETL is performed on-the-fly as data passes through adapters. This form of ETL process is discussed in previous section under the name of the speed layer.

- **Scheduled Jobs:** This form of ETL process is performed either one for a large static database or periodically for large dynamic database. This also enables reflection of any changes in the original data source, in terms of addition, deletion, or update, in the Linked data. This form of ETL is essentially the traditional batch layer.

3.5 Conclusion

In this section, we have provided an overview of the Linked Water Dataspace. The architecture shows how the dataspace is realised. We follow the Lambda architecture and extend it to facilitate the Linked Data approach. The dataspace is used for ingesting, managing, and publishing water data that includes both sensor data and entity data. First, a simple entity management process is used that exploits non-technical experts. Second, an event processing approach is used for real-time sensor data. This approach enables semantic processing of in-flow data enrichment and approximate matching. In short, this section provided a view of the implementation of Linked Water Dataspace with respect to the requirements introduced in Section 2 as shown in Table 6.

Table 6: Summary of Requirements and Approaches

3.5.1.1 Requirement	3.5.1.2 Approach
Standardisation	Standard compliant ontologies are used for describing sensors and their data.
Consuming Open Data	One of the support services that is included in the dataspace consists of crawling weather data an integrating it in prediction services.
Publishing Linked Data	The produced data by some support services is published in RDF using linked data principles.
Data Linking	Relevant open data sources are integrated in the dataspace

	through explicit links between entities.
Real-time data / events	The dataspace is designed with respect to the lambda architecture that covers real-time event processing.
Real-time Analytics	Real-time analytics of data is considered in the platform as part of the speed layer of the lambda architecture.
Data integration	Lambda architecture in essence deals with both historical and real-time data. This data integration is insured by relevant support services.
Heterogeneity of Sensor Data Events	The platform is designed to handle a variety of sensor types. Consequently for each sensor type, a dedicated collector is designed for collecting data and converting it into RDF for further processing.
Enrichment of Sensor Data Events with Open Data	Sensor readings are further enriched by dedicated support services using open data.

4 Concrete Pilot Cases in the Dataspace

In this section, we provide a concrete realization of the Linked Water Dataspace using the tools and techniques discussed in previous sections. We have implemented the Linked Water Dataspace, for the WATERNOMICS project, as a realization of the Lambda architecture. Our implementation departs from the original Lambda architecture due to the central role of WKAN catalog service in speed, batch, and serving layers. We call the addition the "catalog layer". The services mainly implemented through customization of open source software including Apache stack of technologies.

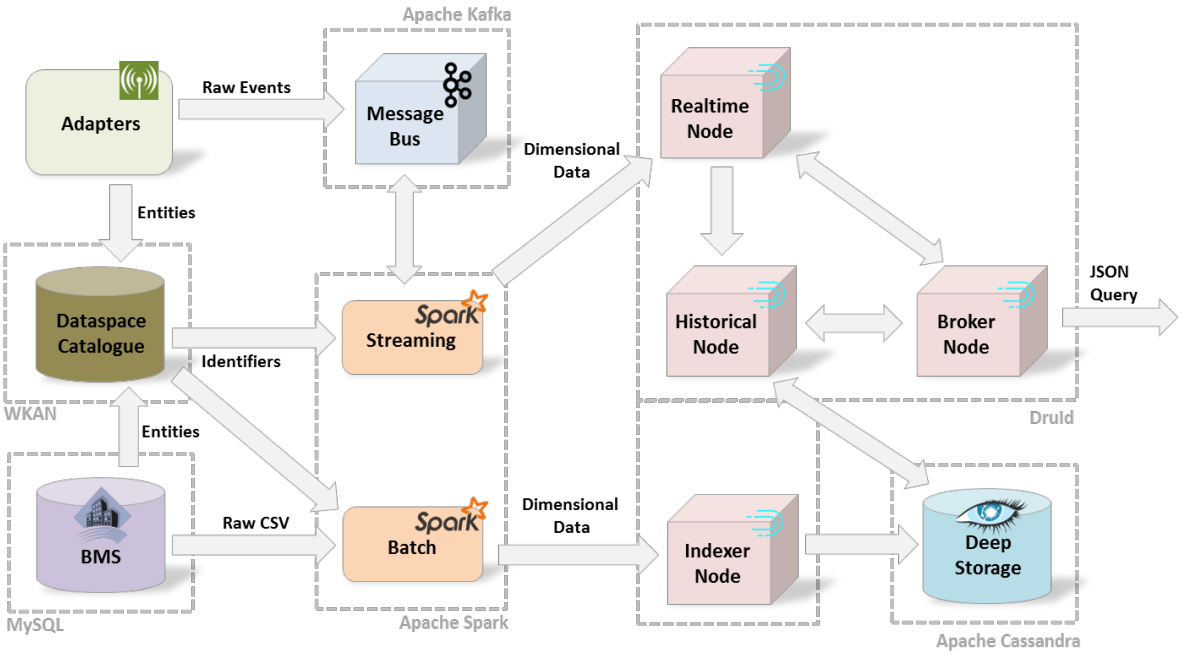

Figure 14: Lambda architecture realization of the Linked Water Dataspace

Figure 14 shows that data from the BMS and water sensor in the Galway pilot being processed in the Linked Water Dataspace. First, entities are defined into the WKAN dataspace catalog as well as entities from other sources such as newly installed sensors. Batch layer is implemented using Spark SQL when historical data from BMS is fed indexer node of Druid. Real-time data from sensors is fed into the Kafka message broker, which provides a high availability integration point for speed layer data from different pilot. Real-time data from Kafka is processed through Spark streaming code to real-time node of Druid. The combined code from Spark streaming and SQL provides a standardized way of generating dimensional data that is served using the Druid.

The Druid nodes use a Cassandra as deep storage for historical data; however, we have implemented the deep storage using traditional file system. The batch data is made available through the historical node and streaming data is made available through real-time node. Periodically, the streaming data is pushed to the historical node as new data arrives. The broker node of Druid seamlessly exposes batch data and real-time data, with the need of writing queries for real-time and batch data separately. The following sections present two concrete cases that have been implemented within the Linked Water Dataspace in Galway pilot: the building management system and a water sensor joining the dataspace. Figure 15 illustrates an example screenshot of the running Druid cluster for the realtime node.

Cluster Servers

Average Server Percent Used: 0%

Show 10 entries Search all columns:

Server host	Server maxSize	Server type	Server tier	Server priority	Server currSize	Server percentUsed
localhost:8025	0	realtime	_default_tier	0	0	0

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

Cluster Segments

Show 10 entries Search all columns:

Server host	Segment dataSource	Segment interval	Segment version	Segment dimensions	Segment metrics
localhost:8025	vtec.neb.consumption15mins	2016-01-13T00:00:00.000Z/2016-01-14T00:00:00.000Z	2016-01-13T00:00:00.000Z		count,mValue
localhost:8025	vtec.neb.consumption15mins	2016-01-12T00:00:00.000Z/2016-01-13T00:00:00.000Z	2016-01-12T00:00:00.000Z		count,mValue
localhost:8025	vtec.neb.consumption15mins	2016-01-10T00:00:00.000Z/2016-01-11T00:00:00.000Z	2016-01-10T00:00:00.000Z		count,mValue
localhost:8025	vtec.neb.consumption15mins	2016-01-14T00:00:00.000Z/2016-01-15T00:00:00.000Z	2016-01-14T00:00:00.000Z		count,mValue
localhost:8025	vtec.neb.consumption15mins	2016-01-11T00:00:00.000Z/2016-01-12T00:00:00.000Z	2016-01-11T00:00:00.000Z		count,mValue

Showing 1 to 5 of 5 entries

First Previous 1 Next Last

Figure 15: Druid Cluster for Realtime nodes

4.1 Galway pilot sites data

Galway pilot has two important sites: the Engineering Building at NUI Galway and Coláiste na Coiribe secondary school.

The Engineering Building at NUI Galway (NEB for short) is a state of the art educational facility designed to be a 'living laboratory' where the building itself is an interactive teaching tool. The Engineering Building opened in 2011, it is the largest engineering building in Ireland and includes lecture halls, classrooms, offices, laboratory facilities, a café, and showers and bathrooms.

Figure 16: NUI Galway Engineering Building

The building accommodates approximately 1,100 students and 100 staff in 14,000 m2 of floor space on four floors. The majority of students are undergraduates aged 18-24 years.Coláiste na Coiribe (CnaC) is an Irish language secondary school with approximately 350 students and 25 teaching and administrative staff. The existing school is housed at a small city centre location. To facilitate the demand for places at the school and to address space pressures, a new 7,400 m^2 school is currently construction at a new sub-urban location in Galway

Figure 17: Coláiste na Coiribe Pilot Site

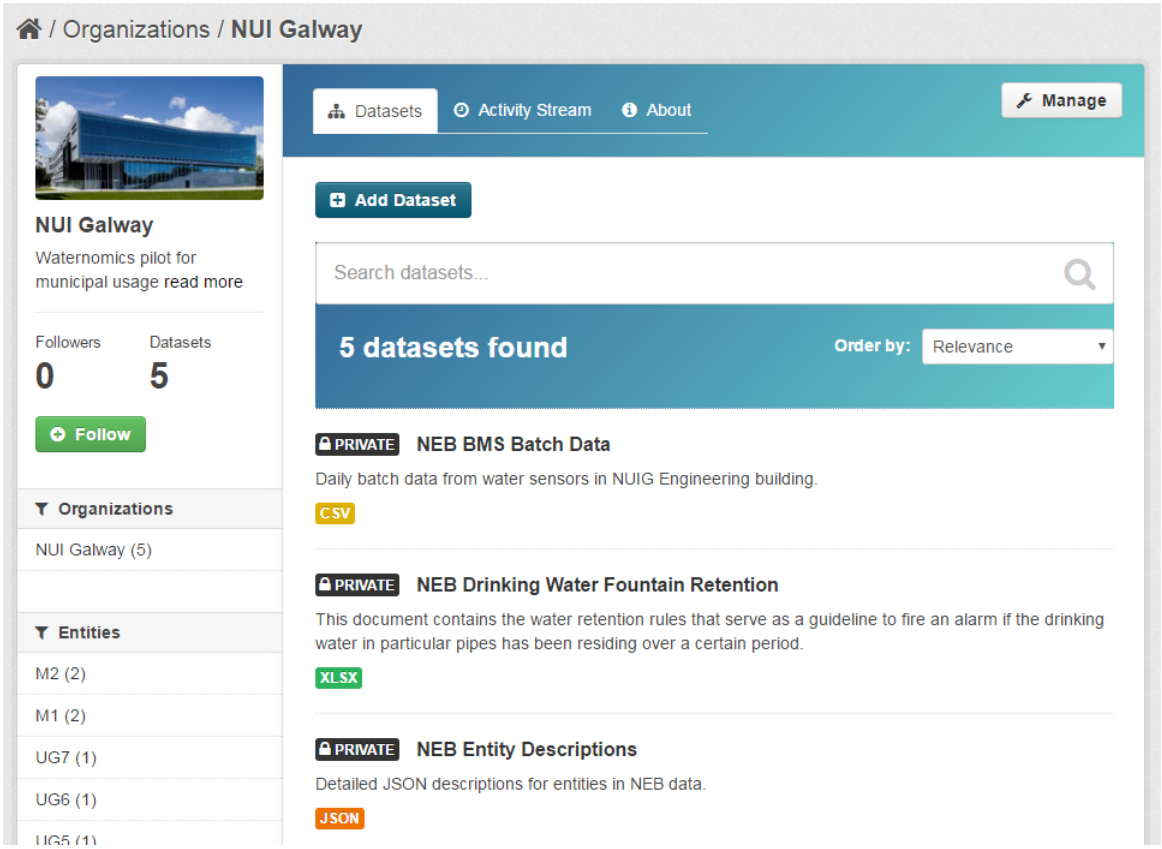

Figure 18: Datasets and data sources in WKAN for Galway pilot

In the following, we present both the Engineering Building at NUI Galway and the CnaC data collection as a single process as these two pilots share the same location and consequently the same required data sources.

4.1.1 Data Sources and Open Data

The NUIG & CnaC pilot aims to collect both real-time and historical data for water management for a large building. This includes following data sources

- Historical and batch data from BMS (Cylon)
- Real-time data from VTEC sensors

A set of relevant open data for both pilot sites are also included in the Waternomics catalogue, i.e., WKAN, this includes[1]:

- Open data from weather prediction and observation services
- Public calendar data used by analytics services for distinguishing between water consumption in working days vs. non-working days
- Drought data in Ireland
- Updates from Irish Water services

[1] Please note that most of these open data sources have already been used by support services an application. This list contains also other relevant open data sources that can be used for future applications.

- Irish water prices
- Water footprints in Ireland
- Water metaphors
- Social media feeds (twitter and reddit) related to Water news in Ireland
- Water quality in the region of west of Ireland
- Water Quality Reports in Ireland
- Statistics about personal water usage quantities

4.1.2 Catalog Meta-data

All of the above mentioned data sources join the Linked Water Dataspace for NUIG & CnaC pilot through definition in WKAN. Figure 18 shows a list of datasets associated with NUIG pilot. It shows summary meta-data for each dataset in form of tags and description. Users can select a data source to reveal further meta-data which includes the location of data. As a convention, all datasets for historical and real-time data from sensors of pilots are tagged as "private". This way there associated meta-data is only visible to authorized users. By comparison, open data sets are defined as public datasets which can be used by everyone.

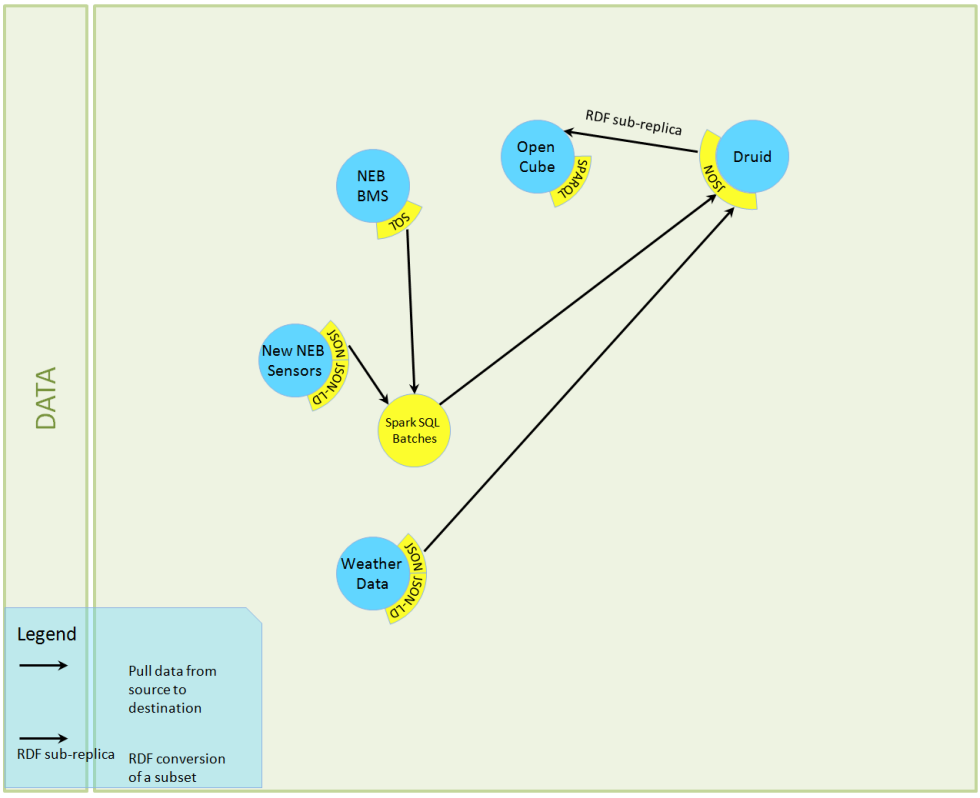

Figure 19: An example realization of Lambda Batch Layer in Linked Water Dataspace

4.1.3 The Building Management System Joins the Dataspace

The New Engineering Building (NEB) at NUIG has a commercial grade Building Management System (BMS) that collects and stores data from existing water sensors. The BMS contains information about water metering and related entities. Entities include such as sensors, rooms, water outlets, etc. Metering data generated by sensors which records information such as water

flow. The BMS joins the Linked Water Dataspace by following the process described previously, which enable the visibility of entities in dataspace. The metering data that is collected over many months is included in the dataspace through batch processing. The batch processing aspect of the dataspace is realized via the following elements:

- A set of SQL scripts are used to extract the historical events data from the database of the BMS.
- A Spark Core map/reduce jobs is used to transform and aggregate low levels events to suitable granularity.
- A Druid indexing node is used to move the dimensional aggregated historical data into the Druid deep storage and make it available for further querying.

The above process is performed once for ingestion of large number of events collected over many months. Afterwards scheduled Spark jobs are deployed to ingest daily events from the BMS.

4.1.4 A Water Sensor Joins the Dataspace

The real-time aspect of the dataspace is realized via the following elements:

- A RESTful API adapter for the sensor.
- The Kafka middleware which forms the backbone of events distribution and reliability guarantee for production/consumption staging.
- The Spark Streaming map/reduce jobs which do the first shot of aggregation and processing.
- The Druid real-time node, which moves the dimensional aggregated stream data from the Kafka bus into the Druid deep storage and make it available for further querying.
- Other components in the dataspace which could consume from the Kafka nodes on the fly such as enrichment of events, RDFization, Collider nodes for matching, etc.

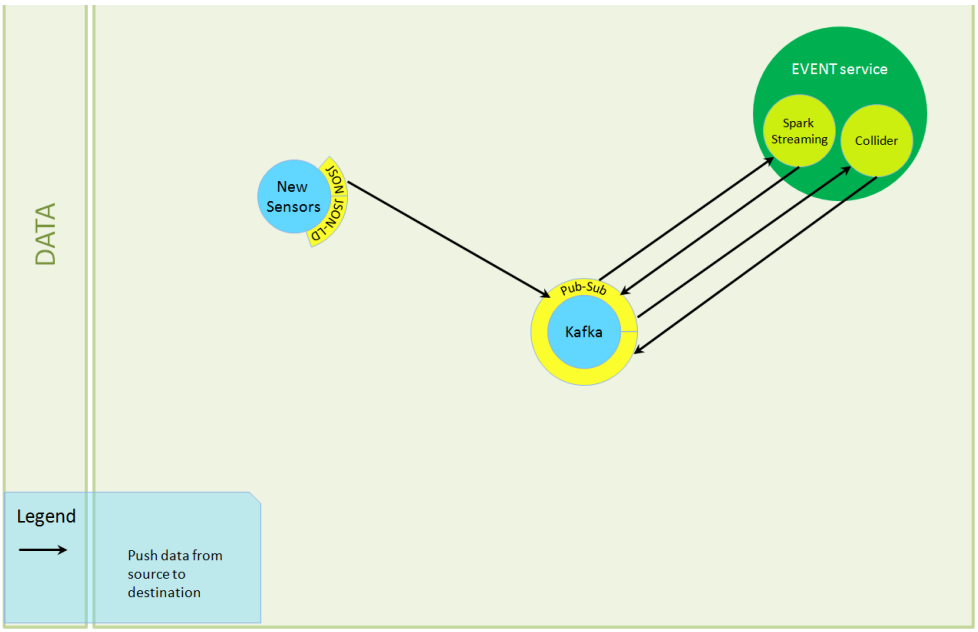

Figure 20: Linked Water Dataspace Lambda Speed Layer

Figure 20 illustrates the basic speed layer of the Linked Water data space which includes the main elements from the ones aforementioned. Figure 21 shows the concrete path of the sensor data coming from the sensors installed in the various pilot sites which as described below.

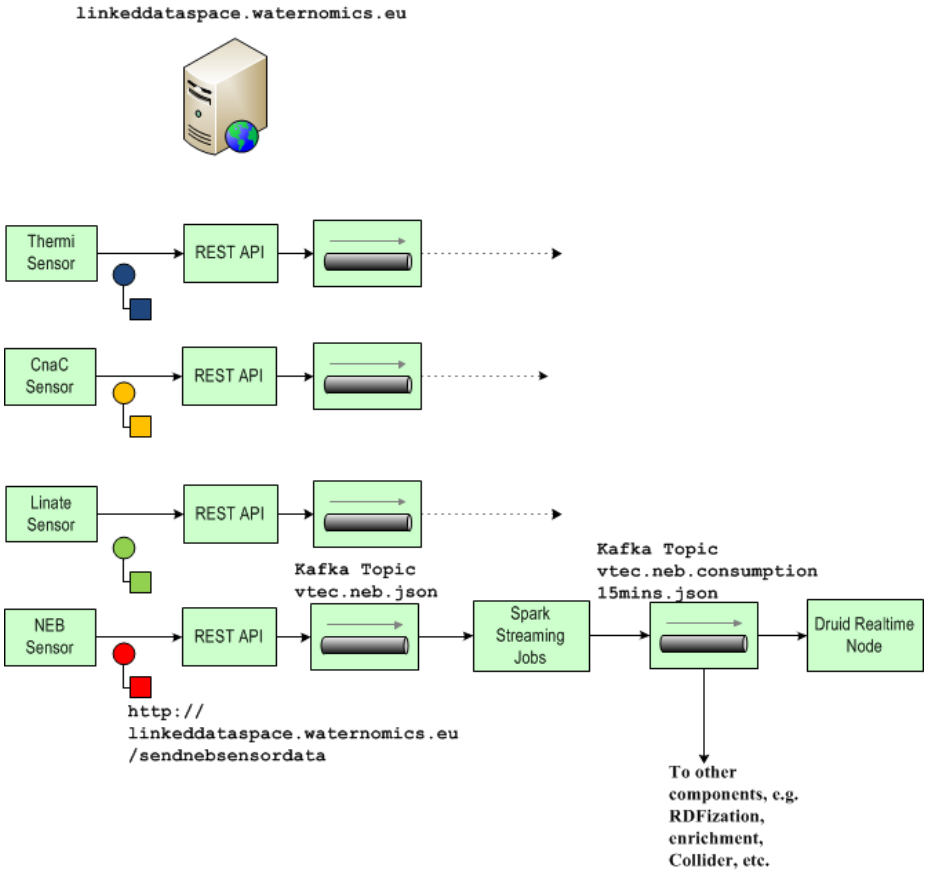

Figure 21: Senor Data Flow

Let us take for instance a VTEC sensor of those installed in the NUIG engineering building. Such a sensor is able to measure four metrics: fvel, vel, flow, today, and net. These correspond to the velocity of sound in water, velocity of water, flow of water, the accumulative today consumption, and the accumulative overall consumption. Readings are sent from the vtec sensor into the RESTful api of the dataspace and looks like the following:

Listing 1: Raw Sensor JSON Data

```
{
   "SensorID":"USF_05",
   "location":"IRE",
   "site":"NEB",
   "sensorType":"USF",
   "date":"16-01-26",
   "time":"17:27:57",
   "description":"CWS to Zinc Canteen",
   "data":[
      {
         "type":"fVel",
         "metric":"m/s",
         "value":" 1220.61"
      },
      {
```

```
            "type":"today",
            "metric":"m3",
            "value":"13.6328"
         },
         {
            "type":"net",
            "metric":"m/s",
            "value":"5116"
         },
         {
            "type":"flow",
            "metric":"m3/h",
            "value":"3.48875"
         },
         {
            "type":"vel",
            "metric":"m/s",
            "value":"0.116306"
         }
      ]
}
```


The JSON data is sent to the Linked Water Dataspace via a RESTful API as HTTP POST to *http://linkeddataspace.waternomics.eu/sendnebsensordata*. The RESTful API implemented with the dropwizard-kafka-http client then forwards the JSON messages into a Kafka topic which is vtec.neb.json.

The Spark streaming jobs take over from here. There are two mains tasks accomplished by Spark streaming jobs:

- Calculating the consumption between two consecutive readings. The reason for that is that the actual reading coming from the sensor only holds information about the accumulative consumption, so the Spark job holds states of last consumption and then calculates the difference. Other sensors such as the mini-water sensors only send flow information rather than accumulative consumption, so the Spark job calculates the consumption based on the flow. The result of this is published into a Kafka topic that represents the firehose.

- Aggregation. That is important to cut down the raw firehose of readings into a suitable granularity for storage by Druid and which also can be used for other applications. The aggregation is done over the firehose topic. It is done of a 15 mins level, but configuration is possible to this. Nonetheless, the firehose keeps available in case an application of a service wants to work directly on it. The resulting aggregated data is dimensional and ready for Data cubes and is published on Kafka topics such as vtec.neb.consumption15mins.json. An example data item in this topic is as follows:

Listing 2: Sensor Dimensional JSON Data


```
{
   "TimeStamp":"2016-01-26T17:30:00+0000",
   "dCountry":"IRE",
   "dPilotSite":"NEB",
   "dSensor":"USF_05",
```

```
    "dSensorType":"USF",
    "dYear":"2016",
    "dMonth":"01",
    "dDay":"26",
    "dHour":"17",
    "dMinute":"30",
    "mValue":"0.00053"
}
```


A Druid real-time node is configured to persist the data coming from the dimensional Kafka topic into the deep storage and make it query-able in a similar way the batch data is, thus applications can query the batch and speed layer via the serving layer and make use of it either for presenting the data visually or conducting various types of analytics over the data.

Other processes can consume from the Kafka topics, the firehose or the aggregated, for various purposes including an adapter process that performs a JSON to JSON-LD RDFization of data according to the ontology. Also the enrichment can take place to add further fields into the data such as location or other metadata. Collider instance can consume from these topics too to do approximate matching.

4.1.5 Querying the NUIG Engineering Building in the Dataspace

In this section, we describe how one of the support services, i.e. the water usage analytics service, is querying the dataspace for serving end-user applications with water usage data.

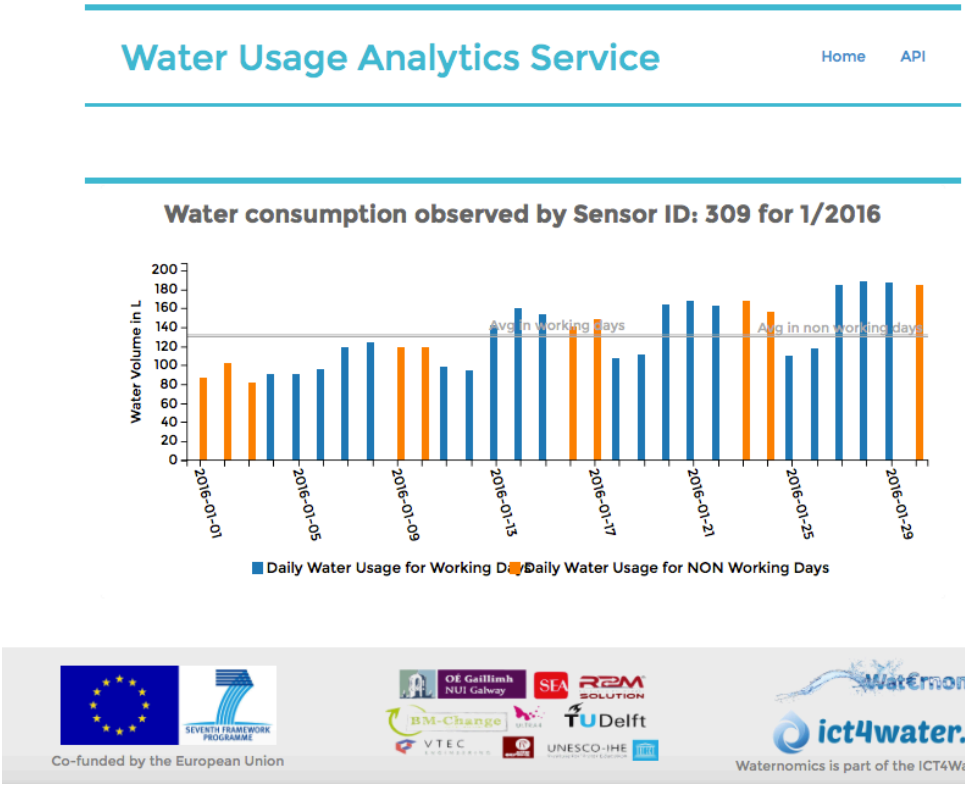

Figure 22: Water Usage Analytics Service

The Water Usage Analytics Service allows to query water data for a particular month, identifies working vs. non-working days using open data sources, then identifies outliers, computes average consumptions etc. This service can also be used as a stand-alone application as it proposes a simple user interface to visualize monthly data as shown in Figure 22. In the following we will discuss how this services gets the data from the datasapce.

The analytics service exposes its results through a RESTful API. In order to get the data for a particular sensor (e.g., 309), in a particular month (e.g., 12) of a year (e.g., 2015) the service can be querried as follows:

`http://vmwaternomics01.deri.ie:8011/RestAPI/getAnalytics?sensorID=309&month=1&year=2015`

In the backend of the analytics service, the parameters of this query are used to query DRUID using Listing 3:

- This listing indicates in line 3 that this query needs access to the data source "nuigBMSneb" as it is the location where DRUID saves Galway pilot data.
- In line 4, both the month and the year are converted into an interval based on the queried month. If the queried month is the current month, the interval becomes from the first day of the month until the day before the system data (in order to make sure that there is data available in the dataspace.)
- In line 7 and 8, the query identifies filters over the data it is querying. In this example, we have the conjunction of two filters: one for the identifier of the sensor and one for the hour. The identifier of the sensor is taken from the initial query. The filter dHour = 23 allows to take only the readings of the last hour of the day. This filter is specific to the data coming from the Building Management System of Galway pilot as it stores the daily water data as cumulative consumption over the entire day. Consequently, the last hour contains the overall consumption of the day.
- Lines 9 to 11 indicate the aggregations that required, in this example, we require the number of readings recorded over the last hour of the day (resulting in "readings") and their sum (resulting in "consumption"). (see Listing 4)

```
1.   {
2.    "queryType": "groupBy",
3.    "dataSource": "nuigBMSneb",
4.    "intervals": [ "2015-01-01/2015-01-31" ],
5.    "granularity": "all",
6.    "dimensions" : ["dDay", "dMonth", "dYear", "dHour"],
7.    "filter": { "type": "and", "fields": [{"type":
"selector","dimension": "dSensor", "value": "309"},
8.                                          {"type": "selector","dimension":
"dHour", "value": "23"}] },
9.     "aggregations": [
10.     {"type": "longSum", "fieldName": "count", "name": "readings"},
11.     {"type": "doubleSum", "fieldName": "added", "name": "consumption"}
12.  ]
13. }
```

Listing 3: DRUID query for the sensor 309, month = 1 and year = 2015

```
[
  {
    "version": "v1",
    "timestamp": "2015-01-01T00:00:00.000Z",
    "event": {
      "readings": 4,
      "dmonth": "1",
      "dday": "1",
      "dyear": "2015",
      "dhour": "23",
      "consumption": 86.5999984741211
    }
  },
  …
]
```


Listing 4: Results of the DRUID query

Results from DRUID are in JSON format as shown in Listing 4. Each daily reading is identified by its timestamp and its resulting value is encapsulated as an event with the dimensions indicated in the query: e.g., readings, dDay, dHour, dMonth, dYear and consumption.

4.2 The Thermi Pilot

The pilot in Thermi targets domestic environment users. The main differentiation in that environment is this between adults and children in a family although different households have different priorities and incentives for participating in the pilot. However, this differentiation is mostly accommodated by the customization of applications towards each specific households needs. In this pilot the objectives of reducing water consumption and raising water awareness are also accompanied by the objective of changing users behavior and promoting education. In order to achieve this multiple data sources are required for building appropriate services and applications.
In the following we list relevant data sources and the process of joining this data into the Waternomics dataspace.

4.2.1 Data Sources and Open Data

Thermi pilot aims to collect both real-time and historical data for water management for each of the housholds. This includes following data sources:

- Historical water usage data from household (CSV files)
- Real-time data from VTEC sensors

A set of relevant open data for this pilot can be accessed via support services or from the the Waternomics catalogue, i.e., WKAN (see Figure 23, this includes[1]:

- Open data from weather prediction and observation services
- Public calendar data used by analytics services for distinguishing between water consumption in working days vs. non-working days

[1] Please note that most of these open data sources have already been used by support services an application. This list contains also other relevant open data sources that can be used for future applications.

- Drought data in Greece
- Water prices in Greece and Europe
- Water footprints in Greece
- Water metaphors
- Social media feeds (twitter and reddit) related to Water news in Greece
- Statistics about personal water usage quantities

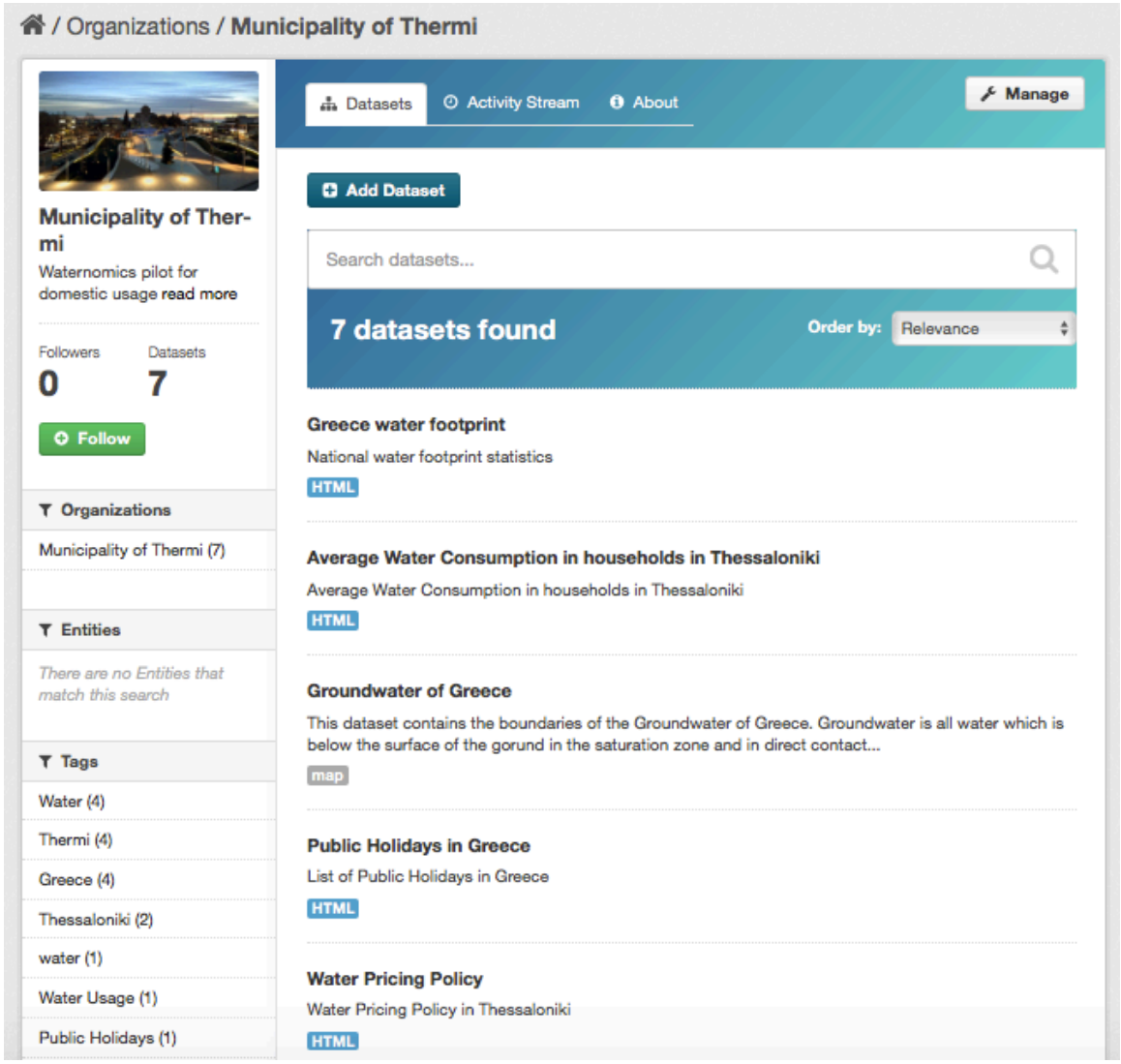

Figure 23: Datasets relevant to Thermi Pilot including Open Data Sources

4.2.2 Catalog Meta-data

In the case of Thermi pilot the meta-data is primarily concerned with historical and real-time usage of households. Figure 24 shows the datasets defined in WKAN catalog for Thermi pilot. For instance, the historical data for households is recorded in an Excel file, which is made available for dataspace users in WKAN for query and analysis. Meta-data for households and related entities is also made available through WKAN.

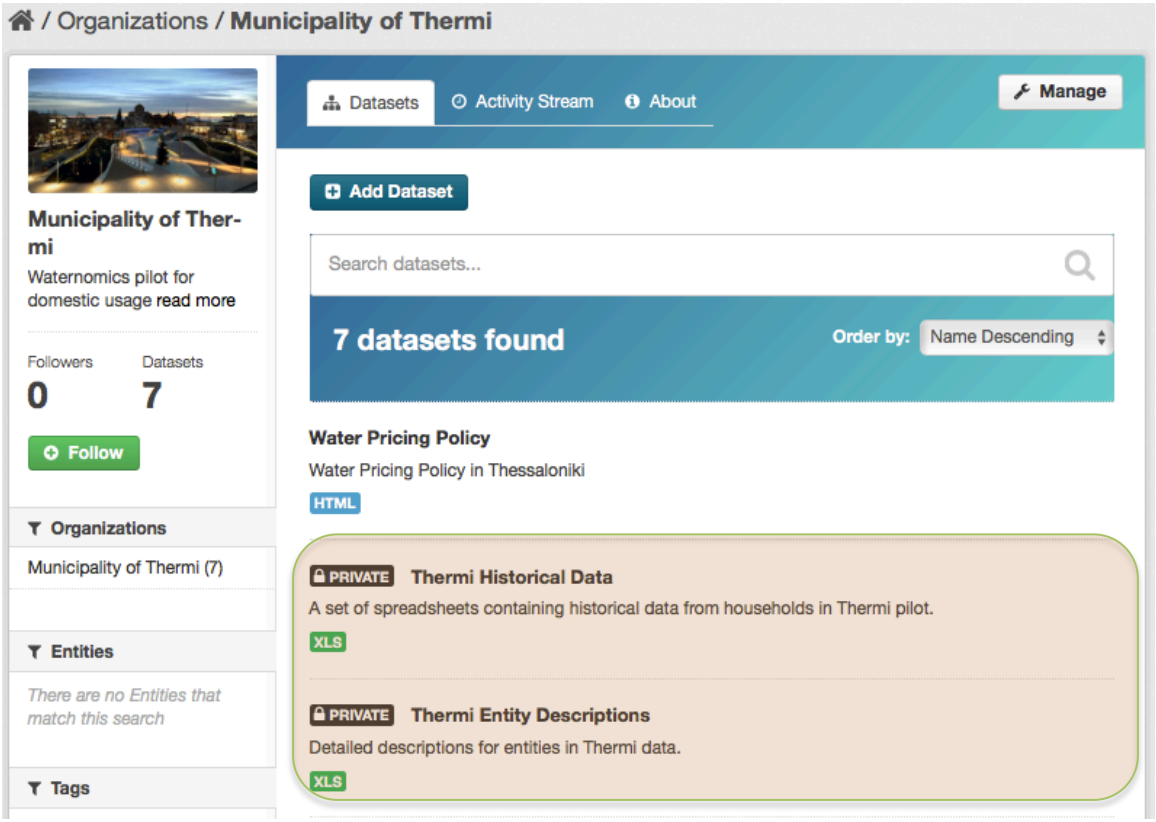

Figure 24: Datasets and data sources in WKAN for Thermi pilot

4.2.3 Thermi Historical Household Dataset Joins the Dataspace

The historical water usage data for Thermi pilot is available in monthly and quarterly reported values in an Excel spreadsheet. This data joins the Linked Water Dataspace as the file is made available through a URL and defined in the WKAN catalog. To further facilitate query services on this dataset, a simple batch process is applied to serve the data using DRUID. The conversion process from Excel to JSON is similar to the process followed for the historical data in Galway pilot. The only difference is that the process is only run once for the Thermi pilot, as opposed to daily batches in case of Galway pilot.

4.2.4 A Household Water Sensor Joins the Dataspace

The household sensors in Thermi join the dataspace in the same way as other VTEC sensors as detailed in Section 4.1.4. The difference here is that the majority of sensors in Thermi are miniwater sensors which produce readings as in the following listing.

```
{
   "location":"GRE",
   "site":"THERMI",
   "sensorType":"MWM",
   "SensorID":"HH_01_QPJFC",
   "description":"Home 5, bathroom data",
   "data":[
      {
         "subID":"mwm_1_LLVNO",
         "description":"Washing Machine",
```

```
            "type":"flow",
            "metric":"m3/h",
            "value":"0.23"
        },
        {
            "subID":"mwm_2_MIDHD",
            "description":"Cold water",
            "type":"flow",
            "metric":"m3/h",
            "value":"0.23"
        },
        {
            "subID":"mwm_3_QOLWD",
            "description":"Hot Water",
            "type":"flow",
            "metric":"m3/h",
            "value":"0.23"
        }
    ]
}
```


Listing 5: Miniwater sensors in Thermi

We notice that the sensors do not report consumption directly, but they rather report flow. To tackle this, the Spark jobs calculate the consumption out of the flow data and then consumption is used for downstream processing.

4.2.5 Querying the Thermi Households in the Dataspace

The process of querying Thermi Households data from the dataspace is similar to the NEB pilot's processes.

The main difference is in the data source that we are querying as well as the aggregation level. An example query for cumulative data over a day is as shown in Listing 6. The parameters of this query:

- This listing indicates in line 3 that this query needs access to the data source "**ThermiHouseoldsSensors**" as it is the location where DRUID will save Linate pilot data.
- Line 4 indicates the querying interval, in this case the interval is for the entire month of January 2015.
- In line 7, the query identifies a filter over the sensor data it is querying (sensorID = HH_01_QPJFC.mwm_3_QOLWD). This sensor observes the hot water in the kitchen tap from household 01.
- Lines 8 to 10 indicate the aggregations that required, in this example; we require the number of readings recorded over the day (resulting in "readings") and their sum (resulting in "consumption").

```
1.   {
2.    "queryType": "groupBy",
3.    "dataSource": " ThermiHouseoldsSensors",
4.    "intervals": [ "2015-01-01/2015-01-31" ],
5.    "granularity": "all",
6.    "dimensions" : ["dDay", "dMonth", "dYear"],
```

```
7.    "filter": { "type":  "selector","dimension": "dSensor", "value": "HH_01_QPJFC.
mwm_3_QOLWD"},
8.     "aggregations": [
9.     {"type": "longSum", "fieldName": "count", "name": "readings"},
10.     {"type": "doubleSum", "fieldName": "added", "name": "consumption"}
11.  ]
12. }
```

Listing 6: DRUID query for Thermi House 01 sensor HH_01_QPJFC.mwm_3_QOLWD, month = 1 and year = 2015

4.3 The Linate Airport

Linate pilot serves a variety of users in a business environment each one having a specific role in the company's structure. For some roles water related information is more crucial than others and in some groups information provided in a timely fashion is crucial while others can check details more casually. Moreover, apart from company's employee a major objective defined in D5.1 is the improvement of SEA's corporate responsibility profile through environment awareness raising actions. In order to satisfy the wide variety of needs for such an environment a wide variety of data sources are required.
Please note that this pilot is experiencing delays in its installation process due to long tendering process and legal issues that are reported in D5.1 and D8.3.2. For this reason, in the following we identify relevant data sources for this pilot as well as our architecture and plan for joining and querying sensor data sources.

4.3.1 Data Sources and Open Data

Linate pilot aims to collect real-time data for water management. This includes following data sources:

- Real-time data from VTEC sensors (Currently collecting data using a standalone ultrasonic meter that is described in D4.1)

- Real-time data from other Ultrasonic sensors (Not yet installed, but in this deliverable we will describe our plan and architecture for joining this data into the dataspace)

A set of relevant open data for this pilot are also included in the Waternomics catalogue, i.e., WKAN, this includes [1]:

- Open data from weather prediction and observation services
- Public calendar data used by analytics services for distinguishing between water consumption in working days vs. non-working days
- Drought monitoring data in Italy, and Water prices in Italy and Europe
- Water footprints in Italy
- Water metaphors
- Social media feeds (twitter and reddit) related to Water news in Italy
- Statistics about personal water usage quantities

[1] Please note that most of these open data sources have already been used by support services an application. This list contains also other relevant open data sources that can be used for future applications.

4.3.2 Catalog Meta-data

At the moment, the Linate pilot has only on data source defined in WKAN catalog as shown in Figure 25. The data source represents the planned real-time data from VTEC sensors to be installed in Linate. In the later part of the project, a small instance of WKAN catalog may be installed within Linate premises to serve as an airport specific dataspace catalog.

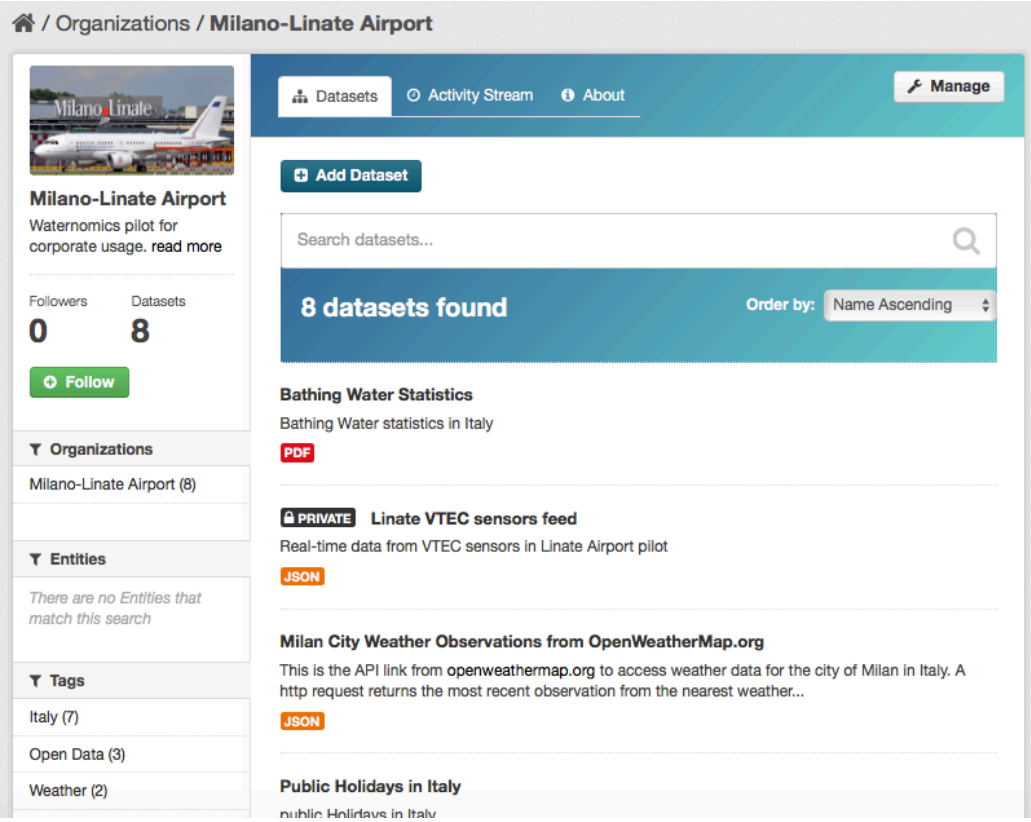

Figure 25: Datasets and data sources in WKAN for Linate pilot

4.3.3 A Water Sensor Joins the Dataspace

The planned VTEC sensors for Linate can join the dataspace in the same way as the sensors in NEB as detailed in Section 4.1.4. In practice, if concerns about the boundaries which can prevent data from leaving the pilot site, a replica of the dataspace machines can be put in place within the pilot site in Linate.

4.3.4 Querying the Linate Airport in the Dataspace

The process of querying Linate data from the dataspace is similar to the other pilots' processes. The main difference is in the data source that we are querying as well as the aggregation level. An example query for cumulative data over a day is as shown in Listing 7. The parameters of this query:

- This listing indicates in line 3 that this query needs access to the data source "**LinateVTECSensors**" as it is the location where DRUID will save Linate pilot data.

- Line 4 indicates the query interval. In this example it queries the entire month of January 2015.
- In line 7, the query identifies a filter over the sensor data it is querying (sensorID = USF_L01).
- Lines 8 to 10 indicate the aggregations that required, in this example; we require the number of readings recorded over the day (resulting in "readings") and their sum (resulting in "consumption").

```
1.   {
2.    "queryType": "groupBy",
3.    "dataSource": "LinateVTECSensors",
4.    "intervals": [ "2015-01-01/2015-01-31" ],
5.    "granularity": "all",
6.    "dimensions" : ["dDay", "dMonth", "dYear"],
7.    "filter": { "type":  "selector","dimension": "dSensor", "value": "USF_L01"},
8.     "aggregations": [
9.     {"type": "longSum", "fieldName": "count", "name": "readings"},
10.     {"type": "doubleSum", "fieldName": "added", "name": "consumption"}
11.  ]
12. }
```

Listing 7: DRUID query for the Linate sensor USF_L01, month = 1 and year = 2015

5 Summary

Water management is a challenging issue with the growing demand on water resources due to the increased urbanization, climate change, economic growth, etc. In this context, Waternomics project aims to develop and introduce ICT as an enabling technology for managing water as a resource and increasing end-user conservation awareness.

A major step towards achieving Waternomics objectives consists of collecting water usage data coming from sensors together with other relevant open data sources for an effective analytics to drive decision making: e.g., planning, adjustments and predictions.

Relevant data collected from various sources need to be standardised, enriched interlinked and shared among services and applications. We call this process: Linked Water Dataspace management. This deliverable reports on the design and implementation efforts towards this Linked Water Dataspace with respect to the requirements defined in Section 2.

After analysing relevant technological contributions that can be used in this context (D3.1.1), we design a dataspace following principles of the lambda architecture that facilitates the management of both historical and real-time data. The result of our research is reported in Section 3.

We reused existing open source tools that have been proven to be effective in practice. Indeed, the availability of high quality Open Source tools that support the process of publishing data as Linked Data, reduces the cost of implementation. We showed in this report how the dataspace has been put into practice in the four pilot sites: Galway New Engineering Building, CnaC school, Thermi, and Linate. We showed what available data sources exist and what joined the dataspace and how. We also showed how such data can be queried.

The primary contribution of this report is an architecture that implements the Linked Water Dataspace. The architecture realizes a dataspace that forms a low-barrier access to data sources and consumers. The architecture features a Lambda architecture which with a novel design that puts it as a part of a dataspace leading to the case where dataspace services such as the catalogue can support various layers in lambda such as the serving layer. The infrastructure has been developed for the Linked Water Dataspace as discussed in this document, and relevant components and services have been reported in D 3.2

6 References

[1] "The Open Definition." [Online]. Available: http://opendefinition.org/.

[2] E. Bruke, "An Autonomic Approach to Real-Time Predictive Analytics using Open Data and the Web of Things," National University of Ireland, Galway, 2013.

[3] T. Berners-Lee, "Linked Data- Design Issues," 2006. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html.

[4] S. Hasan and E. Curry, "Thingsonomy: Tackling Variety in Internet of Things Events," *IEEE Internet Comput.*, vol. 19, no. 2, pp. 10–18, 2015.

[5] S. Hasan and E. Curry, "Approximate Semantic Matching of Events for the Internet of Things," *ACM Trans. Internet Technol.*, vol. 14, no. 1, pp. 1–23, Aug. 2014.

[6] S. Hasan and E. Curry, "Thematic event processing," in *Proceedings of the 15th International Middleware Conference on - Middleware '14*, 2014, pp. 109–120.

[7] S. Hasan, E. Curry, M. Banduk, and S. O'Riain, "Toward Situation Awareness for the Semantic Sensor Web: Complex Event Processing with Dynamic Linked Data Enrichment," in *The 4th International Workshop on Semantic Sensor Networks 2011 (SSN11), a Workshop of ISWC 2011*, 2011, pp. 60–72.

[8] S. Hasan, K. Gunaratna, Y. Qin, and E. Curry, "Demo: approximate semantic matching in the collider event processing engine," in *Proceedings of the 7th ACM international conference on Distributed event-based systems - DEBS '13*, 2013, p. 337.

[9] S. Hasan, S. O'Riain, and E. Curry, "Towards unified and native enrichment in event processing systems," in *Proceedings of the 7th ACM international conference on Distributed event-based systems - DEBS '13*, 2013, p. 171.

[10] S. Hasan, S. O'Riain, and E. Curry, "Approximate Semantic Matching of Heterogeneous Events," in *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems (DEBS 2012)*, 2012, pp. 252–263.

[11] S. Hasan and E. Curry, "Tackling Variety in Event-based Systems," in *Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems*, 2015, pp. 256–265.

[12] M. Franklin, A. Halevy, and D. Maier, "From Databases to Dataspaces : A New Abstraction for Information Management," *Data Manag.*, vol. 34, no. 4, 2005.

[13] A. Halevy, M. Franklin, and D. Maier, "Principles of dataspace systems," *Proc. twenty-fifth ACM SIGMOD-SIGACT-SIGART Symp. Princ. database Syst. - Pod. '06*, pp. 1–9, 2006.

[14] N. Marz and J. Warren, *Big Data: Principles and best practices of scalable realtime data systems*. O'Reilly Media, 2013.