A Real-time Linked Dataspace for the Internet of Things: Enabling "Pay-As-You-Go" Data Management in Smart Environments

Edward Curry [a],[*], Wassim Derguech [a], Souleiman Hasan [a], Christos Kouroupetroglou [b], Umair ul Hassan [a]

[a] *Insight Centre for Data Analytics, National University of Ireland Galway, Ireland*
[b] *Ultra4, Admitou 27, Thessaloniki, 56224, Greece*

HIGHLIGHTS

- High-level design of real-time linked dataspace for IoT-enabled smart environments.
- Principled "Pay-As-You-Go" data management using tiered dataspace support services.
- Dataspace query service enabling unified queries across live, historical, & entity data.
- Demonstrates the use of a real-time linked dataspace to create analytics & decision support apps.
- Experiences and lessons from 5 real-world smart environments.

ARTICLE INFO

ABSTRACT

As smart environments move from a research vision to concrete manifestations in real-world enabled by the Internet of Things, they are encountering a number of very practical challenges in data management in terms of the flexibility needed to bring together contextual and real-time data, the interface between new digital infrastructures and existing information systems, and how to easily share data between stakeholders in the environment. Therefore, data management approaches for smart environments need to support flexibility, dynamicity, incremental change, while keeping costs to a minimum. A Dataspace is an emerging approach to data management that has proved fruitful for personal information and scientific data management. However, their use within smart environments and for real-time data remains largely unexplored.

This paper introduces a Real-time Linked Dataspace (RLD) as an enabling platform for data management within smart environments. This paper identifies common data management requirements for smart energy and water environments, details the RLD architecture and the key support services and their tiered support levels, and a principled approach to "Pay-As-You-Go" data management. The paper presents a dataspace query service for real-time data streams and entities to enable unified entity-centric queries across live and historical stream data. The RLD was validated in 5 real-world pilot smart environments following the OODA (Observe, Orient, Decide, and Act) Loop to build real-time analytics, decisions support, and smart apps for energy and water management. The pilots demonstrate that the RLD enables incremental pay-as-you-go data management with support services that simplify the development of applications and analytics for smart environments. Finally, the paper discusses experiences, lessons learnt, and future directions.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

Driven by a desire to use ICT to manage resources more effectively and efficiently, *smart environments* emerged in the form of smart cities, smart buildings, smart grids, smart water, and smart mobility [1]. A key driver in the development of smart environments is the convergence of technologies including the Internet of Things, Big Data [2], Middleware [3], Mobile computing [4] and the digitisation of traditional physical infrastructure with sensors and network connectivity [5]. In this paper, we focus on the middleware and application layers of a smart environment, specifically investigating the key data management requirements for supporting the creation of applications and decision support

tools. We are interested in the interface between new digital infrastructures and existing information systems, how to bring together contextual and real-time data [6,7], and how to easily share data between stakeholders in the environment.

Dataspaces are an emerging approach to pay-as-you-go data management that has proved fruitful for personal information and scientific data management. However, their use within Smart Environments and for Real-Time data remains largely unexplored and a number of open questions exist: What are the pay-as-you-go requirements of a smart environment? What real-time support services are needed by a dataspace? Can a set of incremental service are levels be defined for pay-as-you-go data management in a smart environment? What are appropriate data model in dataspace for a smart environment? How does a dataspace perform in a real-world deployment?

In this work, we explore the use of the dataspace paradigm within smart environments. The paper introduces the Real-time Linked Dataspace (RLD) as an enabling platform for data management within smart environments. At a high-level, the RLD combines the "Pay-As-You-Go" paradigm of dataspaces and linked data with entity-centric real-time query capabilities. The contributions of this paper are:

- It defines the requirements, and high-level design of a real-time linked dataspace for Internet-of-Things enabled smart energy and water environments.
- A principled incremental approach to "pay-as-you-go" data management with dataspace services providing tiered levels of support.
- A dataspace query service for real-time data streams and entities that enables unified queries across live streams, historical streams, and entities.
- Demonstrates the use of the approach within the OODA (Observe, Orient, Decide, and Act) Loop to build real-time analytics, decisions support, and smart apps for smart energy and water management.
- Experiences and lessons from the 5 real-world pilot smart environments

The structure of the paper is as follows: in section 2, it details the motivation for new forms of data management within smart environments by identifying common data requirements of 5 smart environments in the smart energy and water management domains. The coverage of these requirements within existing approaches is then analysed in section 3. Section 4 details the fundamentals of the RLD including the main components of the architecture. Section 5 describes the key dataspace support services and their tiered service levels. Section 6 discusses a principled approach to Pay-As-You-Go data management based on the tiered levels of the support services. In Section 7 the RLD is demonstrated in smart environments following the OODA (Observe, Orient, Decide, and Act) Loop to build real-time analytics, decisions support tools, and smart apps for energy and water management. Finally, the paper discusses the results of the pilots including experiences and lessons learnt.

2. IoT-enAbled Smart Environments

Mark Weiser et al. defined a smart environment as "a physical world that is richly and invisibly interwoven with sensors, actuators, displays, and computational elements, embedded seamlessly in the everyday objects of our lives, and connected through a continuous network" [8]. In the past decade, smart environments have moved from a research vision to concrete manifestations in real-world deployments. Internet of Things (IoT) projects such as the SmartSantander smart city project [9] are enabling smart environments by deploying tens of thousands of Internet-connected
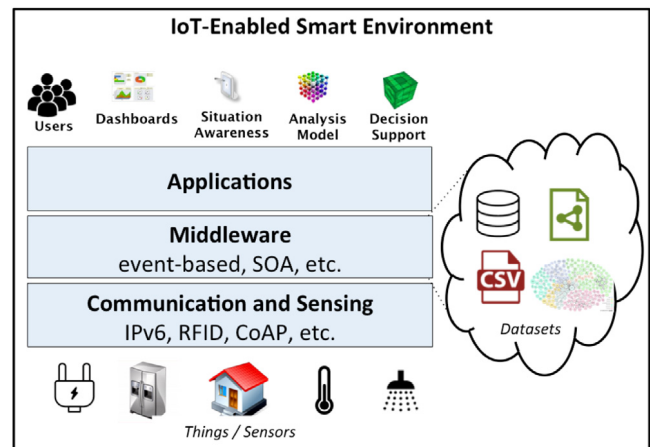

Fig. 1. Three-layered framework for an internet of things enabled smart environment.
Source: Adapted from [11].

sensor devices monitoring everything from solar radiation and temperature to flows of traffic and water. The sensor devices provide a digital representation of the state of the real-world, in the case of SmartSantander a digital representation of the city, which enables visibility into processes and operations of the city that can be analysed and optimised.

Real-world smart environments are encountering many challenges including the interoperability of diverse technologies (including legacy systems) [10], meeting the needs of diverse stakeholders with extensive goals and expectations, and working within the limited budgets available to invest in infrastructure. As illustrated in Fig. 1, the range of IoT technical challenges within a smart environment can be studied using a three-layered framework [11]:

- **Layer 1- Communication and Sensing:** A basic requirement is an infrastructure of communication and sensing that maps the world of physical things into the world of computationally processable data.
- **Layer 2- Middleware:** There is a need for middleware that can abstract the underlying technologies for the application developers. Data distribution, processing, and integration with legacy information systems take place at this layer.
- **Layer 3- Applications:** IoT-enabled applications will need to present the data gathered and analysed by the middleware layer in an intuitive and user-friendly manner using new visualisations and user experiences to ensure cognitive-friendly smart environments.

While significant efforts in the area of IoT come from communication and sensing levels, there has been a growing realisation that the challenges of the IoT will be more prevalent at the middleware layer [12] including data collection, management, analytics, and sharing. IoT-enabled smart environments are presenting new challenges for existing approaches to data management concerning their need for flexibility, dynamicity [1], and an ability to deal with incremental change while keeping costs to a minimum. Understanding these challenges in more detail requires an understanding of how the Internet of Things is enabling smart environments. In the scope of this paper, we examine the data management requirements of smart energy and water management systems as a study of the practical data requirements of smart environments.

2.1. Smart energy and water management

Managing energy and water holistically within a smart environment requires decision support tools that present meaningful

and contextual information about usage, price, and availability of energy and water intuitively and interactively to users. Users will need different forms of information to manage their energy and water consumption, from home users managing their water usage, business users managing the water consumption of their commercial activities, to municipalities managing regional distribution and consumption at the city level. In order to develop smart apps for these diverse users, it is necessary to leverage knowledge from a number of different domains including metering (household, neighbourhood, etc.), collection and catchment management, energy generation, environmental impacts, water quality, energy usage for pumping water, distribution networks, end-user feedback, occupancy patterns, meteorological data, etc. The design of next-generation smart energy and water management systems poses significant technical challenges regarding information management, integration of heterogeneous data, and real-time processing of dynamic data generated in smart environments.

2.2. Five IoT-enabled smart environments

Over the past three years, we have been involved in numerous projects [13–15] concerned with next-generation information platforms for smart energy and water management systems. As detailed in Fig. 2, the five pilots are:

- **Smart Airport:** Linate airport in Milan represents a large-scale commercial energy and water consumer with uses from washing activities, toilets, restaurants, and irrigation, flight operations, to safety critical infrastructure for emergency response. Linate targets a variety of users from the company's employees (including executives, operational managers and technical staff), to passengers. The variety of sensors used in the airport require the management of different events and their availability for applications in near-real-time. Significant contextual data from the Airport's operational legacy systems are needed to process the events for decision-making.
- **Smart Office**: The Insight Building was built in the 1990s without a building management system and was retrofitted with energy sensors. As typically in an organisation, Insight has many information systems that run its operations, including finance and enterprise resource planning, budgeting, and office IT assets. These enterprise systems can help in identifying energy wastage and promoting conservation actions within the office.
- **Smart Homes:** The Municipality of Thermi in Greece provides a residential smart water pilot with a representative sample of 10 domestic residences. The target users are the residents (both adults and children), municipality management, a developer community for smart home "Apps", research scientists, and the local water utility. Data from IoT devices in each home needs to be managed in a near-*real-time* manner to provide feedback to users on their water consumption. Secure sharing of datasets with both the research and developer community is needed.
- **Mixed Use:** The Engineering Building at NUI Galway is a state-of-the-art smart building with significant numbers of sensors and actuators. Target users include academic staff, managers, technicians, researchers, and students. This smart environment is designed to be a 'living laboratory' where the building itself is an interactive teaching tool where students can utilise data from the environment in their projects and research works.
- **Smart School:** Coláiste na Coiribe is a newly constructed Irish language secondary school. The school accommodates students aged 12–18, together with teaching and operational staff. The school has been fitted with a commercial state-of-the-art building management system to manage its energy and water consumption. A key challenge is to customise the communication of energy and water data for the diverse range of school stakeholders.

2.3. Common data requirements

For each of the five smart environments detailed, a system analysis was performed to identify the functional and non-functional information processing requirements that were needed to develop smart applications. The shared data requirements identified across the pilots are:

Pay-as-you-go data integration, accessibility, and sharing: Each smart environment contains potentially thousands of data sources from sensors and things to legacy information systems. Harnessing this data is critical to enabling the smart environment, challenges include the integration of multiple formats and semantics, discoverability and access, and data re-use and sharing in a low-cost and incremental manner [9,16–19]. This high-level requirement can be broken down into a set of technical requirements:

- *Standard Data Syntax, Semantics, and Linkage:* Facilitate integration and sharing, ideally with open standards and non-proprietary approaches.
- *Single-Point Data Discoverability and Accessibility:* Allow the organisation and access of datasets and metadata through a single location.
- *Incremental Data Management:* Enable a low barrier to entry and a pay-as-you-go paradigm to minimise costs.

Secure access control: Support data access rights to preserve the security of data and privacy of users in the smart environment. Access control is needed at both the level of the dataset and at the level of the data-item (e.g. entity-value).

Real-time data processing and historical querying: Each environment requires support for the real-time processing of data generated from sensors and things within the environments. This requirement can be broken down into two technical requirements:

- *Real-time Data Processing:* Including ingestion, aggregation, and pattern detection within event streams originating from sensors and things in the smart environment.
- *Unified Querying of Real-time Data and Historical Data:* Provide applications and end-users with a holistic queryable state of the smart environment at a latency suitable for user interaction.

Entity-centric data views: Applications and end-users need to be able to explore and query the data from an entity perspective, such as the energy or water usage in a specific building zone. The raw data generated by things (e.g. a Smart Tap) within the environments often only report on the observed values of a particular property (e.g. water consumption). Thus, the raw sensor/thing data may require additional contextual information, such as the location of the sensor [16–18]. This high-level requirement can be broken down into two technical requirements:

- *Entity Management:* The storage, linkage, curation, and retrieval of entity data, such as users, zones, and locations.
- *Event Enrichment:* Enhancement of sensor/things streams with contextual data (e.g. entities) to make the stream data more encapsulated and useful in downstream processing.

The level of importance of these shared data requirements varied within each pilot as detailed in Table 1. Many other requirements were also identified, in particular, interoperability of devices and network protocols, user profiling, the resilience of remote sensors, and advanced privacy-preserving analytics. However, these issues are beyond the scope of this work.

	Airport	**Office**	**Home**	**Mixed Use**	**School**
LOCATION	LINATE AIRPORT, MILAN, ITALY	INSIGHT, GALWAY, IRELAND	HOUSES, THERMI, GREECE	ENGINEERING, NUI GALWAY	COLÁISTE NA COIRIBE, IRELAND
TARGET USERS	• Corporate users • ~9.5 million passengers • Utilities mgmt . • Maintenance staff • Environmental managers	• 130 staff • Office consumers • Operations managers • Utility mroviders • Building managers	• Domestic consumers (adults, young adults, and children) • Utility providers	• Mixed/Public consumers • Building managers • 100 staff • 1000 students (ages 18-24)	• 500 students (ages 12-18) • 40 teachers • School mgmt. • Maintenance staff
INFRASTRUCTURE	• Safery critical • 10 km water network • Multiple buildings • Water meters • Energy meters • Legacy systems	• 2190 sq.m space • 22 offices + 160 open plan spaces • Conference room • 4 meeting rooms • 3 kitchens • Data centre • 30 person café • Energy meters	• 10 households • Typical variety of domentic settings including kitchen, showers, baths, living room, bedrooms, and garden • Water meters	• Water meters • Energy meters • Rainwater harvesting • Café • Weather station • Wet labs • Showers	• Water meters • Energy meters • Rainwater harvesting

Fig. 2. Five smart environments used to identify data management requirements.

Table 1
Level of importance of common data requirements within pilot smart environments.

Requirements	Smart Airport	Smart Office	Smart Homes	Mixed Use	Smart School
Standard Data Syntax, Semantics, and Linkage	High	Medium	Low	Medium	Medium
Single-Point Data Discoverability and Accessibility	High	Medium	High	High	Medium
Incremental Data Management	High	High	Low	High	Medium
Secure Access Control	High	High	High	High	Medium
Real-time Data Processing	High	High	Medium	High	High
Unified Querying of Real-time Data and Historical Data	High	High	High	High	High
Entity Management	High	High	Medium	High	Medium
Event Enrichment	High	High	High	High	Medium

3. Related work

Related works to data sharing and management within smart environments can cover a broad range of topics that touch on issues from the high-level visions of Cyber–Physical Social systems [20], to policy perspectives in service integration in smart cities [21]. This section focuses on the common requirements identified in Section 2.3 to survey the capabilities of existing approaches and highlight the main contribution of this paper. A summary of the analysis is in Table 2.

The CityPulse [17] project provides a distributed system for semantic discovery, data analytics, and interpretation of large-scale and near-real-time Internet of Things data and social media data streams [20]. In addition to providing unified views of the data, CityPulse also provides data analytics modules that perform intelligent data aggregation, event detection, quality assessment, contextual filtering, and decision support. CityPulse supports open standards for semantics, real-time stream processing, and partial entity management. However, no support exists for single-point data access, a pay-as-you-go data management paradigm, unified views over real-time and historical data, security, and event streams enrichment.

The OpenIoT [18] platform enables the semantic interoperability of IoT services in the cloud through the use of the W3C Semantic Sensor Networks (SSN) ontology [22], which provides a common standards-based model for representing physical and virtual sensors. OpenIoT provides a middleware for uniform access to IoT data and support for the development and deployment of IoT applications. OpenIoT supports open standards for semantics, real-time stream processing, security, and entity management. However, it lacks support for single-point data access, a pay-as-you-go data management paradigm, unified views over live and historical data, and event streams enrichment.

The SmartSantander project developed the City Data and Analytics Platform (CiDAP) [9] a centralised platform to access data generated from multiple heterogeneous sensors installed in a city. The platform can deal with historical data and near real-time information in an architecture similar to Lambda [23]. CiDAP provides limited support for data management beyond the low-level sensor streams and pushes these concerns to the application level. The result is applications duplicating common data management functionalities. SmartSantander follows open standards for semantics, single-point data access, security, real-time stream processing, and partial unified queries over streams and datasets. However, it lacks support for an incremental data management paradigm, entity management, or event streams enrichment.

The Spitfire [16] project uses semantic technologies to provide a uniform way to search, interpret and transform sensor data. Spitfire works towards a Semantic Web of Things, by providing abstractions for things, basic services for search and annotation, as well as by integrating sensors and things into the Linked Open Data (LOD) cloud. Spitfire mainly adopts semantic web standards for describing data, partial secure access control, entity management and event enrichment. It does not support single point access for data, incremental data management, real-time data processing, or unified queries for real-time and legacy data.

Table 2
Comparison of related frameworks to common data requirements.

Requirements	CityPulse [17]	Open IOT [18]	Smart Santander [9]	Spitfire [18]	Thing Store [24]	Real-time Linked Dataspace
Standard Data Syntax, Semantics, and Linkage	Yes	Yes	Partial	Yes	No	Yes
Single-Point Data Discoverability and Accessibility	No	No	Partial	No	No	Yes
Incremental Data Management	No	No	No	No	No	Yes
Secure Access Control	No	Yes	Yes	Partial	Partial	Yes
Real-time Data Processing	Yes	Yes	Yes	No	Yes	Yes
Unified Querying of Real-time Data and Historical Data	No	No	Partial	No	No	Yes
Entity Management	Partial	Yes	No	Yes	No	Yes
Event Enrichment	No	No	No	Partial	No	Yes

ThingStore [24] provides a "marketplace" for IoT applications development with the ability to deploy and host. The platform provides support for event detection, service discovery, an Event Query Language together with event notification and management. The architecture of ThingStore is a computation hub to connect things, software and end-users. Thingstore mainly supports secure and real-time data processing and lacks support for open standards to describe data, single-point access for data, entity management and event enrichment, incremental data management, and unified queries for real-time and legacy data.

From the analysis in Table 2, we note that existing works mainly support semantic descriptions of data according to open standards such as Semantic Web and Linked Data. However, they lack an incremental data management paradigm and do not support a single access point to discover and access datasets. Most related works address the real-time processing of data but do not provide unified access to it along with historical data. Half of the works provide some support for entity management. However, streams are not typically enriched with contextual data. Based on the analysis of existing work concerning the requirements identified we can see there is a clear need for an incremental pay-as-you-go data management, a single point of data/stream access, support for entity-centric views of real-time and historical data, and streams enrichment for better entity-centric and contextual data retrieval.

4. Real-time linked dataspaces

In this section, we introduce a Real-time Linked Dataspace (RLD) to meet the key data requirements identified for smart environments. The RLD adopts the pay-as-you-go paradigm of dataspaces and linked data with support for entity-centric real-time query capabilities. This section details the fundamentals of the dataspace approach and describes the architecture for the Real-time Linked Dataspace.

4.1. Dataspaces

A Dataspace is an emerging data management architecture that is very distinct from current approaches to data management. The dataspace approach recognises that in large-scale integration scenarios, involving thousands of data sources, it is difficult and expensive to obtain an upfront unifying schema across all sources [25]. Dataspaces are not a data integration approach [25], they shift the emphasis to providing support for the co-existence of heterogeneous data that does not require a significant upfront investment into a unifying schema. Data is integrated on an "as-needed" basis with the labour-intensive aspects of data integration postponed until they are required. Dataspaces reduce the initial effort required to set up data integration by relying on matching and mapping generation techniques. This results in a loosely integrated set of data sources. When tighter semantic integration is required, it can be achieved in an incremental "pay-as-you-go" fashion by more closely integrating the required data sources.

The goal of a Dataspace Support Platform (DSSP), as detailed in [25], is to provide a set of common related support services to all data sources within the dataspace (e.g. keyword search). The DSSP provides a base functionality needed for data integration that enables developers to focus on application-specific challenges instead of the common data integration tasks faced when working with multiple data sources. To achieve this goal, the DSSP must support all of the data in the dataspace requiring it to work with a large variety of data formats and system interfaces. A dataspace does not host data, the data resides in their native systems, as such it is not in full control of the data and may only provide weak guarantees of consistency and durability. When stronger guarantees are desired, more effort can be put into making agreements among the various systems. To this end, a DSSP must provide tools to support the tighter integration of data in a pay-as-you-go manner. As a result of the varying levels of data integration, the DSSP offers varying levels of service and often will only be able to provide best-effort or approximate results using the data accessible at the time of the query [25]. A comparison of the dataspace paradigm to traditional DBMS is provided in Table 3.

The usage of Dataspaces has been considered in a number of different contexts including managing Personal Information [26, 27], Astronomical data [28], and Biomedical data [29], while research into dataspace support services includes Integration and Curation [30], context-based query [31], data modelling [32,33], data mining [34], and user feedback [35,36].

Dataspaces can provide an approach to enable information management in smart environments that would help to overcome technical and conceptual barriers to information interoperability. However, there has been limited work on the use of the dataspace approach within smart environments and the investigation of relevant support services necessary for real-time data sources. This work builds on past efforts to use dataspaces in Building Data Management [13], Energy Data Management [14], and System of Systems [37]. However, these efforts do not cover the full range of requirements identified for smart environments identified in Section 2.3. In particular, they do not support a principled approach to incremental data management based on a set of support services with tiered levels of support. Finally, current works lack support for a unified entity-centric query framework over real-time and historical data streams in the smart environment.

4.2. Architecture

The Real-time Linked Dataspace (RLD) contains all the relevant information sources within a smart environment including things, sensors, and datasets, and has the responsibility for managing the relationships between these participants. The RLD goes beyond a traditional dataspace approach [25] by supporting the management of entities within the smart environment as first-class citizens along with data sources, and it extends the dataspace support platform with real-time processing and querying capabilities. Fig. 3 illustrates the architecture of the RLD with the following central concepts:

Table 3
DBMS vs. Real-time linked dataspace comparison.
Source: Adapted from [37].

	Model	Formats	Control	Query	Integration	Data processing
DBMS	Relational	Homogeneous	Complete	Exact	Upfront	None
RLD	All	Heterogeneous	Partial	Approx.	Incremental	Real-time & Batch
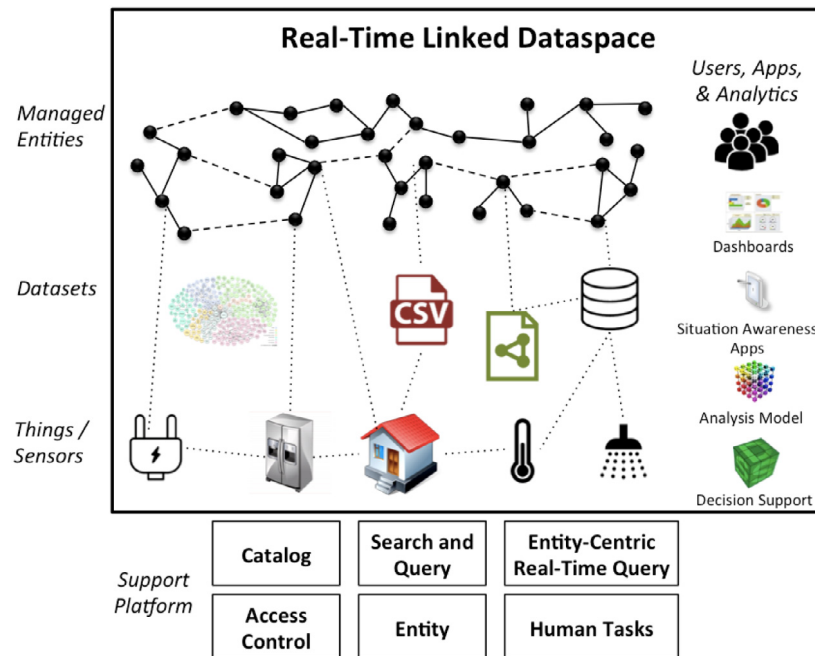
Fig. 3. Real-time linked dataspace architecture.

- **Things/Sensors:** Produce real-time data streams that need to be processed and managed. Things in a smart environment include connected devices, energy and water sensors, and occupant sensors.
- **Datasets**: Available in a wide variety of formats and accessible through different systems interfaces. Example datasets include building management systems, energy and water management systems, passenger information systems, financial data, weather, and (linked) open datasets.
- **Managed Entities:** Actively managed entities within the smart environment including their relationship to participating things, data sources, and other entities in the RLD.
- **Support Platform**: Responsible for providing the functionalities and services essential for managing the dataspace. Key services are the catalog, and the entity-centric real-time query service discussed further in Section 5.
- **Users, Apps, and Analytics:** Interact with the RLD and leverage its data and services to provide data analytics, decision support tools, user interfaces, and data visualisations. Apps/Users can query the RLD in an entity-centric manner. Users can be enlisted in the curation of the data and entities via the Human Task service.

5. Support platform services

The Real-time Linked Dataspace-Support Platform (RLD-SP) provides a set of core services to support developers with a base functionality when working with sources in the RLD. Each of the services in the RLD-SP has been designed to follow the Pay-As-You-Go paradigm to support varying levels of service offerings to the participants in the smart environment. This section details these

services and their tiered-levels of support including the catalog, access control, search and query, entity management, entity-centric real-time query, and the human task service.

5.1. Catalog service

The catalog service plays a crucial role in providing information services for participants in the dataspace including search, browse, and query services. The catalog service extends the original CKAN[1] portal with functionality for an entity-centric view of the dataspace.[2] The catalog provides a registry of:

- **Datasets:** May contain contextual information about a building or thing within a smart environment, real-time sensors data, enterprise data (e.g. customer data, enterprise resource planning systems, etc.), and open data such as weather forecast data.
- **Managed Entities:** An entity defines a concrete instance of a concept within the smart environment (e.g. a sensor or a water outlet). The catalog tracks critical entities in the smart environment and links those entities with the datasets and streams that contain further information about the entities. Metadata about an entity includes the identifier, entity type, and associated datasets. Fig. 4 shows an example of entities defined in the catalog.
- **Applications/Users:** Applications are the descriptions of software that utilises datasets from the RLD users. For example mobile applications, public displays, data services, analytics, web applications, and interactive dashboards. Users

---

1   https://ckan.org.
2   Demo Video available at this link: https://www.youtube.com/watch?v=fHcL-bOREIU.
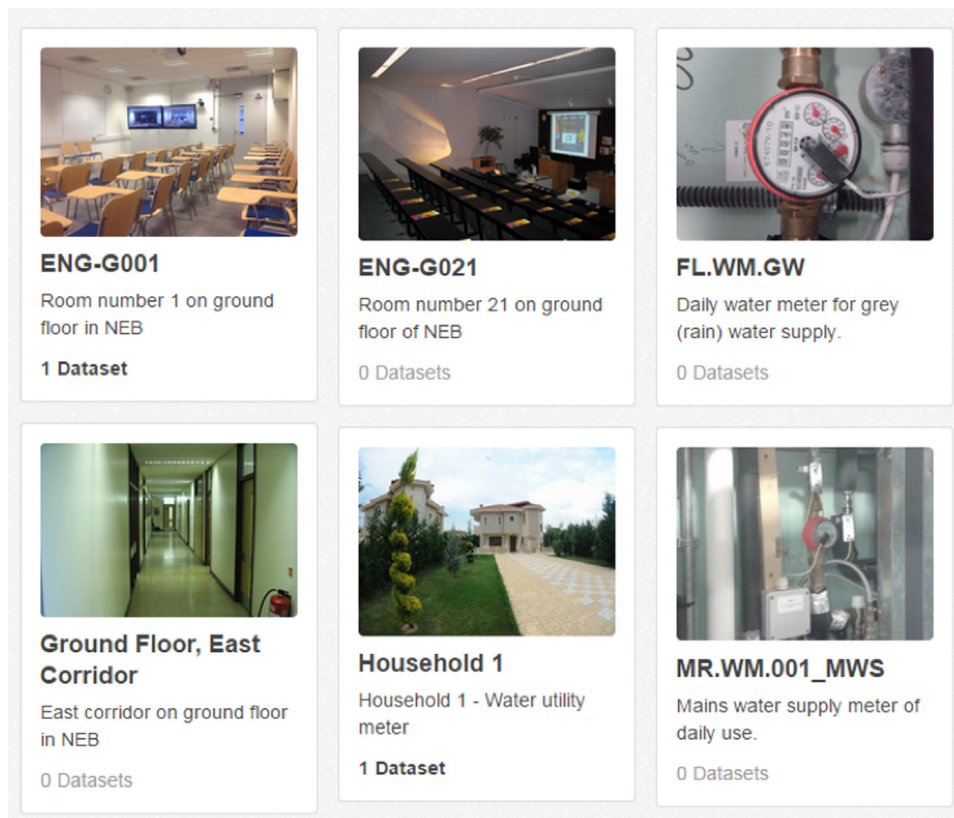
Fig. 4. Example of managed entities in the catalog.

and their role within the RLD are also managed by the catalog.

Within the catalog all datasets and entities are declared along with relevant metadata. The tiered level of service provided by the catalog increases as follows:

- **Registry:** A simple registry of datasets and streams, only pointing to the interfaces available for access.
- **Metadata:** Describing datasets and streams in terms of schema and entities in a non-machine-readable format (e.g. PDF document).
- **Machine-readable:** Machine-readable metadata and simple equivalence mappings between dataset schemas to facilitate queries across the dataspace.
- **Relationships**: Relations between schema and concepts across the dataspace.
- **Semantic Mapping:** Semantic mappings and relationships between domains of different datasets; thus, supporting reasoning and schema agnostic queries.

5.2. Access control service

The access control service ensures secure access to the data sources defined in the catalog. Access is managed by defining access roles for applications/users to the data source/entity that are declared in the catalog. As illustrated in Fig. 5, the secure query service is an intermediary between the applications/users and the dataspace by using the catalog as a reference to verify applications/users to the actual data sources. The advantage of this approach is to keep the applications/users' profiles centrally managed by the catalog under the governance of the dataspace managers. Within the pilot deployments we defined 3 types of

roles for users/applications Dataspace managers, App developers, and End-users. In terms of tiered levels of support the access control service can limit access as follows:

- **None:** The source is not managed by the access control service.
- **Coarse-grained:** Access is limited to the user at the dataset level.
- **Fine-grained:** Access is limited to users at the entity-level.
- **Data anonymisation:** Access to sanitised data for privacy protection. (Not supported in pilots.)

5.3. Search and query service

The objective of the Search and Query service is to help developers and users to find relevant datasets within the dataspace. Users can navigate the dataspace by entities (if supported), or by performing a search or query on the datasets. A key challenge in developing search and query services over heterogeneous sources in a dataspace is the expressivity–usability trade-off. An ideal dataspace query mechanism must provide both high expressivity and high usability [38]. As data sources are more closely integrated into the dataspace the search and query service can offer the following level of functionality:

- **Browsing:** Browsing of the datasets available in the dataspace catalog.
- **Keyword Search:** Basic keyword search of the sources within the dataspace.
- **Structured Queries:** Structured query using SPARQL where the user understands the underlying schema of the data.
- **Question Answering:** A best-effort entity-centric natural language interface to the dataspace that allows users to ask
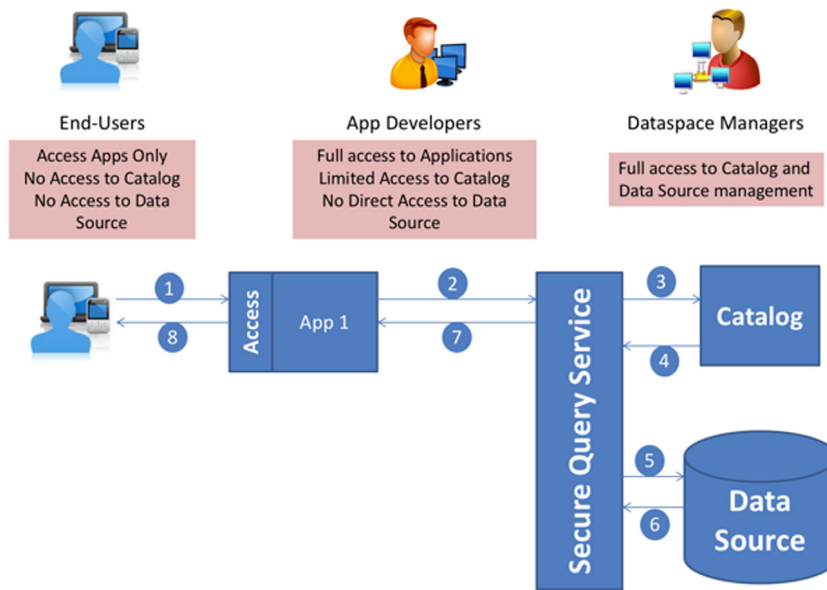
Fig. 5. Query workflow using the secure query service.

questions without understanding the underlying schema [39]. Note, in the pilots we did not implement a question answering system as it was not an explicit requirement.

5.4. Entity management service

Managing information about the critical entities (e.g. real-world objects) in a smart environment is an essential requirement for decision-making applications that rely on accurate entity information. An essential aspect of the RLD is the treatment of entities as first-class citizens. This section details how the RLD uses Semantic Web and Linked Data techniques to manage entities and provides an example of a managed entity from a pilot site.

5.4.1. Semantic web and linked data

Semantic Web and Linked Data leverage open protocols and W3C standards of the Web architecture for sharing structured data on the web. Semantic Web provides a set of standards, tools, and techniques to facilitate sharing and reuse of data across domains. It primarily uses a graph-based representation framework for structuring data and uses standard ontology languages for defining the semantics of data. Ontologies and vocabularies provide a shared understanding of concepts and entities within a domain of knowledge which supports automated processing for data using Semantic Web tools.

The fundamental concept of Linked Data is that data is created with the mindset of sharing and reuse. Linked Data proposes an approach for information interoperability based on the creation of a global information space [40]. The main components of this approach are: (1) Universal Resource Identifiers (URIs) to name things, (2) Resource Description Framework (RDF) for representing data, (3) Linked Data principles for publishing, linking, and integration, (4) Vocabularies to establish and share understanding, and (5) Bottom-up incremental agreement. Linked Data uses open web standards in conjunction with four basic principles for publishing data. These principles are:

- **Naming:** Use URIs as names for things — the use of a Uniform Resource Identifier (URI) (similar to URLs) to identify things such as a person, a building, a device, an organisation, an event or even concepts such as risk exposure or energy

and water consumption, simplifies reuse and the integration of data.
- **Access:** Use URIs based on HyperText Transfer Protocol (HTTP) so that people can look up those names — URIs are used to retrieve data about objects using standard web protocols. For an employee, this could be their organisation and job classification, for an event this may be its location time and attendance, for a device this may be its specification, availability, price, etc.
- **Format:** When a URI is looked up (dereferenced) to retrieve data, provide useful information using a standardised format. Ideally, in Web standard such as RDF.
- **Contextualisation:** Include links to other URIs so that more information can be discovered. Retrieved data may link to other data sources, thus creating a data network, e.g., data about a product may link to all the components it is made of, which may link to their supplier.

In terms of implementing a dataspace for Smart Environments, Semantic Web and Linked Data have three advantages: (a) Separate systems that are designed independently can be later joined/linked at the edges, (b) Interoperability is added incrementally when needed and where it is cost-effective, and (c) Data is expressed in a mixture of vocabularies.

5.4.2. Managed entities

The Entity Management Service (EMS) is concerned with the maintenance of information about entities within the smart environment and together with the catalog service acts as the canonical source of entity (meta)data. Each managed entity within the EMS can be accessed via a URI which can be used to retrieve detailed entity data. The URI also serves as a canonical identifier for the entity. Each entity is linked with datasets that contain information related to it. The relationship between entities and datasets can quickly become complicated within a smart environment. This is a significant challenge within traditional data integration approaches and requires significant upfront investment. The RLD follows the incremental dataspace philosophy. In practice, you only connect data sources related to an entity on an as-needed basis. The approach encourages that entities should be as minimal as possible to achieve the desired results. Fig. 6 describes a minimal data model for entities in one of the smart water pilots.
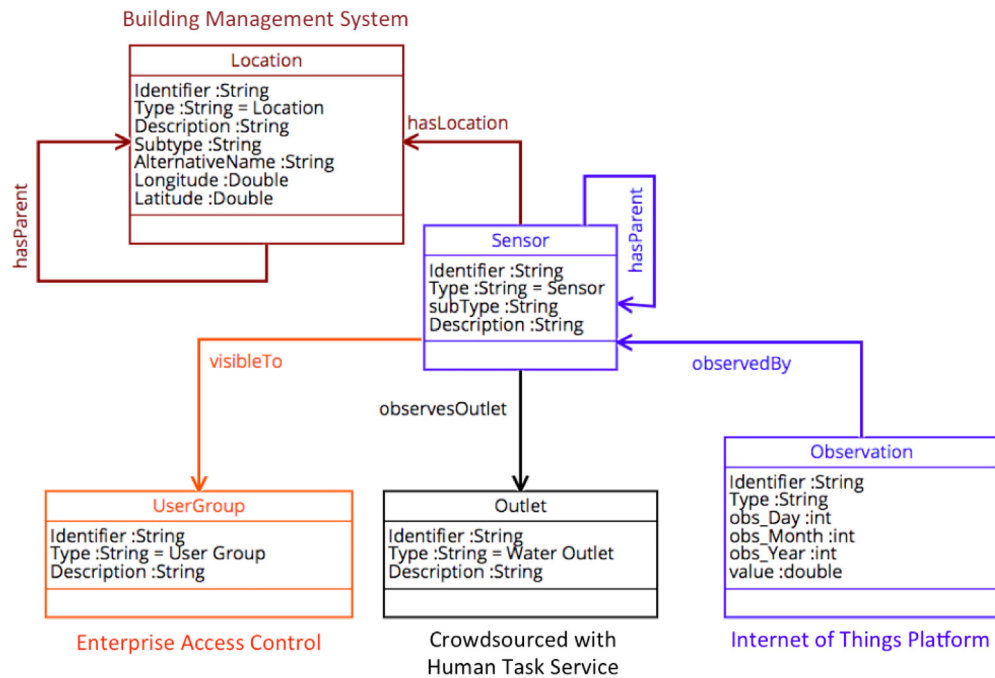
Fig. 6. Minimal data model for entities in a smart water management environment.

The key components and entities of the data model and the sources they originate from are:

- **Sensor:** Measures the flow of water and generates a stream of data to calculate the water consumption levels of the area covered by the sensor (from an Internet of Things Platform).
- **Observation:** Various types of sensors may be installed for metering water consumption; therefore, it is necessary to precisely describe the sensor capabilities including the units and rate of measurement (from an Internet of Things Platform).
- **Outlet:** Information on the actual physical water outlet is necessary 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 (crowdsource using human task service).
- **Location:** Information on the associated spatial locations serviced by the water pipe (from Building Management System).
- **User Group:** Each sensor is associated with a set of users who have permission to access the data (from Enterprise Access Control). This information is used by the access control service described earlier in Section 5.2.

Active entity management can follow the incremental improvement philosophy of dataspaces. The entity management service has the following levels of incremental support:

- **None:** Entities are not managed.
- **Documented:** Entity description (e.g. schema and identifiers) are documented in a non-machine-readable format (e.g. PDF document).
- **Source-level:** Machine-readable entity at the source level.
- **Multi-source mapping:** Canonical identifiers for entities in the dataspace and mapping across sources.
- **Entity Knowledge Graphs:** Entities are semantically linked to other related entities, data, and streams across the dataspace.

5.5. Entity-centric real-time query service

A key requirement in a smart environment is to support the querying of real-time data streams. Within the RLD this is achieved by the entity-centric real-time query service that enables unified queries across live streams, historical streams, and entity data to enable full entity-centric views of the current and past state of the smart environment. This section first discusses the Lambda Architecture and then details how it has been extended in the RLD-SP to support entity-centric real-time queries.

5.5.1. Lambda architecture

The Lambda Architecture is a frequently used Big Data processing architecture that realises the need for real-time data analytics crucial to support data analysis within smart environments. Rather than using two different systems for processing real-time data and historical data, the Lambda Architecture allows seamless ingestion and processing of live and historical streaming data [23] within a single architecture. Streams of events can be sourced from a variety of systems such as sensors, database logs, and website logs. All data entering the system are processed by both the batch layer and the speed layer. The batch layer pre-computes batch views of the stored raw data. The serving layer indexes the batch views for low-latency fast-access queries by applications. The speed layer deals with high-velocity updates by providing real-time append-only views of recent data. Queries are answered by merging results from both batch views (data-at-rest) and real-time views (data-in-motion). The Lambda Architecture has proved very useful for data management within smart environments [9].

5.5.2. Entity-centric real-time index and architecture

Concerning real-time data processing, the Lambda approach meets many of the requirements defined in Section 2.3. However, Lambda does not natively support the inclusion of entity and contextual data within the indexing and querying process. This means that applications need to maintain the relationship between the Lambda index and the entities in the dataspace by themselves. Ideally, an entity-centric real-time query service would be provided
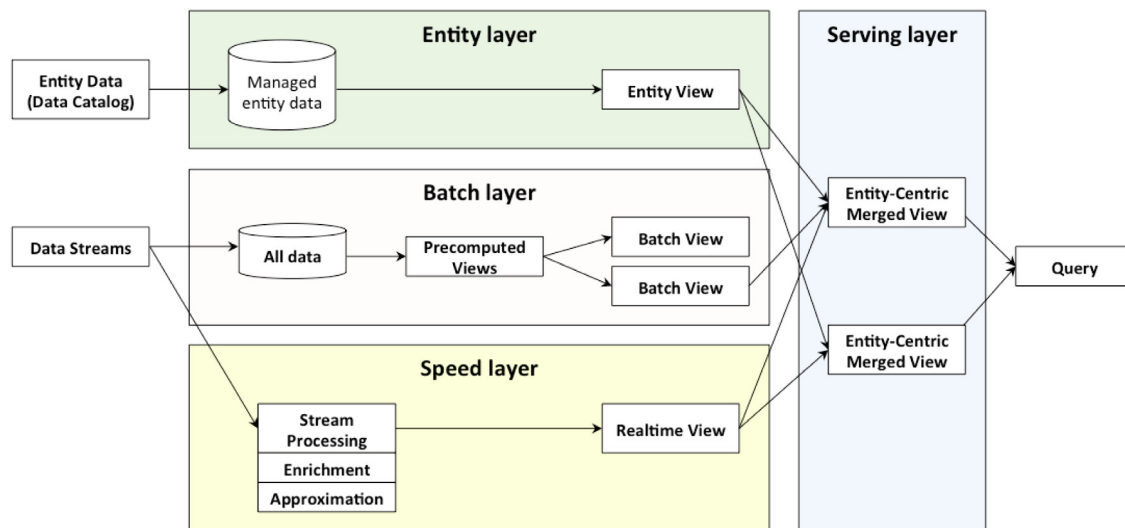
Fig. 7. The four layers of the entity-centric real-time query service.

by the RLD-SP to remove the need for applications to manage the entity/stream relationship.

To meet this requirement, we designed an extension of the Lambda architecture that includes the addition of an entity layer for the indexing of entity data alongside historical and live streams. The approach enables the serving layer to provide merged views across all three layers, removing the need for applications to maintain the entity/stream relationships. The entity-centric real-time query service is part of the RLD-SP and is tightly integrated with other support services such as the catalog, entity management, and access control services.

Fig. 7 illustrates the design of the entity-centric real-time query service. The main components are:

- **Entity Data (Data Catalog):** Provides entity data from the catalog and entity management services.
- **Data Streams:** Produced by the "things" and sensors within the smart environment.
- **Batch Layer:** Provides batch-based processing for accurate, but delayed views of historical data.
- **Speed Layer:** Provides real-time processing for data with low latency processing requirements. Streams in the speed layer are not stored but instead processed on-the-fly to guarantee low-latency approximate views of the data to complement the older views achieved by the batch layer. The speed layer provides a number of support services for processing event data such as approximate event matching and event enrichment.
- **Entity Layer:** Provides a view of the managed entities within the RLD working closely with the catalog and entity management service.
- **Serving layer:** Provides applications and users with a single entity-centric interface for data access. This layer transparently splits queries to the batch, speed, and entity layers to combine pre-computed views over the three layers.
- **Query:** Request for entity-centric views from applications, analytics, and users.

The entity-centric real-time query service has the following levels of incremental tired support:

- **None:** Streams are not managed in the service.
- **Basic Processing:** Basic real-time stream processing in the speed layer only.

- **Historical Views:** Streams are stored in the batch layer for historical views.
- **Enrichment:** Streams are enriched with context and entity data from the catalog and entity management service.
- **Entity-Centric:** Streams are processed in all three layers to provide entity-centric real-time queries.

5.5.3. Approximation and enrichment

The speed layer has two event processing services for event approximation and event enrichment. The approximation service allows automatic semantic matching of events based on user-defined rules using a semantic matching model [41,42]. The semantic event matcher simplifies the task of things/sensors management as it allows the system to match semantically equivalent events across heterogeneous event streams automatically. This reduces the number of event processing rules which need to be written by users of the RLD. Further details on the semantic event matcher are available in [41] and [42].

Data streams can become difficult to process if forwarded beyond a system's boundary. The RLD is fundamentally a System of Systems approach that reflects the reality of multiple systems operating in a smart environment. Thus, it is crucial for the RLD-SP to support the processing of events and streams between systems in the dataspace. Raw data streams usually have a very minimal amount of data, such as sensor ID and the value of the sensor's reading. Enrichment of event data by adding extra contextual information such as where the sensors are located, or what entity it monitors, can result in a more complete event description making it easier for other systems to process the event. The enrichment service in the speed layer performs this task by enhancing the event description with data from related entities in the RLD. The entity management service plays a key role in the enrichment process as the mappings between managed entities are used to determine relevant contextual data for an event. Further details on event enrichment are available in [43].

5.6. Human task service

The Human Task service is concerned with the collaborative aspect of data management [35,44] within the RLD by enabling small data management tasks to be distributed among willing users in the smart environment [45]. The inclusion of users in
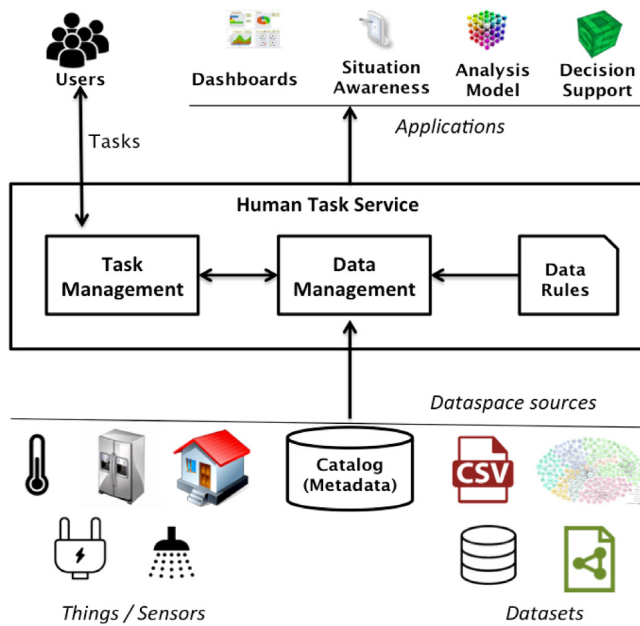
Fig. 8. Overview of the human task service for real-time linked dataspace.

the data management process not only helps in managing data but may help in building user trust and a sense of ownership of the dataspace. Fig. 8 shows a simple architecture for the human task service that includes (1) *Task Assignment:* matching between tasks and users in the smart environment [44] based on characteristics of tasks or the specific requirements of tasks in terms of human capabilities [35,45] and (2) *Quality assurance:* to ensure truthful and correct responses of tasks. The human task service has the following levels of incremental support:

- **None:** No human tasks are used.
- **Schema:** Tasks are used to map schemas between sources.
- **Entity:** Tasks are used to map entities between sources.
- **Enrichment:** Entities are enriched with contextual data.
- **Data Quality:** Tasks are used to improve the quality of data (e.g. verification).

The Human Task service may be called by apps using the RLD, or by other support services within the RLD-SP. A good illustrative example of the service in action is human tasks for entity enrichment. Based on their knowledge and understanding of the environment, users can help with enrichment for important entities. Fig. 9(a) shows an example enrichment task that is associated with an IoT device (e.g. CoAP sensor). The human task can be retrieved by scan the QR code on the device with a mobile phone or tablet device. The task is retrieved from the task service and asks the user a set of simple questions about the surroundings of the sensor to enrich the description of an entity in a smart building (e.g. Fig. 9(b) what are the features of the room where the sensor is located?). Further details on the Human Task service are available in [44–46].

Each of the support services described in this section play a central role in the incremental management of data within the RLD. Users of the RLD need to be aware of the different levels of data management available from each of these support services. In order to make it easier to understanding the incremental nature of the RLD, we have developed a rating framework for articulating the tiered data management within the RLD.

6. Five star pay-as-you-go data management

In contrast to the classical one-time integration of datasets that causes a significant upfront overhead, the RLD adopts a principled pay-as-you-go paradigm for supporting an incremental approach to data management. At the foundation of the approach is the principle that the publisher of the data is responsible for paying the cost of joining the dataspace. This pragmatic decision allows the RLD to grow and enhance gradually with participants joining or leaving the dataspace at any time. The next principle is that data is managed following a tiered approach where an increase in the level of active data management has a corresponding increase in associated costs. The tiered approach to data management provides flexibility by reducing the initial cost and barriers to joining the dataspace. The tiers are described using a variation of the 5-stars scheme defined by Tim Berners-Lee for publishing open data on the Web [47]. Berners-Lee's 5 star scheme is a rating system to determine how accessible, reusable, and interconnected data is on the web. The 5 star scheme has been extended to consider the level of integration of the data source with the RLD support services. At the lowest level, data only needs to be made available at the minimum cost. Over time the level of integration with the RLD-SP can be improved in an incremental manner on an as-needed basis. The more investment is made to work with the support services, the more integrated into the RLD they become. The 5-Stars Pay-As-You-Go model for the RLD is detailed in Table 4.

7. OODA apps with a real-time linked dataspace

John Boyd hypothesised that individuals and organisations undergo a continuous cycle of interaction with their environment. Boyd developed the OODA loop [48] as a decision process by which an entity (either an individual or an organisation) reacts to an event by breaking the decision cycle down to four interrelated and overlapping processes: Observe, Orient, Decide, and Act (OODA), through which one cycles continuously. Boyd initially applied the OODA loop to military operations, and it was later applied to enterprise operations, more recently it has been considered as an approach for processing observations within cyber–physical systems [20]. In this latter context we apply the OODA loop as a high-level design guide for smart energy and water systems within smart environments. As illustrated in Fig. 10, the four OODA processes applied to smart environments are:

- **Observation:** The gathering of data from the smart environment to understand its state.
- **Orientation:** The analysis and synthesis of data to form an assessment of the circumstances within the smart environment. Moving from Data to Information, Knowledge, and Insights.
- **Decision:** Consideration of the options to determine an appropriate course of action. The goal is to optimise the operation of the smart environment. The use of predictive modelling can play a significant role.
- **Action:** The physical execution of decisions via actuation (both automated and human). Once the result of the action is observed, the loop starts over again.

To validate the RLD approach it has been used in the development of OODA smart applications and decision support for the five smart energy and water environments in Section 2. The remainder of this section details the role of the RLD at each phase in the loop.
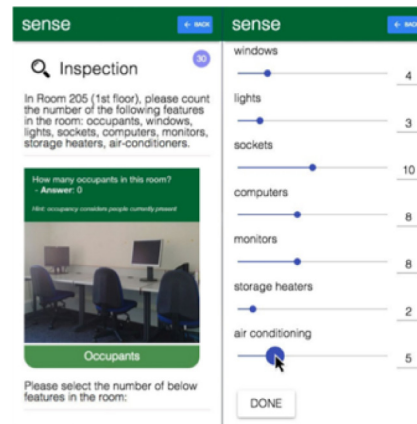
Table 4
5-star pay-as-you-go scheme for the real-time linked dataspace support services.

Star rating	Data format	Catalog	Access control	Search and query	Entity	Real-time	Human
* Basic	Any format (e.g. PDF).	Registry of Datasets, and Streams	None	None	None	None	None
** Machine-readable	Machine-readable structured data (e.g. Excel) Documentation to understand the data/stream structure, format and characteristics.	Non-machine-readable Metadata document (e.g. PDF)	Coarse-grained (Dataset level)	Browsing	Entities identifiers in documenta-tion	Stream processing	Schema Mapping
*** Basic Integration	Non-proprietary format (e.g. CSV, JSON, XML)	Machine-readable metadata Equivalence between Schema Concepts	Fine-grained (Entity-level) Secure query service	Keyword search	Source level (siloed)	Historical views of streams	Entity Mapping
**** Advanced Integration	Open standards (RDF, JSON-LD) to identify things/entities using the first two principles of linked data	Relationships between Schemas (dataspace level)	Fine-grained (Entity-level) Data anonymisation	Structured queries	Canonical identifiers and entity mappings across sources	Stream enrichment with context and entity data	Entity Enrichment
***** Full Semantic Integration, search, and query	Follows all publishing principles of linked data	Full Semantic Mappings	Fine-grained (Entity-level) Data anonymisation	Domain-agnostic question answering	Knowledge graphs semantically link entities to related entities, data, and streams	Entity-centric real-time query	Data Quality improvement

(a) Sensor Metadata Enrichment   (b) Entity Enrichment (e.g. Room features)

Fig. 9. Examples of a human task to enrich entities.

7.1. Observation

The RLD-SP services support the observation phase by min-imising the amount of effort required for a data source to join the RLD. The incremental approach of the RLD made it easier to gradually improve the collection of observations from the smart environment by adding a new sensor or dataset to the RLD. The 5-star schema was useful for specifying and planning the level of service needed for each data source.

The human task service enables the engagement of users in maintaining a high-quality catalog of managed entities. Active participation of users in a smart environment improves their en-gagement and sense of ownership while supporting a high-level of accuracy in the data maintained by the dataspace. In the Insight

pilot, we noticed a direct benefit of using the human task service for the collaborative management of entities in the environment to provide a more accurate and rich understanding of the environ-ment's state [44].

7.2. Orientation

The primary objective of the orientation phase is to support situational awareness of the smart environment. The entity man-agement service builds awareness regarding the entities in the environment through entity linking and enrichment. The search and query service and real-time query service enable users to un-derstand the current and historical state of the smart environment.

Within all of the pilots, a key goal is to increase the visibil-ity, understanding, and awareness of energy and water use. RLD
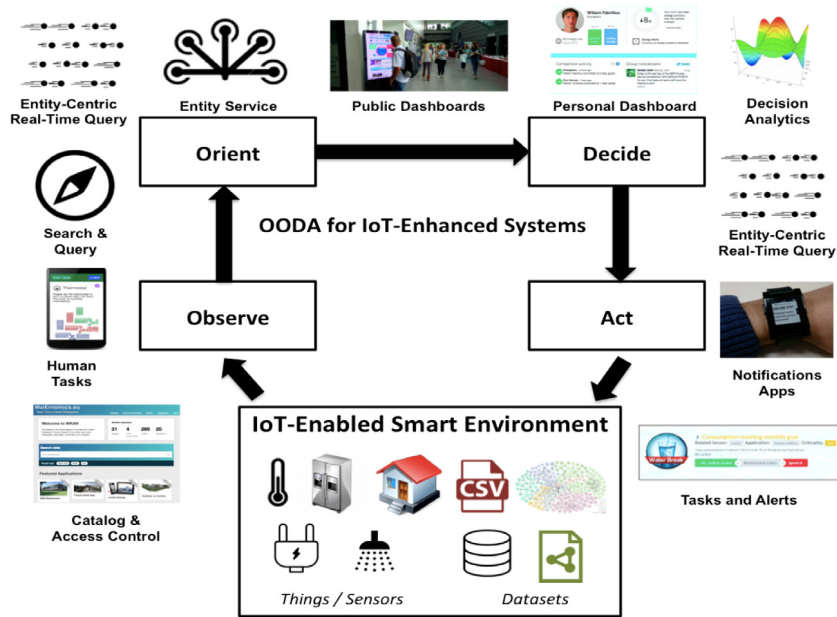
Fig. 10. Role of RLD-support services and application across OODA Loop phases in smart environments.

(a) Linate Airport (b) Smart School Display [3] (c) Smart Building

Fig. 11. Public interactive displays and dashboards.

services can be used to build dashboards to provide situational awareness and orientation to the users within each environment with targeted information on their environment's energy and water consumption. Within the different pilots, this is manifested in a variety of ways and at different time-frames, from informing the residents in the smart home as they live their day, the detailed analysis required by building managers and operational staff, to brief encounters with "frequent-flyer" passengers as they pass through the airport. User orientation in the pilots was driven by public displays, interactive touch screen displays and tablet apps (see Fig. 11). These user interfaces communicate current and historical energy and water usage within the environment, convey information about the importance of the resource, footprint, tips on how to improve consumption, and games to calculate the users' footprint in real time. The displays are also personalised to target different users by using metaphors so that they can use them in communicating relevant messages to the target user. The applications in the orientation phase make extensive use of real-time, historical, and contextual data sources.

7.3. Decision

Once users have built a certain level of awareness of the energy or water consumption of their environment, they can use their expertise to start taking decisions towards more sustainable behaviour. In the decision phase, a critical aspect of the dashboards is to provide users with targeted information on usage, goal setting, targets for conservation, and tips to improve their consumption behaviour. For example, managers can define consumption thresholds to serve as triggers for "alerts", notifying them of excessive usage, goals attained, or the detection of possible faults. Developing decision support apps is simplified by using the entity-centric real-time query service to analyse data from the environment.

A specific example of decision support is the Water Retention Time Observer application that determines the residence time of drinking water pipes and creates alerts in case of potential issues. In this context, the water retention time observer application can assist managers to provide timely notifications regarding low water quality in drinking water pipes. This is achieved by detecting inactivity in specific measurement points in the water network and sending a notification if stagnant water is detected.

(a) Actuation Task (b) Notification on SmartWatch

Fig. 12. Example of tasks and notifications within the smart environment.

7.4. Action

Applications in the action phase of the OODA Loop help users in smart environments meet their goals for energy and water consumption by taking appropriate actions. Real-time event processing is used to express these goals as a set of rules which can generate alerts and suggested preventive actions. Actions are then communicated to the users in the smart environment using appropriate means of communication: emails, notifications on dashboards, messages on smart devices, and human tasks. The occupants of the environment can participate in taking energy or water saving actions. In the Smart Building pilot, we implemented a collective energy management system where the RLD was used for the identification of energy saving tasks that were routed to the building occupants using the human task service to take energy conservation actions such as turning off the light in empty rooms, or closing a window when an air conditioner is on in a room. Fig. 12(a) shows an example of these "Citizen" actuation tasks.

The role of a building manager is a demanding one that often has personnel away from their desk working in the field. An anytime, anywhere, notification mechanism was needed for managers. The wearable info-centre application was developed to enable notification of high-priority alertst. Fig. 12(b) shows an example notification using the wearable info-centre.

8. Pilot results and lessons learned

This section presents the results and insights gained from deploying the RLD in the smart environments described in Section 2. Each pilot followed a similar methodology for design, deployment, and evaluation [15]. In this section, we detail the energy and water savings of the pilots, the performance of the entity-centric real-time query service, and a set of experiences with lessons learnt from deploying the RLD in the pilots.

8.1. Deployments

The RLD has been implemented mainly through an open source stack of technologies. As shown in Fig. 13 entities from data sources (e.g. BMS, sensors) are exported into the CKAN-based dataspace catalog. Batch data is fed into map/reduce jobs in the Apache Spark SQL node, while real-time data from sensors are fed into

the Apache Kafka message bus for distribution and then into map/reduce jobs in the Apache Spark Streaming node. Results from the batch nodes are then fed into a Druid indexer node as dimensional data, while the streaming data goes into Kafka and then into a Druid real-time node. The Druid nodes use the Apache Cassandra deep storage data store. Both batch data and real-time data are exposed transparently via a Druid broker node which can be queried by applications in JSON format.

8.2. Energy and water savings

During the initial period of the pilots, energy and water metering data was collected from existing monitoring systems to establish baselines for consumption across all pilots. During the control period, the users within the pilots had access to the data generated by the metering infrastructure system through traditional information systems (e.g. Building Management System, and basic public dashboards within the airports, office building, and school). The data collection period for each pilot spanned between 6 to 16 months which also included a range of user interventions such as pre-surveys, focus groups, interviews, and feedback cycles. The RLD was used to develop Smart Energy and Water Applications and decision support analytics across the pilot smart environments. Table 5 detail the characteristics of the pilots during the study period, number of events generated in the environment, number of apps deployed, and savings achieved in terms of energy and water. Regarding energy and water savings, the RLD supports these impacts in 3 fundamental ways:

- **Entity-Centric Views:** Connecting data across silos provided "big picture" entity-centric views of the resource consumption within the smart environments. Entity-centric views made it easier for the users within the smart environments (e.g. building managers) to identify waste and efficiency opportunities as the data was structured and organised around the real-world entities they work with everyday.
- **Quick Wins:** The Pay-As-You-Go approach was useful for building the business case and getting "buy-in" from users by enabling quick wins to demonstrate the benefit of the approach. Quick wins that clearly demonstrated energy and water savings encouraged non-technical business users to more actively engage with the RLD. The project team could
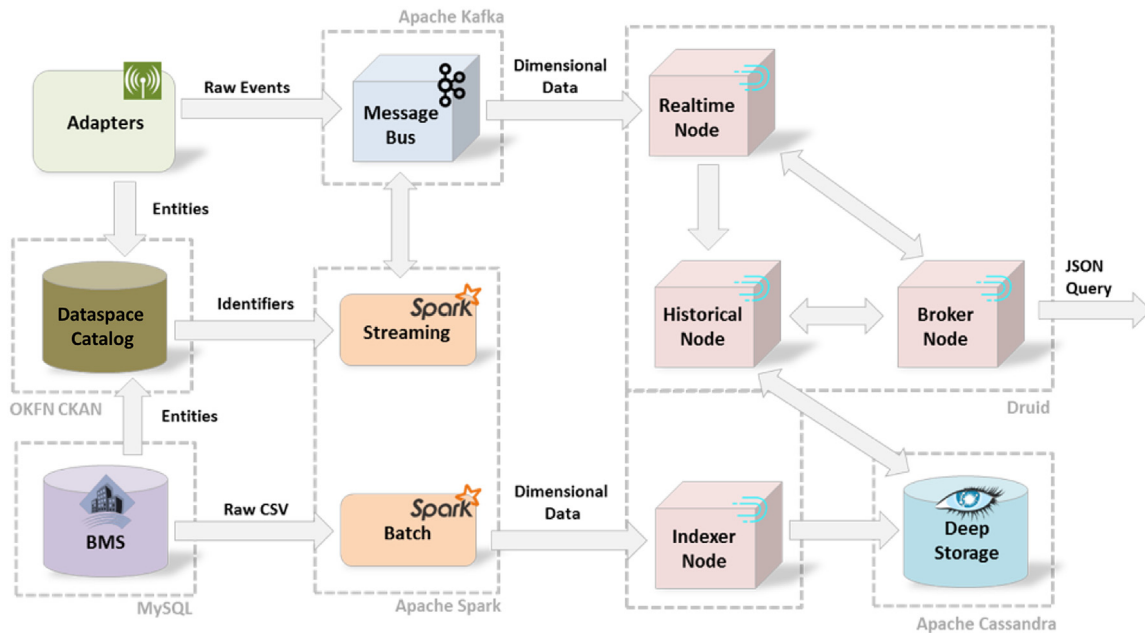
Fig. 13. Real-time linked dataspace deployment.

Table 5
Summary of impact of energy and water management in smart environments.

Pilot site	Study period	Events per year	Apps deployed	Actual savings measured	Est. annual savings
Linate Airport, Italy	10 Mts	∼11.5 million	8	2954 cubic m^3 3013 kg CO_2	54,000 m^3 55,080 kg CO_2
Thermi, Greece	16 Mts	∼2.3 million	11	30% water reduction	–
Engineering Building, NUI Galway	16 Mts	∼36 million	8	174 m^3 177 kg CO_2	8089 m^3 8251 kg CO_2
Coláiste na Coiribe, Galway, Ireland	12 Mts	∼1 million	5	2179 m^3 2223 kg CO_2	9306 m^3 9492 kg CO_2
Insight, Ireland	6 Mts	∼8 million	4	24% energy reduction	–

build a business case around apps and decision support tools that would reduce resource usage and its associated economic costs. The savings identified can be used to justify the necessary investment needed in data integration.

- **Actionable Notifications:** The RLD enabled highly specialised decision analytics that provided action and notification alerts for each of the pilot smart environments including leak detection, fault detection and abnormal usage patterns. These actionable alerts and notifications were crucial for building managers and operational staff who do not have time to study and analyse the data generated in the smart environment.

8.3. Entity-centric real-time query service performance

A key contribution of this paper is the entity-centric real-time query service for the RLD. The query latency for the service was evaluated within each environment to ensure it could support interactive user querying [49,50]. We evaluated seven common queries within the developed applications to determine the level of query interactivity of the service. Table 6 presents these results based on the average of 5 runs for each query. The majority of queries have an "instantaneous response" of under 0.1 s, and all queries are responsive under 1 s which is needed for "good navigation". This initial evaluation demonstrates the suitability of the query service for apps within the smart environments.

8.4. Experiences and lessons learnt

Based on a reflection of our experience using the RLD in several pilot environments, the following lessons were identified as key learnings to inform the design of future smart environments using the RLD.

Developer education: Across the pilots, we worked with a diverse set of developer teams with different backgrounds from embedded devices to web front-ends. The dataspace concept was new to most of them, and they were accustomed to working in an environment where they have full-control with the expectation of exact results. Also, the store-and-query culture is more common among developers and users. The processing of data on-the-fly and detecting only data of interest in real time, without storage in many cases, can be challenging (aka. event processing) for some developers to understand. Embracing the dataspace took time and required us to demonstrate both the benefits and limitations of the paradigm. Developer education was critical to the adoption of the dataspace. Workshops and tutorials held at pilot sites proved to be an effective mechanism of engaging developers in order to educate them on the capabilities of the platform and the dataspace data management approach.

Incremental data management can support agile software development: The project teams for each pilot operated using an agile software development methodology. The incremental dataspace approach proved useful during the design and development phase. The RLD enabled the teams to work at the pace suitable

Table 6
Query latency (seconds) of entity-centric real-time query service in selected pilots.

Query type	Airport A	Mixed Use A	Home A	School A	Airport B	Mixed Use B
timeBoundary	0.266415	0.076204	0.078805	0.076004	0.080605	0.078605
dataSourceMetadata	0.091405	0.078005	0.080805	0.153609	0.137808	0.092205
segmentMetadata	0.073804	0.084405	0.074004	0.140008	0.077405	0.077604
search	0.162609	0.142808	0.085805	0.076404	0.136808	0.204812
timeseries	0.072404	0.080005	0.077204	0.072404	0.134008	0.083605
groupBy	0.073404	0.078205	0.075604	0.081605	0.078605	0.072204
topN	0.078405	0.086805	0.072604	0.073004	0.077804	0.076604

to the stakeholders and data owners involved. The RLD allowed the project team to include new data sources during an iteration, or to increase the level of integration of an existing source. The decoupling achieved via the catalog and the use of event and streams, removed dependencies between parties. It enabled the team to work with participants in the pilots in an incremental manner where we could quickly demonstrate value with a low upfront investment in data integration. As the pilots progressed, more and more data became available in the RLD and enabled the creation of sophisticated data-intensive applications and analytics.

Build the business case for data-driven innovation: It is important to clearly articulate the business case for the RLD to justify the necessary investment in data infrastructure. Within our pilots, we discovered a strong business case for data-driven innovation by justifying the investment based on the resulting cost savings achieved by improving resource efficiency (e.g., energy and water savings). A key challenge was to bring together the different stakeholders to support and deliver the project. For example, within our pilot, the IT organisations had the data, but the savings resulting from the system benefit the operations teams of the organisations (e.g. water and energy). Thus, operations have a clear motivation to invest, but IT does not. By bringing these stakeholders together, we were able to build a holistic business case.

Integration with legacy data is a significant cost in smart environments: While sensors and connected devices are an essential source of data in a smart environment, they are not the only source of data necessary to make an environment "smart". In our pilots, a considerable number of different legacy data sources needed to be integrated to collect the necessary information to make informed and intelligent decisions. While the RLD provided an effective incremental approach that integrates legacy data at a minimum cost, it is not a silver bullet to data integration costs in smart environments and the cost of integrating with legacy data should not be underestimated. This is of particular relevance within enterprise settings where the non-technical challenges (e.g. sharing data between departments) can be as significant as the technical ones.

The 5-star pay-as-you-go schema simplified communication with non-technical users: The Pay-As-You-Go star schema was very useful regarding communicating both enhanced functionality and the additional costs of tighter integration with the RLD-SP services. Within the pilots, it was widespread to integrate data to the 3-star level on most services. The investment to bring a source to 4 and 5 stars was only made for core datasets within a pilot, and not for each service. Interestingly, many datasets that were initially identified in the early design phases as of high-importance (e.g. sensor specifications, detailed infrastructure schematics) remained at the 1-star level as they were not needed by the final application developed. This resulted in significant savings by avoiding unnecessary integration costs. Within the commercial pilots the 5-stars model supported the articulation of the business case for the investment necessary to include data sources in the dataspace.

A secure canonical source for entity data simplifies application development: Programmable access to the catalog by enabling queries over machine-readable metadata and entities was crucial to facilitate application development in the dataspace. The role of the catalog as a canonical source for identifiers for entities was critical to manage the entities in the dataspace. Demonstrating the secure query service was essential to get "buy-in" and build trust with the pilot data owners. For example, sensor data within the domestic pilot were of a sensitive nature, and we needed to assure residents that access was restricted to privileged users.

Data quality with Things and Sensors is challenging in an operational environment: Data quality challenges are further complicated as participating data sources, and things within the RLD are not under its full control. Data quality issues included incorrect file formats, incorrect timestamps, unusual sensor usage values, multiple and conflicting values, and missing data. Specifically, concerning the timestamps, the different time zones of pilot sites in different countries posed a challenge, as well as the time changes due to the Daylight Savings Time. Keeping raw data where possible allowed these issues to be addressed and for the analysis to be rerun with the data quality issues resolved. Finally, physical access to the infrastructure can be a significant challenge within operational smart environments. At Linate, the infrastructure was often underground within secured areas of the airport. One cannot rely on having physical access to restart or update infrastructure. As a result, the system design must be fault tolerant and adapt to operating conditions.

Working with three pipelines adds overhead: The complexity of maintaining three (batch, real-time, and entity) different processing pipelines was challenging in terms of the engineering and operational overhead involved. Diagnosing problems and fault required the workflow of all pipelines to be check for issues, and this can increase the time needed to resolve an issue. One possible future direction is to look at end-to-end exactly-once stream processing technologies (Kappa Architectures). However, the highly decentralised nature of a smart environment and the lack of end-to-end control within dataspaces may not be suitable to the additional coordination/control overhead of exactly-once stream processing.

9. Conclusion and future work

As smart environments become a reality, they need to cope with the inherent complexity of the data management challenges that they face. In this paper, we identified the shared data management requirements for Internet-of-Things (IoT) enabled smart energy and water environments. The paper explores the use of the Dataspace data management approach within Smart Environments for real-time IoT and contextual data. The paper introduces a Real-time Linked Dataspace (RLD) that goes beyond traditional data management approaches by combining the pay-as-you-go paradigm of dataspaces and linked data with real-time query capabilities. The RLD Support Platform includes a number of tiered services to support the management of data and users in smart environments including catalog, entity management, and human task services. An entity-centric real-time query service is also provided to support unified queries across entities and both live and historical stream data.

The RLD has been validated within five real-world smart environment pilot deployments to build real-time analytics, decisions

support, and smart apps for smart energy and water management. The pilots demonstrate that the RLD provides effective support services for the development of smart applications and decision support analytics at each stage of the "OODA" loop within smart environments. They show that the entity-centric real-time query service of the RLD is suitable to meet the interactive query latency requirements within smart environments. Finally, based on the results of the pilots the RLD will be further exploited by our industrial partners, including a licensing deal for the platform, a start-up leveraging it within fitness centres, SEA are investigating deployment at Malpensa Airport, and regional governments in Ireland and Greece are exploring additional deployments.

In our future work we plan to investigate:

Large-scale Deployments and Reduced Cost of Operation: We will investigate the performance of the RLD in larger-scale deployments (e.g. city-wide data ecosystems [51]) and within different types of smart environments including mobility and marine. In particular, we will explore ways to improve the maintenance and operational costs of the platform within large-scale deployments.

Enhanced Supported Services: Many enhancements are possible for the support services of the RLD including the use of Natural Language Interfaces to improve the user experience, decentralised support services for large-scale deployments, and further automation to support incremental data integration.

Privacy-by-design: With the increase in personal information captured in smart environments and introduction of the European Union's new General Data Protection Regulation, there is a clear need for additional protection of personal data within the dataspace. Enhanced dataspace support services are needed that leverage privacy-by-design approaches for storing and analysing personal data using processes for anonymisation or pseudonymization.

Scaling Entity Management: Within larger-scale deployments, it will be necessary to enhance the entity services to support both the increase in data and users involved. The use of summarisation techniques for entities in the dataspace could significantly simplify entity management and improve performance. Furthermore, we will investigate the more extensive usage of the human service for human-in-the-loop entity curation within smart environments.

Support Services for Multimedia Data: As multimedia data becomes more common in smart environments through the Internet of Multimedia Things there will a direct need for support within dataspaces. We are investigating support services for rich content types including text and multimedia streams within the dataspace that leverage advances in deep learning for image processing (e.g. object detection) [52].

Acknowledgements

References

[1] E. Ahmed, I. Yaqoob, A. Gani, M. Imran, M. Guizani, Internet-of-things-based smart environments: State of the art, taxonomy, and open research challenges, IEEE Wirel. Commun. 23 (5) (2016).

[2] J.M. Cavanillas, E. Curry, W. Wahlster, New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe, Springer, 2016.

[3] P.A. Bernstein, Middleware: A model for distributed system service, Commun. ACM 39 (2) (1996).

[4] G. Forman, J. Zahorjan, The challenges of mobile computing, Computer 27 (4) (1994).

[5] E. Curry, S. Dustdar, Q.Z. Sheng, A. Sheth, Smart cities –enabling services and applications, J. Internet Serv. Appl. 7 (1) (2016).

[6] A.K. Dey, G.D. Abowd, D. Salber, A context-based infrastructure for smart environments, in: Managing Interactions in Smart Environments, Springer, London, 2000, pp. 114–128.

[7] G.D. Abowd, Beyond Weiser: From ubiquitous to collective computing, Computer 49 (1) (2016).

[8] M. Weiser, R. Gold, J.S. Brown, The origins of ubiquitous computing research at PARC in the late 1980s, IBM Syst. J. 38 (4) (1999).

[9] B. Cheng, S. Longo, F. Cirillo, M. Bauer, E. Kovacs, Building a big data platform for smart cities: Experience and lessons from Santander, in: 2015 IEEE International Congress on Big Data, pp. 592–599.

[10] L. Baresi, L. Mottola, S. Dustdar, Building software for the internet of things, IEEE Internet Comput. 19 (2) (2015).

[11] L. Atzori, A. Iera, G. Morabito, The internet of things: A survey, Comput. Netw. 54 (15) (2010) 2787–2805.

[12] C.C. Aggarwal, N. Ashish, A. Sheth, The internet of things: A survey from the data-centric perspective, in: Managing and Mining Sensor Data, Springer, US, 2013, pp. 383–428.

[13] E. Curry, J. O'Donnell, E. Corry, S. Hasan, M. Keane, S. O'Riain, Linking building data in the cloud: Integrating cross-domain building data using linked data, Adv. Eng. Inform. 27 (2) (2013).

[14] E. Curry, S. Hasan, S. O'Riain, Enterprise energy management using a linked dataspace for energy intelligence, in: The Second IFIP Conference on Sustainable Internet and ICT for Sustainability (SustainIT 2012), 2012, pp. 1–6.

[15] E. Curry, V. Degeler, E. Clifford, D. Coakley, Linked water data for water information management, in: 11th International Conference on Hydroinformatics, HIC, 2014.

[16] D. Pfisterer, et al., SPITFIRE: Toward a semantic web of things, IEEE Commun. Mag. 49 (11) (2011).

[17] D. Puiu, et al., CityPulse: Large scale data analytics framework for smart cities, IEEE Access 4 (2016).

[18] J. Soldatos, et al., OpenIoT: Open source internet-of-things in the cloud, in: Interoperability and Open-Source Solutions for the Internet of Things, Springer International Publishing, 2015, pp. 13–25.

[19] J.M. Schleicher, M. Vogler, S. Dustdar, C. Inzinger, Enabling a smart city application ecosystem: Requirements and architectural aspects, IEEE Internet Comput. 20 (2) (2016).

[20] A. Sheth, P. Anantharam, C. Henson, Physical-cyber-social computing: An early 21st-century approach, IEEE Intell. Syst. 28 (1) (2013).

[21] T. Nam, et al., Smart cities and service integration, in: Proceedings of the 12th Annual International Digital Government Research Conference on Digital Government Innovation in Challenging Times - dg.o '11, 2011, p. 333.

[22] Report Work on the SSN ontology - Semantic Sensor Network Incubator Group, 2016. [Online]. Available: https://www.w3.org/2005/Incubator/ssn/wiki/Report_Work_on_the_SSN_ontology. (Accessed 20 November 2016).

[23] N. Marz, J. Warren, Big Data: Principles and Best Practices of Scalable Realtime Data Systems, first ed., Manning Publications Co., Greenwich, CT, USA, 2015.

[24] K. Akpınar, K.A. Hua, K. Li, ThingStore: A platform for internet-of-things application development and deployment, in: Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems - DEBS '15, 2015, pp. 162–173.

[25] M. Franklin, A. Halevy, D. Maier, From databases to dataspaces: A new abstraction for information management, ACM SIGMOD Rec. 34 (4) (2005).

[26] J. Dittrich, M. Antonio, V. Salles, iDM : A unified and versatile data model for personal dataspace management, in: Proc. 32nd Int. Conf. Very Large Data Bases - VLDB'06, 2006, pp. 367–378.

[27] L. Blunschi, J. Dittrich, O.R. Girard, S. Kirakos, K. Marcos, A.V. Salles, A dataspace odyssey: The iMeMex personal dataspace management system, in: Conf. Innov. Data Syst. Res., 2007, pp. 114–119.

[28] R. Grossman, E. Creel, M. Mazzucco, R. Williams, A dataspace infrastructure for astronomical data, in: Data Mining for Scientific and Engineering Applications, Springer US, 2001, pp. 115–123.

[29] A. Hasnain, et al., Linked biomedical dataspace: Lessons learned integrating data for drug discovery, in: The Semantic Web – ISWC 2014, Springer International Publishing, 2014, pp. 114–130.

[30] D.W. Archer, L.M.L. Delcambre, D. Maier, A Framework for Fine-grained Data Integration and Curation, with Provenance, in a Dataspace, in: TAPP'09 First Work. Theory Pract. Proven., 2009.

[31] Y. Li, X. Meng, Supporting context-based query in personal dataspace, in: Proceeding of the 18th ACM Conference on Information and Knowledge Management - CIKM '09, 2009, p. 1437.

[32] M. Zhong, M. Liu, Q. Chen, Modeling heterogeneous data in dataspace, in: 2008 IEEE Int. Conf. Inf. Reuse Integr., 2008, pp. 404–409.

[33] A.D. Sarma, X. Dong, A.Y. Halevy, Data modeling in dataspace support platforms, in: Conceptual Modeling: Foundations and Applications, Springer, Berlin, Heidelberg, 2009, pp. 122–138.

[34] R. Grossman, M. Mazzucco, DataSpace: A data web for the exploratory analysis and mining of data, Comput. Sci. Eng. 4 (4) (2002).

[35] U. ul Hassan, S. O'Riain, E. Curry, Leveraging matching dependencies for guided user feedback in linked data applications, in: Proceedings of the Ninth International Workshop on Information Integration on the Web - IIWeb '12, 2012, pp. 1–6.

[36] S.R. Jeffery, M.J. Franklin, A.Y. Halevy, Pay-as-you-go user feedback for dataspace systems, in: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008, p. 847.

[37] E. Curry, System of systems information interoperability using a linked dataspace, in: 2012 7th International Conference on System of Systems Engineering, SoSE, pp. 101–106.

[38] A. Freitas, E. Curry, J.G. Oliveira, S. O'Riain, Querying heterogeneous datasets on the linked data web: challenges, approaches, and trends, IEEE Internet Comput. 16 (1) (2012).

[39] A. Freitas, E. Curry, Natural language queries over heterogeneous linked data graphs: A distributional-compositional semantics approach, in: Proceedings of the 19th International Conference on Intelligent User Interfaces - IUI '14, 2014, pp. 279–288.

[40] T. Heath, C. Bizer, Linked data: Evolving the web into a global data space, Synth. Lect. Semantic Web: Theory Technol. 1 (1) (2011).

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

[42] S. Hasan, E. Curry, Curry, Thingsonomy: Tackling variety in internet of things events, IEEE Internet Comput. 19 (2) (2015).

[43] S. Hasan, E. Curry, M. Banduk, S. O'Riain, Toward situation awareness for the semantic sensor web: Complex event processing with dynamic linked data enrichment, in: Proceedings of the 4th International Workshop on Semantic Sensor Networks 2011 - SSN11, 2011.

[44] U. ul Hassan, M. Bassora, A. Vahid, S. O'Riain, E. Curry, A collaborative approach for metadata management for internet of things, in: Proceedings of the 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, 2013.

[45] U. ul Hassan, A. Zaveri, E. Marx, E. Curry, J. Lehmann, ACRyLIQ: Leveraging DBpedia for adaptive crowdsourcing in linked data quality assessment, in: Knowledge Engineering and Knowledge Management, Springer International Publishing, 2016, pp. 681–696.

[46] U. ul Hassan, E. Curry, Efficient task assignment for spatial crowdsourcing: A combinatorial fractional optimization approach with semi-bandit learning, Expert Syst. Appl. 58 (2016).

[47] T. Berners-Lee, Linked data-design issues, 2006, https://www.w3.wrg/Design Issues/LinkedData.Html.

[48] R. Boyd John, A discourse on winning and losing, in: Air University Document MU43947, Briefing, vol. 1, 1987.

[49] R.B. Miller, Response time in man-computer conversational transactions, in: Proceedings of the Fall Joint Computer Conference, Part I on - AFIPS '68 (Fall, Part I), 1968, p. 267.

[50] J. Nielsen, Usability Engineering, vol. 2, Kindle ed., Morgan Kaufmann, 1994.

[51] E. Curry, A. Sheth, Next-generation smart environments: From system of systems to data ecosystems, IEEE Intell. Syst. 33 (2018) 69–76.

[52] A. Aslam, E. Curry, Towards a generalized approach for deep neural network based event processing for the internet of multimedia things, IEEE Access 6 (2018) 25573–25587.

Wassim Derguech is currently a technical manager at Derilinx Ltd., a company that specializes in the development of Linked and Open Data Solutions for publishing, analyzing, and visualizing enterprise and open data using Linked Data principles. Prior to this role, Wassim was a researcher at the Insight Centre for Data Analytics at the National University of Ireland, Galway. He obtained his Ph.D. degree from the same university in 2016. In his research, he experienced semantic web and related technologies for managing services and business processes applied to various domains such as e-Government, customs clearance procedures, Internet of Things, Green and Sustainable IT and e-Learning analytics. Particularly, he was involved in SENSE and Waternomics projects in data modelling, quality assessment, enrichment, publishing, access control and analytics.

Souleiman Hasan received the Ph.D. degree in Computer Science from the National University of Ireland, Galway (NUIG) in 2016. He was a Ph.D. researcher and then a postdoctoral researcher at the Insight Centre for Data Analytics at NUIG where he investigated the data coupling problem in distributed event processing systems within heterogeneous environments, such as the Internet of Things, using advanced data analytics techniques for data representation. Applications of the research in the smart cities, power management, and water management domains have been among the SFI, Enterprise Ireland, and European projects he was involved in including DERI Energy, SENSE, and Waternomics. He has served as a program committee member and reviewer for several scientific venues and his work has been published in various journals and international conferences. Dr. Hasan is currently a Lecturer at Maynooth University, Ireland.

Christos Kouroupetroglou was awarded his Ph.D. in 2010 and his thesis subject was "Semantically enhanced web browsing interfaces". His latest research and development efforts include his participation in two EU funded projects for two Greek SMEs dealing with the use of ICT to raise awareness for water conservation issues (http://www.waternomics.eu) and the use of robotics to combat loneliness and isolation of people with dementia (http://www.mario-porject.eu) . He is teaching as a laboratory associate at the ATEI of Thessaloniki and at the Mediterranean college of Thessaloniki in modules related to Web Applications Development and as a distance learning lecturer for the University of Nicosia in the module of "Application of Technology in Special Education". He is the author of "Enhancing the Human Experience through Assistive Technologies and E-Accessibility" and has chaired a number of conference sessions and online symposia.

Edward Curry is currently a research leader at the Insight Centre for Data Analytics and at LERO The Irish Software Research Centre. His research interests are predominantly in open distributed systems, particularly in the areas of incremental data management, approximation, and unstructured event processing, with a special interest in applications for smart environments and data ecosystems. His research work is currently focused on engineering adaptive systems that are a foundation of smart and ubiquitous computing environments. He has authored or co-authored over 150 scientific articles in journals, books, and international conferences. He has presented at numerous events and has given invited talks at Berkeley, Stanford, and MIT. He is a Vice President of the Big Data Value Association—a non-profit industry-led organization with the objective of increasing the competitiveness of European Companies with data-driven innovation. He is a Lecturer in informatics with the National University of Ireland Galway.

Umair ul Hassan is currently a Research Fellow at the Insight Centre of Data Analytics After completing his Ph.D. at the National University of Ireland Galway on the topic of assignment problems in spatial crowdsourcing, he worked as a postdoctoral researcher on WATERNOMICS project. He has extensive work experience in the industry, government, and academia; furthermore, he has been directly involved in Big Data & Analytics projects funded by Science Foundation Ireland and European Commission in telecom, energy, transport, and environment sectors. He has served on the technical committee of several international conferences and workshops. He is the winner of the Best Paper Award at IEEE International Conference on Ubiquitous Intelligence and Computing in 2014. His general interests include collaborative data platforms, human computation and crowdsourcing, smart environments, and software ecosystems for addressing societal challenges.